Provable Privacy with Non-Private Pre-Processing

Yaxi Hu [1] **Amartya Sanyal** [1] **Bernhard Schölkopf** [1]

Abstract

When analyzing Differentially Private (DP) machine learning pipelines, the potential privacy cost of data-dependent pre-processing is frequently overlooked in privacy accounting. In this work, we propose a general framework to evaluate the additional privacy cost incurred by non-private data-dependent pre-processing algorithms. Our framework establishes upper bounds on the overall privacy guarantees by utilising two new technical notions: a variant of DP termed Smooth DP and the bounded sensitivity of the pre-processing algorithms. In addition to the generic framework, we provide explicit overall privacy guarantees for multiple data-dependent pre-processing algorithms, such as data imputation, quantization, deduplication, standard scaling and PCA, when used in combination with several DP algorithms. Notably, this framework is also simple to implement, allowing direct integration into existing DP pipelines.

1. Introduction

With the growing emphasis on user data privacy, Differential Privacy (DP), has become the preferred solution for safeguarding training data in Machine Learning (ML) and data analysis (Dwork et al., 2006). DP algorithms are designed to ensure that individual inputs minimally affect the algorithm's output, thus preserving privacy. This approach is now widely adopted by various organizations for conducting analyses while maintaining user privacy (AppleDP, 2017; U.S. Census Bureau, 2020).

Pre-processing data is a standard practice in data analysis and machine learning. Techniques such as data imputation for handling missing values (Anil Jadhav & Ramanathan, 2019), deduplication for reducing memorization and eliminating bias (Kandpal et al., 2022; Lee et al., 2022), standard scaling for reducing the impact of outliers, and dimensionality reduction for denoising or visualization (Abadi et al., 2016; Zhou et al., 2021; Pinto et al., 2024) are commonly used. These pre-processings are also prevalent prior to applying DP algorithms, a pipeline that we term *pre-processed DP pipeline*. Among other uses, this has been shown to improve privacy-accuracy trade-off (Tramer & Boneh, 2021; Ganesh et al., 2023).

A fundamental assumption, required by DP, is that individual data points are independent. However, if training data is used in pre-processing, this assumption is compromised. For example, when deduplicating a dataset, whether a point remains after the pre-processing is dependent on the presence of other points in its vicinity. Similarly, for mean imputation, the imputed value depends on the values of other data points. These dependencies, also evident in PCA and pre-training, can undermine the privacy guarantee of the pre-processed DP pipeline.

There are multiple strategies to address this. A straightforward method to derive privacy guarantees for this pipeline is to use group privacy where the size of the group can be as large as the size of the dataset, thereby resulting in weak privacy guarantees. This idea, albeit not in the context of pre-processing, was previous explored under the name Dependent Differential Privacy (DDP) (Zhao et al., 2017; Liu et al., 2016). Another approach is to use public data for the pre-processing. Broadly referred to as semi-private learning algorithms, examples of these methods include pre-training on public data and learning projection functions using the public data e.g. (Pinto et al., 2024; Li et al., 2022b;a; Yu et al., 2021). Despite their success, these methods crucially rely on the availability of high-quality public data.

In the absence of public data, an alternative approach is to privatise the pre-processing algorithm. However, designing new private pre-processing algorithms complicate the process, increasing the risk of privacy breaches due to errors in implementation or analysis. Moreover, private pre-processing can be statistically or computationally more demanding than private learning itself. For example, the sample complexity of DP-PCA(Chaudhuri et al., 2012) is dependent on the dimension of the training data (Liu et al., 2022), implying that the costs associated with DP-PCA

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany. Correspondence to: Yaxi Hu <yaxi.hu@tuebingen.mpg.de>, Amartya Sanyal <amsa@di.ku.dk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

An implementation for our framework is available at https://github.com/yaxihu/privacy-non-private-preprocessing.

could surpass the benefits of private learning in a lower-dimensional space.

To circumvent these challenges, an alternative approach is non-private pre-processing with a more rigorous analysis of the entire pipeline. This method is straightforward, circumvents the need for modifying existing processes, and avoids the costs associated with private pre-processing. Naturally, this raises the important question,

> *What is the price of non-private pre-processing in differentially private data analysis ?*

Our work shows that the overall privacy cost of pre-processed DP pipeline can be bounded with minimal degradation in privacy guarantee. To do this, we rely on two new technical notions: sensitivity of pre-processing functions (Definition 4) and Smooth-DP (Definition 6). In short,

1. We introduce a generic framework to quantify the privacy loss in the *pre-processed DP pipeline*. Applying this framework, we evaluate the impact of commonly used pre-processing techniques such as deduplication, quantization, data imputation, standard scaling, and PCA on the overall privacy guarantees.
2. We base our analysis on a novel variant of differential privacy, termed smooth DP, which may be of independent interest. We demonstrate that smooth DP retains essential DP characteristics, including post-processing and composition capabilities, as well as a slightly weaker form of amplification by sub-sampling.
3. We propose an algorithm to balance desired privacy levels against utility in the *pre-processed DP pipeline*. This approach is based on the Propose-Test-Release Mechanism, allowing the user to choose desired privacy-utility trade-off.

Related Work Closely related to our work, Debenedetti et al. (2023) also studies the necessity of conducting privacy analysis across the entire ML pipeline, rather than focusing solely on the training process. They identify that pre-processing steps, such as deduplication, can unintentionally introduce correlations into the pre-processed datasets, leading to privacy breaches. They show that the empirical privacy parameter of DP-SGD (Abadi et al., 2016) can be deteriorated by over five times with membership inference attack designed to exploit the correlation introduced by deduplication. While they show privacy attacks using deduplication with DP-SGD that can maximise the privacy loss, our work quantifies the privacy loss in deduplication as well as other pre-processing algorithms with several privacy preserving mecahnisms, thereby presenting a more holistic picture of this problem.

Another line of research, focusing on privacy in correlated datasets (Liu et al., 2016; Zhao et al., 2017; Humphries et al., 2023), shows that correlations in the datasets can increase privacy risk of ML models. In response, Pufferfish Differential Privacy (Kifer & Machanavajjhala, 2014; Song et al., 2017) and Dependent Differential Privacy (Liu et al., 2016; Zhao et al., 2017) were proposed as privacy notions tailored for datasets with inherent dependencies. However, these definitions usually require complete knowledge of the datasets' dependency structure or the data generating process and sometimes lead to vacuous privacy guarantees. Moreover, the application of these privacy notions usually complicates the privacy analysis, as many privacy axioms, such as composition, do not hold under these more general notions.

2. Preliminaries in Differential Privacy

Before providing the main results of our work, we first introduce some common definitions and mechanisms in Differential Privacy.

2.1. Rényi Differential Privacy

Differential privacy restricts the change in the output distribution of a randomized algorithm by altering a single element in the dataset. Formally, for $\varepsilon, \delta > 0$, a randomized algorithm $\mathcal{A}$ satisfies (ε, δ)-DP if for any two dataset S, S' that differ by exactly one element and any possible output of the algorithm $\mathcal{O}$,

$$\mathbb{P}\left[\mathcal{A}(S) \in \mathcal{O}\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{A}(S') \in \mathcal{O}\right] + \delta. \tag{1}$$

In this work, we mainly focus on a stronger notion of DP, known as Rényi Differential Privacy (RDP)(Mironov, 2017), based on the Rényi divergence between two distributions.

Definition 1 (Rényi Divergence)**.** *Let P, R be two probability distributions with $\mathrm{supp}(P) \subseteq \mathrm{supp}(R)$. Let $\alpha > 1$. The Rényi divergence with order α between P and R is defined as*

$$D_\alpha(P||R) = \frac{1}{\alpha - 1}\log \mathbb{E}_{x\sim R}\left(\frac{P(x)}{R(x)}\right)^{\alpha}.$$

Let $d_H(\cdot, \cdot)$ denote the Hamming distance between two datasets, we formally define RDP as follows.

Definition 2 ($(\alpha, \varepsilon(\alpha))$-RDP)**.** *Let $\varepsilon(\alpha)$ be a function that maps each α to a positive real number. A randomized algorithm $\mathcal{A}$ is $(\alpha, \varepsilon(\alpha))$-RDP if for all $\alpha > 1$, for any datasets S, S' differing at a single point, it holds that*

$$D_\alpha(\mathcal{A}(S)||\mathcal{A}(S')) \leq \varepsilon(\alpha).$$

This definition of RDP can be easily converted to standard DP via Lemma M. While Definition 2 is unconditional on the possible set of datasets, a relaxed version of conditional RDP can be defined over a given dataset collection $\mathcal{L}$ such that the neighboring datasets $S, S' \in \mathcal{L}$. When the dataset collection is known in advance, the conditional definition

of RDP allows for tighter privacy analysis. We use this conditional version in some of our analysis[1].

2.2. Private mechanisms

There are several ways to make non-private algorithms private. All of them implicitly or explicitly add carefully calibrated noise to the non-private algorithm. Below, we briefly define the three most common ways in which DP is injected in data analysis tasks and machine learning algorithms.

Output perturbation The easiest way to inject DP guarantees in an estimation problem is to perturb the output of the non-private estimator with appropriately calibrated noise. Two most common ways to do so are *Gaussian Mechanism* and *Laplace Mechanism*. For any deterministic estimator f, both mechanisms add noise proportional to the global sensitivity Δ_f, defined as the maximum difference in f over all pairs of neighboring datasets. For a given privacy parameter $\varepsilon > 0$, both the Gaussian mechanism, denoted $\mathcal{M}_G$, and the Laplace mechanism, denoted $\mathcal{M}_L$, produce an output of the form $f(S) + \xi$. Here, ξ follows a Gaussian distribution $\mathcal{N}(0, {}^{\Delta_f^2}/_{\varepsilon^2})$ for the Gaussian mechanism, and a Laplace distribution $\mathrm{Lap}({}^{\Delta_f}/_{\varepsilon})$ for the Laplace mechanism.

Random sampling While Output perturbation is naturally suited to privatising the output of non-private estimators, it is less intuitive when selecting discrete objects from a set. In this case, a private mechanism can sample from a probability distribution defined on the set of objects. The *Exponential Mechanism*, denoted as $\mathcal{M}_E$, falls under this category and is one of the most fundamental private mechanisms. Given a score function Q with global sensitivity Δ_Q, it randomly outputs an estimator w with probability proportional to $\exp\left(\frac{Q(w,S)\varepsilon}{2\Delta_Q}\right)$.

Gradient perturbation Finally, most common ML applications use gradient-based algorithms to minimize a loss function on a given dataset. A common way to inject privacy in these algorithms is to introduce Gaussian noise into the gradient computations in each gradient descent step. This is referred to as Differential Private Gradient Descent (DP-GD) denoted as $\mathcal{A}_{\mathrm{GD}}$ (Bassily et al., 2014; Song et al., 2021). Other variants that are commonly used are Differential Private Stochastic Gradient Descent (DP-SGD) with subsampling (Bassily et al., 2014; Abadi et al., 2016), denoted as $\mathcal{A}_{\mathrm{SGD-samp}}$, and DP-SGD with iteration (Feldman et al., 2018), denoted as $\mathcal{A}_{\mathrm{SGD-iter}}$. We include the detailed description of each of these methods in Appendix A.2.2.

[1]To avoid confusion, we note that our privacy guarantees are not *conditional*, in the sense that it does not suffer a catastrophic failure under any dataset.

3. Main Results

We first introduce a norm-based privacy notion, called Smooth RDP, that allows us to conduct a more fine-grained analysis on the impact of pre-processing algorithms. Using this definition, we establish our main results on the privacy guarantees of a pre-processed DP pipeline.

3.1. Smooth RDP

Our analysis on the privacy guarantees of pre-processed DP pipelines relies on a privacy notion that ensures indistinguishability between two datasets with a bounded L_{12} distance. Here, the L_{12} distance between two datasets S and S' of size n is defined as $d_{12}(S, S') = \sum_{i=1}^{n} \|S_i - S'_i\|_2$. We introduce this privacy notion as Smooth Rényi Differential Privacy (SRDP), defined as follows:

Definition 3 ($(\alpha, \varepsilon(\alpha, \tau))$-smooth RDP)**.** *Let $\varepsilon(\alpha, \tau)$ be a function that maps each α, τ pair to a real value. A randomized algorithm $\mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP if for each $\alpha > 1$ and $\tau > 0$,*

$$\sup_{S,S':d_{12}(S,S')\leq\tau} D_\alpha(\mathcal{A}(S)||\mathcal{A}(S')) \leq \varepsilon(\alpha, \tau).$$

SRDP shares similarities with distance-based privacy notions (Lecuyer et al., 2019; Epasto et al., 2023), but they differ in a key aspect: the distance-based privacy considers neighboring datasets with bounded distance, while SRDP allows for comparison over two datasets differing in every entry.

Similar to conditional RDP over a set $\mathcal{L}$, we define conditional SRDP over a set $\mathcal{L}$ by imposing the additional assumption that $S, S' \in \mathcal{L}$ in Definition 3.

While conditional SRDP over a set $\mathcal{L}$ can be considered as a special case of Pufferfish Rényi Privacy (Kifer & Machanavajjhala, 2014), we show that it satisfies desirable properties, such as sequential composition and privacy amplification by subsampling, which are not satisfied by the more general Pufferfish Renyi Privacy (Pierquin & Bellet, 2023).

Properties of SRDP Similar to RDP, SRDP satisfies (sequential) composition and closure under post-processing.

Lemma 1. *Let $\mathcal{L}$ be any dataset collection. Then, the following holds.*

- ***Composition*** *Let k be a positive integer. Let $\alpha > 0, \varepsilon_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \forall i \in [k]$. For any $i \in [k]$, if the randomized algorithms $\mathcal{A}_i$ is $(\alpha, \varepsilon_i(\alpha, \tau))$-SRDP over $\mathcal{L}$, then the composition of the k algorithms $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is $(\alpha, \sum_{i=1}^{k} \varepsilon_i(\alpha, \tau))$-SRDP over $\mathcal{L}$.*
- ***Post-processing*** *Let $\alpha > 0, \varepsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and f be an arbitrary algorithm. For any $\tau > 0$, if $\mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP over $\mathcal{L}$, then $f \circ \mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP over $\mathcal{L}$.*

Table 1. RDP and SRDP parameters of DP mechanisms. We let γ and κ_τ denote the inverse pointwise divergence and maximum divergence between two datasets.

Notation	Meaning	Mechanism	Assumptions	RDP	SRDP
f	Output function	$\mathcal{M}_G$	f is L-Lipschitz	$\frac{\alpha\varepsilon^2}{2}$	$\frac{\alpha L^2\tau^2\varepsilon^2}{2\Delta_f^2}$
		$\mathcal{M}_L$	f is L-Lipschitz	ε	$\frac{L\tau}{\Delta_f}\varepsilon$
Q	Score function	$\mathcal{M}_E$	Q is L-Lipschitz	ε	$\frac{L\tau}{\Delta_Q}\varepsilon$
ℓ	Loss function	$\mathcal{A}_{\mathrm{GD}}$	ℓ is L-Lipschitz and μ-smooth, $\sigma = \frac{L\sqrt{T}}{\varepsilon n}$	$2\alpha\varepsilon^2$	$\frac{\alpha\mu^2\tau^2\varepsilon^2}{2L^2}$
T	Number of iterations	$\mathcal{A}_{\mathrm{SGD-samp}}$	ℓ is L-Lipschitz and μ-smooth, $\sigma = \Omega(L\sqrt{T}/\varepsilon n)$, inverse point-wise divergence γ, $1 \le \alpha \le \min\left\{\frac{\sqrt{T}}{\varepsilon}, \frac{L^2T}{\varepsilon^2 n^2}\log\frac{n^2\varepsilon}{L\sqrt{T}}\right\}$	$\frac{\alpha^2\varepsilon^2}{2}$	$\frac{\alpha\mu^2\tau^2\varepsilon^2\gamma^2}{2L^2}$
σ	Variance of gradient noise	$\mathcal{A}_{\mathrm{SGD-iter}}$	ℓ is convex, L-Lipschitz and μ-smooth, $\sigma = \frac{8\sqrt{2\log n}\eta L}{\varepsilon\sqrt{n}}$, $\varepsilon = O(1/n\alpha^2)$, maximum divergence κ_τ, $L\sqrt{2\alpha(\alpha-1)} \le \sigma$	$\frac{\alpha\varepsilon^2}{2}$	$\frac{\alpha\tau^2\mu^2 n\log(n-\kappa_\tau+2)}{2(n-\kappa_\tau+1)L^2\log n}$
η	Learning rate				

SRDP also satisfies a form of Privacy amplification by subsampling. We state the weaker version without additional assumptions in Appendix A.1 and use a stronger version for $\mathcal{A}_{\mathrm{SGD-iter}}$ in Theorem 1 with some light assumptions.

SRDP parameters for common private mechanisms In Theorem 1, we present the RDP and SRDP parameters for the private mechanisms discussed above. The SRDP parameter usually relies on the Lipschitzness and Smoothness (see Appendix A.1 for definitions) of the output or objective function.

Theorem 1 (Informal)**.** *The DP mechanisms discussed in Section 2.2 satisfy RDP and SRDP under assumptions on the output or the objective functions. We summarize the parameters and their corresponding assumptions in Table 1.*

Table 1 demonstrates that the RDP parameter ε of most private mechanisms increases to $O(\tau\varepsilon)$ for SRDP. In contrast, a naive analysis using group privacy inflates the privacy parameter to $O(n\varepsilon)$, as any two datasets with d_{12} distance τ can differ by at most n entries. Hence, SRDP always leads to tighter privacy parameter than group privacy for $\tau = o(n)$, which we are able to exploit later.

For DP-SGD by subsampling and iteration, the SRDP guarantee is dependent on two properties: the inverse pointwise divergence and maximum divergence of the datasets, detailed in Appendix A.2.2. Briefly, the inverse pointwise divergence γ measures the ratio between the d_{12} distance and the maximum pointwise distance between two datasets. The maximum divergence κ_τ measures the number of points that the two datasets differ. DP-SGD with subsampling benefits from a small γ, while DP-SGD with iteration benefits from a small κ_τ. As we will show in the later sections, at least one of these conditions are usually met in practice.

3.2. Privacy of Pre-Processed DP Pipelines

Before stating the main result, we introduce the term *data-dependent pre-processing algorithm*. Let $\mathcal{X} \in \mathbb{R}^d$ be the instance space with Euclidean norm bounded by 1, i.e.$\forall x \in \mathcal{X}$, $\|x\|_2 \le 1$. A deterministic *data-dependent pre-processing algorithm* $\pi : \mathcal{X}^n \to \mathcal{F}$ takes a dataset S as input and returns a *pre-processing function* $\pi_S : \mathcal{X} \to \mathcal{X} \cup \emptyset$ in a function space $\mathcal{F}$. Intuitively, privacy of a private algorithm is retained under non-private data-dependent pre-processing, if a single element in the dataset has bounded impact on the output of the pre-processing algorithm. We define the sensitivity of a pre-processing algorithm to quantify the impact of a simgle element below.

Definition 4. *Let $S_1 = S \cup \{z_1\}$ and $S_2 = S \cup \{z_2\}$ be two arbitrary neighboring datasets, we define the L_∞ and L_2 sensitivity of a pre-processing function π as*[2]

$$\Delta_\infty(\pi) = \sup_{S_1,S_2:d_H(S_1,S_2)=1} d_H(\pi_{S_1}(S), \pi_{S_2}(S)),$$
$$\Delta_2(\pi) = \sup_{S_1,S_2:d_H(S_1,S_2)=1} \max_{x\in S} \|\pi_{S_1}(x) - \pi_{S_2}(x)\|_2 . \tag{2}$$

When the neighboring datasets S, S' are from a dataset collection $\mathcal{L}$, we define conditional sensitivity of the pre-processing algorithm π as $\Delta_\infty(\mathcal{L}, \pi)$ and $\Delta_\infty(\mathcal{L}, \pi)$ in a similar manner. The conditional sensitivity is non-decreasing as the size of the set of $\mathcal{L}$ increases, i.e. if $\mathcal{L} \subset \mathcal{L}'$, then $\Delta_\infty(\mathcal{L}, \pi) \le \Delta_\infty(\mathcal{L}', \pi) \le \Delta_\infty(\pi)$ and $\Delta_2(\mathcal{L}, \pi) \le \Delta_2(\mathcal{L}', \pi) \le \Delta_2(\pi)$ for all π.

In Theorem 2, we present the privacy guarantees of pre-processed DP pipeline in terms of the RDP and SRDP parameters of the private algorithm and the sensitivity of the

[2]When one of $\pi_{S_1}(x)$ and $\pi_{S_2}(x)$ is $\emptyset$, we define $\|\pi_{S_1}(x) - \pi_{S_2}(x)\|_2 = 1$.

pre-processing algorithms. This is a meta theorem, which we then refine to get specific guarantees for different combinations of pre-processing and private mechanisms in Theorem 7. With a slight abuse of notation, we denote the output of a private algorithm $\mathcal{A}$ on $\pi_S(S)$ as $\mathcal{A} \circ \pi(S)$.

Theorem 2. *For a set of datasets $\mathcal{L}$ and for any $\alpha \geq 1$, $\tau > 0$, consider an algorithm $\mathcal{A}$ that is $(\alpha, \varepsilon(\alpha))$-RDP and $(\alpha, \tilde{\varepsilon}(\alpha, \tau))$-SRDP over the set $\mathcal{L}$. For a pre-processing algorithm π with L_∞ sensitivity Δ_∞ and L_2 sensitivity Δ_2, $\mathcal{A} \circ \pi$ is $(\alpha, \widehat{\varepsilon})$-RDP over $\mathcal{L}$ for all $c_1, c_2 \geq 1$, where*

$$\widehat{\varepsilon} \leq \max\left\{ \frac{\alpha c_1 - 1}{c_1(\alpha - 1)} \tilde{\varepsilon}(\alpha c_1, \Delta_2 \Delta_\infty) + \varepsilon\left(\frac{c_1 \alpha - 1}{c_1 - 1}\right), \frac{\alpha c_2 - 1}{c_2(\alpha - 1)} \varepsilon(\alpha c_2) + \tilde{\varepsilon}\left(\frac{c_2 \alpha - 1}{c_2 - 1}, \Delta_2 \Delta_\infty\right) \right\}. \tag{3}$$

Proof sketch. Consider two neighboring datasets S_1 and S_2, where $S_1 = S \cup \{z_1\}, S_2 = S \cup \{z_2\}$. Let π_1 and π_2 be the output functions of the pre-processing algorithm π on S_1 and S_2 respectively. Our objective is to upper bound the Rényi divergence between the output distribution of $\mathcal{A}$ on the pre-processed datasets $\pi_1(S_1)$ and $\pi_2(S_2)$.

We proceed by constructing a new dataset $\tilde{S}$ that consists of $\pi_2(S)$ and the point $\pi_1(z_1)$, as indicated in Figure 4. The construction ensures that $\tilde{S}$ and $\pi_2(S_2)$ are neighboring datasets, and that the L_{12} distance between $\tilde{S}$ and $\pi_2(S_2)$ is upper bounded by $\Delta_2 \Delta_\infty$. We then apply the RDP property of $\mathcal{A}$ to upper bound the divergence between $\mathcal{A}(\tilde{S})$ and $\mathcal{A}(\pi_2(S_2))$. Then, we employ the SRDP property of $\mathcal{A}$ to upper bound the divergence between $\mathcal{A}(\tilde{S})$ and $\mathcal{A}(\pi_1(S_1))$. Finally, we establish the desired upper bound in Equation (3) by combining the previous two divergences using the weak triangle inequality of Rényi divergence (Lemma J). □

While the privacy guarantee provided by Theorem 2 is conditional over a dataset collection $\mathcal{L}$, it can be extend to an unconditional privacy guarantee over all possible datasets using the Propose-Test-Release (PTR) framework (Dwork & Lei, 2009). We show an example of the application of PTR and its guarantees in Section 5.2.

Comparison with the group privacy or DDP analysis A naive analysis on the privacy guarantee of $\mathcal{A} \circ \pi$ using group privacy or DDP (Zhao et al., 2017; Liu et al., 2016) provides an upper bound that grows polynomially[3] with Δ_∞, which can be as large as poly(n). In contrast, Theorem 2 implies a tighter bound on privacy for common DP mechanisms, especially when $\Delta_\infty \Delta_2 = o(n)$.[4] Next, we present various examples of pre-processing algorithms with small sensitivities $\Delta_\infty \Delta_2 = O(1)$. In Figure 2, we illustrate the improvement of the privacy analysis in Theorem 2 over the conventional analysis via group privacy for numerous pre-processing algorithms.

[3]Linearly if we consider Approximate DP.

[4]Theorem 2 is also applicable for other distance metrics, as long as Smooth RDP and the sensitivity of pre-processing are defined under comparable distance metric.

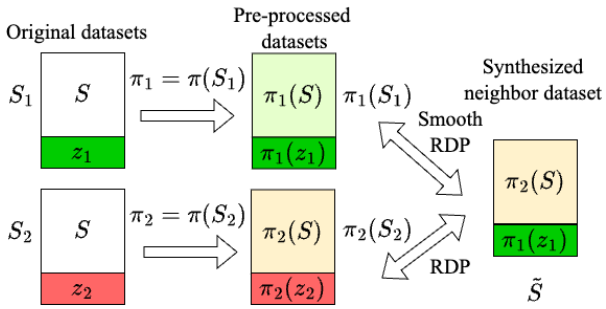


Figure 1. Illustration of the privacy analysis: For two neighboring datasets S_1, S_2, a pre-processing algorithm π yields the pre-processed datasets $\pi_1(S_1)$ and $\pi_2(S_2)$ respectively. A synthetic dataset $\tilde{S}$ is constructed by combining the pre-processed datasets, ensuring that $\tilde{S}$ and $\pi_2(S_2)$ are neighboring datasets and that $\tilde{S}$ and $\pi_1(S_1)$ have bounded L_{12} distance.

4. Privacy Guarantees of Common Pre-Processing Algorithms

In this section, we use Theorem 2 to provide overall privacy guarantees for several common pre-processing algorithms. First, in Section 4.1, we define the pre-processing algorithms π and bound their L_2 and L_∞ sensitivities. Then, in Section 4.2 we provide the actual privacy guarantees for all combinations of these pre-processing algorithms and privacy mechanisms defined in Section 2.2. We assume the instance space $\mathcal{X}$ to be the Euclidean ball with radius 1.

4.1. Sensitivity of Common Pre-Processing Algorithms

Approximate deduplication Many machine learning models, especially Large Language Models (LLMs), are trained on internet-sourced data, often containing many duplicates or near-duplicates. These duplicates can cause issues like memorization, bias reinforcement, and longer training times. To mitigate this, approximate deduplication algorithms are used in preprocessing, as discussed in Rae et al. (2021); Touvron et al. (2023). We examine a variant of these algorithms termed η-approximate deduplication.

The concept of η-approximate deduplication involves defining a *good cluster*. For a dataset S and a point $x \in S$, consider $B(x, \eta; S) = \{\tilde{x} \in S : \|\tilde{x} - x\|_2 \leq \eta\}$, a ball of radius η around x. This forms a *good cluster* if $B(x, \eta; S) = B(x, 3\eta; S)$. Essentially, this means that any point within a good cluster is at least 2η distant from all other points outside the cluster in the dataset. We also define $B(S) = \{B_i = B(x_i, \eta; S)\}_{i=1}^m$ as the set of all good

clusters in a dataset S.

The η-approximate deduplication process, π_η^d, identifies and retains only the center of each good cluster, removing all other points. Points are removed in reverse order of cluster size, prioritizing those with more duplicates.

Quantization Quantization, another pre-processing algorithm for data compression and error correction, is especially useful when the dataset contains measurement errors. We describe a quantization method similar to η-approximate deduplication, denoted as π_η^q: it identifies all *good* clusters in the dataset, and replaces all points within each good cluster with the cluster's centroid in the reverse order of the size of the good clusters. The difference between de duplication and quantization is that while quantization replaces near duplicates with a representative value, deduplication removes them entirely. We discuss the L_2 and L_∞ sensitivity of deduplication and quantization in Proposition 3.

Proposition 3. *For a dataset collection $\mathcal{L}$, the L_2 and L_∞ sensitivities*[5] *of η-approximate deduplication π_η^d and quantization π_η^q are $\Delta_2(\mathcal{L}, \pi_\eta^d) = 1$ and $\Delta_2(\mathcal{L}, \pi_\eta^q) = \eta$, and*

$$\Delta_\infty(\mathcal{L}, \pi_\eta^d) = \Delta_\infty(\mathcal{L}, \pi_\eta^q) = \max_{S\in\mathcal{L}} \max_{B\in B(S)} 2\,|B|\,.$$

While the L_2 sensitivity of deduplication is a constant 1, Δ_∞ is usually upper bounded by a small number. This is because in realistic datasets, the number of near duplicates is generally small for sufficiently small η. This leads to upper bounding the product $\Delta_2\Delta_\infty$ by a small number. For example, in text datasets, the fraction of near duplicates is typically smaller than 0.1, as demonstrated in Table 2 and 3 in Lee et al. (2022).

Model-based imputation Survey data, such as US census data, often contains missing values, resulting from the participants unable to provide certain information, invalid responses, and changing questionnaire over time. Hence, it is crucial to process the missing values in these datasets with data imputation methods, to make optimal use of the available data for analysis while minimizing the introduction of bias into the results.

We consider several imputation techniques which use the values of the dataset to impute the missing value. This can involve training a regressor to predict the missing feature based on the other feature, or simply imputing with dataset-wide statistics like mean, median or trimmed mean. For the sake of clarity, we only discuss mean imputation in the main text but provide the guarantees for other imputations in Proposition 16 and Table 3 in Appendix C.2.

[5]When the definition of neighboring dataset is refined to addition and deletion of a single data point, the L_∞ sensitivity of both deduplication and quantization on a set $\mathcal{L}$ can be reduced to $\max_{S\in\mathcal{L}, B\in B(S)} |B|$.

Corollary 4. *For a dataset collection $\mathcal{L}$ with maximum p missing values in any dataset, the L_∞ sensitivity of mean imputation π_{mean} over $\mathcal{L}$ is p and the L_2-sensitivity of π_{mean} is upper bounded by $\frac{2}{n-p}$.*

Principal Component Analysis Principal Component Analysis (PCA) is a prevalent pre-processing algorithm. It computes a transformation matrix $A_k^\top \in \mathbb{R}^{k\times d}$ using the top k eigenvalues of the dataset S's covariance matrix. PCA serves two main purposes: dimension reduction and rank reduction. For dimension reduction, denoted as $\pi_{\text{PCA}-\text{dim}}$, PCA projects data into a lower-dimensional space (typically for high-dimensional data visualization), using the pre-processing function $\pi_{S,\text{dim}}(x) = A_k^\top x$. For rank reduction, represented as $\pi_{\text{PCA}-\text{rank}}$, PCA leverages the low-rankness of the dataset with the function $\pi_{S,\text{rank}}(x) = A_k A_k^\top x$.

The primary difference between these two PCA applications is in the output dimensionality. Dimension reduction yields data of dimension k, while rank reduction maintains the original dimension d, but with a low-rank covariance matrix of rank k. We detail the L_∞ and L_2 sensitivity for both PCA variants in Proposition 5.

Proposition 5. *For a dataset collection $\mathcal{L}$, the L_∞ sensitivity of $\pi_{\text{PCA}-\text{dim}}$ and $\pi_{\text{PCA}-\text{rank}}$ is the size of the datasets in $\mathcal{L}$, i.e. n. The L_2-sensitivity of $\pi_{\text{PCA}-\text{dim}}$ and $\pi_{\text{PCA}-\text{rank}}$ is bounded by $2\Delta_2$ and Δ_2 respectively, where*

$$\Delta_2 = \frac{4(3n+2)}{n(n-1)\min\{\delta_{\min}^k(\mathcal{L}), \delta_{\min}^1(\mathcal{L})\}},$$

where $\delta_{\min}^k(\mathcal{L}) = \min_{S\in\mathcal{L}} \lambda_k(S) - \lambda_{k+1}(S)$ is the minimum gap between the kth and $(k+1)^{th}$ eigenvalue over any covariance matrix of $S \in \mathcal{L}$.

Standard Scaling Scaling is one of the most common pre-processing methods. In Proposition 6, we provide the sensitivity results of standard scaling which scales each feature to have mean 0 and standard deviation 1, and min max scaling, which scales each feature to the interval between 0 and 1.

Proposition 6. *For a dataset collection $\mathcal{L}$, the L_∞ sensitivity of standard scaling and min max scaling is the size of the datasets in $\mathcal{L}$, i.e. n. The L_2 sensitivity of standard scaling is*

$$\Delta_2 = \frac{2}{\sigma_{\min}^3 n} + \frac{2}{n\sigma_{\min}},$$

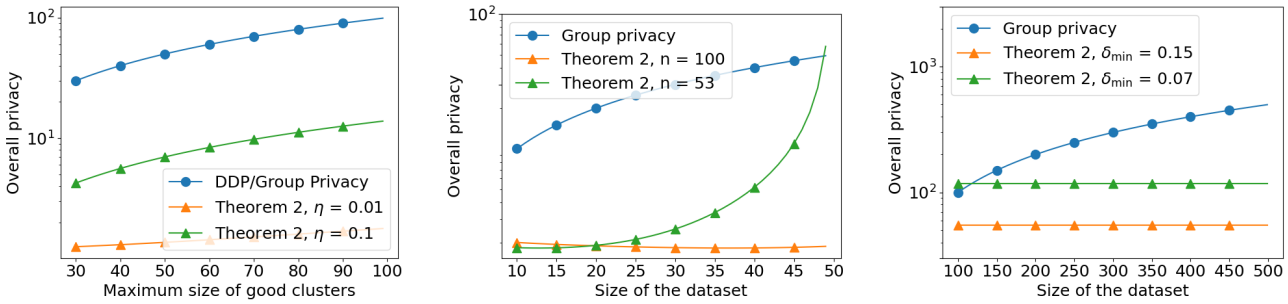
where $\sigma_{\min}$ is the minimum standard deviation over datasets in $\mathcal{L}$.

4.2. Privacy Analysis for Pre-Processing Algorithms

After establishing the necessary elements of our analysis, including the sensitivities of pre-processing algorithms (Section 3) and the SRDP parameters of private mechanisms (Table 1), we are ready to present the exact overall privacy guarantees for specific pre-processed DP pipelines. In Table 2,

Table 2. Overall privacy guarantees $\widehat{\varepsilon}$ of pre-processed DP pipelines with $\alpha \geq 11$. Let p represent the L_∞ sensitivity of deduplication, quantization and mean imputation. We also assume the size of the dataset $n \geq 101$ for PCA, and the Lipschitz and smoothness parameters, along with global sensitivity, are set to 1. We omit the privacy guarantees of deduplication due to space constraint. However, we note that the privacy guarantees for deduplication are the same as those for quantization with $\eta = 1$. See Appendix D for details.

	Quantization	Mean imputation	PCA	Standard Scaling
$\mathcal{M}_G$	$1.05\alpha\varepsilon^2\left(1+\eta^2p^2\right)$	$1.05\alpha\varepsilon^2\left(1+\frac{4p^2}{(n-p)^2}\right)$	$1.05\alpha\varepsilon^2\left(1+\frac{12.2^2}{(\delta^k_{\min})^2}\right)$	$1.05\alpha\varepsilon^2\left(1+\frac{4}{\sigma^3_{\min}}\right)$
$\mathcal{A}_{\mathrm{GD}}$	$1.05\alpha\varepsilon^2\left(4+\eta^2p^2\right)$	$4.2\alpha\varepsilon^2\left(1+\frac{p^2}{(n-p)^2}\right)$	$1.05\alpha\varepsilon^2\left(4+\frac{12.2^2}{(\delta^k_{\min})^2}\right)$	$4.2\alpha\varepsilon^2\left(1+\frac{1}{\sigma^3_{\min}}\right)$
$\mathcal{M}_L/\mathcal{M}_E$	$\varepsilon\left(1+\eta p\right)$	$\varepsilon\left(1+\frac{2p}{n-p}\right)$	$\varepsilon\left(1+\frac{12.2}{\delta^k_{\min}}\right)$	$\varepsilon\left(1+\frac{4}{\sigma^3_{\min}}\right)$
$\mathcal{A}_{\mathrm{SGD-samp}}$	–	–	$1.05\alpha\varepsilon^2\left(2\alpha+\frac{12.2^2}{(\delta^k_{\min})^2}\right)$	$2.1\alpha\varepsilon^2\left(\alpha+\frac{8}{\sigma^6_{\min}}\right)$
$\mathcal{A}_{\mathrm{SGD-iter}}$	$1.1\alpha\varepsilon^2\left(1+\frac{\frac{\eta^2p^2n}{\log n}}{\frac{n-p}{\log n-p}}\right)$	$1.1\alpha\varepsilon^2\left(1+\frac{\frac{4p^2\frac{n}{\log n}}{(n-p)^3}}{\log n-p}\right)$	–	–



(a) Quantization (b) Mean Imputation (c) PCA for rank reduction

Figure 2. Visualization of the overall privacy of pre-processed Gaussian mechanism analysed with group privacy and our bound from Theorem 2. Here, η is the distance threshold of the quantization algorithm, n is the size of the possible datasets, and $\delta_{\min}$ is the minimum gap between the kth and the $k+1^{th}$ eigenvalue of all possible datasets.

we present the privacy guarantees for various combinations of pre-processing methods and private mechanisms.

Theorem 7 (Informal)**.** *Pre-processed DP pipelines comprised of all combinations of private mechanisms in Section 2.2 and pre-processing algorithms in Section 4.1 are $(\alpha, \widehat{\varepsilon})$-RDP where $\widehat{\varepsilon}$ is specified in Table 2.*

Table 2 shows that the pre-processing methods discussed in Section 4.1 typically lead to a minimal, constant-order increase in the privacy cost for common DP mechanisms, depending on the datasets in $\mathcal{L}$. Specifically, deduplication results in a constant increase in the privacy cost when the datasets in $\mathcal{L}$ do not contain large clusters, i.e. $\Delta_\infty = o(n)$, whereas quantization can handle larger clusters of size $\Delta_\infty = o(n/\eta)$. The privacy cost of mean imputation remains constant for datasets with few missing values ($p = o(n)$). Standard scaling leads to constant amplification in privacy cost when all features in dataset collection $\mathcal{L}$ have non-vanishing standard deviation. For PCA, we only present the results for rank reduction and the results for dimension reduction follows similarly. The additional privacy cost remains small for low rank datasets with $\delta_{\min}$ bounded away from 0.

Moreover, Figure 2 demonstrates a comparison between our SRDP-based analysis (Theorem 2) and the naive group privacy/DDP analysis for pre-processed Gaussian mechanism. This comparison illustrates the advantage of our SRDP-based analysis over group privacy/DDP and how the privacy guarantees vary with properties of the dataset collections $\mathcal{L}$. For quantization, the SRDP-based privacy analysis provides significantly stronger privacy guarantees for smaller η. For mean imputation, the difference between SRDP-based analysis and group pirvacy analysis is large for smaller number of missing values. However, in the extreme scenario where over 90% of the data points are missing, group privacy achieves a similar guarantee as our analysis with SRDP. In the case of PCA for rank reduction, the privacy parameter obtained with SRDP-based privacy analysis remains constant, while it increases for group privacy analysis with the size of the dataset.

5. Unconditional Privacy Guarantees in Practice

The previous section provides a comprehensive privacy analysis for pre-processed DP pipelines where the resulting

privacy guarantee depends on properties of the dataset collection. In this section, we first illustrate how conditional privacy analysis can become ineffective for pathological datasets (Section 5.1). Then, we introduce a PTR-inspired framework to address this and establish unconditional privacy guarantees in Section 5.2. Using PCA for rank reduction as an example, we provide convergence guarantee for excess empirical loss for generalized linear models and validate our results with synthetic experiments.

5.1. Limitations of conditional privacy guarantee

The previous analysis in Section 3 and 4 rely on the chosen dataset collection $\mathcal{L}$. Typically, for dataset collections $\mathcal{L}$ that are well-structured, non-private pre-processing leads to only a slight decline in the privacy parameters as discussed in Section 4.2. However, this degradation can become significant for datasets that exhibit pathological characteristics. In the following, we present several examples where our guarantees in Theorem 2 become vacuous due to the pathological nature of the dataset.

Imputation If the number of missing points is comparable to the size of the dataset i.e. $p \approx n$, the privacy guarantee in Theorem 2 deteriorates to the level of those obtained from group privacy or DDP. This deterioration is also reflected in Figure 2b.

PCA When performing PCA with a reduction to rank k, the privacy guarantee can worse and scale with n for very small $\delta^k_{\min}$, in particular when $\delta^k_{\min} = O\left(1/n\right)$. This situation can occur naturally when the data is high rank or when k is chosen "incorrectly".

Deduplication Consider a dataset $S = \{z_1, ..., z_n\}$, where each pair of distinct data points are at least η distance apart, yet each point are no more than η distance from a specific reference point x, i.e.$\|z_i - z_j\|_2 \geq \eta$, $\|z_i - x\|_2 \leq \eta$, for $i \neq j \in [n]$.Under these conditions, applying deduplication on S leaves the dataset unchanged as all points are uniformly η distance from each other. However, if x is added to S, then deduplication would eliminate all points except x. As a result, the L_∞ sensitivity of deduplication becomes n, resulting in the same privacy analysis as group privacy. Interestingly, a similar example was leveraged by Debenedetti et al. (2023) in their side-channel attack.

It is important to recognise that datasets exhibiting pathologucal characteristics, like those above, are usually not of practical interest in data analysis. For instance, applying mean imputation to a dataset where the number of missing points is nearly equal to the size of the dataset or implementing approximate deduplication that results in the removal of nearly all data are not considered sensible practice. A possible solution to this problem is to potentially refuse to provide an output when the input dataset is deemed pathological, provided that the decision to refuse is made in a manner that preserves privacy. This approach is adopted in the following section, where we employ the Propose-Test-Release framework proposed by Dwork & Lei (2009) to establish unconditional privacy guarantees over all possible datasets, at the expense of accuracy on datasets that are "pathological".

5.2. Unconditional privacy guarantees via PTR

We present Algorithm 1 that applies the PTR procedure to combine the non-private PCA for rank reduction with DP-GD. It exhibits unconditional privacy guarantee over all possible datasets (Theorem 8). While this technique is specifically described for combining PCA with DP-GD, it's worth noting that the same approach can be applied to other pairings of DP mechanisms and pre-processing algorithms, though we do not explicitly detail each combination since their implementations follow the same basic principles.

Algorithm 1 PTR for $\pi_{\mathrm{PCA-rank}}$ on DP-GD

Input: Dataset S, estimated lower bound β of $\delta_k(S)$, privacy parameters ε, δ, Lipschitzness L

Set $\Gamma = \min_{S':\delta_k(S')<\beta} d_H(S, S') + Lap\left(\frac{1}{\varepsilon}\right)$ $\delta_k(S') = \lambda_k\left(S'\right) - \lambda_{k+1}\left(S'\right)$
if $\Gamma \leq (\log \frac{2}{\delta})/\varepsilon$ **then**
  Return $\perp$
else
  Compute the pre-processing function $\pi^S_{\mathrm{PCA-rank}}$ using the dataset S and set $\sigma = \frac{2L\sqrt{T}}{\varepsilon n}$.
  Return $\mathcal{A}_{\mathrm{DP-GD}}(\pi^S_{\mathrm{PCA-rank}}(S))$ with parameter σ.
end if

Theorem 8. *For any L-Lipschitz and μ-smooth loss function ℓ, $\varepsilon > 0$ and $\delta \leq \exp\left(-1.05\varepsilon^2\left(1 + \frac{12.2^2\mu^2}{L^2\beta^2}\right)\right)$, Algorithm 1 with privacy parameters ε, δ, and estimated lower bound β is $(\widehat{\varepsilon} + \varepsilon, \delta)$-DP on a dataset of size $n \geq 101$, where $\widehat{\varepsilon} = 3\varepsilon\sqrt{1.05\left(1 + \frac{12.2^2\mu^2}{L^2\beta^2}\right)\log\frac{1}{\delta}}$.*

The RDP parameter on PCA with DP-GD in Table 2 with the same parameter σ can be converted to $(\widehat{\varepsilon}, \delta)$- DP if it holds that for all $S \in \mathcal{L}$, $\delta_k\left(S\right) \geq \beta$. In comparison, the guarantee in Theorem 8 is marginally worse by an additive ε, but it remains applicable even when $\delta_k\left(S\right) < \beta$ for some $S \in \mathcal{L}$. However, when $\delta_k\left(S\right) < \beta$ for some $S \in \mathcal{L}$, the bound in Table 2 degrades to the same as obtained via group privacy/DDP.

For generalized linear models, we present an high probability upper bound on the excess empirical loss of Algorithm 1 conditional on the properties of the private dataset. Given a dataset $S = \{(x_i, y_i)\}_{i=1}^n$ and a loss function ℓ, let $\hat{\ell}_S(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\langle\theta, x_i\rangle, y_i)$ denote the empirical loss of a generalized linear model on the dataset S and let

$\theta^\star = \arg\min_{\theta\in\mathbb{R}^d} \hat{\ell}_S(\theta)$. For simplicity, we assume S is centered and ℓ is L-Lipschitz.

Proposition 9. *For δ, ε defined in Algorithm 1, $n \geq 101$, and any $\beta \leq \delta_k(S)$, with probability at least $1-\xi$, Algorithm 1 outputs $\hat{\theta}$ such that the excess empirical risk*

$$\mathbb{E}\left[\hat{\ell}_S(\hat{\theta})\right] - \hat{\ell}_S(\theta^\star) = O\left(\frac{L\sqrt{1+k\log 1/\delta}}{\varepsilon n}\right) + 2L\Lambda, \quad (4)$$

where $\xi = \frac{1}{2\delta}\exp\left(-\frac{(\delta_k(S)-\beta)n\varepsilon}{12.2}\right)$, $\Lambda = \sum_{i=k+1}^{d}\lambda_i(S)$ where the high probability is over the randomness in Step 1 and the expectation is over the randomness of Step 6.

The results follow a similar analysis as (Song et al., 2021), which shows that the convergence bound of DP-GD scales with the rank of the dataset. However, when the dataset remains full rank but the first k eigenvalues dominates the rest, e.g. $\Lambda = O(\sqrt{k})$, the convergence bound following (Song et al., 2021) is $O(\sqrt{d})$. In contrast, Proposition 9 leads to a dimension-independent convergence bound of order $O(\sqrt{k})$, with a slight degradation in the privacy guarantee. Here, β introduces the trade-off between privacy and utility. Large β leads to tighter privacy guarantee (small effective ε in Theorem 8) at the risk of worse utility (large ξ in Proposition 9).

Experimental results We conducted experiments on a synthetic approximately low rank dataset to corroborate our results in Proposition 9, and summarised the results in Figure 3. We generate a 2-class low rank dataset consisting of 1000 data points with dimension 6000 and approximately rank 50. The synthetic dataset has positive yet small eigenvalues for the kth eigenvectors for $k \geq 50$, ensuring $2L\Lambda$ in Equation (4) is small but positive.

For each overall privacy parameter ε, we evaluated the excess empirical risk of DP-SGD with: a) non-private PCA with an adjusted privacy parameter from Table 2, b) DP-PCA, with half of the privacy budget to allocated to DP-PCA, and c) no pre-processing.

Figure 3 shows that pre-processed DP pipeline outperforms logistic regression without pre-processing. This is because the dataset's approximate low rankness, indicated by $2L\Lambda > 0$, prevents logistic regression from leveraging the dimension-independent optimization guarantees using the original high dimensional datasets (Song et al., 2021). However, non-private PCA, while incurring a minor constant-order privacy cost, effectively utilizes the data's low rankness and offers significant benefits, especially at smaller ε values where optimization error is dominated by the first term in Equation (4). Figure 3 also demonstrates that DP-PCA exhibits the worst performance among the three methods, beccause the inherent error of order $\Omega(d)$ of DP-PCA dominates given a high dimensional dataset with $d \ll n$(Liu et al., 2022).

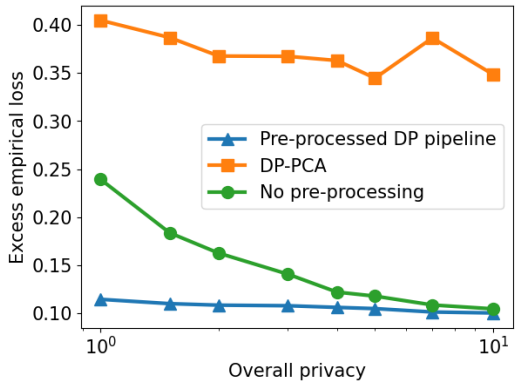


Figure 3. Comparison of excess empirical loss of private logistic regression: for each level of overall privacy ε, pre-processed DP pipeline consistently outperforms other methods.

In Appendix F, we provide the details on experimental setups and a discussion on the practical modification of Algorithm 1 with clipping.

6. Conclusion

In this paper, we investigate the often-neglected impact of pre-processing algorithms in private ML pipelines. We propose a framework to assess the additional privacy cost from non-private pre-processing steps using two new technical notions: Smooth RDP and sensitivities of pre-processing algorithms. Finally, we propose a PTR-based procedure to relax some of the necessary assumptions in our framework and make it practically usable. Several interesting directions of future work remain unexplored, including handling more complex pre-processing algorithms, such as pre-trained deep neural feature extractors and private algorithms like private data synthesis.

Acknowledgements

We thank Francesco Pinto for the initial discussion of the project. We also thank Alexandru Ţifea, Jiduan Wu and Omri Ben-Dov for their useful feedback. Additionally, we extend our gratitude to the reviewers of ICML for their careful reviews and insightful suggestions.

Impact Statement

Our work addresses a common practice, non-private pre-processing, that leads to additional privacy cost in ML applications. We identify the importance of extending the privacy guarantees from the standalone ML model to the whole ML pipeline and provide simple tools to achieve it. Our work contributes to the safer use of machine learning in society.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Alabi, D., McMillan, A., Sarathy, J., Smith, A., and Vadhan, S. Differentially private simple linear regression. *arXiv:2007.05157*, 2022.

Anil Jadhav, D. P. and Ramanathan, K. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 2019.

AppleDP. Learning with privacy at scale, 2017.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 2014.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research (JMLR)*, 2011.

Chaudhuri, K., Sarwate, A., and Sinha, K. Near-optimal differentially private principal components. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2012.

Debenedetti, E., Severi, G., Carlini, N., Choquette-Choo, C. A., Jagielski, M., Nasr, M., Wallace, E., and Tramèr, F. Privacy side channels in machine learning systems. *arXiv:2309.05610*, 2023.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Annual ACM Symposium on Theory of Computing (STOC)*, 2009.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 2006.

Epasto, A., Mirrokni, V., Narayanan, S., and Zhong, P. k-means clustering with distance-based privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018.

Ganesh, A., Haghifam, M., Nasr, M., Oh, S., Steinke, T., Thakkar, O., Thakurta, A., and Wang, L. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning (ICML)*, 2023.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. 1985.

Humphries, T., Oya, S., Tulloch, L., Rafuse, M., Goldberg, I., Hengartner, U., and Kerschbaum, F. Investigating membership inference attacks under data dependencies. In *IEEE Computer Security Foundations Symposium (CSF)*, 2023.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

Kifer, D. and Machanavajjhala, A. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 2014.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, 2019.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Li, T., Zaheer, M., Reddi, S., and Smith, V. Private adaptive optimization with side information. In *International Conference on Machine Learning (ICML)*, 2022a.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022b.

Liu, C., Chakraborty, S., and Mittal, P. Dependence makes you vulnerable: Differential privacy under dependent tuples. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2016.

Liu, F. Statistical properties of sanitized results from differentially private laplace mechanism with univariate bounding constraints. *arXiv:1607.08554*, 2016.

Liu, X., Kong, W., Jain, P., and Oh, S. DP-PCA: Statistically optimal and differentially private pca. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Mironov, I. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

Nesterov, Y. E. *Introductory Lectures on Convex Optimization - A Basic Course*. 2004. ISBN 978-1-4613-4691-3.

Pierquin, C. and Bellet, A. Rényi pufferfish privacy: General additive noise mechanisms and privacy amplification by iteration. *arXiv:2312.13985*, 2023.

Pinto, F., Hu, Y., Yang, F., and Sanyal, A. PILLAR: How to make semi-private learning more effective. *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., and et al., Y. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv:2112.11446*, 2021.

Schindler, D. *A Variant of Weyl's Inequality for Systems of Forms and Applications*. 2015.

Song, S., Wang, Y., and Chaudhuri, K. Pufferfish privacy mechanisms for correlated data. In *ACM International Conference on Management of Data*, 2017.

Song, S., Steinke, T., Thakkar, O., and Thakurta, A. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., and et al., B. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*, 2023.

Tramer, F. and Boneh, D. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations (ICLR)*, 2021.

U.S. Census Bureau. Understanding differential privacy, 2020.

van Erven, T. and Harremoës, P. Rényi divergence and Kullback-Leibler divergence. *arXiv:1206.2459*, 2012.

Yu, D., Zhang, H., Chen, W., and Liu, T.-Y. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 2014.

Zhao, J., Zhang, J., and Poor, H. V. Dependent differential privacy for correlated data. In *IEEE Globecom Workshops (GC Wkshps)*, 2017.

Zhou, Y., Wu, S., and Banerjee, A. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *International Conference on Learning Representations (ICLR)*, 2021.

Zwald, L. and Blanchard, G. On the convergence of eigenspaces in kernel principal component analysis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2005.

We present the detailed proof of our results and some additional findings in the appendix. To distinguish the lemmas from existing literature and new lemmas used in our proofs, we adopt alphabetical ordering for established results and numerical ordering for our contributions.

A. Proofs regarding Smooth RDP (Section 3.1)

In this section, we present the proofs regarding SRDP properties. In Appendix A.1, we introduce SRDP satisfies some desired properties, and present their proofs. Appendix A.2 provides the exact SRDP parameters of common DP mechanisms.

A.1. Properties of SRDP and their proofs

While conditional SRDP over a set $\mathcal{L}$ can be considered as a special case of Pufferfish Rényi Privacy (Kifer & Machanavajjhala, 2014), we show that it satisfies desirable properties, such as sequential composition and privacy amplification by subsampling, which are not satisfied by the more general Pufferfish Renyi Privacy (Pierquin & Bellet, 2023).

Properties of SRDP Similar to RDP, SRDP satisfies (sequential) composition and closure under post-processing.

Lemma 2. *Let $\mathcal{L}$ be any dataset collection. Then, the following holds.*

- ***Composition*** *Let $\alpha > 0, \varepsilon_i : \mathbb{R}\times\mathbb{R} \to \mathbb{R}, \forall i \in [k]$. For any $i \in [k]$, if the randomized algorithms $\mathcal{A}_i$ is $(\alpha, \varepsilon_i(\alpha, \tau))$-SRDP over $\mathcal{L}$, then the composition $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is $(\alpha, \sum_{i=1}^k \varepsilon_i(\alpha, \tau))$-SRDP over $\mathcal{L}$.*
- ***Post-processing*** *Let $\alpha > 0, \varepsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and f be an arbitrary algorithm. For any $\tau > 0$, if $\mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP over $\mathcal{L}$, then $f \circ \mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP over $\mathcal{L}$.*

SRDP also satisfies a form of Privacy amplification by subsampling. We state the weaker version without additional assumptions in Appendix A.1 and use a stronger version for $\mathcal{A}_{\mathrm{SGD-iter}}$ in Theorem 1 with some light assumptions.

Lemma 2. *Let $\mathcal{L}$ be any dataset collection. Then, the following holds.*

- ***Composition*** *Let $\alpha > 0, \varepsilon_i : \mathbb{R}\times\mathbb{R} \to \mathbb{R}, \forall i \in [k]$. For any $i \in [k]$, if the randomized algorithms $\mathcal{A}_i$ is $(\alpha, \varepsilon_i(\alpha, \tau))$-SRDP over $\mathcal{L}$, then the composition $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is $(\alpha, \sum_{i=1}^k \varepsilon_i(\alpha, \tau))$-SRDP over $\mathcal{L}$.*
- ***Post-processing*** *Let $\alpha > 0, \varepsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and f be an arbitrary algorithm. For any $\tau > 0$, if $\mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP over $\mathcal{L}$, then $f \circ \mathcal{A}$ is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP over $\mathcal{L}$.*

Proof. **Composition:** For $i \in [k]$ and any two datasets $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, let ν_i and ν_i' be the distribution of $\mathcal{A}_i(S)$ and $\mathcal{A}_i(S')$ respectively. Then, using to the independence between $\mathcal{A}_i(S)$ and $\mathcal{A}_{i+1}(S)$, denote the joint distribution of $\mathcal{A}_i(S)$ and $\mathcal{A}_{i+1}(S)$ as $\nu_i \times \nu_{i+1}$ and the joint distribution of $\mathcal{A}_i(S')$ and $\mathcal{A}_{i+1}(S')$ as $\nu_i' \times \nu_{i+1}'$.

Then, we prove the composition property of by upper bounding $D_\alpha\left(\mathcal{A}_1(S), \ldots, \mathcal{A}_k(S) || \mathcal{A}_1(S'), \ldots, \mathcal{A}_k(S')\right)$ for $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$. To achieve this, we employ the following lemma on additivity of Rényi divergence (Lemma A).

Lemma A (Additivity of Renyi divergence (van Erven & Harremoës, 2012))**.** *For $\alpha > 1$ and distributions $\nu_1, \nu_2, \nu_1', \nu_2'$,*

$$D_\alpha\left(\nu_1 \times \nu_2 || \nu_1' \times \nu_2'\right) = D_\alpha\left(\nu_1 || \nu_1'\right) + D_\alpha\left(\nu_2 || \nu_2'\right).$$

Applying Lemma A at step (a),

$$\begin{aligned} D_\alpha\left(\mathcal{A}_1(S), \ldots, \mathcal{A}_k(S) || \mathcal{A}_1(S'), \ldots, \mathcal{A}_k(S')\right) &= D_\alpha(\nu_1 \times \ldots \times \nu_k || \nu_1' \times \ldots \times \nu_k') \\ &\overset{(a)}{\leq} D_\alpha\left(\nu_1 || \nu_1'\right) + D_\alpha\left(\nu_2 \times \ldots \times \nu_k || \nu_2' \times \ldots \times \nu_k'\right) \\ &\overset{(b)}{=} \varepsilon_1(\alpha, \tau) + D_\alpha\left(\nu_2 \times \ldots \times \nu_k || \nu_2' \times \ldots \times \nu_k'\right) \\ &\overset{(c)}{=} \varepsilon_1(\alpha, \tau) + \varepsilon_2(\alpha, \tau) + \ldots \\ &\leq \sum_{i=1}^k \varepsilon_i(\alpha, \tau) \end{aligned} \tag{5}$$

where step (b) follows from the fact that $\mathcal{A}_1$ is SRDP with parameter $\varepsilon_1(\alpha, \tau)$, and step (c) is obtained by decomposing $D_\alpha(\nu_2 \times \ldots \times \nu_k || \nu'_2 \times \ldots \times \nu'_{i+1})$ in the similar manner as step (a) and applying Lemma A iteratively. This completes the proof for composition of SRDP.

Post-processing: For any $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, let $X = \mathcal{A}(S)$, $X' = \mathcal{A}(S')$ be random variables with probability distribution ν_X, ν'_X, and $Y = f(\mathcal{A}(S)), Y' = f(\mathcal{A}(S'))$ be random variables with distributions ν_Y and ν'_Y. For any value x, we denote the conditional distribution of $Y = f(x)$ and $Y' = f(x)$ by $\nu_{Y|x}$ and $\nu'_{Y|x}$ respectively. We note that $\nu_{Y|x} = \nu'_{Y|x}$ by definition.

We will upper bound the Rényi divergence between $f(\mathcal{A}(S))$ and $f(\mathcal{A}(S'))$,

$$\begin{aligned} D_\alpha\left(f(\mathcal{A}(S)) || f(\mathcal{A}(S'))\right) &= \frac{1}{\alpha - 1} \log \mathbb{E}_{y \sim \nu'_Y} \left[\frac{\nu_Y(y)}{\nu'_Y(y)}\right]^\alpha \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{y \sim \nu'_Y} \left[\mathbb{E}_{x \sim \nu'_X} \left(\frac{\nu_{Y|x}(y)\nu_X(x)}{\nu'_{Y|x}(y)\nu'_X(x)}\right)\right]^\alpha \\ &\overset{(a)}{\leq} \frac{1}{\alpha - 1} \log \mathbb{E}_{y \sim \nu'_Y} \left[\mathbb{E}_{x \sim \nu'_X} \left(\frac{\nu_{Y|x}(y)\nu_X(x)}{\nu'_{Y|x}(y)\nu'_X(x)}\right)^\alpha\right] \\ &\overset{(b)}{=} \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim \nu'_X} \left[\frac{\nu_X(x)}{\nu'_X(x)}\right]^\alpha = D_\alpha\left(\mathcal{A}(S) || \mathcal{A}(S')\right) \end{aligned} \tag{6}$$

where (a) follows from Jensen's inequality and the convexity of $h(x) = x^\alpha$ for $\alpha \geq 2$ and $x \geq 0$, and (b) follows from the fact that $\nu_{Y|x} = \nu'_{Y|x}$ for fixed x. This completes the proof. □

We present a general version of privacy amplification by subsampling for SRDP in Lemma 3. We consider a subsampling mechanism that uniformly selects B elements from a dataset with replacement. Lemma 3 implies that the SRDP parameter of a SRDP algorithm decreases by $O(1/\sqrt{n})$ when $B = 1$ for datasets of size n.

Lemma 3 (Privacy amplification by subsampling of SRDP)**.** *For $B \geq 1$, let $\pi : \mathcal{X}^n \to \mathcal{X}^B$ be the subsampling mechanism that samples B elements from the dataset S of size n uniformly at random. Let $\mathcal{A}$ be a randomized algorithm that is $(\alpha, \varepsilon(\alpha, \tau))$-SRDP. For α, τ such that $\varepsilon(\alpha, \tau) \leq \frac{1}{\alpha - 1}$ and for any integer $\frac{1}{B} \leq k \leq \frac{n-1}{B}$, $\mathcal{A} \circ \pi$ satisfies $(\alpha, \varepsilon'(\alpha, \tau))$-SRDP, where*

$$\varepsilon'(\alpha, \tau) = 2\left(1 - \frac{kB - 1}{n}\right)^B \varepsilon\left(\alpha, \frac{\tau}{k}\right) + 2\left(1 - \left(1 - \frac{kB - 1}{n}\right)^B\right) \varepsilon\left(\alpha, \tau\right).$$

Proof. Let $S, S' \subset \mathcal{L}$ with $d_{12}(S, S') \leq \tau$. First, order the points in S, S' as $S = \{z_i\}_{i=1}^n$ and $S' = \{z'_i\}_{i=1}^n$ such that $\sum_{i=1}^n \|z_i - z'_i\|_2 \leq \tau$.

Let m be a fixed integer. Let $\mathcal{I} = \{i_1, i_2, \ldots\}$ be the set of indices such that for any $j \in \mathcal{I}$, $\left\|z_j - z'_j\right\| \geq \frac{\tau}{m}$ holds. For $1 \leq m \leq n - 1$, if $d_{12}(S, S') \leq \tau$, there can be at most $m - 1$ indices in $\mathcal{I}$.

Thus, for an index i sampled from $[n]$ uniformly at random, we have

$$\mathbb{P}_{i \sim \text{Unif}([n])}\left[\|z_i - z'_i\| \leq \frac{\tau}{m}\right] = 1 - \mathbb{P}_{i \sim \text{Unif}([n])}\left[\|z_i - z'_i\| \geq \frac{\tau}{m}\right] \geq 1 - \frac{k - 1}{n}. \tag{7}$$

Let $J = \{i_1, ..., i_B\}$ be a set where each index i_j is sampled independently and uniformly from $[n]$. For any integer $1 \leq m \leq n - 1$, we compute the probability that $d_{12}(S_J, S'_J) \leq \frac{\tau}{k}$.

$$\mathbb{P}_{J \sim \text{Unif}([n])^B}\left[\sum_{i \in J} \|z_i - z'_i\|_2 \leq \frac{B\tau}{m}\right] \overset{(a)}{\geq} \mathbb{P}_{J \sim \text{Unif}([n])^B}\left[\forall i \in J, \|z_i - z'_i\|_2 \leq \frac{\tau}{m}\right] \overset{(b)}{\geq} \left(1 - \frac{m - 1}{n}\right)^B \tag{8}$$

where step (a) follows as $\forall i \in J, \|z_i - z'_i\|_2 \leq \frac{\tau}{m}$ implies $\sum_{i \in J} \|z_i - z'_i\|_2 \leq \frac{B\tau}{m}$, step (b) follows by Equation (7) and the independence of each index in J.

Replacing $m = kB$, we have for $1 \leq kB \leq n-1$, i.e.for $\frac{1}{B} \leq k \leq \frac{n-1}{B}$,

$$\mathbb{P}_{J\sim\text{Unif}([n])^B}\left[\sum_{i\in J}\|z_i - z_i'\|_2 \leq \frac{\tau}{k}\right] \geq \left(1 - \frac{kB-1}{n}\right)^B \tag{9}$$

Next, we apply the weak convexity of Rényi divergence (Lemma B) to get the desired bound.

Lemma B (Weak convexity of Renyi divergence, Lemma 25 in Feldman et al. (2018))**.** *Let μ_1, μ_2 and υ_1, υ_2 be probability distributions over the same domain such that $D_\alpha(\mu_i||\upsilon_i) \leq \frac{c}{\alpha-1}$ for $i \in \{1,2\}$ and $c \in (0,1]$. For $\beta \in (0,1)$,*

$$D_\alpha(\beta\mu_1 + (1-\beta)\mu_2||\beta\upsilon_1 + (1-\beta)\upsilon_2) \leq (1+c)\left[\beta D_\alpha(\mu_1||\upsilon_1) + (1-\beta)D_\alpha(\mu_2||\upsilon_2)\right].$$

Let $\mathcal{E}$ denote the event that the L_{12} distance between S_J and S_J' is smaller than $\frac{\tau}{k}$, i.e.$\sum_{i\in J}\|z_i - z_i'\|_2 \leq \frac{\tau}{k}$, for $J \sim \text{Unif}([n])^B$. Then, Equation (9) implies that the probability of $\mathcal{E}$ is at least $\left(1 - \frac{kB-1}{n}\right)^B$.

Let μ_1, υ_1 be the distribution of $\mathcal{A}(S_J)$ and $\mathcal{A}(S_J')$ conditional on $\mathcal{E}$ occurring, and let μ_2, υ_2 be the distribution of $\mathcal{A}(S_J)$ and $\mathcal{A}(S_J')$ conditional on $\mathcal{E}$ does not occur. By our assumption on τ, α, $\varepsilon(\alpha,\tau) \leq \frac{1}{\alpha-1}$ and $c = 1$. Applying Lemma B with $c = 1$ in step (a),

$$\begin{aligned}
D_\alpha(\mathcal{A}(S_J)||\mathcal{A}(S_J')) &= D_\alpha(\mathbb{P}[\mathcal{E}]\,\mathcal{A}(S_J|\mathcal{E}) + \mathbb{P}[\neg\mathcal{E}]\,\mathcal{A}(S_J|\neg\mathcal{E})||\mathbb{P}[\mathcal{E}]\,\mathcal{A}(S_J'|\mathcal{E}) + \mathbb{P}[\neg\mathcal{E}]\,\mathcal{A}(S_J'|\neg\mathcal{E})) \\
&\overset{(a)}{\leq} 2\mathbb{P}[\mathcal{E}]\,\varepsilon\left(\alpha, \frac{\tau}{k}\right) + 2\mathbb{P}[\neg\mathcal{E}]\,\varepsilon(\alpha,\tau) \\
&\overset{(b)}{\leq} 2\left(1 - \frac{kB-1}{n}\right)^B \varepsilon\left(\alpha, \frac{\tau}{k}\right) + 2\left(1 - \left(1 - \frac{kB-1}{n}\right)^B\right)\varepsilon(\alpha,\tau),
\end{aligned}$$

In step (b), we note that $2\mathbb{P}[\mathcal{E}]\,\varepsilon\left(\alpha, \frac{\tau}{k}\right) + 2\mathbb{P}[\neg\mathcal{E}]\,\varepsilon(\alpha,\tau)$ increases with $\mathbb{P}[\mathcal{E}]$ by the fact that $\varepsilon\left(\alpha, \frac{\tau}{k}\right) \leq \varepsilon(\alpha,\tau)$. Then, we obtain the upper bound by substitute in the lower bound of $\mathbb{P}[\mathcal{E}]$ from Equation (9). This completes the proof. □

However, under additional structural assumption on the pair of datasets that needs to remain indistinguishable, we can establish a more effective amplification of smooth RDP through subsampling. In Appendix A.2.2, we present the additional assumption with Definition 7. We also apply the stronger amplification in the proof of overall privacy guarantees for DP-SGD with subsampling.

A.2. Proof of Theorem 1(Derivation of Table 1)

In this section, we explore results regarding the SRDP parameters for various DP mechanisms. We start with the formal definitions of two fundamental assumptions, Lipschitzness and smoothness, in Definition 5 and 6. Then, we provide the proof of Theorem 1 (SRDP parameters in Table 1). For clarity, we present the proof for each private mechanism separately in Appendices A.2.1 and A.2.2.

Definition 5 (Lipschitzness)**.** *A function $f : \mathcal{K} \to \mathbb{R}$ is L-Lipschitz over the domain $\mathcal{K} \subset \mathcal{X}$ with respect to the distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ if for any $X, X' \in \mathcal{K}$, $|f(X) - f(X')| \leq Ld(X, X')$.*

Definition 6 (Smoothness)**.** *A loss function $\ell : \Theta \times \mathcal{K} \to \mathbb{R}_+$ is μ-smooth if for all $\theta \in \Theta$ and $x, x' \in \mathcal{K}$, $\|\nabla_\theta \ell(\theta, x) - \nabla_\theta \ell(\theta, x)\|_2 \leq \mu\|x - x'\|_2$.*

We note that Definition 6 is different from the usual definition of smoothness in the literature that requires $\|\nabla_\theta \ell(\theta, x) - \nabla_\theta \ell(\theta', x)\|_2 \leq \mu\|\theta - \theta'\|_2$ for all $x \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ (Chaudhuri et al., 2011; Feldman et al., 2018). However, many loss functions, including square loss and logistic loss, are smooth on common model classes based on Definition 6.

A.2.1. Output Perturbation and Random Sampling methods

This section provides the detailed theorems and proofs for Smooth RDP parameters of output perturbation methods (Gaussian and Laplace mechanism) and random sampling methods (Exponential mechanism) in Theorem 10 to 12 respectively, as stated in Table 1.

Theorem 10 (Theorem 1 for Gaussian mechanism)**.** *For an L-Lipschitz function f with global sensitivity Δ_f and the privacy parameter $\varepsilon > 0$, let $\mathcal{M}_G(S) = f(S) + \mathcal{N}\left(0, \Delta_f^2/\varepsilon^2\right)$ denote the Gaussian mechanism. For a dataset collection $\mathcal{L}$ and for $\tau > 0$, $\mathcal{M}_G$ is $\left(\alpha, \frac{\alpha\varepsilon^2}{2}\right)$-RDP and $\left(\alpha, \frac{\alpha L^2\tau^2\varepsilon^2}{2}\right)$-SRDP for $\alpha \geq 1$.*

Proof. We first show that $\mathcal{M}_G$ is $\left(\alpha, \frac{\alpha\varepsilon^2}{2}\right)$-RDP. For any neighboring datasets S, S' with $d_H(S, S') = 1$,

$$\begin{aligned} D_\alpha\left(\mathcal{M}_G(S)||\mathcal{M}_G(S')\right) &= D_\alpha\left(f(S) + \mathcal{N}\left(0, \Delta_f^2/\varepsilon^2\right)||f(S') + \mathcal{N}\left(0, \Delta_f^2/\varepsilon^2\right)\right) \\ &= D_\alpha\left(\mathcal{N}\left(f(S), \Delta_f^2/\varepsilon^2\right)||\mathcal{N}\left(f(S'), \Delta_f^2/\varepsilon^2\right)\right) \end{aligned}$$

We apply Lemma C to upper bound the Rényi divergence between two Gaussian random variable with same variance.

Lemma C (Corrolary 3 in Mironov (2017))**.** *For any two Gaussian distributions with the same variance σ^2 but different means μ_0, μ_1, denoted by $\mathcal{N}\left(\mu_0, \sigma^2\right)$ and $\mathcal{N}(\mu_1, \sigma^2)$, the following holds,*

$$D_\alpha\left(\mathcal{N}(\mu_0, \sigma^2)||\mathcal{N}(\mu_1, \sigma^2)\right) \leq \frac{\alpha\left(\mu_0 - \mu_1\right)^2}{2\sigma^2}.$$

Hence,

$$D_\alpha\left(\mathcal{M}_G(S)||\mathcal{M}_G(S')\right) \overset{(a)}{\leq} \sup_{(S,S')\in\mathcal{S}} \frac{\alpha\left(f(S) - f(S')\right)^2}{2\Delta_f^2/\varepsilon^2} \overset{(b)}{\leq} \frac{\alpha\varepsilon^2}{2},$$

where (a) follows from Lemma C and (b) follows due to the fact that $|f(S) - f(S')| \leq \Delta_f$ by the definition of global sensitivity.

Then, we show that $\mathcal{M}_G$ is $\left(\alpha, \frac{\alpha L^2\tau^2\varepsilon^2}{2}\right)$-SRDP over the dataset collection $\mathcal{L}$. Specifically, we will show that for any $\tau > 0$, for any two datasets $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, $D_\alpha\left(\mathcal{A}(S)||\mathcal{A}(S')\right) \leq \frac{\alpha L^2\tau^2\varepsilon^2}{2\Delta_f^2}$. Let $S, S' \in \mathcal{L}$ such that $d_{12}(S, S') \leq \tau$, we have

$$\begin{aligned} D_\alpha\left(\mathcal{M}_G(S)||\mathcal{M}_G(S')\right) &= D_\alpha\left(\mathcal{N}\left(f(S), \Delta_f^2/\varepsilon^2\right)||\mathcal{N}\left(f(S'), \Delta_f^2/\varepsilon^2\right)\right) \\ &\overset{(a)}{\leq} \frac{\alpha\left(f(S) - f(S')\right)^2}{2\Delta_f^2/\varepsilon^2} \\ &\overset{(b)}{\leq} \frac{\alpha\left(L\left\|S - S'\right\|_F\right)^2\varepsilon^2}{2\Delta_f^2} \overset{(c)}{\leq} \frac{\alpha L^2\tau^2\varepsilon^2}{2\Delta_f^2} \end{aligned} \tag{10}$$

where step (a) follows by Lemma C, step (b) follows by the Lipschitzness of the function f with respect to the Frobenius norm, and step (c) follows by the fact that Frobenius norm is always smaller than L_{12} norm, i.e.$\left\|S - S'\right\|_F \leq d_{12}(S, S')$. □

Theorem 11 (Theorem 1 for Laplace mechanism)**.** *For an L-Lipschitz function f with global sensitivity Δ_f and the privacy parameter ε, let $\mathcal{M}_L(S) = f(S) + Lap\left(\Delta_f/\varepsilon\right)$ denote the Laplace mechanism. For a dataset collection $\mathcal{L}$ and for $\tau > 0$, $\mathcal{M}_L$ is (α, ε)-RDP and $\left(\alpha, \frac{L\tau}{\Delta_f}\varepsilon\right)$-SRDP for $\alpha \geq 1$.*

Proof. Laplace mechanism $\mathcal{M}_L$ is $(\varepsilon, 0)$-DP (Dwork et al., 2006). By the equivalence of $(\varepsilon, 0)$-DP and (∞, ε)-RDP (Mironov, 2017), $\mathcal{M}_L$ is (∞, ε)-RDP. This implies that $\mathcal{M}_L$ is (α, ε)-RDP for any $1 \leq \alpha \leq \infty$ by the monotonicity of Rényi divergence (Lemma D).

Lemma D (Monotonicity of Rényi divergence (Mironov, 2017))**.** *For $1 \leq \alpha < \beta$, for any two distributions P, Q, the following holds*

$$D_\alpha(P||Q) \leq D_\beta(P||Q).$$

Next, we will show $\mathcal{M}_L$ is $(\alpha, {}^{L\tau\varepsilon}/_{\Delta_f})$-SRDP over the set $\mathcal{L}$. Specifically, we first show that $\mathcal{M}_L$ is $(\infty, {}^{L\tau\varepsilon}/_{\Delta_f}))$-SRDP. Using the monotonicity of Rényi divergence (Lemma D), we can propagate this property to any $\alpha < \infty$.

Let h be any output in the output space of f. For any two datasets $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$,

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}_L(S) = h)}{\mathbb{P}(\mathcal{M}_L(S') = h)} &= \frac{\frac{\varepsilon}{\Delta_f} \exp\left(-\frac{|h - f(S)|\varepsilon}{\Delta_f}\right)}{\frac{\varepsilon}{\Delta_f} \exp\left(-\frac{|h - f(S')|\varepsilon}{\Delta_f}\right)} \\ &\leq \exp\left(\frac{\varepsilon}{\Delta_f}\left(|-f(S') + f(S)|\right)\right) \\ &\overset{(a)}{\leq} \exp\left(\frac{\varepsilon}{\Delta_f} L \left\|S - S'\right\|_F\right) \overset{(b)}{\leq} \exp\left(\frac{L\tau\varepsilon}{\Delta_f}\right).\end{aligned} \tag{11}$$

where (a) follows by the L-Lipschitzness of the f with respect to Frobenius norm and (b) follows from the fact that $\left\|S - S'\right\|_F \leq d_{12}(S, S') \leq \tau$.

This implies the output distributions have bounded infinite Rényi divergence, i.e.

$$D_\infty(\mathcal{M}_L(S) || \mathcal{M}_L(S')) = \max_{h \in \mathcal{H}} \log \frac{\mathcal{P}(\mathcal{M}_L(S) = h)}{\mathcal{P}(\mathcal{M}_L(S') = h)} \leq \frac{L\tau\varepsilon}{\Delta_f}, \tag{12}$$

where the last inequality follows from Equation (11).

By Monotocity of Rényi divergence (Lemma D), for any $\alpha \geq 1$, the α-Rényi divergence between the output distribution of Laplace mechanism on any two datasets $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$ is also bounded by $L\tau\varepsilon/\Delta_f$. This concludes the proof.

□

Theorem 12 (Theorem 1 for Exponential mechanism)**.** *For an L-Lipschitz score function Q with global sensitivity Δ_Q, privacy parameter ε, and a dataset collection $\mathcal{L}$, the exponential mechanism, and for any $\tau > 0$, $\mathcal{M}_E$ is (α, ε)-RDP and $(\alpha, L\tau\varepsilon/\Delta_Q)$-SRDP over $\mathcal{L}$ for any $\alpha \geq 1$.*

Proof. The proof for RDP of Exponential mechanism is similar to that of Laplace mechanism. As the exponential mechanism $\mathcal{M}_E$ is $(\varepsilon, 0)$-DP (Dwork et al., 2006). By the equivalence of $(\varepsilon, 0)$-DP and (∞, ε)-RDP (Mironov, 2017), $\mathcal{M}_E$ is (∞, ε)-RDP. By the monotonicity of Rényi divergence (Lemma D), $\mathcal{M}_E$ is also (α, ε)-RDP for any $\alpha \geq 1$.

To show that $\mathcal{M}_E$ is $\left(\alpha, \frac{L\tau\varepsilon}{\Delta_Q}\right)$-SRDP over the set $\mathcal{L}$, we first show that $\mathcal{M}_E$ is $(\infty, L\tau\varepsilon/\Delta_Q)$-SRDP over $\mathcal{L}$ and then apply the monotonicity of Rényi divergence (Lemma D).

For any $S, S' \in \mathcal{L}$ such that $d_{12}(S, S') \leq \tau$ and any output h in the output space,

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}_E(S) = h)}{\mathbb{P}\left(\mathcal{M}_E(S') = h\right)} &= \frac{\exp\left(\frac{Q(S,h)\varepsilon}{2\Delta_Q}\right)}{\exp\left(\frac{Q(S',h)\varepsilon}{2\Delta_Q}\right)} \frac{\int_{h \in \mathcal{H}} \exp \frac{Q(S',h)\varepsilon}{2\Delta_Q} dh}{\int_{h \in \mathcal{H}} \exp \frac{Q(S,h)\varepsilon}{2\Delta_Q} dh} \\ &\leq \frac{\exp\left(\frac{Q(S,h)\varepsilon}{2\Delta_Q}\right)}{\exp\left(\frac{Q(S',h)\varepsilon}{2\Delta_Q}\right)} \frac{\int_{h \in \mathcal{H}} \exp \frac{Q(S',h) - Q(S,h)\varepsilon}{2\Delta_Q} \exp \frac{Q(S,h)\varepsilon}{2\Delta_Q} dh}{\int_{h \in \mathcal{H}} \exp \frac{Q(S,h)\varepsilon}{2\Delta_Q} dh} \\ &\leq \exp\left(\frac{2 \max_{h \in \mathcal{H}} |Q(S,h) - Q(S',h)|\, \varepsilon}{2\Delta_Q}\right) \overset{(a)}{\leq} \exp\left(\frac{L\tau\varepsilon}{\Delta_Q}\right)\end{aligned} \tag{13}$$

where step (a) follows by the L-Lipschitzness of the score function $Q(S, h)$ for all $h \in \mathcal{H}$ and the fact that $\left\|S - S'\right\|_F \leq d_{12}(S, S') \leq \tau$.

This implies bounded infinite order Rényi divergence between the two output distributions, i.e.

$$D_\infty(\mathcal{M}_E(S) || \mathcal{M}_E(S')) = \max_{h \in \mathcal{H}} \log \frac{\mathcal{P}(\mathcal{M}_E(S) = h)}{\mathcal{P}(\mathcal{M}_E(S') = h)} \leq \frac{L\tau\varepsilon}{\Delta_Q}, \tag{14}$$

where the last inequality follows from Equation (13).

As shown in Equation (14), $\mathcal{M}_E$ is $(\infty, {}^{L\tau\varepsilon}/_{\Delta_Q})$-SRDP. This implies that $\mathcal{M}_E$ is $(\alpha, {}^{L\tau\varepsilon}/_{\Delta_Q})$-SRDP for any $\alpha \geq 1$ by monotonicity of Rényi divergence (Lemma D). This completes the proof. □

A.2.2. Gradient-based methods

In this section, we start with providing the formal definitions of gradient-based methods mentioned in Section 2.2 and evaluated in Table 1, including DP-GD, DP-SGD with subsampling, and DP-SGD with iteration. Then, we provide the detailed theorems and proofs for their Smooth RDP parameters, as stated in Table 1, in Theorem 13 to 15.

DP-GD Given a L-Lipschitz loss function ℓ, DP-GD (Bassily et al., 2014; Song et al., 2021), denoted by $\mathcal{A}_{\mathrm{DP-GD}}$ starts from some random initialization w_0 in the parameter space $\mathcal{H}$ and conducts projected gradient descent for T iterations as $w_{t+1} = \Pi_{\mathcal{H}}(w_t - \eta\tilde{g}_t)$ with the noisy gradient on the whole dataset $\tilde{g}_t$ defined as $\tilde{g}_t = \frac{1}{n}\sum_{i=1}^{n} \nabla_w \ell(w_{t-1}, x) + \mathcal{N}(0, \sigma^2)$, and outputs the average parameter over the T iterations $\frac{1}{T}\sum_{t=1}^{T} w_t$.

DP-SGD with subsampling Another commonly used method is DP-SGD with subsampling (Abadi et al., 2016; Bassily et al., 2014), denoted as $\mathcal{A}_{\mathrm{SGD-samp}}$. In each gradient descent step, DP-SGD with subsampling first draws a uniform subsample of size B from the dataset without replacement. The gradient update is then performed by adding noise to the average gradient derived from this subsample. In contrast, DP-GD computes the average gradient of the entire dataset for its updates.

DP-SGD with iteration Differentially Private Fixed Gradient Descent (DP-FGD), denoted by $\mathcal{A}_{\mathrm{FGD}}^{(T)}$, is a variant of DP-SGD. It processes the data points in a fixed order — the gradient at the tth step is calculates using the tth point in the dataset— and outputs the parameter obtained after the Tth iteration. *DP-SGD with iteration* (Feldman et al., 2018), denoted as $\mathcal{A}_{\mathrm{SGD-iter}}$, uses DP-FGD as a base procedure. It takes an extra parameter κ_τ and first uniformly samples an integer T from $[n - \kappa_\tau + 1] = \{1, \ldots, n - \kappa_\tau + 1\}$ and releases the output from $\mathcal{A}_{\mathrm{FGD}}^{(T)}$, i.e., it releases the result from the Tth iteration. While DP-SGD with iteration cannot take advantage of privacy amplification by subsampling for its privacy analysis, it relies on privacy amplification with iteration (Feldman et al., 2018) to achieve a comparable privacy guarantee to that of $\mathcal{A}_{\mathrm{SGD-samp}}$.

Theorem 13 (Theorem 1 for DP-GD)**.** *For an L-Lipschitz and μ-smooth loss function ℓ, privacy parameter ε, dataset collection $\mathcal{L}$, and $\sigma = \frac{L\sqrt{T}}{\varepsilon n}$, $\mathcal{A}_{\mathrm{GD}}$ is $\left(\alpha, \frac{\alpha\varepsilon^2}{2}\right)$-RDP and $\left(\alpha, \frac{\alpha\mu^2\tau^2\varepsilon^2}{2L^2}\right)$-SRDP over $\mathcal{L}$ for any $\alpha \geq 1, \tau > 0$.*

Proof. We denote each noisy gradient descent step as $\theta_{t+1} = \theta_t - \eta g(\theta_t, S)$ where $g(\theta_t, S) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\theta, S_i) + \mathcal{N}(0, {}^{L\sqrt{T}}/_{\varepsilon n})$ is the noisy gradient. To show that each gradient descent step is RDP, it suffices to show the gradient operator g is $(\alpha, {}^{\alpha\varepsilon^2}/_2)$-RDP by the fact that RDP is preserved by post-processing (Lemma E).

Lemma E (Post-processing of RDP (Mironov, 2017))**.** *Let $\alpha > 0, \varepsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and f be an arbitrary algorithm. For any $\ell > 0$, if $\mathcal{A}$ is $(\alpha, \varepsilon(\alpha))$-RDP, then $f \circ \mathcal{A}$ is $(\alpha, \varepsilon(\alpha))$-RDP.*

Now, we show the gradient operator g is $\left(\alpha, \frac{\alpha\varepsilon^2}{2}\right)$- RDP. By the Lipschitzness of the loss function, for all θ, S_i, we have $\nabla_\theta(\ell(\theta, S_i)) \leq L$. Thus, the global sensitivity of the average gradient $\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\theta, S_i)$ is upper bounded by $\frac{L}{n}$. For any neighboring datasets S, S' such that $d_H(S, S') = 1$, for any θ,

$$D_\alpha\left(g(\theta, S) || g(\theta, S')\right) \overset{(a)}{\leq} \frac{\alpha\left(\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\theta, S_i) - \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\theta, S_i')\right)^2}{2\frac{L^2 T}{\varepsilon^2 n^2}} \overset{(b)}{\leq} \frac{2\alpha\varepsilon^2}{T} \tag{15}$$

where step (a) follows by the Rényi divergence of Gaussian mechanism (Lemma C), and step (b) follows by the fact that the sensitivity of each gradient estimation is $\frac{L}{n}$. This implies that each gradient estimation g and thus, each gradient descent step, are $(\alpha, {}^{2\alpha\varepsilon^2}/_{2T})$-RDP. Applying the composition theorem of RDP (Lemma F), we can show that DP-GD with T gradient descent step is $\left(\alpha, 2\alpha\varepsilon^2\right)$-RDP.

Lemma F (Composition of RDP, Proposition 1 in Mironov (2017))**.** *For $\alpha \geq 1$, let f be (α, ε_1)-RDP and g be (α, ε_2)-RDP. Then, the mechanism (f, g) is $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP.*

Then, for any $S, S' \in \mathcal{L}$ such that $d_{12}(S, S') \leq \tau$, we show that the Rényi divergence between the gradient estimate with S and S', $D_\alpha\left(g(\theta, S)||g(\theta, S')\right)$, is upper bounded. By the definition of gradient operator g and $\sigma = \frac{L\sqrt{T}}{\varepsilon n}$, we have

$$\begin{aligned} D_\alpha\left(g(\theta, S)||g(\theta, S')\right) &= D_\alpha\left(\mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ell(\theta, S_i), \frac{L^2 T}{\varepsilon^2 n^2}\right)||\mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ell(\theta, S_i'), \frac{L^2 T}{\varepsilon^2 n^2}\right)\right) \\ &\overset{(a)}{\leq} \frac{\alpha\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ell(\theta, S_i) - \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ell(\theta, S_i')\right\|_2^2}{2\frac{L^2 T}{n^2\varepsilon^2}} \\ &\overset{(b)}{\leq} \frac{\alpha\varepsilon^2\mu^2\left(\sum_{i=1}^{n}\|S_i - S_i'\|_2\right)^2}{2L^2 T} \overset{(c)}{=} \frac{\alpha\varepsilon^2\mu^2\tau^2}{2L^2 T} \end{aligned} \tag{16}$$

where step (a) follows from Lemma C, step (b) follows from the smoothness assumption of the loss function, i.e.$\|\nabla_\theta \ell(\theta, S_i) - \nabla_\theta \ell(\theta, S_i')\|_2 \leq \mu\|S_i - S_i'\|_2$, and step (c) follows from the definition of L_{12} distance and the fact that $d_{12}(S, S') \leq \tau$.

Applying composition of SRDP (Lemma 2) over the T gradient descent steps concludes the proof. □

For establishing the overall privacy guarantees for DP-SGD with subsampling and iteration, we introduce two properties of the dataset collection: the inverse point-wise distance (Definition 7) and the maximum distance (Definition 8). Then, we present the overall privacy guarantee for DP-SGD with subsampling and iteration in Theorem 13 and Theorem 15, with additional assumptions on the inverse point-wise distance and τ-constrained maximum distance of the dataset collection.

Definition 7 (Inverse point-wise distance of a dataset collection $\mathcal{L}$)**.** *Let k be the maximum integer such that for every pair of datasets $S_1 = \left\{S_1^i\right\}_{i=1}^n \in \mathcal{L}$ and $S_2 = \left\{S_2^i\right\}_{i=1}^n \in \mathcal{L}$ and for all $i \in [n]$, $\left\|S_1^i - S_2^i\right\|_2 \leq \frac{d_{12}(S_1, S_2)}{k}$. The inverse point-wise distance γ of a dataset collection $\mathcal{L}$ is defined as $\gamma = \frac{n}{k}$.*

Definition 8 (τ-constrained maximum distance of a dataset collection $\mathcal{L}$)**.** *For $\tau > 0$, the τ-constrained maximum distance κ_τ of a dataset collection $\mathcal{L}$ is defined as the maximum Hamming distance between any two datasets $S_1, S_2 \in \mathcal{L}$ such that $d_{12}(S_1, S_2) \leq \tau$.*

Theorem 14 (Theorem 1 for DP-SGD with subsampling)**.** *For any L-Lipschitz and μ-smooth loss function ℓ, privacy parameter ε, dataset collection $\mathcal{L}$ with inverse point-wise distance γ, and $\sigma = \Omega\left(\frac{L\sqrt{T}}{\varepsilon n}\right)$, $\mathcal{A}_{\mathrm{SGD-samp}}$ is $\left(\alpha, \frac{\alpha^2\varepsilon^2}{2}\right)$-RDP and $\left(\alpha, \frac{\alpha\mu^2\tau^2\varepsilon^2\gamma^2}{2L^2}\right)$-SRDP over $\mathcal{L}$ for any $\tau > 0$ and $1 \leq \alpha \leq \min\left\{\frac{\sqrt{T}}{\varepsilon}, \frac{L^2 T}{\varepsilon^2 n^2}\log\frac{n^2\varepsilon}{L\sqrt{T}}\right\}$.*

Proof. We let π_u be a subsampling mechanism that uniformly samples one point from the dataset. We denote each gradient descent step as $\theta_{t+1} = \theta_t - \eta g(\theta_t, S)$ where $g(\theta_t, S) = \nabla_\theta \ell(\theta, \pi_u(S)) + \mathcal{N}(0, \sigma^2)$ represents the gradient estimate.

We will show that each gradient estimate g is $(\alpha, {}^{\alpha^2\varepsilon^2}/_{2T})$-RDP and thus, each gradient descent step satisfies RDP with the same parameters by post-processing theorem of RDP (Lemma E). Then, by the composition theorem for RDP (Lemma F) over the T gradient descent steps, we can conclude that $\mathcal{A}_{\mathrm{SGD-samp}}$ is $(\alpha, {}^{\alpha^2\varepsilon^2}/_{2})$-RDP.

First, for each gradient estimate $g(\theta, \pi_u(S))$, we apply Lemma G to upper bound the Rényi divergence for each gradient step in Equation (18).

Lemma G (Lemma 3 in Abadi et al. (2016))**.** *Suppose that $f : \mathcal{X} \to \mathbb{R}^p$ is a function with $\|f(\cdot)\|_2 \leq L$. Assume $\sigma \geq 1$, and let $i \sim \mathrm{Unif}([n])$ represent a uniform random variable over the integers in the set $[n] = 1, 2, \ldots, n$. Then, for any positive integer $1 \leq \alpha \leq \sigma^2 \ln\frac{n}{\sigma}$ and any pair of neighboring datasets S, S', the mechanism $\mathcal{M}(S) = f(S_i) + \mathcal{N}(0, \sigma^2 I)$ satisfies*

$$D_\alpha\left(\mathcal{M}(S)||\mathcal{M}(S')\right) \leq \underbrace{\frac{L^2\alpha(\alpha+1)}{n^2(1 - {}^1/_n)\sigma^2}}_{\text{(I)}} + \underbrace{O\left(\frac{\alpha^3 L^3}{n^3\sigma^3}\right)}_{\text{(II)}}. \tag{17}$$

For neighboring datasets S, S' and for any θ, each gradient estimate satisfies,

$$D_\alpha\left(g(\theta, \pi_u(S))||g(\theta, \pi_u(S'))\right) \overset{(a)}{\leq} \frac{c\alpha^2 L^2}{n^2\sigma^2} \overset{(b)}{\leq} \frac{\alpha^2\varepsilon^2}{2T}, \tag{18}$$

Note that part (II) is smaller than part (I) in Equation (17) for $\alpha \leq \frac{\sqrt{T}}{\varepsilon}$. Hence, step (a) follows by application of Lemma G for some positive constant $c \geq 1$. Step (b) follows by choosing $\sigma = \frac{L\sqrt{cT}}{\varepsilon n}$.

Next, we will show that $\mathcal{A}_{\mathrm{SGD-samp}}$ is $\left(\alpha, \frac{\alpha\tau^2\varepsilon^2\mu^2\gamma^2}{2L^2}\right)$-SRDP. It suffices to show that each gradient estimate is $\left(\alpha, \frac{\alpha\tau^2\varepsilon^2\mu^2\gamma^2}{2TL^2}\right)$-SRDP by the post-processing and composition theorem of SRDP (Lemma 2).

Let k be the maximum integer such that for every pair $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, the point-wise distance is less than τ/k. We note that for any $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, then for any $i \in [n]$,

$$d_{12}(\pi_u(S), \pi_u(S')) = \|S_i - S'_i\|_2 \leq \frac{\tau}{k}. \tag{19}$$

Next, we will upper bound the Rényi divergence between two gradient estimates. Let $i \sim \mathrm{Unif}\left([n]\right)$ be the sampled index. Then, for any $S, S' \in \mathcal{L}$ such that $d_{12}(S, S') \leq \tau$, for any θ,

$$\begin{aligned} D_\alpha\left(g(\theta, \pi_u(S))||g(\theta, \pi_u(S'))\right) &\overset{(a)}{\leq} D_\alpha\left(g(\theta, S_i)||g(\theta, S'_i)\right) \\ &\overset{(b)}{\leq} \frac{\|\nabla_\theta \ell(\theta, S_i) - \nabla_\theta \ell(\theta, S'_i)\|_2^2}{2\sigma^2} \\ &\overset{(c)}{\leq} \frac{\tau^2\alpha\mu^2}{2k^2\sigma^2} \overset{(d)}{\leq} \frac{\alpha\tau^2\varepsilon^2\mu^2}{2TL^2}\left(\frac{n}{k}\right)^2 \overset{(e)}{=} \frac{\alpha\tau^2\varepsilon^2\mu^2\gamma^2}{2TL^2} \end{aligned} \tag{20}$$

In step (a), $i \sim \mathrm{Unif}([n])$ is as defined above. Step (b) follows by the Rényi divergence of Gaussian distributions (Lemma C), and step (c) follows by the smoothness assumption of ℓ, i.e.$\|\nabla_\theta \ell(\theta, S) - \nabla_\theta \ell(\theta, S')\|_2 \leq \mu \|S - S'\|_2$ and Equation (19). Step (d) follows by the substitution of $\sigma = \frac{L\sqrt{cT}}{\varepsilon n}$ and $c \geq 1$, and step (e) follows by the definition of inverse point-wise distance $\gamma = \frac{n}{k}$ as specified in Definition 7.

This completes the proof. □

Remark 1. *Without Definition 7, we can employ the weaker subsampling results for SRDP (Lemma 3). However, that would results in an extra $n^{3/2}$ factor in the SRDP parameter, as Equation (20) will be replaced with Equation (21).*

$$\begin{aligned} D_\alpha\left(g(\theta, \pi_u(S))||g(\theta, \pi_u(S'))\right) &\leq \min_{k \geq 1}\left[\frac{1}{k^2}\left(1 - \frac{k-1}{n}\right) + \frac{k-1}{n}\right]\frac{\alpha\mu^2\tau^2\varepsilon^2 n^2}{L^2T} \\ &\overset{(a)}{\leq} \left(\frac{1}{2n-3} + \frac{(2n-4)\left(\sqrt{2n-3}-1\right)}{n(2n-3)}\right)\frac{\alpha\mu^2\tau^2\varepsilon^2 n^2}{L^2T} \\ &\overset{(b)}{\leq} \frac{2\mu^2\tau^2\varepsilon^2 n^{3/2}\alpha}{L^2T} \end{aligned} \tag{21}$$

where (a) is obtained by choosing $k = \sqrt{2n-3}$ and (b) follows by $\left(\frac{1}{2n-3} + \frac{(2n-4)\left(\sqrt{2n-3}-1\right)}{n(2n-3)}\right) \leq \frac{2}{\sqrt{n}}$.

Theorem 15 (Theorem 1 for DP-SGD with iteration)**.** *For an L-Lipschitz and μ-smooth convex loss function ℓ, let $\beta = \sup_{z \in \mathcal{X}\times\mathcal{Y}} \left\|\nabla_\theta^2 \ell(\theta; z)\right\|$. For any privacy parameter ε, learning rate $\eta \leq 2/\beta$, $\tau \geq 0$, dataset collection $\mathcal{L}$ with τ-constrained maximum distance κ_τ, and $\alpha > 1$ such that $\max\left\{L\sqrt{2(\alpha-1)\alpha}, \tau L\sqrt{2(\alpha-1)\alpha}\right\} < \sqrt{\frac{2\log n}{n}}\frac{8L}{\varepsilon}$, assume $\mathcal{L}$ satisfies that for any $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, the differing points in S, S' are consecutive, then $\mathcal{A}_{\mathrm{SGD-iter}}$ with parameter κ_τ and $\sigma = \sqrt{\frac{2\log n}{n}}\frac{8L}{\varepsilon}$ is $\left(\alpha, \frac{\alpha\varepsilon^2}{2}\right)$-RDP and $\left(\alpha, \frac{\alpha\tau^2\mu^2 n\log(n-\kappa_\tau+2)}{2(n-\kappa_\tau+1)L^2\log n}\right)$-SRDP.*

Proof. We first note that $\mathcal{A}_{\mathrm{SGD-iter}}$ is $(\alpha, \frac{\alpha\varepsilon^2}{2})$-RDP following Theorem 26 in Feldman et al. (2018) (see Lemma H below for completeness).

Lemma H (Privacy guarantee of SGD by iteration, Theorem 26 in Feldman et al. (2018))**.** *Let ℓ be an convex L-Lipschitz and β-smooth loss function over $\mathbb{R}^d$. Then, for any learning rate $\eta \leq \frac{2}{\beta}$, $\alpha > 1$, $\sigma \geq L\sqrt{2(\alpha-1)\alpha}$, $\mathcal{A}_{\mathrm{SGD-iter}}$ satisfies $\left(\alpha, \frac{4\alpha L^2\log n}{n\sigma^2}\right)$-RDP.*

For SRDP, we consider datasets $S, S' \in \mathcal{L}$ such that $d_{12}(S, S') \leq \tau$ and with all differing points appearing consecutively. Without loss of generality, we assume the first differing point has index t. By the assumption on the τ-constrained maximum distance of $\mathcal{L}$, there are in total κ_τ consecutive differing points in S, S', $t \leq n - \kappa_\tau + 1$. In the following, we consider two cases: i) $t > T$, and ii) $t \leq T$.

As the gradient descent step with S and S' are exactly the same before t, in the first case, with $t > T$, we have

$$D_\alpha\left(\mathcal{A}_{\mathrm{SGD-iter}}(S)||\mathcal{A}_{\mathrm{SGD-iter}}(S')\right) = D_\alpha\left(\mathcal{A}^T_{FGD}(S)||\mathcal{A}^T_{FGD}\right) \leq D_\alpha\left(\mathcal{A}^t_{FGD}(S)||\mathcal{A}^t_{FGD}\right) = 0.$$

In the second case, we first employ Lemma 4 to upper bound the Rényi divergence of the output of DPFGD at some fixed step T.

Lemma 4. *For $\tau \geq 0$, let $\mathcal{L}$ be a dataset collection with dataset size n and τ-constrained maximum distance κ_τ. Let $\mathcal{L}$, τ and σ be parameters that satisfy the same assumptions as in Theorem 15. For any two datasets $S, S' \in \mathcal{L}$ with $d_{12}(S, S') \leq \tau$, the algorithm $\mathcal{A}^T_{\mathrm{FGD}}$ satisfies*

$$D_\alpha\left(\mathcal{A}^T_{\mathrm{FGD}}(S)||\mathcal{A}^T_{\mathrm{FGD}}(S')\right) \leq \frac{2\alpha\tau^2 L}{\sigma^2(n - \tilde{t} + 1)},$$

where $\tilde{t} \leq n - \kappa_\tau + 1$ being the index of the first pair of differing points.

For some fixed T, by Lemma 4 the Rényi divergence between the outputs of $\mathcal{A}_{\mathrm{SGD-iter}}$ on datasets S and S' is upper bounded by

$$D_\alpha\left(\mathcal{A}^T_{FGD}(S)||\mathcal{A}^T_{FGD}(S')\right) \leq \frac{2\alpha\tau^2 L}{\sigma^2(T - t + 1)}. \tag{22}$$

Then, as T is a uniform random variable in $\mathcal{A}_{\mathrm{SGD-iter}}(S)$, we can upper bound the Rényi divergence at some random time T by the weak convexity of Rényi divergence (Lemma B). We note that $t \leq T$ in case (ii), then for all T, as α satisfies $\sigma \geq \tau L\sqrt{2(\alpha - 1)\alpha}$,

$$D_\alpha\left(\mathcal{A}^T_{FGD}(S)||\mathcal{A}^T_{FGD}(S')\right) \leq \frac{2\alpha\tau^2 L^2}{\sigma^2\left(T - t + 1\right)} \leq \frac{2\alpha\tau^2 L^2}{\sigma^2} \leq \frac{1}{\alpha - 1}.$$

Therefore, we can apply Lemma B with $c = 1$. For any $t \leq T$,

$$\begin{aligned}
D_\alpha\left(\mathcal{A}_{\mathrm{SGD-iter}}(S)||\mathcal{A}_{\mathrm{SGD-iter}}(S')\right) &\overset{(a)}{\leq} \frac{2}{n - \kappa_\tau + 1} \sum_{T \in [n - \kappa_\tau + 1]} D_\alpha\left(\mathcal{A}^T_{\mathrm{FGD}}(S)||\mathcal{A}^T_{\mathrm{FGD}}(S')\right) \\
&\overset{(b)}{\leq} \frac{2}{n - \kappa_\tau + 1} \sum_{T=t}^{n - \kappa_\tau + 1} \frac{2\alpha\tau^2\mu^2}{\sigma^2\left(T - t + 1\right)} \\
&\leq \frac{4\alpha\tau^2\mu^2}{(n - \kappa_\tau + 1)\sigma^2} \log\left(n - \kappa_\tau - t + 2\right) \\
&\leq \frac{4\alpha\tau^2\mu^2}{\sigma^2} \frac{\log(n - \kappa_\tau + 2)}{n - k + 1}
\end{aligned} \tag{23}$$

where (a) follows applying Lemma B and (b) follows from Equation (22) for $T \geq t$. Substituting σ concludes the proof.

□

Proof of Lemma 4. Consider two datasets $S, S' \in \mathcal{L}$ that satisfy the assumptions in Lemma 4. Define the point-wise distance between these datasets as $d_i = \|S_i - S'_i\|_2$. Let $\tilde{t}$ be the first index where $d_i > 0$. By the assumption that differing points in S, S' are consecutively ordered, it follows that $\tilde{t} \leq n - \kappa_\tau + 1$. By the definition of τ-constrained maximum distance (Definition 8) and the assumptions on the dataset collection $\mathcal{L}$, we have,

$$d_i = \frac{\tau}{\kappa_\tau}, \quad \forall i \in \{\tilde{t}, \tilde{t} + 1, \ldots, \tilde{t} + \kappa_\tau\} \tag{24}$$

and $d_i = 0$ otherwise.

We will define Contractive Noisy Iteration (CNI) (Definition 9) and construct a CNI that outputs $\mathcal{A}^T_{\mathrm{FGD}}$ after T steps.

Definition 9 (Contractive Noisy Iteration (CNI)). *Given an initial state $\theta_0 \in \Theta$, a sequence of contractive functions $\psi_t : \Theta \to \Theta$, and a noise parameter $\sigma > 0$, the Contractive Noisy Iteration $(\theta_0, \{\theta_t\}, \mathcal{N}(0, \sigma^2))$ is defined by the following update rule:*

$$\theta_{t+1} = \psi_{t+1}(\theta_t) + Z_t,$$

where $Z_t \sim \mathcal{N}(0, \sigma^2)$.

For $S, S' \in \mathcal{L}$, we construct two series of contractive function $\{\psi_i\}$ and $\{\psi'_i\}$ as the gradient descent on the ith data point of S and S' respectively. Formally,

$$\begin{aligned} \psi_i(\theta) &= \theta - \eta \nabla_\theta(\theta, S_i) \\ \psi'_i(\theta) &= \theta - \eta \nabla_\theta(\theta, S'_i) \end{aligned} \tag{25}$$

The functions ψ_i and ψ'_i are contractive functions for $\eta \leq 2/\beta$ (Nesterov, 2004). It follows by the definition of DPFGD that $\theta_T = \mathcal{A}^T_{\mathrm{FGD}}(S)$ and $\theta'_T = \mathcal{A}^T_{\mathrm{FGD}}(S')$ are the Tth outputs of the CNIs $(\theta_0, \{\psi_t\}, \mathcal{N}(0, (\eta\sigma)^2))$ and $(\theta_0, \{\psi'_t\}, \mathcal{N}(0, (\eta\sigma)^2))$ respectively.

For these two CNIs, we can apply Lemma I to upper bound the Rényi divergence between their T^{th} outputs.

Lemma I (Theorem 22 in Feldman et al. (2018) with fixed noise distribution). *Let X_T, X'_T denote the output of two Contractive Noisy Iteration $(X_0, \{\psi_t\}, \mathcal{N}(0, (\eta\sigma)^2))$ and $(X_0, \{\psi'_t\}, \mathcal{N}(0, (\eta\sigma)^2))$ after T steps. Let $s_t = \sup_x \|\psi_t(x) - \psi'_t(x)\|_2$. Let $a_1, \ldots, a_T$ be a sequence of reals such that $z_t = \sum_{i \leq t} s_i - \sum_{i \leq t} a_i \geq 0$ for all $t < T$ and $z_T = 0$. Then,*

$$D_\alpha\left(X_T || X'_T\right) \leq \sum_{i=1}^{T} D_\alpha\left(\mathcal{N}(0, \eta^2\sigma^2) || \mathcal{N}(a_i, \eta^2\sigma^2)\right)$$

Following the definition in Lemma I, for the two contractive noisy maps $(\theta_0, \{\psi_t\}, \mathcal{N}(0, (\eta\sigma)^2))$ and $(\theta_0, \{\psi'_t\}, \mathcal{N}(0, (\eta\sigma)^2))$, we define s_i as

$$\begin{aligned} s_i &= \sup_\theta \|\psi_i(\theta) - \psi'_i(\theta)\|_2 \\ &\overset{(a)}{\leq} \sup_\theta \|\theta - \eta g(\theta, S_i) - \theta + \eta g(\theta, S'_i)\|_2 \\ &\overset{(b)}{\leq} \eta\mu d_i \overset{(c)}{=} \begin{cases} \frac{\eta\mu\tau}{\kappa_\tau} & i \in \{\tilde{t}, \tilde{t}+1, \ldots, \tilde{t}+\kappa_\tau\} \\ 0 & \text{Otherwise} \end{cases} \end{aligned} \tag{26}$$

where step (a) follows by Equation (25), step (b) follows by the smoothness assumption on the loss function ℓ, i.e.$\|\nabla_\theta \ell(\theta, S_i) - \nabla_\theta \ell(\theta, S'_i)\|_2 \leq \mu \|S_i - S'_i\|_2$ for any θ, S_i, S'_i. Step (c) follows by Equation (24).

By choosing $a_i = 0$ for all $i < \tilde{t}$ and $a_i = \frac{\eta\tau L}{T - \tilde{t} + 1}$ for $i \geq \tilde{t}$, we have that for all $t \leq T - \kappa_\tau + 1$, $z_t = \sum_{i \leq t} s_i - \sum_{i \leq t} a_i \geq 0$ and $z_T = 0$. This allows for the application of Lemma I,

$$\begin{aligned} D_\alpha\left(\mathcal{A}^T_{\mathrm{FGD}}(S) || \mathcal{A}^T_{\mathrm{FGD}}(S)\right) &\overset{(a)}{\leq} \frac{2\alpha}{\eta^2\sigma^2} \sum_{i=1}^{T} a_i^2 \\ &\overset{(b)}{\leq} \frac{2\alpha}{\eta^2\sigma^2} \sum_{i=\tilde{t}}^{T} \frac{\tau^2\eta^2\mu^2}{(T - \tilde{t} + 1)^2} \\ &= \frac{2\alpha\tau^2\mu^2}{\sigma^2\left(T - \tilde{t} + 1\right)} \end{aligned} \tag{27}$$

where step (a) follows from the application Lemma I and Rényi divergence of two Gaussian distributions (Lemma C). Step (b) follows by the definition of a_i. □

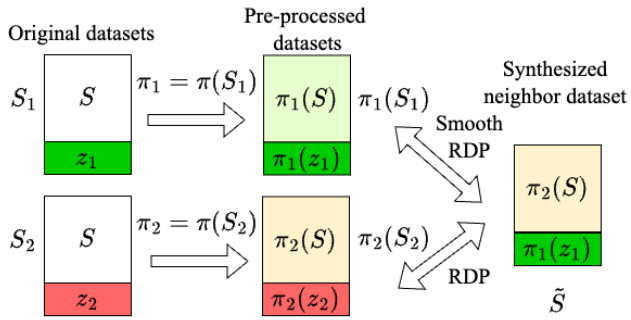


Figure 4. Illustration of the privacy analysis

B. Proof of Meta-Theorem (Theorem 2)

Theorem 2. *For a set of datasets $\mathcal{L}$ and for any $\alpha \geq 1$, $\tau > 0$, consider an algorithm $\mathcal{A}$ that is $(\alpha, \varepsilon(\alpha))$-RDP and $(\alpha, \tilde{\varepsilon}(\alpha, \tau))$-SRDP over the set $\mathcal{L}$. For a pre-processing algorithm π with L_∞ sensitivity Δ_∞ and L_2 sensitivity Δ_2, $\mathcal{A} \circ \pi$ is $(\alpha, \widehat{\varepsilon})$-RDP over $\mathcal{L}$ for all $c_1, c_2 \geq 1$, where*

$$\widehat{\varepsilon} \leq \max\left\{ \frac{\alpha c_1 - 1}{c_1(\alpha - 1)} \tilde{\varepsilon}(\alpha c_1, \Delta_2 \Delta_\infty) + \varepsilon\left(\frac{c_1 \alpha - 1}{c_1 - 1}\right), \right. \\ \left. \frac{\alpha c_2 - 1}{c_2(\alpha - 1)} \varepsilon(\alpha c_2) + \tilde{\varepsilon}\left(\frac{c_2 \alpha - 1}{c_2 - 1}, \Delta_2 \Delta_\infty\right) \right\}. \tag{3}$$

Proof. Consider two neighboring datasets S_1 and S_2, where $S_1 = S \cup \{z_1\}$ and $S_2 = S \cup \{z_2\}$. Let π_1 and π_2 be the pre-processing functions output by the pre-processing algorithm π on S_1 and S_2 respectively. Our objective is to establish an upper bound on the Rényi divergence between the output distribution of $\mathcal{A}$ on the pre-processed dataset $\pi_1(S_1)$ and $\pi_2(S_2)$, i.e. $D_\alpha\left(\mathcal{A}(\pi_1(S_1)) || \mathcal{A}(\pi_2(S_2))\right)$ and $D_\alpha\left(\mathcal{A}(\pi_2(S_2)) || \mathcal{A}(\pi_1(S_1))\right)$.

In the following, we first derive an upper bound on the Rényi divergence between $\mathcal{A}(\pi_1(S_1))$ and $\mathcal{A}(\pi_2(S_2))$. To do so, we construct a new dataset $\tilde{S}$ using the components of $\pi_1(S_1)$ and $\pi_2(S_2)$ as indicated in Figure 4, i.e.$\tilde{S} = \pi_1(S_1) \cup \pi_2(S_2)$. Then, using the same approach, we will upper bound the Rényi divergence between $\mathcal{A}(\pi_2(S_2))$ and $\mathcal{A}(\pi_1(S_1))$.

By contruction, $\tilde{S}$ and $\pi_2(S_2)$ are neighboring datasets, and that the L_{12} distance between $\tilde{S}$ and $\pi_2(S_2)$ is upper bounded by $\Delta_2 \Delta_\infty$. Using the RDP property of algorithm $\mathcal{A}$ we upper bound the divergence between $\mathcal{A}(\tilde{S})$ and $\mathcal{A}(\pi_2(S_2))$,

$$D_\alpha\left(\mathcal{A}(\pi_1(S_1)) || \mathcal{A}(\tilde{S})\right) \leq \tilde{\varepsilon}(\alpha, \Delta_\infty \Delta_2). \tag{28}$$

Similarly, using the SRDP property of the algorithm $\mathcal{A}$ over $\mathcal{L}$, we upper bound the divergence between $\mathcal{A}(\tilde{S})$ and $\mathcal{A}(\pi_1(S_1))$,

$$D_\alpha\left(\mathcal{A}(\tilde{S}) || \mathcal{A}(\pi_2(S_2))\right) \leq \varepsilon(\alpha). \tag{29}$$

Now, we combine Equations (28) and (29) using the weak triangle inequality of Rényi divergence (Lemma J), to upper bound the Rényi divergence between $\mathcal{A}(\pi_1(S_1))$ and $\mathcal{A}(\pi_2(S_2)$.

Lemma J (Triangle inequality of Rényi divergence (Mironov, 2017))**.** *Let μ_1, μ_2, μ_3 be distributions with the same support. Then, for $\alpha > 1$, $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, it holds that*

$$D_\alpha(\mu_1 || \mu_2) \leq \frac{\alpha - 1/p}{\alpha - 1} D_{p\alpha}(\mu_1 || \mu_3) + D_{q(\alpha - 1/p)}(\mu_3 || \mu_2).$$

Using Lemma J, for any $c_1 > 1$, we have

$$\begin{aligned} D_\alpha(\mathcal{A}(\pi_1(S_1))||\mathcal{A}(\pi_2(S_2))) &\leq \frac{\alpha - \frac{1}{c_1}}{\alpha - 1} D_{c_1\alpha}(\mathcal{A}(\pi_1(S_1))|\mathcal{A}(\tilde{S})) + D_{\frac{c_1\alpha-1}{c_1-1}}(\mathcal{A}(\tilde{S})||\mathcal{A}(\pi_2(S_2))) \\ &\overset{(a)}{\leq} \frac{\alpha c_1 - 1}{c_1(\alpha-1)} \tilde{\varepsilon}(\alpha c_1, \Delta_\infty \Delta_2) + \varepsilon\left(\frac{c_1\alpha - 1}{c_1 - 1}\right) \end{aligned} \tag{30}$$

where step (a) follows from Equations (28) and (29).

Similarly, by constructing a dataset $\tilde{S}$ consisting of $\pi_1(S)$ and $\pi_2(z_2)$, we upper bound $D_\alpha\left(\mathcal{A}(\pi_2(S_2))||\mathcal{A}(\tilde{S})\right)$ and $D_\alpha\left(\mathcal{A}(\tilde{S})||\mathcal{A}(\pi_1(S_1))\right)$ with the SRDP and RDP property in a similar manner as Equations (28) and (29). Applying Lemma J, we can show that for any $c_2 > 1$,

$$D_\alpha(\mathcal{A}(\pi_2(S_2))||\mathcal{A}(\pi_1(S_1))) \leq \frac{\alpha c_2 - 1}{c_2(\alpha-1)} \varepsilon(\alpha c_2) + \tilde{\varepsilon}\left(\frac{c_2\alpha - 1}{c_2 - 1}, \Delta_\infty \Delta_2\right) \tag{31}$$

Combining Equation (30) and Equation (31) concludes the proof.

□

C. Proofs for the sensitivity of different pre-processing algorithms (Section 4.1)

In this section, we bound the sensitivity of pre-processing algorithms discussed in Section 4.1.

C.1. Sensitivity analysis of deduplication and quantization

Proposition 3. *For a dataset collection $\mathcal{L}$, the L_2 and L_∞ sensitivities*[6] *of η-approximate deduplication π_η^d and quantization π_η^q are $\Delta_2(\mathcal{L}, \pi_\eta^d) = 1$ and $\Delta_2(\mathcal{L}, \pi_\eta^q) = \eta$, and*

$$\Delta_\infty(\mathcal{L}, \pi_\eta^d) = \Delta_\infty(\mathcal{L}, \pi_\eta^q) = \max_{S \in \mathcal{L}} \max_{B \in B(S)} 2\,|B|\,.$$

Proof. Consider two neighboring datasets $S_1, S_2 \in \mathcal{L}$. Without loss of generality, let $S = S_1 \cap S_2$, $z_1 = S_1 \setminus S$ and $z_2 = S_2 \setminus S$, similar to the notations of original datasets in Figure 4.

Sensitivity analysis of deduplication As the data space $\mathcal{X}$ is bounded by 1, it is obvious that the upper bound on L_2 sensitivity of π_η^d is 1.

To bound the L_∞ sensitivity, we first recall the definition of a "good" cluster. For a dataset S and a point $x \in S$, define $B(x, \eta; S) = \{\tilde{x} \in S : \|\tilde{x} - x\|_2 \leq \eta\}$, a ball of radius η around x. A point x is the centroid of a *good cluster* if $B(x, \eta; S) = B(x, 3\eta; S)$. The set of all good clusters in a dataset S is denoted by $B(S) = \{B_i\}_{i=1}^m$ where $B_i := B(x_i, \eta; S)$ for all $x_i \in S$ satisfying $B(x_i, \eta; S) = B(x_i, 3\eta; S)$.

We will first prove that the difference between the datasets $\pi_{\eta,S_1}^d(S)$ and $\pi_{\eta,S}^d(S)$ is the ball $B(z_1, \eta; S_1)$. Similarly, we show that the maximum difference between $\pi_{\eta,S_2}^d(S)$ and $\pi_{\eta,S}^d(S)$ is $B(z_2, \eta; S_2)$. Taking supremum over all neighboring datasets in $\mathcal{L}$, these two results imply that the L_∞ sensitivity of deduplication is upper bounded by twice the size of the largest good cluster in any dataset $S \in \mathcal{L}$.

To calculate that the maximum difference between the datasets $\pi_{\eta,S_1}^d(S)$ and $\pi_{\eta,S}^d(S)$, we assume without loss of generality that there exists $x_1, \ldots, x_k \in S$ satisfying $\|z_1 - x_i\| \leq \eta$ for all $i \in [k]$. We consider the following cases:

Case I: $x_1, \ldots x_k$ are not in any good cluster centered at some $c \in S$. We will discuss the two sub-cases: one where the point z_1 forms the centroid of a good cluster, and another where it does not.

If the point z_1 is the centroid of a good cluster, then $x_1, \ldots x_k$ are the only points in the good cluster $B(z_1, \eta; S)$ and will be removed by π_{η,S_1}^d. In contrast, in Case I, $x_1, \ldots, x_k$ will not be removed by $\pi_{\eta,S}^d$. Hence, the difference between $\pi_{\eta,S_1}^d(S)$ and $\pi_{\eta,S}^d(S)$ is $\{x_1, \ldots, x_k\} \subset B(z_1, \eta; S)$.

If the point z_1 is not the centroid of a good cluster, then z_1 is not in any good cluster. We will prove this claim by contradiction. Assume z_1 is in a good cluster centered at some $c \in S$. Then, for any x_i, $i \in [k]$,

$$\|x_i - c\|_2 \leq \|x_i - z_1\|_2 + \|z_1 - c\|_2 \leq 2\eta. \tag{32}$$

If $\|x_i - c\|_2 \leq \eta$, then x_i is also in the good cluster around $c \in S$, contradicting the assumption that none of $\{x_1, \ldots x_k\}$ is in any good cluster. On the other hand, if $\eta \leq \|x_i - c\|_2 \leq 2\eta$, then $B(c, 3\eta; S)$ cannot be a good cluster.

Therefore, we have shown that when $x_1, \ldots x_k$ are not in any good cluster and the point z_1 is not the centroid of a good cluster, then none of $\{x_, \ldots, x_k, z_1\}$ is in a good cluster. In this case, $\pi_{\eta,S}^d(S) = \pi_{\eta,S_1}^d(S)$.

Case II: There exists some point $x \in \{x_1, \ldots, x_k\}$ in a good cluster centered at $c \in S$, i.e.$x \in B(c, \eta; S)$ and $B(c, \eta; S)$ is a good cluster. We first consider the effect of $\pi_{\eta,S}^d$ and π_{η,S_1}^d on the single point x. Note that

$$\|z_1 - c\|_2 \leq \|z_1 - x\|_2 + \|x - c\|_2 \leq 2\eta. \tag{33}$$

If $\|z_1 - c\|_2 \leq \eta$, $B(c, \eta; S_1) = B(c, \eta; S \cup \{z_1\})$ is also a good cluster. Then, π_{η,S_1}^d and π_{η,S_1}^d has the same effect on x.

[6] When the definition of neighboring dataset is refined to addition and deletion of a single data point, the L_∞ sensitivity of both deduplication and quantization on a set $\mathcal{L}$ can be reduced to $\max_{S \in \mathcal{L}, B \in B(S)} |B|$.

If $\eta \leq \|z_1 - c\|_2 \leq 2\eta$, then $z_1 \in B(c, 3\eta; S_1)$ but $z_1 \notin B(c, \eta; S_1)$. This implies $B(c, 3\eta; S_1) \neq B(c, \eta; S_1)$, and $B(c, \eta; S_1)$ is not a good cluster. Therefore, $\pi^d_{\eta,S}$ removes the point x, while π^d_{η,S_1} does not. In this case, x is a different point between $\pi^d_{\eta,S}(S)$ and $\pi^d_{\eta,S_1}(S)$.

We note that there are at most k different points between $\pi^d_{\eta,S}(S)$ and $\pi^d_{\eta,S_1}(S)$, when all points $\{x_1, \ldots, x_k\}$ are in some good cluster $B(c_i, \eta; S)$ for $c_i \in S$ and z_1 is selected such that none of $B(c_i, \eta; S_1)$ remains to be a good cluster. In this case, the difference between $\pi^d_{\eta,S}(S)$ and $\pi^d_{\eta,S_1}(S)$ is $\{x_1, \ldots, x_k\} \subseteq B(z_1, \eta; S)$.

Following a similar argument, we can show that the maximum number of different points between $\pi^d_{\eta,S_2}(S)$ and $\pi^d_{\eta,S}(S)$ is k and this maximum set of different points is a subset of $B(z_2, \eta; S_2)$. This concludes the proof for the sensitivity of deduplication.

Senstivity analysis of quantization The analysis of L_∞ sensitivity of quantization is the same as that of deduplication. To get the L_2 sensitivity of quantization, we consider two cases. If a point is in a good cluster, then quantization process change this point to the centroid of the cluster. The L_2 distance incurred by the pre-processing is upper bounded by η by the definition of η-quantization. If a point is not in a good cluster, it remains unchanged after quantization. Combining the two cases, the L_2 sensitivity of quantization is upper bounded by η.

□

C.2. Sensitivity analysis of model-based imputation

In this section, we first provide a general result on the sensitivity for any model-based imputation method. We then introduce specific imputation methods, including mean imputation (Corollary 4 in the main text), median imputation, trimmed mean imputation, and linear regression. We summarize their sensitivity results in Table 3.

For a given model f, the imputation algorithm π_f first generates an imputation function f_S by fitting the model f to a dataset S. Then, it replaces each missing value in the dataset with the prediction based on the imputation function f_S. The L_2 and L_∞ sensitivity of model-based imputation is presented in Proposition 16.

Several widely used models for imputation include mean, median, trimmed mean, and linear regression, described as follows. Mean imputation replaces the missing values in the jth feature with the empirical mean of the available data for that feature. On the other hand, median imputation replaces each missing value with the median of the non-missing points in the corresponding feature. The trimmed mean estimator with parameter m is an interpolation between the mean and median. It estimates each missing value by computing the mean after removing the m smallest and largest points from the remaining points of the feature that is not missing. The above methods address the missing values at a feature j using data from that feature alone. In contrast, linear regression also employs the information of the other features of the missing point. Specifically, it estimates the missing value by the prediction of a linear regression on the jth feature using some or all other features in the dataset.

Many imputation methods, such as median and linear regression, do not have bounded global sensitivity even when the instance space is bounded (Alabi et al., 2022). However, the local sensitivity of these methods are usually bounded on well-behaved datasets. Given a collection of well-behaved datasets $\mathcal{L}$, the L_2 sensitivity over the collection is the upper bound on the local sensitivity of all datasets $S \in \mathcal{L}$. We exploit this property to find the L_2 sensitivity of the aforementioned imputation methods. We introduce additional notation and provide a summary of the L_2 sensitivity for different methods in Table 3.

Proposition 16 (Sensitivity of model-based imputation)**.** *For a dataset collection $\mathcal{L}$, the L_∞ sensitivity of model-based imputation π_f over $\mathcal{L}$ is the maximum number of missing values present in any dataset in $\mathcal{L}$. Furthermore, the L_2-sensitivity of π_f over $\mathcal{L}$ is given by*

$$\Delta_2(\mathcal{L}, \pi_f) = \max_{\substack{S', S \in \mathcal{L} \\ d_H(S,S')=1}} \max_{x \in S \cup S'} \sqrt{\sum_{j=1}^{d} \left(f_S(x_j) - f_{S'}(x_j)\right)^2},$$

where x_j denotes the jth feature of x. Specifically, for mean imputation, median imputation, trimmed mean imputation, and regression, we present their sensitivities and corresponding assumptions in Table 3.

Notation	Meaning	Imputation Model f	L_2-Sensitivity $\Delta_2(\mathcal{L}, \pi_f)$
p	Maximum number of missing points in any $S \in \mathcal{L}$	Mean	$\frac{2}{n-p}$
$x_{(m)}^{\min}$	Minimum m^{th} ordered statistics of any feature in any $S \in \mathcal{L}$	Median	$x_{((n+1)/2)}^{\max} - x_{(n/2)}^{\min}$
$x_{(m)}^{\max}$	Maximum m^{th} ordered statistics of any feature in any $S \in \mathcal{L}$	m-Trimmed Mean	$x_{(m)}^{\max} - x_{(n-m)}^{\min} / n-2m-p$
$\lambda_{\max}$ $\lambda_{\min}$	Maximum eigenvalue of $X^T X$, for any $S = (X, Y) \in \mathcal{L}$ Minimum eigenvalue of $X^T X$, for any $S = (X, Y) \in \mathcal{L}$	Linear regression	$\frac{\lambda_{\max}^2}{(\lambda_{\max}+1)\lambda_{\min}^2} + \frac{1}{\lambda_{\min}}$

Table 3. Notations and L_2 sensitivity of different imputation models

Proof. It is obvious that the L_∞ sensitivity of model-based imputation is upper bounded by the number of entries with missing values in any of the dataset $S \in \mathcal{L}$. The L_2 sensitivity is upper bounded by the sensitivity of the model f over the set $\mathcal{L}$. This concludes the proof.

L_2 **Sensitivity of mean over** $\mathcal{L}$ For neighboring datasets $S_1, S_2 \in \mathcal{L}$, write $S_1 = S \cup \{z_1\}$ and $S_2 = S \cup \{z_2\}$. Without loss of generality, denote $S = \{s_i\}_{i=1}^{n-1}$. For each data point s_i, we denote its jth feature as s_{ij}. Also, we denote the number of available data points for the jth feature as n_j. In the following, we derive an upper bound on the L_2 sensitivity of mean imputation $\Delta_2(\mathcal{L}, \pi_{\text{mean}})$,

$$\begin{aligned}
\Delta_2(\mathcal{L}, \pi_{\text{mean}})^2 &\leq \sum_{j=1}^{d} \left(\frac{1}{n_j} \left(\sum_{i=1}^{n-1} s_{ij} + z_{1j} \right) - \frac{1}{n_j} \left(\sum_{i=1}^{n-1} s_{ij} + z_{2j} \right) \right)^2 \\
&\leq \sum_{j=1}^{d} \left(\frac{1}{n_j} \left(z_{1j} - z_{2j} \right) \right)^2 \\
&\leq \frac{1}{n-p} \sum_{j=1}^{d} \left(z_{1j} - z_{2j} \right)^2 = \frac{4}{(n-p)^2}
\end{aligned} \tag{34}$$

where the last inequality follows by the fact that the instance space is bounded with diameter 1. Taking square root from both side of Equation (34) completes the proof.

L_2 **Sensitivity of median and** m**-trimmed mean over** $\mathcal{L}$ The L_2-sensitivity of median and m-trimmed mean over $\mathcal{L}$ follows directly by the definition.

L_2 **Sensitivity of linear regression** π_{LR} **over** $\mathcal{L}$ For linear regression, we present the L_2 sensitivity by considering two datasets $X, X' \in \mathcal{L}$ where $X' = X \cup \{z\}$. For any $j \in [d]$, imputation with linear regression of the feature X_j looks at the a submatrix of X that does not include the jth feature. Denote the submatrix that is used for linear regression as X_r (X_r includes a subset of features, specified by the index r, of the original dataset X). Let $\Sigma = \frac{1}{n} X^\top X$, and let $\Sigma_r = \frac{1}{n} X_r^\top X_r$ be a principal matrix of Σ with submatrix X_r. We calculate the L_2 sensitivity of linear regression on imputing the point x_j below,

$$\begin{aligned}
\Delta_2(\mathcal{L}, \pi_{\text{LR}}) &= \left\| x_r^\top (X_r^\top X_r)^{-1} X_r^\top X_j - x_r^\top (X_r^\top X_r + zz^\top)^{-1} (X_r^\top X_j + zz_j) \right\| \\
&\leq \|x_r\| \left\| (X_r^\top X_r)^{-1} X_r^\top X_j - (X_r^\top X_r + zz^\top)^{-1} (X_r^\top X_j + zz_j) \right\| \\
&\leq \left\| \left((X_r^\top X_r)^{-1} - (X_r^\top X_r + zz^\top) \right)^{-1} X_r^\top X_j - (X_r^\top X_r + zz^\top)^{-1} zz_j \right\| \\
&\leq \left\| \left((X_r^\top X_r)^{-1} - (X_r^\top X_r + zz^\top) \right)^{-1} X_r^\top X_j \right\| + \left\| (X_r^\top X_r + zz^\top)^{-1} zz_j \right\| \\
&\overset{(a)}{\leq} \left\| \frac{(X_r^\top X_r)^{-1} zz^\top (X_r^\top X_r)^{-1}}{(1 + z^\top (X_r^\top X_r)^{-1} z)} X_r^\top X_j \right\| + \frac{1}{\lambda_{\min}} \\
&\overset{(b)}{\leq} \frac{\lambda_{\max}^2}{(\lambda_{\max} + 1)\lambda_{\min}^2} + \frac{1}{\lambda_{\min}}
\end{aligned} \tag{35}$$

where step (a) follows from Sherman-Morrison-Formula and that the eigenvalue of a principal submatrix is always larger than the smallest eigenvalue of the original matrix following Theorem 4.3.15 in (Horn & Johnson, 1985) and then taking supremum over all $X \in \mathcal{L}$. Similarly, step (b) follows from the fact that the eigenvalue of a principal submatrix is always smaller than the largest eigenvalue of the original matrix following Theorem 4.3.15 in (Horn & Johnson, 1985) and then taking supremum over all $X \in \mathcal{L}$. □

C.3. Sensitivity analysis of PCA

Proposition 5. *For a dataset collection $\mathcal{L}$, the L_∞ sensitivity of $\pi_{\mathrm{PCA-dim}}$ and $\pi_{\mathrm{PCA-rank}}$ is the size of the datasets in $\mathcal{L}$, i.e. n. The L_2-sensitivity of $\pi_{\mathrm{PCA-dim}}$ and $\pi_{\mathrm{PCA-rank}}$ is bounded by $2\Delta_2$ and Δ_2 respectively, where*

$$\Delta_2 = \frac{4(3n+2)}{n(n-1)\min\{\delta^k_{\min}(\mathcal{L}), \delta^1_{\min}(\mathcal{L})\}},$$

where $\delta^k_{\min}(\mathcal{L}) = \min_{S\in\mathcal{L}} \lambda_k(S) - \lambda_{k+1}(S)$ is the minimum gap between the kth and $(k+1)^{th}$ eigenvalue over any covariance matrix of $S \in \mathcal{L}$.

Proof. **PCA for dimension reduction:** For any two neighboring datasets $S, \tilde{S}$, without loss of generality, we denote $S = \{x_i\}_{i=1}^n$, and $\tilde{S} = \{x_i\}_{i=1}^{n-1} \cup \{x'_n\}$. Let $\Sigma, \tilde{\Sigma}$ denote their empirical covariance matrices and $\hat{\mu}, \hat{\mu}'$ denote the empirical mean. Let A_k and $\tilde{A}_k$ be the matrix consisting of first k eigenvectors of Σ and $\tilde{\Sigma}$ respectively.

First, for any $x \in \mathcal{X}$, we upper bound the $\left\|A_k^\top x - \tilde{A}_k^\top x\right\|_F$ by a linear function of $\left\|\tilde{\Sigma} - \Sigma\right\|_F$ using properties of the dataset collection $\mathcal{L}$,

$$\left\|A_k^\top x - \tilde{A}_k^\top x\right\|_F \overset{(a)}{\leq} \left\|A_k - \tilde{A}_k^\top\right\|_F \overset{(b)}{=} \mathrm{Tr}\left(2(I - A_k^\top \tilde{A}_k)\right) \tag{36}$$

where step (a) follows from Cauchy-Schwarz inequality and bounded instance space and step (b) follows from the definition of Frobenius norm $\|A\|_F = \mathrm{Tr}(A^\top A)$ and the orthonormality of $A_k, \tilde{A}_k$.

Let σ_i denote the ith singular value of $A_k^\top \tilde{A}_k$, and let θ_i be the ith canonical angle of $A_k^\top \tilde{A}_k$, i.e. $\cos\theta_i = \sigma_i$. We can write $\mathrm{Tr}\left(2(I - A_k^\top \tilde{A}_k)\right)$ as follows,

$$\begin{aligned}\mathrm{Tr}\left(2(I - A_k^\top \tilde{A}_k)\right) &\overset{(a)}{=} 2\left(k - \sum_{i=1}^k \sigma_i^2\right) \overset{(b)}{=} 2\left(k - \sum_{i=1}^k (\cos\theta_i)^2\right) \\ &\overset{(c)}{=} 2\left(k - k + \sum_{i=1}^k (\sin\theta_i)^2\right) \overset{(d)}{\leq} \frac{4\left\|\tilde{\Sigma} - \Sigma\right\|_F}{\min\{\delta^k_{\min}(\mathcal{L}), \delta^1_{\min}(\mathcal{L})\}}.\end{aligned} \tag{37}$$

where step (a) and (b) are due to the definition of singular value and canonical angle, step (c) follows from the fact that $(\sin\theta)^2 + (\cos\theta)^2 = 1$ for any θ, and step (d) follows by Davis-Kahan Theorem (Lemma K) stated below.

Lemma K (Davis-Kahan Theorem (Yu et al., 2014))**.** *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric and positive definite, with eigenvalues $\lambda_1, \geq \ldots \geq \lambda_d$ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_d$ respectively. For $k \leq d$, let V and $\hat{V}$ be the dataset whose matrices consisting of the first k eigenvectors of Σ and $\hat{\Sigma}$ respectively. Then,*

$$\left\|\sin\Theta(V, \hat{V})\right\|_F \leq \frac{2\min\left\|\Sigma - \hat{\Sigma}\right\|_F}{\min\{\lambda_k - \lambda_{k+1}, \lambda_1 - \lambda_2\}},$$

where $\Theta(V, \hat{V})$ denotes the $k \times k$ diagonal matrix of the principal angles between two subspaces V and $\hat{V}$.

It remains to upper bound the Frobenius norm of $\left\|\Sigma - \tilde{\Sigma}\right\|_F$. By the definition of the empirical covariance matrix, we

decompose $\tilde{\Sigma}$ as

$$\begin{aligned}\tilde{\Sigma} =& \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\hat{\mu}')(x_i-\hat{\mu}')^\top + \frac{1}{n-1}(x_n'-\hat{\mu}')(x_n'-\hat{\mu}')^\top \\ \overset{(a)}{=}& \underbrace{\frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\mu-\frac{1}{n}(x_n'-x_n))(x_i-\mu-\frac{1}{n}(x_n'-x_n))^\top}_{\text{part I}} \\ &+ \underbrace{\frac{1}{n-1}(x_n'-\mu-\frac{1}{n}(x_n'-x_n))(x_n'-\mu-\frac{1}{n}(x_n'-x_n))^\top}_{\text{part II}}\end{aligned} \tag{38}$$

where (a) follows from $\mu' = \mu + \frac{1}{n}(x_n' - x_n)$.

Part (I) can be written as

$$\begin{aligned}\text{Part I} =& \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\mu-\frac{1}{n}(x_n'-x_n))(x_i-\mu-\frac{1}{n}(x_n'-x_n))^\top \\ =& \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\mu)(x_i-\mu)^\top - \frac{1}{n(n-1)}\sum_{i=1}^{n-1}\left((x_i-\mu)(x_n'-x_n)^\top + (x_n'-x_n)(x_i-\mu)^\top\right) \\ &+ \frac{1}{n^2}(x_n'-x_n)(x_n'-x_n)^\top \\ =& \Sigma - \frac{1}{n-1}(x_n-\mu)(x_n-\mu)^\top + \frac{1}{n(n-1)}\left((x_n-\mu)(x_n'-x_n)^\top + (x_n'-x_n)(x_n-\mu)^\top\right) \\ &+ \frac{1}{n^2}(x_n'-x_n)(x_n'-x_n)^\top\end{aligned} \tag{39}$$

Similarly, we can write part (II) as

$$\begin{aligned}\text{Part II} =& \frac{1}{n-1}(x_n'-\mu-\frac{1}{n}(x_n'-x_n))(x_n'-\mu-\frac{1}{n}(x_n'-x_n))^\top \\ =& \frac{(x_n'-\mu)(x_n'-\mu)^\top}{n-1} - \frac{(x_n'-\mu)(x_n'-x_n)^\top}{n(n-1)} - \frac{(x_n'-x_n)(x_n'-\mu)^\top}{n(n-1)} \\ &+ \frac{(x_n'-x_n)(x_n'-x_n)^\top}{(n-1)n^2}\end{aligned} \tag{40}$$

Substituting Equation (39) and Equation (40) into Equation (38), and by the fact that $\left\|xx^T\right\|_F \leq 1$ for $\|x\|_2 \leq 1$, we get

$$\left\|\tilde{\Sigma} - \Sigma\right\|_F \leq \frac{2(3n+2)}{n(n-1)}. \tag{41}$$

Finally, substituting Equation (41) into Equation (37), and then Equation (37) into Equation (36) completes the proof.

PCA for rank reduction: For any two neighboring datasets $S, S' \in \mathcal{L}$, we define the notations of $\tilde{\Sigma}, \Sigma, A_k, \tilde{A}_k$ similarly as in the proof for $\pi_{\text{PCA-dim}}$. We state Lemma L, which is used to upper bound $\left\|A_kA_k^\top x - \tilde{A}_k\tilde{A}_k^\top x\right\|_F$.

Lemma L (Simplified version of Theorem 3 in (Zwald & Blanchard, 2005))**.** *Let A be a symmetric positive definite matrix with eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d$. Let B be a symmetric positive matrix. For an integer $k > 0$, let A_k be the matrix consisting the first k eigenvectors of A and $\tilde{A}_k$ be the matrix consisting of the first k eigenvectors of $A + B$. Then, A_k and $\tilde{A}_k$ satisfy that*

$$\left\|A_kA_k^\top - \tilde{A}_k\tilde{A}_k^\top\right\| \leq \frac{2\|B\|}{\lambda_k - \lambda_{k+1}}.$$

Applying Lemma L with $A + B = \tilde{\Sigma}$ and $A = \Sigma$, we can show an upper bound on the term $\left\| A_k A_k^\top x - \tilde{A}_k \tilde{A}_k^\top x \right\|_F$ for any $x \in \mathcal{X}$.

$$\left\| A_k A_k^\top x - \tilde{A}_k \tilde{A}_k^\top x \right\|_F \leq \frac{\left\| \tilde{\Sigma} - \Sigma \right\|_F}{\lambda_k(S) - \lambda_{k+1}(S)} \leq \frac{4(3n+2)}{n(n-1)\delta_{\min}(S)} \tag{42}$$

where the last inequality follows by Equation (41) and the definition of $\delta_{\min}(S) = \min\{\delta^k_{\min}(S), \delta^1_{\min}(S)\}$. Taking the supremum over all dataset $S \in \mathcal{L}$ concludes the proof. □

C.4. Sensitivity Analysis of Scaling

Proposition 6. *For a dataset collection $\mathcal{L}$, the L_∞ sensitivity of standard scaling and min max scaling is the size of the datasets in $\mathcal{L}$, i.e. n. The L_2 sensitivity of standard scaling is*

$$\Delta_2 = \frac{2}{\sigma^3_{\min} n} + \frac{2}{n\sigma_{\min}},$$

where $\sigma_{\min}$ is the minimum standard deviation over datasets in $\mathcal{L}$.

Proof. The L_∞ sensitivity for both scaling methods is trivially upper bounded by the size of the datasets in $\mathcal{L}$. In the following, we prove the L_2 sensitivity for standard scaling and min max scaling respectively.

Proof of L_2 sensitivity for standard scaling For any two neighboring datasets S, S', let μ, μ' denote the mean of S, S' respectively and σ, σ' denote the standard deviation of S, S' respectively. Then, the L_2 sensitivity of standard scaling is

$$\begin{aligned}
\Delta_2 &= \max_{S,S',x} \left\| \frac{x-\mu}{\sigma} - \frac{x-\mu'}{\sigma'} \right\|_2 \\
&= \max_{S,S',x} \left\| \frac{\sigma' x - \sigma'\mu - \sigma x + \sigma\mu'}{\sigma\sigma'} \right\| \\
&= \max_{S,S',x} \left\| \frac{(\sigma' - \sigma)(x-\mu)}{\sigma\sigma'} \right\|_2 + \max_{S,S'} \left\| \frac{\mu - \mu'}{\sigma'} \right\| \\
&\overset{(a)}{\leq} \frac{2\max_{S,S'} \|\sigma' - \sigma\|}{\sigma^2_{\min}} + \frac{\max_{S,S'} \|\mu - \mu'\|}{\sigma_{\min}}
\end{aligned} \tag{43}$$

where step (a) follows from $\|x - \mu\|_2 \leq 2$ for any x and S and the definition of $\sigma_{\min}$.

From Liu (2016), the global sensitivity of sample mean and variance are $\frac{2}{n}$. We then show that for a dataset collection $\mathcal{L}$, the global sensitivity conditional on $\mathcal{L}$ is upper bounded by $1/\sigma_{\min} n$.

For any S, S' with sample variances $\sigma^2, (\sigma')^2$, we have

$$|\sigma - \sigma'| \leq \frac{\left|\sigma^2 - (\sigma')^2\right|}{2\sigma_{\min}} \leq \frac{1}{\sigma_{\min} n}, \tag{44}$$

where the first inequality follows by rearranging $\left|\sigma^2 - (\sigma')^2\right| = |(\sigma - \sigma')(\sigma + \sigma')| \geq 2\sigma_{\min} |\sigma - \sigma'|$, and the second inequality by substituting the global sensitivity of sample covariance from Liu (2016).

Substituting the sensitivity of sample mean $2/n$ and sample standard deviation $\frac{1}{\sigma_{\min} n}$ into Equation (43), we conclude the proof. □

D. Proofs for the overall privacy guarantees (Section 4.2)

In this section, we provide the proofs of the overall privacy guarantees for specific pre-processed DP pipelines, as stated in Table 2 in Section 4.2. For clarity, we restate the full version of Theorem 7 and specify the privacy guarantee for each category of privacy mechanisms (each row in Table 2) in Theorem 17. Then, we present the proof for each category separately. As the proof of DP-GD is the same as that of Gaussian mechanism, we omit the proof for DP-GD for simplicity.

Theorem 17 (Full version of Theorem 7)**.** *Let p denote the L_∞ sensitivity of deduplication, quantization and mean imputation. Let $n \geq \min\{101, p\}$ be the size of any dataset in the dataset collection $\mathcal{L}$. Let ℓ be a 1-Lipschitz and 1-smooth loss function. For an output function f and a score function Q, assume their Lipschitz parameter and global sensitivity are both 1. Then,*

(i) Gaussian mechanism with output function f satisfy $\left(\alpha, 1.05\alpha\varepsilon^2(1+p^2)\right)$-RDP, $\left(\alpha, 1.05\alpha\varepsilon^2(1+\eta^2p^2)\right)$-RDP, $\left(\alpha, 1.05\alpha\varepsilon^2\left(1+\frac{4p^2}{(n-p)^2}\right)\right)$-RDP, $\left(\alpha, 1.05\alpha\varepsilon^2\left(1+\frac{12.2^2}{\delta_{\min}^2}\right)\right)$-RDP and $1.05\alpha\varepsilon^2\left(1+\frac{4}{\sigma_{\min}^3}\right)$-RDP when coupled with deduplication, quantization, mean imputation, PCA for rank reduction and standard scaling respectively.

(ii) Exponential mechanism with score function Q and Laplace mechanism with output function f satisfy $(\alpha, \varepsilon(1+p))$-RDP, $(\alpha, \varepsilon(1+\eta p))$-RDP, $\left(\alpha, \varepsilon\left(1+\frac{2p}{n-p}\right)\right)$-RDP, $\left(\alpha, \varepsilon\left(1+\frac{12.2}{\delta_{\min}}\right)\right)$-RDP and $4.2\alpha\varepsilon^2\left(1+\frac{1}{\sigma_{\min}^3}\right)$-RDP when coupled with deduplication, quantization, mean imputation, PCA and standard scaling for rank reduction respectively.

(iii) DP-SGD with subsampling with loss function ℓ satisfies $\left(\alpha, 1.05\alpha\varepsilon^2\left(2\alpha+\frac{12.2^2}{\delta_{\min}^2}\right)\right)$-RDP and $2.1\alpha\varepsilon^2\left(\alpha+\frac{8}{\sigma_{\min}^6}\right)$-RDP when coupled with PCA for rank reduction and standard scaling respectively.

(iv) DP-SGD with iteration with loss function ℓ coupled with deduplication, quantization and mean imputation satisfy $\left(\alpha, 1.1\alpha\varepsilon^2\left(1+\frac{p^2n\log(n-p)}{(n-p)\log n}\right)\right)$-RDP, $\left(\alpha, 1.1\alpha\varepsilon^2\left(1+\frac{\eta^2p^2n\log(n-p)}{(n-p)\log n}\right)\right)$-RDP, and $\left(\alpha, 1.1\alpha\varepsilon^2\left(1+\frac{4p^2n\log(n-p)}{(n-p)^3\log n}\right)\right)$-RDP respectively.

Proof of Theorem 17 (i). We apply Theorem 2 with $c_1 = c_2 = 2$,

$$\begin{aligned}
\widehat{\varepsilon}(\alpha) &\leq \max\left\{\frac{2\alpha-1}{2(\alpha-1)}\tilde{\varepsilon}(2\alpha, \Delta_2\Delta_2) + \varepsilon(2\alpha-1), \frac{2\alpha-1}{2(\alpha-1)}\varepsilon(2\alpha) + \tilde{\varepsilon}(2\alpha-1, \Delta_2\Delta_\infty)\right\} \\
&\overset{(a)}{\leq} \max\left\{\frac{2\alpha-1}{2(\alpha-1)}\tilde{\varepsilon}(2\alpha, \Delta_2\Delta_\infty) + \frac{2\alpha-1}{2(\alpha-1)}\varepsilon(2\alpha), \frac{2\alpha-1}{2(\alpha-1)}\varepsilon(2\alpha) + \frac{2\alpha-1}{2(\alpha-1)}\tilde{\varepsilon}(2\alpha, \Delta_2\Delta_\infty)\right\} \\
&= \frac{2\alpha-1}{2(\alpha-1)}\left(\tilde{\varepsilon}(2\alpha, \Delta_2\Delta_2) + \varepsilon(2\alpha)\right) \\
&\overset{(b)}{\leq} 1.05\left(\tilde{\varepsilon}(2\alpha, \Delta_2\Delta_\infty) + \varepsilon(2\alpha)\right).
\end{aligned} \tag{45}$$

where step (a) follows from the monotonicity of Renyi Divergence (Lemma D) and step (b) follows from $\frac{2\alpha-1}{2(\alpha-1)} \leq 1.05$ for $\alpha > 11$.

By substituting the expression of RDP and SRDP parameter from Table 1 for Gaussian mechanism, the L_2 and L_∞ sensitivity (Δ_2 and Δ_∞) for deduplication, quantization, PCA, standard scaling and mean imputation from Proposition 3, 5 and 6 and Corollary 4, and the Lipschitz parameter and global sensitivity of the output function f into Equation (45), we complete the proof of the overall privacy guarantees for Gaussian mechanism.

□

Proof of Theorem 17 (ii). We apply Theorem 2 with $c_1 = c_2 = 1$,

$$\begin{aligned}
\widehat{\varepsilon}(\alpha) &\leq \max\left\{\frac{\alpha-1}{\alpha-1}\tilde{\varepsilon}(\alpha, \Delta_2\Delta_2) + \varepsilon(\infty), \frac{\alpha-1}{\alpha-1}\varepsilon(\alpha) + \tilde{\varepsilon}(\infty, \Delta_2\Delta_\infty)\right\} \\
&= \max\{\tilde{\varepsilon}(\alpha, \Delta_2\Delta_\infty) + \varepsilon(\infty), \varepsilon(\alpha) + \tilde{\varepsilon}(\infty, \Delta_2\Delta_\infty)\}
\end{aligned} \tag{46}$$

We then derive an upper bound on $\widehat{\varepsilon}(\alpha)$ by monotonicity of RDP,

$$\widehat{\varepsilon}(\alpha) \overset{(a)}{\leq} \widehat{\varepsilon}(\infty) \overset{(b)}{\leq} \tilde{\varepsilon}(\infty, \Delta_2\Delta_\infty) + \varepsilon(\infty) \tag{47}$$

where step (a) follows from Lemma D and step (b) follows by setting $\alpha = \infty$ in Equation (46).

By substituting the expression of RDP and SRDP parameter from Table 1 for Laplace mechanism, the L_2 and L_∞ sensitivity (Δ_2 and Δ_∞) for deduplication, quantization, imputation and PCA from Proposition 3, Corollary 4 and Proposition 5, and the Lipschitz parameter and global sensitivity of the output function f into Equation (45), we complete the proof of the overall privacy guarantees for Laplace mechanism.

Similarly, for exponential mechanism, we substitute the expression of RDP and SRDP parameter from Table 1 for exponential mechanism, the L_2 and L_∞ sensitivity (Δ_2 and Δ_∞) for deduplication, quantization, PCA, standard scaling and mean imputation from Proposition 3, 5 and 6 and Corollary 4, and the Lipschitzness and the global sensitivity of the score function Q, into Equation (45). This completes the proof of the overall privacy guarantees for exponential mechanism.

□

In the following, we provide the proof of combining DP-SGD with subsampling with PCA for rank reduction because datasets preprocessed by PCA for rank reduction satisfies Definition 7 automatically.

Proof of Theorem 17 (iii). We first note that for PCA, for any two neighboring datasets, S_1, S_2, by the definition of sensitivity of PCA, $\pi_{\mathrm{PCA-rank}}(S_1), \pi_{\mathrm{PCA-rank}}(S_2)$ and $\pi_{\mathrm{PCA-dim}}(S_1), \pi_{\mathrm{PCA-dim}}(S_2)$ has inverse point-wise divergence 1 because $\Delta_\infty = n$.

Following a similar argument as in the proof Gaussian mechanism (Theorem 17 (i)), we set the parameters in Theorem 2 $c_1 = c_2 = 2$ and obtain Equation (45). Then, we substitute in Equation (45) the following parameters: a)RDP and SRDP parameters for DP-SGD with subsampling from Table 1, b)$\gamma = 1$, c) Lipschitzness and smoothness parameter of the loss function ℓ, and d) L_2 and L_∞ sensitivity (Δ_2 and Δ_∞) for PCA and standard scaling from Proposition 5 and 6. This completes the proof. □

For DP-SGD with iteration, the pre-processing mechanisms achieve a tighter bound while satisfying $\Delta_\infty \ll n$ by Definition 8. Therefore, this method will not provide tighter bound for PCA and standard scaling. However, it improves the privacy analysis for imputation, deduplication and quantization, as proved below.

Proof of Theorem 17 (iv). We first note that the maximum divergence κ_τ of datasets pre-processed by imputation, deduplication or quantization is upper bounded by $\Delta_\infty = p$.

Following a similar argument as in the proof of Gaussian mechanism (Theorem 17 (i)), we set the parameters in Theorem 2 $c_1 = c_2 = 2$. We substitute in Equation (45) the following parameters: a)RDP and SRDP parameters for DP-SGD with iteration from Table 1, b)$\kappa_\tau = \Delta_\infty = p$, c) Lipschitzness and smoothness parameter of the loss function ℓ, and d) L_2 and L_∞ sensitivity (Δ_2 and Δ_∞) for quantization, deduplication and imputation from from Proposition 3 and Corollary 4. This completes the proof. □

E. Privacy and accuracy guarantees of Algorithm 1 (Section 5.2)

In this section, we present the proofs for the theoretical guarantees of Algorithm 1, as stated in Section 5.2. Specifically, we provide the proofs for the privacy guarantee (Theorem 8) and the accuracy guarantee (Proposition 9) of Algorithm 1.

Theorem 8. *For any L-Lipschitz and μ-smooth loss function ℓ, $\varepsilon > 0$ and $\delta \leq \exp\left(-1.05\varepsilon^2\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)\right)$, Algorithm 1 with privacy parameters ε, δ, and estimated lower bound β is $(\widehat{\varepsilon}+\varepsilon, \delta)$-DP on a dataset of size $n \geq 101$, where $\widehat{\varepsilon} = 3\varepsilon\sqrt{1.05\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)\log\frac{1}{\delta}}$.*

Proof. For brevity, we denote Algorithm 1 as $\mathcal{A}_{\mathrm{PTR}}$. For any two neighboring datasets $S_1, S_2 \in \mathcal{L}$, we consider two cases: i) the minimum eigen-gap of either S_1 or S_2 is smaller than or equal to β, and ii) the minimum eigen-gap of both S_1 and S_2 is greater than β.

Case (i) Without loss of generality, we assume the eigen-gap of S_1 is smaller than or equal to β. We will show that for any neighboring dataset S_2 of S_1, i.e. $d_H(S_1, S_2) = 1$, $\mathcal{A}_{\mathrm{PTR}}$ satisfies the following inequality, for any output set $O \subset \mathcal{H} \cup \{\perp\}$

$$\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_1) \in O\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_2) \in O\right] + \frac{\delta}{2}. \tag{48}$$

First consider the case $O \subset \mathcal{H}$. Then, assuming that the output is $\perp$ with high probability i.e. $\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_1) \neq \perp\right] \leq \delta/2$, we have that

$$\mathbb{P}[\mathcal{A}_{\mathrm{PTR}}(S_1) \in O] \leq \mathbb{P}[\mathcal{A}_{\mathrm{PTR}}(S_1) \neq \perp] \leq \frac{\delta}{2} \leq e^{\epsilon}\mathbb{P}[\mathcal{A}_{\mathrm{PTR}}(S_2) \in O] + \frac{\delta}{2}. \tag{49}$$

Now we show $\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_1) \neq \perp\right] \leq \delta/2$. Given a input dataset S, we denote the Γ in step 1 of Algorithm 1 by $\Gamma(S)$.

$$\mathbb{P}(\mathcal{A}_{\mathrm{PTR}}(S_1) \neq \perp) \overset{(a)}{\leq} \mathbb{P}\left(\Gamma(S_1) \leq \frac{\log\frac{2}{\delta}}{\varepsilon}\right) \overset{(b)}{=} \mathbb{P}\left(\mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log\frac{2}{\delta}}{\varepsilon}\right) \overset{(c)}{\leq} \frac{\delta}{2}. \tag{50}$$

where step (a) follows because the algorithm does not return $\perp$ only if $\Gamma(S_1) \geq \frac{\log\frac{2}{\delta}}{\varepsilon}$(Line 5 in Algorithm 1), and step (b) follows from $\min_{S':\delta_{\min}(S')\leq\beta} d_H(S', S_1) = 0$ as S_1 itself satisfies $\delta_{\min}(S_1) \leq \beta$. Step (c) follows by the tail bound of Laplace distributions: for a positive real number $t > 0$, $\mathbb{P}_{z\sim\mathrm{Lap}(0,b)}(z \geq t) \leq e^{-t/b}$.

Now consider the only other possible output $O = \{\perp\}$. We note that for any neighboring dataset S_2 of S_1,

$$\min_{S':\delta_{\min}(S')\leq\beta} d_H(S', S_1) = 0, \quad \min_{S':\delta_{\min}(S')\leq\beta} d_H(S', S_2) \leq 1 \tag{51}$$

This implies

$$\mathbb{P}\left(\Gamma(S_2) \leq \frac{\log\frac{2}{\delta}}{\varepsilon}\right) \geq \mathbb{P}\left(\mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log\frac{2}{\delta}}{\varepsilon} - 1\right). \tag{52}$$

Define $J(S) = \min_{S':\delta_{\min}(S')<\beta} d_H(S, S')$. It is simple to note that J has global sensitivity 1 due to Equation (51). Thus, Γ is essentially Laplace mechanism on J and thereby $(\varepsilon, 0)$-DP. This implies that for any neighboring datasets S_2 of S_1,

$$\frac{\mathbb{P}(\mathcal{A}_{\mathrm{PTR}}(S_1) = \perp)}{\mathbb{P}(\mathcal{A}_{\mathrm{PTR}}(S_2) = \perp)} = \frac{\mathbb{P}\left(\Gamma(S_1) \leq \frac{\log\frac{2}{\delta}}{\varepsilon}\right)}{\mathbb{P}\left(\Gamma(S_2) \leq \frac{\log\frac{2}{\delta}}{\varepsilon}\right)} \overset{(a)}{\leq} \frac{\mathbb{P}\left(\mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log\frac{2}{\delta}}{\varepsilon}\right)}{\mathbb{P}\left(\mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log\frac{2}{\delta}}{\varepsilon} - 1\right)} \overset{(b)}{\leq} e^{\varepsilon} \tag{53}$$

where the numerator of step (a) follows by Equation (50) and the denominator of step (a) follows by Equation (52). Step (b) follows by the tail bound of Laplace distributions. Combining Equation (50) and Equation (53), we show that Equation (48) holds in case (i).

Case (ii) Consider S_1, S_2 whose k^{th}-eigengaps are greater than β. We will show that the following holds, for any $O \subset \mathcal{H} \cup \{\perp\}$

$$\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_1) \in O\right] \leq e^{\widehat{\varepsilon}}\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_2) \in O\right] + \frac{\delta}{2} \tag{54}$$

where $\widehat{\varepsilon} = 3\varepsilon\sqrt{1.05\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)\log\frac{1}{\delta}}$. First consider the case $O = \{\perp\}$. Following a similar argument as Equation (53), we have

$$\frac{\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_1) = \perp\right]}{\mathbb{P}\left[\mathcal{A}_{\mathrm{PTR}}(S_2) = \perp\right]} \leq e^{\varepsilon} \leq e^{\widehat{\varepsilon}}. \tag{55}$$

Now consider the case when the output is not $\perp$. Then, choosing any $\alpha \geq 11$ yields that the privacy parameter of DP-GD combined with non-private PCA is $\left(\alpha, 1.05\alpha\varepsilon\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)\right)$-RDP from Table 2. Then, invoking Lemma M converts the RDP parameter to Approximate DP.

Lemma M (RDP to Approximate DP (Mironov, 2017))**.** *If $\mathcal{A}$ is an (α, ε)-RDP algorithm, then for $0 < \delta < 1$, it satisfies $\left(\varepsilon + \frac{\log\frac{1}{\delta}}{\alpha-1}, \delta\right)$-differential privacy.*

In particular, we choose $\alpha = \sqrt{\frac{\log\frac{1}{\delta}}{1.05\varepsilon^2\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)}} + 1 \geq 11$ and obtain that the output in step 6 of Algorithm 1 is $(\widehat{\varepsilon}, \delta)$-DP, where

$$\widehat{\varepsilon} \overset{(a)}{\leq} 2\varepsilon\sqrt{1.05\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)\log\frac{1}{\delta}} + 1.05\varepsilon\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right) \overset{(b)}{\leq} 3\varepsilon\sqrt{1.05\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)\log\frac{1}{\delta}}$$

Here, step (a) follows by Lemma M with chosen α and step (b) utilizes the fact that δ satisfies $\log\frac{1}{\delta} \geq 1.05\varepsilon^2\left(1+\frac{12.2^2\mu^2}{L^2\beta^2}\right)$.

This algorithm also discloses information about the dataset regarding its minimum eigen-gap. Specifically, when the minimum eigen-gap of the private dataset is smaller than β, then with high probability the output is $\perp$. However, the additional privacy cost incurred is smaller than that of releasing Γ directly. As Γ is a Laplace mechanism with global sensitivity 1, releasing Γ satisfies $(\varepsilon, 0)$-DP. By basic composition of approximate differential privacy (Lemma N), we can combine the privacy cost for releasing the results from DP-GD and releasing Γ. Thus, Algorithm 1 is $(\widehat{\varepsilon} + \varepsilon, \delta)$-DP under case (ii).

Lemma N (Basic composition of differential privacy (Dwork & Roth, 2014))**.** *If algorithm $\mathcal{A}_1$ is $(\varepsilon_1, \delta_1)$-DP and $\mathcal{A}_2$ is $(\varepsilon_2, \delta_2)$-DP, then $(\mathcal{A}_1, \mathcal{A}_2)$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.*

Combining the two cases concludes the proof. □

Proposition 9. *For δ, ε defined in Algorithm 1, $n \geq 101$, and any $\beta \leq \delta_k(S)$, with probability at least $1 - \xi$, Algorithm 1 outputs $\hat{\theta}$ such that the excess empirical risk*

$$\mathbb{E}\left[\hat{\ell}_S(\hat{\theta})\right] - \hat{\ell}_S(\theta^\star) = O\left(\frac{L\sqrt{1 + k\log 1/\delta}}{\varepsilon n}\right) + 2L\Lambda, \tag{4}$$

where $\xi = \frac{1}{2\delta}\exp\left(-\frac{(\delta_k(S)-\beta)n\varepsilon}{12.2}\right)$, $\Lambda = \sum_{i=k+1}^{d}\lambda_i(S)$ where the high probability is over the randomness in Step 1 and the expectation is over the randomness of Step 6.

Proof. We first prove that the probability that Algorithm 1 outputs $\perp$ is small. Then, we show that in the even that it does not output $\perp$ but a real vector, then the excess empirical risk is small.

Let $\mathcal{S}_{d_H} = \{S, S' \mid d_H(S, S') = 1 \wedge \delta_{\min}(S), \delta_{\min}(S') > 0\}$. Let $\Sigma_S, \Sigma_{S'}$ denote the covariance matrices of S and S' respectively and $\delta_k(S)$ denote the kth eigen-gap of Σ_S. Then, for any $S, S' \in \mathcal{S}_{d_H}$, we first upper bound $|\delta_k(S) - \delta_k(S')|$

$$\begin{aligned} |\delta_k(S') - \delta_k(S)| &= \left|\lambda_k(\tilde{\Sigma}) - \lambda_{k+1}(\tilde{\Sigma}) - \lambda_k(\Sigma) + \lambda_{k+1}(\Sigma)\right| \\ &\leq \left|\lambda_k(\tilde{\Sigma}) - \lambda_k(\Sigma)\right| + \left|\lambda_{k+1}(\Sigma) - \lambda_{k+1}(\tilde{\Sigma})\right|. \end{aligned} \tag{56}$$

To bound this term, we upper bound $\left|\lambda_k(\tilde{\Sigma}) - \lambda_k(\Sigma)\right|$ for any k. Using Weyl's inequality (Schindler, 2015)

$$\max_{j\in[d]}\left|\lambda_j(\Sigma) - \lambda_j(\tilde{\Sigma})\right| \leq \left\|\tilde{\Sigma} - \Sigma\right\|_{\mathrm{op}} \leq \left\|\tilde{\Sigma} - \Sigma\right\|_F \leq \frac{6.1}{n}, \tag{57}$$

where the last inequality follows from a similar argument as Equation (41) in the proof of Proposition 5 for $n \geq 101$. This yields

$$|\delta_k(S') - \delta_k(S)| \leq \frac{12.2}{n}. \tag{58}$$

Thus, for any S_0, S_m with $d_H(S_0, S_m) = m$, we can construct a series of datasets $S_1, \ldots, S_{m-1}$ such that $d_H(S_j, S_{j+1}) = 1$ for $j \in \{0, ..., m-1\}$. By applying Equation (58) iteratively over each i, we have

$$\delta_k(S_0) \leq \delta_k(S_m) + \frac{12.2m}{n}. \tag{59}$$

More generally, for any pairs of dataset S, S', we have

$$d_H(S, S') \geq \frac{|\delta_k(S) - \delta_k(S')|\, n}{12.2} \tag{60}$$

Hence,

$$\min_{S':\delta_k(S')\leq\beta} d_H(S, S') \geq \frac{(\delta_k(S) - \beta)\, n}{12.2}. \tag{61}$$

Now, we upper bound the probability that Algorithm 1 does not output $\perp$.

$$\begin{aligned}
\mathbb{P}\left[\Gamma(S) \geq \frac{\log\frac{1}{\delta}}{\varepsilon}\right] &= \mathbb{P}\left[\min_{S':\delta_k(S')\leq\beta} d_H(S, S') + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \geq \frac{\log\frac{1}{\delta}}{\varepsilon}\right] \\
&\geq \mathbb{P}\left[\frac{(\delta_k(S) - \beta)\, n}{12.2} + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \geq \frac{\log\frac{1}{\delta}}{\varepsilon}\right] \\
&\overset{(a)}{=} 1 - \frac{1}{2}\exp\left(\varepsilon\left(\frac{\log\frac{1}{\delta}}{\varepsilon} - \frac{(\delta_k(S) - \beta)\, n}{12.2}\right)\right) \\
&= 1 - \frac{1}{2\delta}\exp\left(\frac{\beta - \delta_k(S)}{12.2} n\varepsilon\right),
\end{aligned} \tag{62}$$

where step (a) follows from the tail bound of Laplace distribution when $\delta_k(S) \geq \beta + \frac{12.2\log\frac{1}{\delta}}{n\varepsilon}$.

We have shown that Algorithm 1 does not output $\perp$ with probability $1 - \frac{1}{2\delta}\exp\left(\frac{\beta-\delta_k(S)}{12.2} n\varepsilon\right)$. It remains to derive the convergence guarantee of Algorithm 1 when the output is not $\perp$. To show this, we utilize the property that PCA transforms the dataset to have a low-rank covariance matrix, which allows us to apply the dimension-independent convergence guarantee of DP-GD following the analysis of Song et al. (2021) and establish the desired convergence guarantee.

Let $A_k^\top \in \mathbb{R}^{k\times d}$ be the matrix consisting of the first k eigenvectors of the covariance matrix of S, $\Sigma = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Let $S_k = \{(A_k A_k^\top x_i, y_i)\}_{i=1}^n$. We can decompose the error into three terms,

$$\mathbb{E}\left[\hat{\ell}_S(\hat{\theta})\right] - \ell_S(\theta^\star) \leq \underbrace{\left|\mathbb{E}\left[\hat{\ell}_S(\hat{\theta})\right] - \mathbb{E}\left[\hat{\ell}_{S_k}(\hat{\theta})\right]\right|}_{(a)} + \underbrace{\left|\mathbb{E}\left[\hat{\ell}_{S_k}(\hat{\theta})\right] - \ell_{S_k}(\theta^\star)\right|}_{(b)} + \underbrace{|\ell_S(\theta^\star) - \ell_{S_k}(\theta^\star)|}_{(c)} \tag{63}$$

By Theorem 3.1 in Song et al. (2021), part (b) is upper bounded by $O\left(\frac{L\sqrt{1+k\log\frac{1}{\delta}}}{\varepsilon n}\right)$.

Lemma O (Theorem 3.1 in Song et al. (2021))**.** *Let $\theta_0 = 0^p$ be the initial point of $\mathcal{A}_{\mathrm{GD}}$. Let the dataset be centered at 0. Let k be the rank of the projector to the eigenspace of the covariance matrix $\sum_{i=1}^n x_i x_i^\top$. For a L-Lipschitz loss function ℓ, $\mathcal{A}_{\mathrm{GD}}$ with $T = n^2\varepsilon^2$ with appropriate learning rate η output $\hat{\theta}$ satisfying*

$$\mathbb{E}\left[\hat{\ell}_S(\hat{\theta})\right] - \hat{\ell}_S(\theta^\star) \leq O\left(\frac{L\sqrt{1 + 2k\log\frac{1}{\delta}}}{\varepsilon n}\right).$$

Next, we show that part (a) and (c) are upper bounded by $L\sum_{i=k+1}^{n}\lambda_i(S) = L\Lambda$. To show this, we prove that for any $\theta \in B_2^d$, $\left|\hat{\ell}_{S_k}(\theta) - \hat{\ell}_S(\theta)\right| \leq L\sum_{i=k+1}^{n}\lambda_i(S)$.

$$\begin{aligned}\left|\hat{\ell}_{S_k}(\theta) - \hat{\ell}_S(\theta)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\ell(\theta^\top A_k A_k^\top x_i, y_i) - \frac{1}{n}\sum_{i=1}^{n}\ell(\theta^\top x_i, y_i)\right| \\ &\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n} L\left\|\theta^\top A_k A_k^\top x_i - \theta^\top x_i\right\|_2 \\ &\overset{(b)}{\leq} \frac{1}{n}\sum_{i=1}^{n} L\left\|A_k A_k^\top x_i - x_i\right\|_2 \overset{(c)}{\leq} L\sum_{i=k+1}^{d}\lambda_i(S) = L\Lambda\end{aligned} \tag{64}$$

where step (a) follows from Lipschitzness of the loss function ℓ, step (b) follows due to the projection step in DP-GD projects θ to the Euclidean ball with radius 1, and step (c) is obtained by substituting the reconstruction error of PCA, which is $\sum_{i=k+1}^{d}\lambda_i(S)$. This concludes the proof. □

F. Experiment setups and discussion

F.1. Experiment setups

Data generation The synthetic data is generated with the `make_classification` function in the `sklearn` library. We generate a 2-class low rank dataset consisting of 1000 data points with dimension 6000 and rank 50. We set the parameter `n_cluster_per_class` in `make_classification` to 1.

Models and allocation of privacy budget We compared the excess empirical loss of three models in Figure 3. We provide the details of the models below.

- **Pre-processed DP pipeline:** We employ non-private PCA to reduce the dimensionality of the original dataset to k and then apply private logistic regression. In particular, we use the `make_private_with_epsilon` method from the `Opacus` library with PyTorch SGD optimizer with learning rate `1e-2`, `max_grad_norm = 10` and `epochs = 10`. The privacy parameters `epsilon` and `delta` are obtained by adjusting the desired overall privacy level with our framework (Theorem 17).
- **No pre-processing:** We directly apply private logistic regression with the same parameters on the original high-dimensional dataset.
- **DP-PCA:** We first implement the DP-PCA in Chaudhuri et al. (2012) and then apply private logistic regression. We allocate half of the privacy budget to DP-PCA and the remaining half to private logistic regression.

F.2. Discussion on clipping

In practice, when the sensitivity of the original function is unbounded, one can apply the technique of clipping to restrict the sensitivity (Abadi et al., 2016; Liu et al., 2022). We can also incorporate clipping into our pre-processed DP pipeline. Specifically, we first non-privately compute the PCA projection matrix A_k with the original dataset S. We clip the original data to S_{clipped} with the clipping threshold $C \in [0.1R, 0.7R, 0.99R]$, where R represents the maximum norm of the original dataset. Then, we apply the projection matrix to the clipped dataset to obtain the pre-processed dataset $S_{\text{preprocessed}} = A_k A_k^\top S_{\text{clipped}}$. Finally, we apply private logistic regression on $S_{\text{preprocessed}}$, with the privacy parameter set by our framework.

As shown in Table 4, the effect of clipping on accuracy depends on the clipping threshold. This is because clipping reduces the L_2 sensitivity of PCA from ${}^{12R}/_{n\delta^k_{\min}}$ to ${}^{12C}/_{n\delta^k_{\min}}$, allowing for selection of a larger privacy parameter during private learning and better accuracy. However, clipping also introduces inaccuracies in the dataset.

Table 4. Excess empirical loss of clipped pre-processed DP-pipeline

	Clipping 0.1	Clipping 0.7	Clipping 0.99	Pre-processed DP pipeline
1.0	0.84	**0.89**	0.88	**0.89**
2.0	0.84	**0.89**	**0.89**	**0.89**
5.0	0.85	**0.90**	**0.90**	**0.90**

However, we note that R is dataset dependent and might lead to additional privacy leakage.