Open problem - Better privacy guarantees for larger groups

Achraf Azize

July 2023

1 Defintion and Problem

Definition 1 (Group-wise zero-concentrated differential privacy). *Assume possible datasets consist of records from domain U, and U can be partitioned into k fixed, disjoint groups $U_1, \ldots, U_k$. Let $v, \xi : \mathcal{D} \to \mathbb{R}^k$ be two functions associating a dataset to a vector of privacy budgets (one per group). We say a mechanism $\mathcal{M}$ satisfies v, ξ-group-wise zero-concentrated differential privacy (zCDP) if, for any two datasets D, D' differing in the addition or removal of a record in U_i, and for all $\alpha > 1$, we have:*

$$D_\alpha\left(\mathcal{M}(D) \| \mathcal{M}(D')\right) \leq \alpha \cdot v(D)_i + \xi(D)_i$$
$$D_\alpha\left(\mathcal{M}(D') \| \mathcal{M}(D)\right) \leq \alpha \cdot v(D)_i + \xi(D)_i$$

where D_α is the Rényi divergence of order α.

Problem: Let $r \in (0, 1]$ be an acceptable level of relative error, and k be the number of distinct, mutually-exclusive partitions of domain X. Given a dataset D, let $x(D)$ be a vector containing the count of records in each partition. The objective is to find a mechanism $\mathcal{M}$ which takes in r, k and D, and outputs $\hat{x}(D)$ such that $\mathbb{E}\left[|x(D)_i - \hat{x}(D)_i|\right] < r \cdot x(D)_i$ for all i, and satisfies v-group-wise zCDP where $v(D)_i$ is as small as possible for all i. The privacy guarantee $v(D)_i$ should only depend on $x(D)_i$, and should be non-increasing with $x(D)_i$.

2 An Example Algorithm

Algorithm 1. Adding data-dependent noise as a post-processing step.
Require: A dataset D where each data point belongs to one of k groups, a privacy parameter ρ, and a relative error rate r.
1. Let $\sigma^2 = 1/(2\rho)$
2. For $i = 1$ to k do:
3. Let x_i be the number of people in D in group i
4. Sample $X_i \sim \mathcal{N}\left(x_i, \sigma^2\right)$

5. Sample $Y_i \sim \mathcal{N}\left(X_i, (rX_i)^2\right)$
6. end for
7. return $Y_1, \ldots, Y_k$

2.1 Accuracy Analysis:

Computing the expectation, we have

$$\begin{aligned}\mathbb{E}[Y_i] &= \mathbb{E}[\mathbb{E}[Y_i \mid X_i]] \\ &= \mathbb{E}[X_i] = x_i\end{aligned}$$

And for the variance, we get

$$\begin{aligned}\mathbb{V}[Y_i] &= \mathbb{E}[\mathbb{V}[Y_i \mid X_i]] + \mathbb{V}[\mathbb{E}[Y_i \mid X_i]] \\ &= \mathbb{E}[r^2 X_i^2] + \mathbb{V}[X_i] \\ &= r^2(x_i^2 + \sigma^2) + \sigma^2 = r^2 x_i^2 + \sigma^2(1 + r^2)\end{aligned}$$

Combining both gives that

$$\begin{aligned}\mathbb{E}\left[|Y_i - x_i|\right] &\leq \sqrt{\mathbb{V}[Y_i]} \\ &= \sqrt{r^2 x_i^2 + \sigma^2(1 + r^2)} \\ &\leq r x_i + \sigma\sqrt{1 + r^2}\end{aligned}$$

2.2 Privacy Analysis:

Attempt 1: We can rewrite $Y_i = X_i Z$ as the product of two i.i.d Gaussian random variables, where $X_i \sim \mathcal{N}(x_i, \sigma^2)$ and $Z \sim \mathcal{N}(1, r^2)$.

Using Theorem 2.1 from [CYIK16], we get the exact PDF distribution of Y_i given by:

$$\begin{aligned}f_{Y_i}(y) &= \exp\left(-\frac{x_i^2}{2\sigma^2} - \frac{1}{2r^2}\right) \times \sum_{n=0}^{\infty} \sum_{m=0}^{2n} \begin{pmatrix} 2n \\ m \end{pmatrix} \frac{x_i^m y^{2n-m} |y|^{m-n}}{\pi (2n)! \sigma^{n+m+1} r^{3n-m+1}} K_{m-n}\left(\frac{|y|}{\sigma r}\right) \\ &\triangleq \exp\left(-\frac{x_i^2}{2\sigma^2} - \frac{1}{2r^2}\right) \times h(x_i, \sigma, r, y)\end{aligned}$$

where K_ν denotes the modified Bessel function of the second kind and order ν.

For a neighbouring dataset, $Y_i' = X_i' Z$ where $X_i' \sim \mathcal{N}(x_i \pm 1, \sigma^2)$ and $Z \sim \mathcal{N}(1, r^2)$.

The PDF distribution of Y_i' is then given by:

$$f_{Y_i'}(y) = \exp\left(-\frac{(x_i \pm 1)^2}{2\sigma^2} - \frac{1}{2r^2}\right) \times h(x_i \pm 1, \sigma, r, y)$$

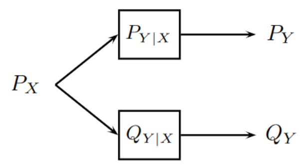
Finally, we get that

$$D_\alpha\left(Y_i \| Y_i'\right) = \frac{1}{\alpha-1}\log\left(\int_{-\infty}^{\infty}\left(\frac{f_{Y_i}(y)}{f_{Y_i'}(y)}\right)^\alpha f_{Y_i'}(y)\,\mathrm{d}y\right)$$
$$= \frac{\alpha}{\alpha-1}\frac{1\pm 2x_i}{2\sigma^2} + \frac{1}{\alpha-1}\log\left(\int_{-\infty}^{\infty}\left(\frac{h(x_i,\sigma,r,y)}{h(x_i\pm 1,\sigma,r,y)}\right)^\alpha f_{Y_i'}(y)\,\mathrm{d}y\right)$$

TO DO: Compute an upper bound of the second term.
Attempt 2: We have $X_i \sim \mathcal{N}(x_i, \sigma^2)$ and $Y_i \mid X_i \sim \mathcal{N}(X_i, r^2 X_i^2)$.

On the other hand, for a neighbouring dataset, $X_i' \sim \mathcal{N}(x_i \pm 1, \sigma^2)$ and $Y_i' \mid X_i' \sim \mathcal{N}(X_i', r^2 X_i'^2)$.

We recall the following theorem:



If $P_X \overset{P_{Y|X}}{\longrightarrow} P_Y$ and $P_X \overset{Q_{Y|X}}{\longrightarrow} Q_Y$, then

$$D_f\left(P_Y \| Q_Y\right) \le \mathbb{E}_{X\sim P_X}\left[D_f\left(P_{Y|X} \| Q_{Y|X}\right)\right].$$

where D_f is an f-divergence.

The idea is to use the theorem, where the input is the iid Gaussian pair $Z_i \triangleq (X_i, X_i')$, the first channel is $P_{Y|Z_i} = \mathcal{N}(X_i, r^2X_i^2)$, the second channel is $Q_{Y|Z_i} = \mathcal{N}(X_i', r^2 X_i'^2)$. The marginals are then resp Y_i and Y_i'. Applying the Theorem gives that

$$D_f(Y_i \| Y_i') \le \mathbb{E}_{(X_i,X_i')}[D_f(\mathcal{N}(X_i, r^2X_i^2) \| \mathcal{N}(X_i', r^2X_i'^2))].$$

Unfortunately, the Renyi divergence is not directly an f-divergence (maybe a link could be found to apply a version of this result).

For now, let us look at just the KL ($\alpha = 1$), we get that

$$D_1(Y_i \| Y_i') \le \mathbb{E}_{(X_i,X_i')}\left[D_1(\mathcal{N}(X_i, r^2X_i^2) \| \mathcal{N}(X_i', r^2X_i'^2))\right]$$
$$= \mathbb{E}_{(X_i,X_i')}\left[2\log\left(\frac{X_i'}{X_i}\right) + \frac{r^2X_i^2 + (X_i - X_i')^2}{2r^2X_i'^2} - \frac{1}{2}\right]$$
$$\le 2\log\left(\mathbb{E}\left[\frac{X_i'}{X_i}\right]\right) + \frac{1}{2}\left(1+\frac{1}{r^2}\right)\mathbb{E}\left[\frac{X_i^2}{X_i'^2}\right] - \frac{1}{r^2}\mathbb{E}\left[\frac{X_i}{X_i'}\right] + \frac{1}{2}\left(\frac{1}{r^2}-1\right)$$

which reduces to computing the expectation of the quotient of two iid Gaussian variables.

Unfortunately, these expectations do not exist as Ordinary integrals but only in a Principal Value sense (linguisticturn comment).

Plugging the formulas of the expectations in the Principal Value sense gives

$$\mathbb{E}\left[\frac{X_i'}{X_i}\right] = \mathbb{E}\left[X_i'\right]\mathbb{E}\left[\frac{1}{X_i}\right] = (x_i \pm 1)\frac{\sqrt{2}}{\sigma}F\left(\frac{x_i}{\sqrt{2}\sigma}\right)$$

$$\mathbb{E}\left[\frac{X_i}{X_i'}\right] = \mathbb{E}\left[X_i\right]\mathbb{E}\left[\frac{1}{X_i'}\right] = (x_i)\frac{\sqrt{2}}{\sigma}F\left(\frac{x_i \pm 1}{\sqrt{2}\sigma}\right)$$

$$\mathbb{E}\left[\frac{X_i^2}{X_i'^2}\right] = \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[\frac{1}{X_i'^2}\right] = \left(\sigma^2 + x_i^2\right)\frac{1}{2\sigma^2}\left(\frac{\sqrt{2}\,(x_i \pm 1)}{\sigma}F\left(\frac{x_i \pm 1}{\sqrt{2}\sigma}\right) - 1\right)$$

where F is the Dawson function $F(x) = e^{-x^2}\int_0^x e^{t^2}$.

Plugging everything in the upper bound gives an upper bound on the KL.

TO DO: Generalise to α Renyi divergence, and see the dependence on x_i (is the upper bound on the KL decreasing in x_i).

3 Another (simpler) algorithm

Algorithm 2. Adding noise directly using $x(D)_i$.
Require: A dataset D where each data point belongs to one of k groups, a privacy parameter ρ, and a relative error rate r.
1. Let $\sigma^2 = 1/(2\rho)$
2. For $i = 1$ to k do:
3. Let x_i be the number of people in D in group i
4. Sample $X_i \sim \mathcal{N}\left(x_i, (rx_i)^2\right)$
6. end for
7. return $X_1, \ldots, X_k$

3.1 Accuracy Analysis

We have directly that

$$\begin{aligned}\mathbb{E}\left[|Y_i - x_i|\right] &\leq \sqrt{\mathbb{V}[Y_i]}\\ &= \sqrt{r^2 x_i^2} = rx_i\end{aligned}$$

3.2 Privacy Analysis

Let $X_i \sim \mathcal{N}\left(x_i, (rx_i)^2\right)$ and for a neighbouring dataset $X_i' \sim \mathcal{N}\left(x_i + 1, r^2\,(x_i + 1)^2\right)$, using the formula for renyi divergence between Gaussian random variables, we get that

$$
\begin{aligned}
D_\alpha(X_i, X_i') &= 2\log\left(\frac{x_i+1}{x_i}\right) + \frac{1}{2\,(\alpha-1)}\log\left(\frac{x_i^2+2x_i+1}{x_i^2+\alpha(2x_i+1)}\right) + \frac{1}{2}\frac{\alpha}{r^2x_i^2+\alpha r^2(2x_i+1)} \\
&\le \frac{1}{2}\frac{\alpha}{r^2x_i^2} + 2\log\left(1+\frac{1}{x_i}\right)
\end{aligned}
$$

using that $\alpha > 1$ and $r, x_i > 0$

This means that Algorithm 2 verifies v, ξ-group-wise approximate zCDP, where

$$
v(D)_i = \frac{1}{2r^2x(D)_i^2}
$$

and

$$
\xi(D)_i = 2\log\left(1+\frac{1}{x(D)_i}\right)
$$

.

Indeed the privacy budgets are non-increasing functions of $x(D)_i$.

Comment. In the original post, one reads "Of course, directly using $x(D)_i$ to determine the scale of the noise for group i leads to a privacy loss which is data dependent, similarly to e.g. PATE [PAEGT17], and as such should be treated as a protected value."

However, any attempt that tries to first estimate $x(D)_i$ and then use the estimated (noisy) counts to add a variance (like in Algorithm 1) will too have a privacy loss that depends on $x(D)_i$ eventually. Thus the comment is not very clear to me.

References

[CYIK16] Guolong Cui, Xianxiang Yu, Salvatore Iommelli, and Lingjiang Kong. Exact distribution for the product of two correlated gaussian random variables. *IEEE Signal Processing Letters*, 23(11):1662–1666, 2016.