DP Unlearning

daniel khen

November 2024

1 Introduction

2 Definitions

2.1 Differential privacy

Let $\varepsilon, \delta \in \mathbb{R}_+$ and let $\mathcal{A}$ be a randomized algorithm that takes a dataset as input. The algorithm $\mathcal{A}$ is said to provide (ε, δ)-differential privacy if, for all datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ on a single element and all subsets $\mathcal{R} \in \operatorname{Im} \mathcal{A}$:

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{R}] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{R}] + \delta$$

2.2 Group privacy

Let $\varepsilon, \delta \in \mathbb{R}_+$ and $n \in \mathbb{N}$, let $\mathcal{A}$ be a randomized algorithm that takes a dataset as input. We say algorithm $\mathcal{A}$ provides (ε, δ, n)-group privacy if, for all datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ on n elements and all subsets $\mathcal{R} \in \operatorname{Im} \mathcal{A}$:

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{R}] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{R}] + \delta$$

2.2.1 Differential privacy versus group privacy

Group privacy will be more helpful with unlearning as unlearning is mostly performed on large forget sets rather then a single datapoint. Thus it will be helpful find a relationship between differential privacy and group privacy. A good intuition is to fine this relation is a series of datasets $\mathcal{D}_n$ such that for every $n \in \mathbb{N}$, $\mathcal{D}_n$ and $\mathcal{D}_{n+1}$ differ by one element. Assuming $\mathcal{A}$ provides (ε, δ)-differential we get a Recurrence relation for $a_n = \Pr[\mathcal{A}(\mathcal{D}_n) \in R]$.

2.2.2 (ε, δ)-differential privacy $\rightarrow (n\varepsilon, \delta\frac{e^{n\varepsilon}-1}{e^{\varepsilon}-1}, n)$-group privacy

Proof is by induction, base case where $n = 1$, $(\varepsilon, \delta, 1)$-group privacy is the same as (ε, δ)-differential privacy. For the n-th case assuming $\mathcal{A}$ provides (ε, δ)-differential privacy and $(n\varepsilon, \delta\frac{e^{n\varepsilon}-1}{e^{\varepsilon}-1}, n)$-group privacy. lets take three databases

$\mathcal{D}, \mathcal{D}', \mathcal{D}''$ such that $\mathcal{D}, \mathcal{D}'$ differ by n elements and $\mathcal{D}', \mathcal{D}''$ differ by one element (and $\mathcal{D}, \mathcal{D}''$ differ by $n+1$ elements), then by group privacy:

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{R}] \leq e^{n\varepsilon}\Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{R}] + \delta\frac{e^{n\varepsilon}-1}{e^{\varepsilon}-1}$$

And by differential privacy:

$$\Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{R}] \leq e^{\varepsilon}\Pr[\mathcal{A}(\mathcal{D}'') \in \mathcal{R}] + \delta$$

Which combines to:

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{R}] \leq e^{n\varepsilon}(e^{\varepsilon}\Pr[\mathcal{A}(\mathcal{D}'') \in \mathcal{R}] + \delta) + \delta\frac{e^{n\varepsilon}-1}{e^{\varepsilon}-1}$$

$$= e^{(n+1)\varepsilon}\Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{R}] + \delta\frac{e^{(n+1)\varepsilon}-1}{e^{\varepsilon}-1}$$

Thus $\mathcal{D}$ provides $((n+1)\varepsilon, \delta\frac{e^{(n+1)\varepsilon}-1}{e^{\varepsilon}-1}, n+1)$-group privacy.

More interestingly, to achieve (ε, δ, n)-group privacy we want $\mathcal{A}$ to provide $(\frac{\varepsilon}{n}, \delta\frac{e^{\varepsilon}-1}{e^{n\varepsilon}-1})$-differential privacy.

2.3 Unlearning

Let $\varepsilon, \delta \in \mathbb{R}_+$, let $\mathcal{A}$ be a randomized learning algorithm that takes a dataset as input and let $\mathcal{U}$ be a randomized unlearning algorithm that takes a model, a dataset and a forget set as input. The algorithm $\mathcal{U}$ is said to provide (ε, δ)-unlearning with respect to the learning algorithm $\mathcal{A}$, the dataset $\mathcal{D}$, and the forget set $\mathcal{S} \subseteq \mathcal{D}$ if, for all $\mathcal{R} \in \operatorname{Im} \mathcal{A}$:

$$\Pr[\mathcal{A}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] \leq e^{\varepsilon}\Pr[\mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] + \delta$$

$$\Pr[\mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] \leq e^{\varepsilon}\Pr[\mathcal{A}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] + \delta$$

2.3.1 Unlearning for fine-tuning

Now let $\mathcal{A}(\mathcal{M}, \mathcal{D})$ be a randomized fine-tuning algorithm that takes both a model and dataset as inputs (i.e. it starts training on a predefined model and weights), we will mark $\mathcal{A}_{\mathcal{M}}(\mathcal{D}) = \mathcal{A}(\mathcal{M}, \mathcal{D})$ a randomized learning algorithm and let U be a randomized unlearning algorithm that takes a model. The algorithm $\mathcal{U}$ is said to provide (ε, δ)-unlearning with respect to $(\mathcal{A}, \mathcal{D}, \mathcal{S})$ if for every model $\mathcal{M}$, $\mathcal{U}$ provides (ε, δ)-unlearning with respect to $(\mathcal{A}_{\mathcal{M}}, \mathcal{D}, \mathcal{S})$.

2.3.2 Unlearning using differntial privacy

Let $\mathcal{B}$ be a randomized learning algorithm that provides (ε, δ, n)-group privacy and let $\mathcal{U}$ be a randomized unlearning algorithm that provides (ε', δ')-unlearning with respect to $(\mathcal{A}, \mathcal{D}, \mathcal{S})$, where $\mathcal{B}$ is a randomized fine-tuning algorithm and $\mathcal{S} \subseteq \mathcal{D}$ is a forget set such that $|\mathcal{S}| \leq n$. Then $\mathcal{U}$ provides $(\varepsilon + \varepsilon', \min(e^{\varepsilon}\delta' + \delta, e^{\varepsilon'}\delta + \delta'))$-unlearning with respect to $(\mathcal{C}, \mathcal{D}, \mathcal{S})$, where $\mathcal{C}$ is a randomized learning algorithm given by $\mathcal{C}(\mathcal{D}) = \mathcal{A}(\mathcal{B}(\mathcal{D}), \mathcal{D})$.

$$\Pr[\mathcal{C}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] = \Pr[\mathcal{A}(\mathcal{B}(\mathcal{D} \setminus \mathcal{S}), \mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}]$$

Lets mark $\mathcal{R}' = \{\mathcal{M} | \mathcal{A}(\mathcal{M}, \mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}\}$, then because of group privacy:

$$\Pr[\mathcal{C}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] = \Pr[\mathcal{B}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}'] \leq e^{\varepsilon}\Pr[\mathcal{B}(\mathcal{D}) \in \mathcal{R}'] + \delta$$
$$\Pr[\mathcal{B}(\mathcal{D}) \in \mathcal{R}'] \leq e^{\varepsilon}\Pr[\mathcal{C}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] + \delta$$

Also:

$$\Pr[\mathcal{B}(\mathcal{D}) \in \mathcal{R}'] = \Pr[\mathcal{A}(\mathcal{B}(\mathcal{D}), (\mathcal{D} \setminus \mathcal{S})) \in \mathcal{R}] = \Pr[\mathcal{A}_{\mathcal{B}(\mathcal{D})}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}]$$

And since for every model $\mathcal{M}$, $\mathcal{U}$ provides (ε, δ)-unlearning with respect to $(\mathcal{A}_{\mathcal{M}}, \mathcal{D}, \mathcal{S})$:

$$\Pr[\mathcal{B}(\mathcal{D}) \in \mathcal{R}'] = \Pr[\mathcal{A}_{\mathcal{B}(\mathcal{D})}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] \leq e^{\varepsilon'}\Pr[\mathcal{U}(\mathcal{A}_{\mathcal{B}(\mathcal{D})}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] + \delta'$$

$$\Pr[\mathcal{U}(\mathcal{C}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] = \Pr[\mathcal{U}(\mathcal{A}_{\mathcal{B}(\mathcal{D})}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] \leq e^{\varepsilon'}\Pr[\mathcal{B}(\mathcal{D}) \in \mathcal{R}'] + \delta'$$

Finally we get:

$$\Pr[\mathcal{C}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] \leq e^{\varepsilon+\varepsilon'}\Pr[\mathcal{U}(\mathcal{C}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] + e^{\varepsilon}\delta' + \delta$$
$$\Pr[\mathcal{U}(\mathcal{C}(\mathcal{D}), \mathcal{D}, \mathcal{S}) \in \mathcal{R}] \leq e^{\varepsilon+\varepsilon'}\Pr[\mathcal{C}(\mathcal{D} \setminus \mathcal{S}) \in \mathcal{R}] + e^{\varepsilon'}\delta + \delta'$$