

Differentially private vector aggregation in the case of multivariate gaussian data

Tim Sehested Poulsen

January 4, 2023

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 referred 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 [6]). *Two dataset $X, X' \in \mathbb{R}^{n \times d}$ are said to be neighbouring 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 [7]). *Let $f(X) : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^k$ be a function. The l_p -sensitivity of f is the maximal possible l_p -norm of the difference between the output of f on two neighbouring dataset. We denote the sensitivity as*

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

and then the total l_2 -sensitivity is then

Throughout the report I will only be working with l_2 -sensitivity and will just denote this as $\Delta(f)$ for ease of notation.

Definition 1.3 ((ε, δ)-Differential Privacy [6]). *A randomized algorithm $\mathcal{M} : \mathbb{R}^{n \times d} \rightarrow \mathcal{R}$ 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*

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

Theorem 1 ((ε, δ)-DP under post-processing [7]). *Let $\mathcal{M} : \mathbb{R}^{n \times d} \rightarrow \mathcal{R}$ be an (ε, δ)-DP algorithm. Let $f : \mathcal{R} \rightarrow \mathcal{R}'$ be an arbitrary mapping, then $f \circ \mathcal{M} : \mathbb{R}^{n \times d} \rightarrow \mathcal{R}'$ is (ε, δ)-DP*

Proof

Fix any pair of neighbouring datasets $X \sim X'$ and let $S \subseteq \mathcal{R}'$ be an arbitrary event. We then define $T = \{r \in \mathcal{R} \mid f(r) \in S\}$. We thus have that

$$\begin{aligned} \Pr [f(\mathcal{M}(X)) \in S] &= \Pr [\mathcal{M}(X) \in T] \\ &\leq e^\varepsilon \cdot \Pr [\mathcal{M}(X') \in T] + \delta = e^\varepsilon \cdot \Pr [f(\mathcal{M}(X')) \in S] + \delta \end{aligned}$$

■

Error Measure As this report concerns itself exclusively with the sum of entries in a dataset, error will be defined as the expected squared l_2 -norm between the true sum and the output of a randomized algorithm. So let $X \in \mathbb{R}^{n \times d}$ be the dataset and $f(X) = \sum_i^n x_i$ be the true sum of all entries. The error of a randomized algorithm $M : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$ which estimates $f(X)$ is then

$$\text{Err}(M) := \mathbb{E} [\|M(X) - f(X)\|^2]$$

extra

1.2 Quadratic forms of random variables

write cohesive text here, and decide what to keep

Lemma 1.1. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a gaussian random variable and let $r \in \mathbb{R}$ be a constant. We then have that*

$$\begin{aligned} rX &\sim \mathcal{N}(\mu, (r\sigma)^2) \\ X - r &\sim \mathcal{N}(\mu - r, \sigma^2) \end{aligned}$$

Quadratics of random variables have been well studied [2, 11], specially in the case of multivariate gaussian variables [8, 11]. Even more research has been done in evaluating the CDF of these quadratic forms for Gaussian random vectors [4, 9].

Theorem 2 (Expectation of a quadratic random variable [2]). *Let X be a d -dimensional random vector with expected value $\mathbb{E}[X] = \boldsymbol{\mu}_X$ and covariance matrix $\text{Var}[X] = \boldsymbol{\Sigma}_X$. Let also A be a constant $d \times d$ symmetric matrix, then*

$$\mathbb{E}[X^T A X] = \text{tr}(A \boldsymbol{\Sigma}_X) + \boldsymbol{\mu}^T A \boldsymbol{\mu}$$

Proof
Blah blah

$$\begin{aligned}\mathbb{E} [X^T A X] &= \text{tr} (\mathbb{E} [X^T A X]) = \mathbb{E} [\text{tr} (X^T A X)] \\ &= \mathbb{E} [\text{tr} (A X X^T)] = \text{tr} (A \mathbb{E} [X X^T]) = \text{tr} (A (\text{Var} [X] + \boldsymbol{\mu} \boldsymbol{\mu}^T)) \\ &= \text{tr} (A \boldsymbol{\Sigma}) + \text{tr} (A \boldsymbol{\mu} \boldsymbol{\mu}^T) = \text{tr} (A \boldsymbol{\Sigma}) + \boldsymbol{\mu}^T A \boldsymbol{\mu}\end{aligned}$$

blah ■

Corollary 1.1. *Let $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_X)$ be a d -dimensional gaussian vector with expected value $\mathbf{0}$, and let σ_j^2 denote the variance of the j 'th dimension where $1 \leq j \leq d$. By theorem 2 we have that the expected l_2 -norm of such a vector is given by*

$$\mathbb{E} [\|X\|^2] = \text{tr}(\boldsymbol{\Sigma}_X) = \sum_{j=1}^d \sigma_j^2$$

2 Algorithms

2.1 The Gaussian Mechanism

One of the most foundational algorithms for achieving (ε, δ) -DP is the Gaussian Mechanism [7]. 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 $g : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$ where the l_2 -sensitivity $\Delta(g)$ is known. It does so by computing the value of $g(X)$ and then adding noise to each dimension drawn from the normal distribution $\mathcal{N}(0, \sigma_{\varepsilon, \delta}^2)$. This can be seen as adding a noise vector η which is then distributed according to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \sigma_{\varepsilon, \delta}^2 I)$. Pseudo code for the algorithm can be seen in Algorithm 1.

Algorithm 1 The Gaussian Mechanism

Input

$\sigma_{\varepsilon, \delta}$ Standard deviation required to achieve (ε, δ) -DP
 $X \in \mathbb{R}^{n \times d}$ Dataset

Output

(ε, δ) -DP estimate of $g(X)$
 $\eta \leftarrow \text{sample from } \mathcal{N}(\vec{0}, \sigma_{\varepsilon, \delta}^2 I)$
return $g(X) + \eta$

The algorithm quite intuitively produces error which is purely given by the norm of the noise added, and the expected error can be calculated to be

$$\mathbb{E} [\|(g(X) + \eta) - g(X)\|^2] = \mathbb{E} [\|\eta\|^2]$$

Which by corollary 1.1 is

$$\mathbb{E} [\|\eta\|^2] = \sum_i^d \sigma_{\varepsilon, \delta}^2 = d \cdot \sigma_{\varepsilon, \delta}^2 \tag{1}$$

It is 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 3. [7] *Let $g : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$ be an arbitrary d -dimensional function with l_2 -sensitivity $\Delta(g) = \max_{X \sim X'} \|g(X) - g(X')\|$, and let $\varepsilon \in (0, 1)$. The Gaussian Mechanism with $\sigma_{\varepsilon,\delta} = \Delta(g) \sqrt{2 \ln(1.25/\delta)}/\varepsilon$ is (ε, δ) -DP.*

The proof is rather long and is therefore omitted here. A few years later it was shown in [1] how to compute the minimal $\sigma_{\varepsilon,\delta}$.

Theorem 4. [1] *The Gaussian Mechanism is differentially private if and only if $\sigma_{\varepsilon,\delta} \geq \Delta(g) \cdot \sigma_{opt}$ where $\sigma_{opt} \in \mathbb{R}$ is the smallest value greater than 0, which satisfies*

$$\Phi\left(\frac{1}{2\sigma_{opt}} - \varepsilon\sigma_{opt}\right) - e^\varepsilon \Phi\left(-\frac{1}{2\sigma_{opt}} - \varepsilon\sigma_{opt}\right) \leq \delta$$

In [1] it is also shown how to compute this value, and since this is of no importance to this project it will therefore not be covered. For the rest of the report I will only be referring to $\sigma_{\varepsilon,\delta}$ as the minimal value given by theorem 4 and not that given by theorem 3.

The main downside of the Gaussian Mechanism is that it adds equal noise to all dimensions, regardless of the sensitivity in that dimension. As this report focuses on giving differentially private estimates of sums of vectors, it is natural to ask whether adding equal noise in all dimensions is optimal in this setting. Let us define the function of interest $f(X) = \sum_{i \in [n]} x_i$, for a dataset $X \in \mathbb{R}^{n \times d}$. The sensitivity of this function must be given by the largest possible norm of any vector. Therefore if the largest difference between any two neighbouring datasets in the j 'th dimension is given by

$$\Delta_j := \max_{X \sim X'} |X^{(j)} - X'^{(j)}| \quad (2)$$

and we say that X and X' differ in the i 'th entry we conclude that

$$\Delta(f) = \max_{X \sim X'} \|f(X) - f(X')\| = \max_{X \sim X'} \|x_i - x'_i\| = \sqrt{\sum_{j \in [d]} \Delta_j^2} = \|\Delta\| \quad (3)$$

This means that by equation 1 the expected error of the Gaussian Mechanism when estimating $f(X)$ is given by

$$\mathbb{E}[\|\eta\|^2] = d \cdot \sigma_{\varepsilon,\delta}^2 = d \cdot (\Delta(f)\sigma_{opt})^2 = d \cdot \sigma_{opt}^2 \cdot \|\Delta\|^2 \quad (4)$$

A logical next step would be to add noise to each dimensions such that it is proportional to the sensitivity of that dimension. This has been studied in [12] and lays the foundation for this report. Their mechanism, appropriately called the Elliptical Gaussian Mechanism works very similarly to the Gaussian Mechanism as described by algorithm 1, though instead of sampling all η_j from $\mathcal{N}(0, \sigma_{\varepsilon,\delta}^2)$, here they are instead drawn from $\mathcal{N}\left(0, \left(\sigma_{opt} \cdot \frac{\Delta_j}{b_j}\right)^2\right)$. The algorithm is described in detail in Algorithm 2

Algorithm 2 The Elliptical Gaussian Mechanism

Input

σ_{opt} Standard deviation as defined by theorem 4
 $X \in \mathbb{R}^{n \times d}$ Dataset
 $\mathbf{b} \in \mathbb{R}^d$ Scaling vector, where $\|\mathbf{b}\| = 1$
 $\Delta \in \mathbb{R}^d$ Sensitivies of all dimensions

Output

(ε, δ) -DP estimate of $f(X)$

for $j \in [d]$ **do**

$\sigma_j \leftarrow \sigma_{opt} \cdot \frac{\Delta_j}{b_j}$

$\eta_j \leftarrow \text{sample from } \mathcal{N}(0, \sigma_j^2)$

end for

return $g(X) + \eta$

add lemma/theorem/proof of why this is (ε, δ) -DP (figure out if needed). It is very similair to why it works for Gaussian data.

The main thing that really differentiates itself from the normal Gaussian Mechanism is the introduction of the scaling vector \mathbf{b} . This is the scaling of how much weight should be attributed to each dimension when adding noise. It is shown how to determine the optimal values for b_j in [12], and what the expected error of the mechanism then is.

Theorem 5 (Optimality and error of the Elliptical Gaussian Mechanism [12]). *The value for b_j which minimizes the expected l_2 error $\mathbb{E}[\|\eta\|^2]$ of the Elliptical Gaussian Mechanism is as follows*

$$b_j = \sqrt{\frac{\Delta_j}{\sum_{j \in [d]} \Delta_j}}$$

which leads the error to be

$$\mathbb{E}[\|\eta\|^2] = \sigma_{opt}^2 \cdot \|\Delta\|_1^2 \quad (5)$$

where $\|\cdot\|_1$ is the l_1 norm.

write the proof or skip it, it is going to be very similair to that which I will make

Comparing the expected error between Algorithm 1 and Algorithm 2 we have the following ratio

$$\frac{d \cdot \sigma_{opt}^2 \cdot \|\Delta\|^2}{\sigma_{opt}^2 \cdot \|\Delta\|_1^2} = \left(\frac{\sqrt{d} \|\Delta\|}{\|\Delta\|_1} \right)^2$$

in which it can be seen that they are equal when all entries of Δ are the same. Otherwise the error for the Elliptical Gaussian Mechanism is lower when Δ is skewed. As argued in [12] Algorithm 2 improves Algorithm 1 by a factor in $[1, d]$.

3 Problem setup

As previously mentioned the problem investigated here consists of realeasing the sum of vectors in a dataset under differential privacy. More formally we wish to release the

value of $f(X) = \sum_{i=1}^n x_i$ under (ε, δ) -DP .

The common factor for achieving (ε, δ) -DP in both the Gaussian Mechanism and the Elliptical Gaussian Mechanism is the knowledge that data lie within some bounds. Specifically for the Elliptical Gaussian Mechanism data is required to lie within some hyperrectangle. It is formally described by equation (2) essentially saying that there is an upper and lower bound on each dimension. This requirement is needed to know the l_2 -sensitivity $\Delta(f)$ as shown in equation (3). In this project I will change this assumption and instead look at the case where each dimension is normally distributed. This means that for each $j \in [d]$ we have that $x_{i,j} \sim \mathcal{N}(\mu_j, \sigma_j^2)$. An equivalent formulation is that the data is multivariately gaussianly distributed but with no correlation between dimensions. This means that $x_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where Σ is a diagonal matrix with the variance of each dimension along its diagonal. It is quite apparent that determining a limit Δ_j is impossible in this setting as the Gaussian distribution is continuously defined on the range $(-\infty, \infty)$. Several recent papers has combatted this by doing something called *clipping* [3, 10]. Clipping is the process of limiting the norm of any one entry to be at most a chosen threshold C . This means that every vector is transformed as such

$$\hat{x}_i := \min \left\{ \frac{C}{\|x_i\|}, 1 \right\} \cdot x_i$$

Clipping entries by a factor C thus means that $\Delta(f) = 2C$ as any one entry cannot have more impact on the summation than C . If the summation $f(X)$ is performed on a clipped dataset \hat{X} it is equivalent to defining the summation function $\hat{f} : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$ as

$$\hat{f}(X) = \sum_i^n \min \left\{ \frac{C}{\|x_i\|}, 1 \right\} \cdot x_i$$

Then by theorem 4 the gaussian mechanism with the function \hat{f} is (ε, δ) -DP with $\sigma_{\varepsilon, \delta} = \Delta(\hat{f}) \sigma_{opt} = 2C \sigma_{opt}$. Though the mechanism is still (ε, δ) -DP it will now have a larger error when regarding the true sum $f(X) = \sum_i^n x_i$ as the actual answer. If the probability of clipping is set so low that we actually don't expect to clip any entries we can use \hat{f} as an approximation of f . Say there are n points in a dataset, I will thus set the probability of clipping to be less than $\frac{1}{n}$ and get that

$$\mathbb{E} \left[\left\| \left(\hat{f}(X) + \eta \right) - f(X) \right\|^2 \right] \approx \mathbb{E} \left[\left\| (f(X) + \eta) - f(X) \right\|^2 \right] = \mathbb{E} \left[\|\eta\|^2 \right]$$

rewrite and introduce better

In this setting, we again have the intuition that adding equal noise in all dimensions is non optimal, and instead the noise in a dimension should be proportional to the variance σ_j^2 in that dimension. In a desire to achieve similar results to that of the Elliptical Gaussian Mechanism, just with normally distributed data, I will use somewhat the same approach to achieve (ε, δ) -DP . They achieve (ε, δ) -DP by finding a transformation of points such that they all lie within the unit ball centered at the origin. There does not exist such a transformation which is linear (I think I mean Continuous/Homeomorphic), as there will always be a non-zero probability of observing points outside the unit ball.

Instead I will introduce the constraint that the expected norm of vectors after the transformation should be 1. I wish to find a scaling of each dimension b_j s.t.

$$\mathbb{E} [\|x_i \odot \mathbf{b}\|] = 1 \iff \mathbb{E} [\|x_i \odot \mathbf{b}\|^2] = \sum_{j \in [d]} (\sigma_j b_j)^2 = 1 \quad (6)$$

where $\mathbf{b} = (b_1, b_2, \dots, b_d)$, \odot is the element-wise product, and σ_j is the standard deviation of the j 'th dimension. Then \hat{f} can be computed on this transformed dataset, where the probability of clipping is less than $\frac{1}{n}$. As Theorem 1 shows, (ε, δ) -DP is preserved under post processing, we can therefore add noise to each coordinate drawn from $\mathcal{N}(0, (2C\sigma_{opt})^2)$ in the transformed space to achieve (ε, δ) -DP. The transformation back to the original space is then done by multiplying each dimension with b_j^{-1} , and due to the linearity of transformation this is also done to the noise added. We end up with the following noise vector being added

$$\boldsymbol{\eta} = (b_1^{-1}\eta_1, b_2^{-1}\eta_2, \dots, b_d^{-1}\eta_d)$$

in which all $\eta_j \sim \mathcal{N}(0, (2C\sigma_{opt})^2)$, and then by lemma 1.1 we have $b_j^{-1}\eta_j \sim \mathcal{N}(0, (b_j^{-1} \cdot 2C\sigma_{opt})^2)$. As the error is approximately given by $\|\boldsymbol{\eta}\|$ we have due to Corollary 1.1 that the expected error is

$$\mathbb{E} [\|\boldsymbol{\eta}\|^2] = \sum_{j=1}^d (b_j^{-1} \cdot 2C\sigma_{opt})^2 = (2C\sigma_{opt})^2 \sum_{j=1}^d b_j^{-2} \quad (7)$$

Give pseudo code for algorithm

Now we are back to a very similar problem to the one solved in [12]. Minimize the error described in equation (7) under the constraint given in equation (6). This leads to the following Lemma

Lemma 3.1. *Let $X \in \mathbb{R}^{n \times d}$ be a dataset where each dimension is independently gaussianly distributed, i.e. $x_{i,j} \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Then the expected error of Algorithm ?? is minimized when*

$$b_j = \frac{1}{\sqrt{\sigma_j} \sqrt{\sum_{i=1}^d \sigma_i}}$$

Proof

Using lagrangian multipliers we find the local maxima or minima of the function subject to equality constraints. To do so we construct the lagrangian function $\mathcal{L} : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ from the optimization problem given by equation (7) and the constraint given by (6) for ease of notation we define $\sigma_{\varepsilon, \delta} := 2C\sigma_{opt}$.

$$\mathcal{L}(\mathbf{b}, \lambda) = \sigma_{\varepsilon, \delta}^2 \sum_{j \in [d]} b_j^{-2} + \lambda \left(\sum_{j=1}^d (\sigma_j b_j)^2 - 1 \right)$$

We then find the stationary points of it, by setting the derivative of it to $\mathbf{0}$. The derivative with respect to b_j is

$$\frac{\partial \mathcal{L}}{\partial b_j}(\mathbf{b}, \lambda) = \frac{\partial}{\partial b_j} (\sigma_{\varepsilon, \delta}^2 \cdot b_j^{-2} + \lambda (\sigma_j b_j)^2) = -2\sigma_{\varepsilon, \delta}^2 b_j^{-3} + 2\lambda \sigma_j^2 b_j$$

I then solve $\frac{\partial \mathcal{L}}{\partial b_j} = 0$ for b_j

$$-2\sigma_{\varepsilon,\delta}^2 b_j^{-3} + 2\lambda\sigma_j^2 b_j = 0 \iff \lambda\sigma_j^2 b_j = \sigma_{\varepsilon,\delta}^2 b_j^{-3} \quad (8)$$

$$\iff b_j^4 = \frac{\sigma_{\varepsilon,\delta}^2}{\lambda\sigma_j^2} \iff b_j = \frac{\sqrt{\sigma_{\varepsilon,\delta}}}{\lambda^{\frac{1}{4}}\sqrt{\sigma_j}} \quad (9)$$

I now have the last partial derivative $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ which I solve for λ using the previous expression for b_j .

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{j=1}^d (\sigma_j b_j)^2 - 1 \\ \sum_{j=1}^d (\sigma_j b_j)^2 - 1 &= 0 \iff \sum_{j=1}^d \sigma_j^2 \left(\frac{\sigma_{\varepsilon,\delta}}{\sqrt{\lambda}\sigma_j} \right) = 1 \iff \\ \frac{\sigma_{\varepsilon,\delta}}{\sqrt{\lambda}} \sum_{j=1}^d \frac{\sigma_j^2}{\sigma_j} &= 1 \iff \sigma_{\varepsilon,\delta} \sum_{j=1}^d \sigma_j = \sqrt{\lambda} \end{aligned}$$

Inserting back into equation 9

$$b_j = \frac{\sqrt{\sigma_{\varepsilon,\delta}}}{\lambda^{\frac{1}{4}}\sqrt{\sigma_j}} = \frac{\sqrt{\sigma_{\varepsilon,\delta}}}{\sqrt{\sigma_{\varepsilon,\delta}} \sum_{i=1}^d \sigma_i \sqrt{\sigma_j}} = \frac{1}{\sqrt{\sigma_j} \sqrt{\sum_{i=1}^d \sigma_i}}$$

show that this stationary point is a minimum

Giving us the value of b_j which minimizes the expected error. ■

Now to evaluate the expected error of the mechanism, which is given by equation (7), it is highly important to determine the value C . As C should be given by the minimal value which satisfies the inequality

$$\Pr [\|x_i \odot \mathbf{b}\| > C] = \Pr [\|x_i \odot \mathbf{b}\|^2 > C^2] < \frac{1}{n}$$

again where n denotes the number of points in the dataset. As x_i is a multivariate gaussian distribution, so is $x_i \odot \mathbf{b}$, and $\|x_i \odot \mathbf{b}\|^2$ is the sum of independent gaussian variables squared. Such a sum is distributed as a generalized Chi-square [5], and neither its PDF nor CDF has a closed form. These quadratic forms of gaussian vectors has been studied well and for decades [11], and some special cases does have closed forms. Some have also studied giving tail bounds on these probabilities [9], unfortunately there was none which were of use in this setting. However, there does exist several numerical algorithms for evaluating the cumulative density function with high precision [4]. I would recommend using one of these algorithms in practice, but for the sake of analysis I will provide an upper bound on C using Bernstein's inequality.

Should I give the general definition of Bernsteins inequality before here? Or just inside the proof

Lemma 3.2. Let X_1, X_2, \dots, X_d be d independent random gaussian variables where for $1 \leq j \leq d$ we have that $X_j \sim \mathcal{N}(0, \sigma_j^2)$. Let thereafter the largest standard deviation be denoted as $\sigma_* := \max_{j \in [d]} \sigma_j$, we can then bound the probability for the sum of variables squared as follows

$$\Pr \left[\sum_{j \in [d]} X_j^2 \geq t \sqrt{8 \sum_{j \in [d]} \sigma_j^4 + \sum_{j \in [d]} \sigma_j^2} \right] < e^{-t^2}$$

for

$$0 \leq t \leq \frac{1}{6\sigma_*^2} \sqrt{2 \sum_{j \in [d]} \sigma_j^4}$$

Proof

At first we define the random variable $Y_j = X_j^2 - \mathbb{E}[X_j^2]$ using the j 'th gaussian random variable. As $\mathbb{E}[Y_j] = \mathbb{E}[X_j^2 - \mathbb{E}[X_j^2]] = \mathbb{E}[X_j^2] - \mathbb{E}[X_j^2] = 0$ we have that Y_j is zero centered. We are thus interested in giving bounds on $\Pr \left[\sum_{j \in [d]} Y_j \right]$. We can use Bernsteins inequality, if the following constraint holds for all $k \in \mathbb{N}$ with $k \geq 2$ and for all $j \in [d]$, and for some $L \in \mathbb{R}$

$$\mathbb{E}[|Y_j^k|] \leq \frac{1}{2} \mathbb{E}[Y_j^2] L^{k-2} k! \quad (10)$$

Initially we have that $\mathbb{E}[|Y_j^k|] = \mathbb{E}[|(X_j^2 - \mathbb{E}[X_j^2])^k|]$ and since $X_j^2 \geq 0$ and therefore also $\mathbb{E}[X_j^2] \geq 0$ we can therefore bound it by

$$\mathbb{E}[|(X_j^2 - \mathbb{E}[X_j^2])^k|] \leq \mathbb{E}[|X_j^{2k}|] = \mathbb{E}[|(\sigma_j^2 Z)^k|] = \sigma_j^{2k} \cdot \mathbb{E}[|Z^k|]$$

Where $Z \sim \chi_1^2$ is a chi square with 1 degree of freedom. The moment generating function of Z is $\mathbb{E}[Z^m] = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2m-1)$. Using this we can get the following bound

$$\begin{aligned} \sigma_j^{2k} \cdot \mathbb{E}[|Z^k|] &= \sigma_j^{2k} \cdot \prod_{c=1}^k (2c-1) = \sigma_j^{2k} \cdot 3 \cdot \prod_{c=3}^k (2c-1) \\ &\leq \sigma_j^{2k} \cdot 3 \cdot \prod_{c=3}^k 2c = \frac{3}{2} \sigma_j^{2k} \cdot 2^{k-2} \cdot k! \end{aligned}$$

Concluding that $\mathbb{E}[|Y_j^k|] \leq \frac{3}{2} \sigma_j^{2k} \cdot 2^{k-2} \cdot k!$.

Secondly I calculate $\mathbb{E}[Y_j^2]$ exactly to be used in equation 10

$$\mathbb{E}[Y_j^2] = \mathbb{E}[(X_j^2 - \mathbb{E}[X_j^2])^2] \quad (11)$$

$$= \mathbb{E}[X_j^4 + \mathbb{E}[X_j^2]^2 - 2X_j^2 \mathbb{E}[X_j^2]] \quad (12)$$

$$= \mathbb{E}[X_j^4] + \mathbb{E}[X_j^2]^2 - 2\mathbb{E}[X_j^2]^2 \quad (13)$$

$$= \mathbb{E}[X_j^4] - \mathbb{E}[X_j^2]^2 = \sigma_j^4 \mathbb{E}[Z^2] - (\sigma_j^2 \mathbb{E}[Z])^2 \quad (14)$$

$$= 3\sigma_j^4 - \sigma_j^4 = 2\sigma_j^4 \quad (15)$$

Equation (10) is therefore rewritten to

$$\frac{3}{2}\sigma_j^{2k} \cdot 2^{k-2} \cdot k! \leq \frac{1}{2} \cdot 2\sigma_j^4 L^{k-2} k! \iff \frac{3}{2}\sigma_j^{2k-4} \cdot 2^{k-2} \leq L^{k-2}$$

This expression is then split into two cases, when $k = 2$, in which it can be seen from equation (10) that this holds for any L . The other case is $k > 2$, where we have that

$$\left(\frac{3}{2}\right)^{\frac{1}{k-2}} \sigma_j^2 \cdot 2 \leq L$$

As $\lim_{k \rightarrow \infty} \left(\frac{3}{2}\right)^{\frac{1}{k-2}} = 1$, and these constraints must hold for all $j \in [d]$, we can define $\sigma_* = \max_{j \in [d]} \sigma_j$ and finally have that

$$L = 3\sigma_*^2$$

We can then use Bernstein's inequality to bound the following

$$\Pr \left[\sum_{j \in [d]} Y_j \geq 2t \sqrt{\sum_{j \in [d]} \mathbb{E}[Y_j^2]} \right] < e^{-t^2} \quad (16)$$

Which in this case can be rewritten to get a tail bound on $\sum_{j \in [d]} X_j^2$, by reusing results from equations (11)-(15)

$$\begin{aligned} \Pr \left[\sum_{j \in [d]} Y_j \geq 2t \sqrt{\sum_{j \in [d]} \mathbb{E}[Y_j^2]} \right] &= \Pr \left[\sum_{j \in [d]} (X_j^2 - \mathbb{E}[X_j^2]) \geq 2t \sqrt{\sum_{j \in [d]} 2\sigma_j^4} \right] \\ &= \Pr \left[\sum_{j \in [d]} X_j^2 \geq 2t \sqrt{\sum_{j \in [d]} 2\sigma_j^4} + \sum_{j \in [d]} \mathbb{E}[X_j^2] \right] = \Pr \left[\sum_{j \in [d]} X_j^2 \geq 2t \sqrt{\sum_{j \in [d]} 2\sigma_j^4} + \sum_{j \in [d]} \sigma_j^2 \right] \\ &= \Pr \left[\sum_{j \in [d]} X_j^2 \geq t \sqrt{8 \sum_{j \in [d]} \sigma_j^4} + \sum_{j \in [d]} \sigma_j^2 \right] \end{aligned}$$

Finally giving us that

$$\Pr \left[\sum_{j \in [d]} X_j^2 \geq t \sqrt{8 \sum_{j \in [d]} \sigma_j^4} + \sum_{j \in [d]} \sigma_j^2 \right] < e^{-t^2}$$

as long as t lies within the following bounds

$$0 \leq t \leq \frac{1}{2L} \sqrt{\sum_{j \in [d]} \mathbb{E}[Y_j^2]} = \frac{1}{6\sigma_*^2} \sqrt{2 \sum_{j \in [d]} \sigma_j^4}$$

■

Combining lemma 3.2 with theorem 3.1 we have can conclude the following:

Introduce variables better

Theorem 6. Let $X \in \mathbb{R}^{n \times d}$ be a dataset in which all dimensions are independently gaussian random variables, i.e. $x_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$. Let $\sigma_* := \max_{j \in [d]} \sigma_j$ be the maximal standard deviation of all dimensions. We then have that when $\ln(n) \leq \frac{\sum_{j \in [d]} \sigma_j^2}{18\sigma_*^2}$ the expected error of algorithm ?? can be upper bounded by

$$\mathbb{E} [\|\boldsymbol{\eta}\|^2] \leq 4\sigma_{opt}^2 \cdot \sum_{j \in [d]} \sigma_j \cdot \left(\sqrt{8 \ln(n) \sum_{i \in [d]} \sigma_i^2} + \sum_{j \in [d]} \sigma_j \right)$$

Proof

By equation (7) the error is given by

$$\mathbb{E} [\|\boldsymbol{\eta}\|^2] = (2C\sigma_{opt})^2 \cdot \sum_{i=1}^d b_i^{-2} \quad (17)$$

When clipping is performed to remove less than n^{-1} we have the following

$$\Pr [\|x_i\| \geq C] = \Pr [\|x_i\|^2 \geq C^2] = \Pr \left[\sum_{j \in [d]} x_{i,j}^2 \geq C^2 \right]$$

clean up sigma notation

Which means I can give an upper bound on C^2 by using lemma 3.2, and inserting that $\sigma_j = \sigma_j \cdot b_j$ where b_i is given by theorem 3.1. I wish the clipping probability to be less than n^{-1} , which implies $t = \sqrt{\ln(n)}$. Combining these results I get that

$$\begin{aligned} C^2 &\leq \sqrt{8 \ln(n) \sum_{j \in [d]} (\sigma_j^4) + \sum_{j \in [d]} \sigma_j^2} \\ &= \sqrt{8 \ln(n) \sum_{j \in [d]} \left(\frac{\sigma_j}{\sum_{i \in [d]} \sigma_i} \right)^2} + \sum_{j \in [d]} \frac{\sigma_j}{\sum_{i \in [d]} \sigma_i} \\ &= \sqrt{8 \ln(n) \sum_{j \in [d]} \sigma_j^2 \cdot \frac{1}{\sum_{i \in [d]} \sigma_i}} + 1 \end{aligned}$$

Inserting this back into equation 17 we conclude

$$\begin{aligned} \mathbb{E} [\|\boldsymbol{\eta}\|^2] &= 4\sigma_{opt}^2 \cdot C^2 \cdot \sum_{i=1}^d b_i^{-2} \\ &\leq 4\sigma_{opt}^2 \cdot \left(\sqrt{8 \ln(n) \sum_{j \in [d]} \sigma_j^2} \cdot \frac{1}{\sum_{i \in [d]} \sigma_i} + 1 \right) \cdot \sum_{i=1}^d \left(\sigma_i \cdot \sum_{j \in [d]} \sigma_j \right) \\ &\leq 4\sigma_{opt}^2 \cdot \left(\sqrt{8 \ln(n) \sum_{j \in [d]} \sigma_j^2} \cdot \frac{1}{\sum_{i \in [d]} \sigma_i} + 1 \right) \cdot \left(\sum_{i=1}^d \sigma_i \right)^2 \\ &= 4\sigma_{opt}^2 \cdot \sum_{i=1}^d \sigma_i \cdot \left(\sqrt{8 \ln(n) \sum_{j \in [d]} \sigma_j^2} + \sum_{i=1}^d \sigma_i \right) \end{aligned}$$

And the constraint on $t = \sqrt{\ln(n)}$ from lemma 3.2 is

$$\begin{aligned}
0 \leq t &\leq \frac{1}{6\sigma_*^2} \sqrt{2 \sum_{j \in [d]} \sigma_j^4} \iff \\
\sqrt{\ln(n)} &\leq \sqrt{\left(\frac{\sum_{j \in [d]} \sigma_j}{\sigma_*} \right)^2 \cdot \frac{2}{36} \cdot \sum_{j \in [d]} \left(\frac{\sigma_j}{\sum_{i \in [d]} \sigma_i} \right)^2} \iff \\
\ln(n) &\leq \frac{1}{18\sigma_*^2} \cdot \sum_{j \in [d]} \sigma_j^2 \iff \\
\ln(n) &\leq \frac{\sum_{j \in [d]} \sigma_j^2}{18\sigma_*^2}
\end{aligned}$$

At this point it would be preferable to compare the error of algorithm 3, to that of doing no transformation, and similarly deciding on a bound C such that the clipping probability was less than n^{-1} . However as I can only give upper bounds on these errors there is little to no value in comparing upper bounds as there is no guarantee of the bounds. So even if I were to determine an upper bound for the error without transformation which was lower than that of algorithm 3, it could very well still be that the error was less in reality. Secondly to produce the upper bound in as in 6 using Bernstein's inequality, I have the constraint that all $\sigma_j \leq \frac{2}{3}$, which is unlikely to happen in reality, whereas $\frac{\sigma_j}{\sum_{j \in [d]} \sigma_j} \leq \frac{4}{9}$ is much more likely to be satisfied for all j . As numerical solutions exist for evaluating $\Pr[\|x_i\| > C^2]$ it is evident (oplagt) that empirically evaluating these errors and comparing them, would yield whether doing the transformation is beneficial. Intuitively it makes sense that allowing different noise to be added for each dimension, can only produce better results. ■

Compare error with upper bound to the one which we be if no transformation was done. i.e clipping in the original space. Impossible due to the constraint on σ_*

Write about alternative ways of giving upper bound on C , and why it is not so important

Write alternative ways of defining the problem. CLip prob less than n^{-1} and such

3.1 Gaussian data

Let $X^{(j)} \sim \mathcal{N}(0, \sigma_j^2)$ As the expected l_2 -norm of x_i is given by

$$\mathbb{E} [\|x_i\|^2] = \sum_{j=1}^d \sigma_j^2$$

To achieve an expected norm of 1 I will scale each dimension by a factor $\frac{1}{b_j}$ which achieves this. If

$$\hat{x}_i = \left(\frac{x_i^{(0)}}{b_0}, \frac{x_i^{(1)}}{b_1}, \dots, \frac{x_i^{(d)}}{b_d} \right)$$

This means that $X^{(j)} \sim \mathcal{N}(0, \frac{\sigma_j^2}{b_j^2})$ and the expected norm is given by

$$\mathbb{E} [\|x_i\|^2] = \sum_{j=1}^d \frac{\sigma_j^2}{b_j^2}$$

and I can introduce that constraint that the expected norm after the transformation should be 1. In such a case when noise is added after the transformation $\hat{X} + \eta$ where $\eta \sim \mathcal{N}(0, t^2)$ and achieves (ε, δ) -DP in this space then then due to linearity of transformation the noise introduced in the original space is then given by then the error is

I desire a transformation of $x_i^{(j)}$ such that the expected norm is 1. Thus I must scale each dimension by $\frac{1}{b_j}$, and have that

$$\mathbb{E} [\|x_i\|^2] = 1$$

Minimize $\|\hat{\eta}\|$ under the constraint that $\mathbb{E} [\|x_i\|^2] = 1$

Lemma 3.3. *Let $X \sim \mathcal{N}(0, \sigma^2)$, and Φ denote the cumulative density function of $\mathcal{N}(0, 1)$, then the cumulative density function of X^2 is given by*

$$F_{X^2}(x) = \Pr [X^2 \leq x] = 2\Phi \left(\frac{\sqrt{x}}{\sigma} \right) - 1$$

Proof

$$\begin{aligned} \Pr [X^2 \leq x] &= \Pr [|X| \leq \sqrt{x}] = 2 \Pr [0 \leq X \leq \sqrt{x}] \\ &= 2 (\Pr [X \leq \sqrt{x}] - \Pr [X \leq 0]) = 2 \left(\Pr [X \leq \sqrt{x}] - \frac{1}{2} \right) \\ &= 2\Phi \left(\frac{\sqrt{x}}{\sigma} \right) - 1 \end{aligned}$$

■

Corollary 3.1. *From lemma 3.3 we can give following bound for $X \sim \mathcal{N}(0, \sigma^2)$.*

$$\Pr [X^2 > (4\sigma)^2] < 10^{-4}$$

Bernsteins inequality then states that

$$\Pr \left[\sum_{j \in [d]} X_j \geq 2t \sqrt{\sum_{j \in [d]} \mathbb{E}[X_j^2]} \right] < e^{-t^2}$$

for all

$$0 \leq t \leq \frac{1}{2L} \sqrt{\sum}$$

I have that \bar{X} is a standard gaussian variable.
Let $Y_j = X_j^2$ I am interested in $\sum_{j \in [d]} Y_j$, then it has

$$\begin{aligned} \mathbb{E}[|Y_j^k|] &= \mathbb{E}[|X_j^{2k}|] = \mathbb{E}[|X_j|^{2k}] = \mathbb{E}[|\sigma_j \bar{X}|^{2k}] = \\ &|\sigma_j|^{2k} \mathbb{E}[|\bar{X}|^{2k}] = \sigma_j^{2k} \prod_{c=1}^k 2c = \sigma_j^{2k} \cdot 2^k \cdot k! \end{aligned}$$

Which means $\mathbb{E}[Y_j^2] = 8\sigma_j^4$. To check the constraint (find value for L)

$$\begin{aligned} \sigma_j^{2k} \cdot 2^k \cdot k! &\leq \frac{1}{2} \cdot 8\sigma_j^4 \cdot L^{k-2} k! \iff \\ \sigma_j^{2k} \cdot 2^k &\leq 4\sigma_j^4 \cdot L^{k-2} \iff \\ \sigma_j^{2(k-2)} \cdot 2^k &\leq 4L^{k-2} \end{aligned}$$

We have that $\sigma_j = \sigma_j b_j = \frac{\sqrt{\sigma_j}}{\sqrt{\sum_{i \in [d]} \sigma_i}} = \sqrt{\frac{\sigma_j}{\sum_{i \in [d]} \sigma_i}}$. Therefore

$$\begin{aligned} \sigma_j^{2(k-2)} \cdot 2^k &= \left(\sqrt{\frac{\sigma_j}{\sum_{i \in [d]} \sigma_i}} \right)^{2(k-2)} \cdot 2^k = \\ &\left(\frac{\sigma_j}{\sum_{i \in [d]} \sigma_i} \right)^{k-2} \cdot 2^k \leq 2^k \end{aligned}$$

This therefore implies

$$2^k \leq 4L^{k-2} \iff 2^{k-2} \leq L^{k-2} \iff 2 \leq L$$

If it is desired that less than 10^{-p} points are removed then $t = \sqrt{p \ln(10)}$ as long as

$$\sqrt{p \ln(10)} \leq \frac{1}{2L} \sqrt{\sum_{j \in [d]} \mathbb{E}[X_j^2]} = \sqrt{4} \cdot \frac{\sqrt{\sum_{i \in [d]} \sigma_i^2}}{\sum_{i \in [d]} \sigma_i}$$

An alternative Again minimize $\|\hat{\eta}\|$, but instead the constraint comes from Chebyshev's inequality, I always have that

$$\Pr \left[\left| \|x_i\|^2 - \mathbb{E}[\|x_i\|^2] \right| \geq k \cdot \sqrt{\text{Var}[\|x_i\|^2]} \right] \leq \frac{1}{k^2}$$

Therefore I can set $\frac{1}{k^2} = 0.05$, and find a transformation where I decide how many standard deviations I must be away from the mean to have norm greater than 1, i.e.

$$\mathbb{E} [\|x_i\|^2] + k \cdot \sqrt{\text{Var} [\|x_i\|^2]} = 1 \implies k = \frac{1 - \mathbb{E} [\|x_i\|^2]}{\sqrt{\text{Var} [\|x_i\|^2]}}$$

I then have that

$$\frac{1}{k^2} = \frac{1}{\left(\frac{1 - \mathbb{E} [\|x_i\|^2]}{\sqrt{\text{Var} [\|x_i\|^2]}} \right)^2} = \left(\frac{\sqrt{\text{Var} [\|x_i\|^2]}}{1 - \mathbb{E} [\|x_i\|^2]} \right)^2 = \frac{\text{Var} [\|x_i\|^2]}{(1 - \mathbb{E} [\|x_i\|^2])^2}$$

I already know that

$$\begin{aligned} \mathbb{E} [\|x_i\|^2] &= \sum_{j=1}^d \frac{\sigma_j^2}{b_j^2} \\ \text{Var} [\|x_i\|^2] &= 2 \sum_{j=1}^d \frac{\sigma_j^4}{b_j^4} \end{aligned}$$

I therefore have the constraint

$$\frac{\text{Var} [\|x_i\|^2]}{(1 - \mathbb{E} [\|x_i\|^2])^2} = \frac{2 \sum_{j=1}^d \frac{\sigma_j^4}{b_j^4}}{\left(1 - \sum_{j=1}^d \frac{\sigma_j^2}{b_j^2} \right)^2} = 0.05$$

Be aware that this could find cases where expected norm is greater than 1 and the constraint then says that they are never less than 1 in norm. Another restriction could be that expected norm is ≤ 1 .

Chebyshev's inequality could also be used when expected norm should be 1 and then put a bound on number of std away one must be to have less than 0.05 fraction of data removed.

Extra

Determining α using Bernsteins inequality

$$\Pr [\|x_i\|^2 - \mathbb{E} [\|x_i\|^2] > t] < 2 \cdot \exp \left(-\frac{t^2/2}{\text{Var} [\|x_i\|^2] + C \cdot t/3} \right)$$

In the case where $\mathbb{E} [\|x_i\|^2] = 1$ and b_i is optimized in this case, we have that

$$P(\|x_i\|^2 > 1 + t) < 2 \exp \left(-\frac{t^2/2}{\text{Var} [\|x_i\|^2] + C \cdot t/3} \right)$$

$$\text{Var} [\|x_i\|^2] = 2 \frac{\sum_i^d \sigma_i^2}{\left(\sum_i^d \sigma_i\right)^2}$$

$$C = \max_{i \in [d]} (\sigma_i) \cdot \frac{16}{\sum_i^d \sigma_i}$$

C is decided such that less than 0.0001 fraction of the data is outside this bound in each dimension. i.e.

$$\forall j \in [d] : \Pr [X^{(j)} > C] < 0.0001$$

Solving for t

$$2 \cdot \exp \left(-\frac{t^2/2}{\text{Var} [\|x_i\|^2] + C \cdot t/3} \right) = 0.0001 \iff \ln\left(\frac{0.0001}{2}\right) = -\frac{t^2/2}{\text{Var} [\|x_i\|^2] + C \cdot t/3} \implies$$
$$t = -\frac{C \ln(\frac{0.0001}{2}) - \sqrt{\ln(\frac{0.0001}{2})(-18 \cdot \text{Var} [\|x_i\|^2] + C^2 \ln(\frac{0.0001}{2}))}}{3}$$

Then we have that $\alpha = 1 + t$.

Determine α for the bound using https://en.wikipedia.org/wiki/Concentration_inequality, or <https://web.stanford.edu/class/cs229t/2017/Lectures/concentration-slides.pdf>

References

- [1] BALLE, B., AND WANG, Y. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. *CoRR abs/1805.06530* (2018).
- [2] BATES, D. *Quadratic Forms of Random Variables*. University of Wisconsin-Madison: STAT 849 lectures, 2010.
- [3] BISWAS, S., DONG, Y., KAMATH, G., AND ULLMAN, J. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems 33* (2020), 14475–14485.
- [4] CHEN, T., AND LUMLEY, T. Numerical evaluation of methods approximating the distribution of a large quadratic form in normal variables. *Computational Statistics & Data Analysis 139* (2019), 75–81.

- [5] DAS, A., AND GEISLER, W. S. A method to integrate and classify normal distributions, 2020.
- [6] 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.
- [7] 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.
- [8] HALLAM, A. *Some Theorems on Quadratic Forms and Normal Variables*. Iowa State University: Econ 671 lectures, 2004.
- [9] HANSON, D. L., AND WRIGHT, F. T. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics* 42, 3 (1971), 1079–1083.
- [10] HUANG, Z., LIANG, Y., AND YI, K. Instance-optimal mean estimation under differential privacy.
- [11] MATHAI, A., AND PROVOST, S. Quadratic forms in random variables.
- [12] PAGH, R., AND LEBEDA, C. Private vector aggregation when coordinates have different sensitivity.