Private and Federated Stochastic Convex Optimization: Efficient Strategies for Centralized Systems

Roie Reshef [*1] **Kfir Yehuda Levy** [*1]

Abstract

This paper addresses the challenge of preserving privacy in Federated Learning (FL) within centralized systems, focusing on both trusted and untrusted server scenarios. We analyze this setting within the Stochastic Convex Optimization (SCO) framework, and devise methods that ensure Differential Privacy (DP) while maintaining optimal convergence rates for homogeneous and heterogeneous data distributions. Our approach, based on a recent stochastic optimization technique, offers linear computational complexity, comparable to non-private FL methods, and reduced gradient obfuscation. This work enhances the practicality of DP in FL, balancing privacy, efficiency, and robustness in a variety of server trust environments.

1. Introduction

Federated Learning (FL) is a novel framework in Machine Learning (ML), enabling collaborative learning among myriad of decentralized devices or systems (McMahan et al., 2017; Kairouz et al., 2021b). Privacy is a paramount issue in FL, underscoring the critical need to prevent the disclosure of private information as a result of the training process.

To address this challenge, Differential Privacy (DP) has emerged as a robust framework for quantifying and managing privacy risks (Dwork et al., 2006a;b). DP offers a formal guarantee that the outcome of a data analysis does not significantly change when any single individual's data is added or removed, thereby ensuring that individual privacy is maintained (Dwork & Roth, 2014). Implementing DP guarantees in FL has been extensively studied e.g. in (Huang et al., 2020; Wei et al., 2020; Girgis et al., 2021; Noble et al., 2022; Lowy & Razaviyayn, 2023; Lowy et al., 2023), necessitating a nuanced equilibrium between data obfuscation and maintaining model accuracy. This intricate balance is crucial to preserve the efficacy of the learning process while ensuring privacy. The critical challenge arising in this context is in ensuring the confidentiality of the model updates transmitted over the distributed network.

FL differentiates between device-based learning, limited by device availability and bandwidth, and silo-based learning, involving full participation from all machines, often across organizations (Kairouz et al., 2021b). This paper focuses on silo-based learning, crucial for achieving accurate and consistent models, and we shall refer to silos as "machines".

In this work, we investigate DP guarantees for FL within the Stochastic Convex Optimization (SCO) framework, a cornerstone in the design and analysis of ML algorithms (Shalev-Shwartz et al., 2009; Shalev-Shwartz & Ben-David, 2014). And we focus on centralized scenarios where a (central) sever employs M machines in parallel. So far, this setting has been mainly investigated in the context of finite-sum problems (ERM), e.g. in (Huang et al., 2020; Wei et al., 2020; Girgis et al., 2021; Noble et al., 2022). Nevertheless, translating ERM guarantees to population loss guarantees leads to sub-optimal bounds, see e.g. (Bassily et al., 2019).

Conversely, optimal guarantees for the population loss in this context were recently substantiated in (Lowy & Razaviyayn, 2023). This work provides different guarantees, depending on the level of trust that machines have in a central server. Concretely, for **(i)** *untrusted* server scenarios, this work substantiates a convergence rate of $O\left(\frac{1}{\sqrt{nM}} + \frac{\sqrt{d}}{\epsilon n\sqrt{M}}\right)$, where M is the number of machines, n is the number of samples used per machine during the training process, d is the dimension of the problem, and ε is the level of ensured DP. Conversely, **(ii)** for the case of a *trusted* server, an improved rate of $O\left(\frac{1}{\sqrt{nM}} + \frac{\sqrt{d}}{\epsilon nM}\right)$ is substantiated. (Lowy & Razaviyayn, 2023) also provide a matching lower bound for these rates. These guarantee is substantiated for the *homogeneous* case, where the data distribution is identical across all machines.

Nevertheless, as we detail in Appendix F, the approach of (Lowy & Razaviyayn, 2023) requires a computational

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Technion, Haifa, Israel. Correspondence to: Roie Reshef <sror@campus.technion.ac.il>, Kfir Yehuda Levy <kfiryehud@gmail.com>.

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

complexity that is proportional to $|\mathcal{S}|^{3/2}$, where $|\mathcal{S}|$ is the size of overall data used by all machines. This is substantially higher compared to standard non-private FL approaches that require only $\propto |\mathcal{S}|$ computations. Moreover, the above bounds are only substantiated for the homogeneous case; and slightly worse bounds are substantiated for the more general heterogeneous case.

Our paper explores methods that ensure DP for centralized FL scenarios, under the setting of SCO. Our contributions:

- **Untrusted Server.** In the case of untrusted server, we provide an approach that ensures an optimal convergence rate rate of $O\left(\frac{1}{\sqrt{nM}} + \frac{\sqrt{d}}{\epsilon n\sqrt{M}}\right)$, with a level ϵ of DP. Our bound applies simultaneously to both homogeneous and heterogeneous cases.
- **Trusted Server.** In the case of trusted server, we provide an approach that ensures an optimal convergence rate rate of $O\left(\frac{1}{\sqrt{nM}} + \frac{\sqrt{d}}{\epsilon nM}\right)$, with a level ϵ of DP. Our bound applies simultaneously to both homogeneous and heterogeneous cases.
- The computational complexity of our approaches is linear in $|\mathcal{S}|$; matching the complexity standard non-private FL approaches (Dekel et al., 2012).

Our results build on a recent stochastic optimization technique named μ^2-SGD (Levy, 2023), that combines two different momentum mechanisms into SGD. We customize this approach to the centralized FL setting, and show it is less sensitive to a change of a single sample in $\mathcal{S}$ compared to standard SGD, which allows to obtain DP with substantially less gradient obfuscation.

Remark. Our paper categorizes the FL environment into trusted or untrusted server scenarios. Our definition of an untrusted server mirrors that in (Lowy & Razaviyayn, 2023). However, this latter work introduces a nuanced concept of a trusted shuffler, capable only of shuffling but not aggregating gradient message identities. We juxtapose this with our trusted server scenario, essentially treating the shuffler and server as a single entity. It is an interesting open question to understand whether one can obtain the same guarantees that we derive for the trusted server scenario, under the more nuanced assumption of a trusted shuffler.

Related Work

SCO with DP guarantees. There exists a rich literature in the context of differential privacy (DP) within stochastic convex optimization (SCO). Foundational work in empirical risk minimization (ERM) within the finite-sum problems has been substantially contributed to by a series of studies, among are (Chaudhuri et al., 2011; Kifer et al., 2012; Thakurta & Smith, 2013; Song et al., 2013; Duchi et al., 2013; Ullman, 2015; Talwar et al., 2015; Wu et al., 2017; Wang et al., 2018; Iyengar et al., 2019; Kairouz et al., 2021a; Avella-Medina et al., 2021; Ganesh et al., 2023). Nevertheless, these works primarily focus on training loss (ERM); and translating these results to population loss guarantees via standard uniform convergence (Shalev-Shwartz et al., 2009) leads to sub-optimal bounds as described in (Bassily et al., 2019; Feldman et al., 2020).

The studies of (Bassily et al., 2014; 2019) have advanced our understanding by establishing population loss guarantees for DP-SCO, and the latter has provided optimal bounds; albeit with super-linear computational complexity $\propto |\mathcal{S}|^{3/2}$. This was later improved by (Feldman et al., 2020), which attained the optimal guarantees with a sample complexity which is linear in $|\mathcal{S}|$, but provides privacy only on the final iterate.

FL with DP guarantees. We have already mentioned previous works that analyze FL with DP guarantees under the SCO setting. Another notable work in this context is (Cheu et al., 2022), which obtains the optimal bounds for the trusted server case, albeit using an expensive vector-shuffling routine, that leads to a computational complexity which is even larger than $|\mathcal{S}|^{3/2}$ (see Cor. B.11 therein).

2. Preliminaries

In this section, we provide the necessary background for analyzing private learning in the context of the standard (single machine) setting, and will touch upon the necessary background towards the more complex federated (or parallel) learning setting.

2.1. Convex Loss Minimization

We focus on Stochastic Convex Optimization (SCO) scenarios where the objective function, $f : \mathcal{K} \mapsto \mathbb{R}$, is convex and takes the following form:

$$f(x) = \mathbb{E}_{z\sim\mathcal{D}}\left[f(x; z)\right] \tag{1}$$

here, $\mathcal{K} \subset \mathbb{R}^d$ is a compact convex set, and $\mathcal{D}$ represents an unknown data distribution from which we may draw i.i.d. samples. A learning algorithm receives a dataset $\mathcal{S} = \{z_1, \ldots, z_n\} \subset \mathcal{Z}^n$ ($\mathcal{Z}$ is the set where the samples reside), of n i.i.d. samples from $\mathcal{D}$, and outputs a solution $x_{\text{output}} \in \mathcal{K}$. Our performance measure is the expected excess loss $\mathcal{R}(x_{\text{output}})$, defined as:

$$\mathcal{R}(x_{\text{output}}) = \mathbb{E}\left[f(x_{\text{output}})\right] - \min_{x\in\mathcal{K}}\{f(x)\} \tag{2}$$

expectation is taken w.r.t. the randomness of the samples, as well as w.r.t. the (possible) randomization of the algorithm.

We focus on first-order optimization methods that iteratively employ the samples in $\mathcal{S}$ to create a sequence of query points, culminating in a solution $x_{\text{output}} \in \mathcal{K}$. To elaborate, at each iteration t, such iterative methods maintain a query point $x_t \in \mathcal{K}$, calculated from previous query points and samples $z_1, \ldots, z_{t-1}$. The subsequent query point x_{t+1} is then determined using x_t and a gradient estimate g_t. This estimate is obtained by drawing a fresh sample $z_t \sim \mathcal{D}$ (taken from $\mathcal{S}$), independently of past samples, and calculating $g_t = \nabla f(x_t; z_t)$, where the derivative is with respect to x.

The independence of the samples ensures that g_t is an unbiased estimator of $\nabla f(x_t)$, in the sense that $\mathbb{E}[g_t | x_t] = \nabla f(x_t)$. It is useful to conceptualize the calculation of $g_t = \nabla f(x_t; z_t)$ as a sort of (noisy) Gradient Oracle. Upon receiving a query point $x_t \in \mathcal{K}$, this Oracle outputs a vector $g_t \in \mathbb{R}^d$, serving as an unbiased estimate of $\nabla f(x_t)$.

Assumptions We will make the following assumptions.
Bounded Diameter: There exists $D > 0$ such that $\max_{x,y \in \mathcal{K}} \|x - y\| \leq D$
Convexity: The function $f(\cdot; z)$ is convex $\forall z \in \textbf{Supp}\{\mathcal{D}\}$.

We also make these assumption about $f(\cdot; z), \forall z \in \mathcal{Z}$:
Lipschitz: There exists $G > 0$ such that:

$$|f(x; z) - f(y; z)| \leq G \|x - y\|, \quad \forall x, y \in \mathcal{K}$$

This also implies that $\|\nabla f(x; z)\| \leq G, \forall x \in \mathcal{K}$.
Smoothness: There exists $L > 0$ such that:

$$\|\nabla f(x; z) - \nabla f(y; z)\| \leq L \|x - y\|, \quad \forall x, y \in \mathcal{K}$$

Since these assumptions hold for $f(x; z)$ for every $z \in \textbf{Supp}\{\mathcal{D}\} \subset \mathcal{Z}$, they also hold for $f(x)$. The above Lipschitz and Smoothness assumptions imply the following:
Bounded Variance: There exist $0 \leq \sigma \leq G$ such that:

$$\mathbb{E} \|\nabla f(x; z) - \nabla f(x)\|^2 \leq \sigma^2, \quad \forall x \in \mathcal{K} \tag{3}$$

Bounded Smoothness Variance: $\exists \sigma_L \in [0, L]$ such,

$$\mathbb{E} \|(\nabla f(x; z) - \nabla f(x)) - (\nabla f(y; z) - \nabla f(y))\|^2 \leq \sigma_L^2 \|x - y\|^2, \quad \forall x, y \in \mathcal{K} \tag{4}$$

For completeness, we provide a proof in Appendix B.1.
Notations: We will employ the following notation, Projection $\Pi_{\mathcal{K}}(x) = \arg\min_{y \in \mathcal{K}} \|x - y\|$, the notation $[n] = \{1, \ldots, n\}$, the notation $\alpha_{1:t} = \sum_{\tau=1}^{t} \alpha_\tau$.

2.2. Differential Privacy

In order to discuss privacy, we need to measure how does private information affects the outputs of an algorithm. The idea behind Differential Privacy (DP) is to look at the output of the algorithm with the private information and without it, and measure the difference between these outputs. The more similar they are, the more private the algorithm is. Concretely, similarity is measured w.r.t. the difference between the probability distributions of the (randomized) outputs.

Rényi Divergence Below is the definition of Rényi divergence, a popular difference measure between probability distributions, which is prevalent in the context of DP.

Definition 2.1 (Rényi Divergence (Rényi, 1961))**.** Let $1 < \alpha < \infty$, and let P, Q be probability distributions such that $\textbf{Supp}\{P\} \subseteq \textbf{Supp}\{Q\}$. The Rényi divergence of order α between P and Q is defined as:

$$\mathbb{D}_\alpha(P \| Q) = \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{X \sim P} \left[\left(\frac{P(X)}{Q(X)} \right)^{\alpha - 1} \right] \right)$$

We follow with the convention that $\frac{0}{0} = 0$. If $\textbf{Supp}\{P\} \nsubseteq \textbf{Supp}\{Q\}$, then the Rényi divergence is defined to be ∞. Divergence of orders $\alpha = 1, \infty$ are defined by continuity.

Notation: If $X \sim P, Y \sim Q$, we will use $\mathbb{D}_\alpha(P \| Q)$ & $\mathbb{D}_\alpha(X \| Y)$ interchangeably.

Adding Gaussian noise is a popular primitive in DP, and we will therefore find the following lemma useful:

Lemma 2.2. *Let $P \sim \mathcal{N}(\mu, I\sigma^2)$ and $Q \sim \mathcal{N}(\mu + \Delta, I\sigma^2)$, two Gaussian distributions. Then, $\mathbb{D}_\alpha(P \| Q) = \frac{\alpha \|\Delta\|^2}{2\sigma^2}$.*

For completeness we provide a proof in Appendix B.2.

Differential Privacy Definitions

Definition 2.3 (Differential Privacy (Dwork et al., 2006a;b))**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-diferentially private, or (ϵ, δ)-DP, if for all neighbouring datasets $\mathcal{S}, \mathcal{S}'$ that differs in a single element, and for all events $\mathcal{O}$, we have:

$$\mathbb{P}\{\mathcal{A}(\mathcal{S}) = \mathcal{O}\} \leq e^\epsilon \mathbb{P}\{\mathcal{A}(\mathcal{S}') = \mathcal{O}\} + \delta$$

The term *neighbouring datasets* refers to $\mathcal{S}, \mathcal{S}'$ being *ordered* sets of data points that only differ on *a single element*.

In this paper, we would adopt a different and prevalent privacy measure that relies on the Rényi divergence:

Definition 2.4 (Rényi Differential Privacy (Mironov, 2017))**.** For $1 \leq \alpha \leq \infty$ and $\epsilon \geq 0$, a randomized algorithm $\mathcal{A}$ is (α, ϵ)-Rényi diferentially private, or (α, ϵ)-RDP, if for all neighbouring datasets $\mathcal{S}, \mathcal{S}'$ that differ in a single element,

$$\mathbb{D}_\alpha(\mathcal{A}(\mathcal{S}) \| \mathcal{A}(\mathcal{S}')) \leq \epsilon$$

Curiously, it is known that RDP implies DP. This is established in the next lemma:

Lemma 2.5 ((Mironov, 2017))**.** *If $\mathcal{A}$ satisfies (α, ϵ)-RDP, then for all $\delta \in (0, 1)$, it also satisfies $\left(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$-*

DP. In particular, if $\mathcal{A}$ satisfies $\left(\alpha, \frac{\alpha\rho^2}{2}\right)$-RDP for every $\alpha \geq 1$, then for all $\delta \in (0,1)$, it also satisfies $\left(\frac{\rho^2}{2} + \rho\sqrt{2\log(1/\delta)}, \delta\right)$-DP.

For completeness we provide a proof in Appendix B.3.

2.3. Federated Learning

In federated learning (FL), multiple machines collaborate together to solve a joint problem. We assume that there exist M machines, and that each machine $i \in [M]$ may independently draw i.i.d. samples from a distribution $\mathcal{D}_i$. Similarly to the standard SCO setting, we can associate an expected convex loss $f_i(\cdot)$ with machine $i \in [M]$, which is defined as $f_i(x) := \mathbb{E}_{z^i \sim \mathcal{D}_i}\left[f_i\left(x; z^i\right)\right]$. We denote the joint objective of all machines by $f : \mathcal{K} \mapsto \mathbb{R}$, where:

$$f(x) := \frac{1}{M}\sum_{i\in[M]} f_i(x) := \frac{1}{M}\sum_{i\in[M]} \mathbb{E}_{z^i\sim\mathcal{D}_i}\left[f_i\left(x; z^i\right)\right]$$

Thus, the objective is an average of M functions $\{f_i : \mathcal{K} \mapsto \mathbb{R}\}_{i\in[M]}$, and each such $f_i(\cdot)$ can be written as an expectation over losses $f_i\left(\cdot, z^i\right)$ where the z^i are drawn from some distribution $\mathcal{D}_i$ which is unknown to the learner. For ease of notation, in what follows we will not explicitly denote $\mathbb{E}_{z^i\sim\mathcal{D}_i}[\cdot]$ but rather use $\mathbb{E}[\cdot]$ to denote the expectation w.r.t. all randomization.

In order to collaboratively minimize $f(\cdot)$, the machines may synchronize and communicate through a central machine called the *Parameter Server* ($\mathcal{PS}$). We will focus on the most common parallelization scheme (Dekel et al., 2012), where at every round t the $\mathcal{PS}$ communicates a query point x_t to all machines. Then, every machine performs a gradient computation based on its local data and communicates gradient estimates back to the $\mathcal{PS}$, which in turn aggregates these estimates and updates the query point. Similarly to the standard SCO setting, our performance measure is the Expected loss w.r.t. $f(\cdot)$ (see Equation (2)). Moreover, we shall assume that each machine $i \in [M]$ maintains and utilizes a dataset $\mathcal{S}_i$ of samples that are drawn i.i.d. from $\mathcal{D}_i$. Finally, we assume that for every $i \in [M]$, the functions $\{f_i(\cdot, z)\}_{z\in\mathcal{Z}_i}$ are convex G-Lipschits and L-smooth, as well as make bounded variance and smoothness variance assumptions (see Equations (3) and (4)).

In the context of privacy, we explore the case where the FL process should ensure DP guarantees individually for every machine $i \in [M]$. The challenge in this context arises due to the fact that each machine communicates *several times* with the $\mathcal{PS}$ which may uncover more information regarding its private dataset. Concretely, we assume that machines do not trust each other, and thus cannot allow other machines to uncover private data. We deal with two cases: **(i) Untrusted Server.** Here the $\mathcal{PS}$ may not uncover private data, and we are therefore required to ensure DP-guarantees w.r.t. information that the machines communicate to the $\mathcal{PS}$ (i.e. gradient estimates). **(ii) Trusted Server.** In this case all machines trust the $\mathcal{PS}$ and are therefore allowed to send information that may expose their privacy. Nevertheless, we still require that machines may not uncover private data of one another from the information that is received from the $\mathcal{PS}$. Thus, for each machine $i \in [M]$, we are required to ensure DP-guarantees w.r.t. information that is distributed by the $\mathcal{PS}$ (i.e. query points).

3. Our Algorithm Mechanisms

Our approach builds on the standard gradient descent template, that we combine with two additional mechanisms, both are crucial to the result that we obtain. Next we elaborate on these mechanisms, and in the next sections discuss and analyze their combination.

3.1. μ^2-SGD

The μ^2-SGD (Levy, 2023) is a variant of standard SGD with several modifications. Its update rule is of the following form: $w_1 = x_1 \in \mathcal{K}$, and $\forall t > 1$:

$$\begin{aligned} w_{t+1} =& \Pi_{\mathcal{K}}\left(w_t - \eta\alpha_t d_t\right) \\ x_{t+1} =& \frac{\alpha_{1:t}}{\alpha_{1:t+1}} x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} w_{t+1} \end{aligned} \tag{5}$$

where $\{\alpha_t > 0\}_t$ are importance weights that may unequally emphasize different update steps. Concretely we will employ $\alpha_t \propto t$, which puts more emphasis on the more recent updates. Moreover, the $\{x_t\}_t$'s are a sequence of weighted averages of the iterates $\{w_t\}_t$, and d_t is an estimate for the gradient at the average point, i.e. of $\nabla f(x_t)$. This is different than standard SGD which employs estimates for the gradients at the iterates, i.e. of $\nabla f(w_t)$. This approach is related to a technique called Anytime-GD (Cutkosky, 2019), which is strongly-connected to the notions of momentum and acceleration (Cutkosky, 2019; Kavis et al., 2019).

While in the natural SGD version of Anytime-GD, one would employ the estimate $\nabla f(x_t; z_t)$, the μ^2-SGD approach suggests to employ a variance reduction mechanism to yield a *corrected momentum* estimate d_t in the spirit of (Cutkosky & Orabona, 2019). This is done as follows: $d_1 := \nabla f(x_1; z_1)$, and $\forall t > 2$:

$$d_t = \nabla f(x_t; z_t) + (1-\beta_t)(d_{t-1} - \nabla f(x_{t-1}; z_t)) \tag{6}$$

where $\beta_t \in [0,1]$ are called *corrected momentum* weights. It can be shown by induction that $\mathbb{E}[d_t] = \mathbb{E}[\nabla f(x_t)]$, nevertheless in general $\mathbb{E}[d_t|x_t] \neq \nabla f(x_t)$ (in contrast to standard SGD estimators). Nevertheless, it was shown in (Levy, 2023) that upon choosing *corrected momentum*

weights of $\beta_t := 1/t$, the above estimates enjoy an error reduction, i.e. $\mathbb{E}\left\|\varepsilon_t\right\|^2 := \mathbb{E}\left\|d_t - \nabla f(x_t)\right\|^2 \leq O(\tilde{\sigma}^2/t)$ at step t, where $\tilde{\sigma}^2 \leq O(\sigma^2 + \sigma_L^2 D^2)$. Implying that the error decreases with t, contrasting with standard SGD where the variance $\mathbb{E}\left\|\varepsilon_t^{\text{SGD}}\right\|^2 := \mathbb{E}\left\|g_t - \nabla f(x_t)\right\|^2$ remains uniformly bounded by σ^2.

3.2. Noisy-μ^2-SGD

An additional mechanism that we utilize consists of adding zero mean noise to the gradients, which in turn adds privacy to the learning algorithm (Abadi et al., 2016). Combining this idea with the μ^2-SGD approach (Equation (5)) induces the following update rule:

$$\begin{aligned} w_{t+1} &= \Pi_{\mathcal{K}}\left(w_t - \eta(\alpha_t d_t + Y_t)\right) \\ x_{t+1} &= \frac{\alpha_{1:t}}{\alpha_{1:t+1}} x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} w_{t+1} \end{aligned} \tag{7}$$

where d_t is a corrected momentum estimate (Equation (6)), and $\{Y_t \sim P_t\}$ is a sequence of independent zero mean noise terms. Curiously, there is a natural tradeoff in choosing the noise magnitude: larger noise degrades the convergence, but improves privacy.

4. Noisy-μ^2-SGD for Differentially-Private FL

In this section we deal with the case of multiple machines, as described in Section 2.3. Our parallelization approach, appearing in Algorithms 1 and 2 is based on the well known "minibacth-SGD" (a.k.a. "parallel-SGD") algorithmic template (Dekel et al., 2012). In this template the server sends a query point to all machines. Then, every machine $i \in [M]$ computes a gradient estimate based on its local data, and communicate it back to the server. And the latter averages these gradient estimates, and updates the models using a gradient step.

Algorithmic Approach Our approach is depicted in Algorithm 1 (untrusted server), and in Algorithm 2 (trusted server). It can be seen that our approach is inspired by Noisy-DP-μ^2(Section 3.2), and therefore differs than the standard minibacth-SGD in three different aspects: **(i)** the parameter server queries the gradients at the averages $\{x_t\}_{t\in[T]}$, which is in the spirit of Anytime-GD; **(ii)** Each machine $i \in [M]$ maintains and updates a (weighted) corrected momentum estimate $q_{t,i} := \alpha_t d_{t,i}$, which is an unbiased estimate of $\alpha_t \nabla f_i(x_t)$, in the spirit of μ^2-SGD (Equations (5) and (6)); **(iii)** A synthetic noise is injected, either to the individual estimates $q_{t,i}$ (in case of untrusted server, Algorithm 1),or to the aggregated estimate $q_t := \frac{1}{M}\sum_{i\in[M]} q_{t,i}$ (in case of trusted server, Algorithm 2).

Remark: Note that for the sake of establishing convergence guarantees we need to assume that the samples in $\mathcal{S}_i$ are drawn i.i.d. from $\mathcal{D}_i$. Nevertheless, our privacy guarantees do not necessitate this requirement, and for the sake of DP guarantees we only assume Lipschitz continuity and smoothness of $f_i(\cdot; z^i)$ for all $i \in [M], z^i \in \mathcal{Z}_i$.

Computational Complexity From the description of Algorithms 1 and 2, it can be directly seen that at every round t, each machine $i \in [M]$ employs a single sample $z_{t,i}$ which is used to compute two noisy gradients estimates $g_{t,i}, \tilde{g}_{t-1,i}$. Thus, the overall computational complexity of our approach is linear in the dataset size $|\mathcal{S}|$, where $\mathcal{S} := \cup_{i\in[M]}\mathcal{S}_i$.

Next we provide a general sensitivity and error analysis that applies to both Algorithms 1 and 2. Then we establish privacy and convergence guarantees for each setting.

4.1. Gradient Error & Sensitivity Analysis

Here we present and discuss an analysis for the error of the $q_{t,i}$ estimates, as well as analyze their sensitivity to a single point in the dataset. This analysis applies to both Algorithms 1 and 2, and will later serve us in deriving privacy and convergence guarantees.

Prior to the analysis, we shall introduce some notation. We will define $g_{t,i} := \nabla f(x_t; z_{t,i})\,, \tilde{g}_{t,i} := \nabla f(x_t, z_{t+1,i})\,, \bar{g}_{t,i} := \nabla f(x_{t,i})$. These notations allow us to write the update rule for $d_{t,i}$ in Algorithm 1 (as well as Algorithm 2) as follows:

$$\begin{aligned} d_{t+1,i} =& g_{t+1,i} + (1-\beta_{t+1})(d_{t,i} - \tilde{g}_{t,i}) \\ =& \beta_{t+1} g_{t+1,i} + (1-\beta_{t+1})(d_{t,i} + g_{t+1,i} - \tilde{g}_{t,i}) \end{aligned}$$

Using the notation $q_{t+1,i} := \alpha_{t+1} d_{t+1,i}$, we can rewrite the above equation:

$$\begin{aligned} q_{t+1,i} :=& \alpha_{t+1} d_{t+1,i} = \alpha_{t+1}\beta_{t+1} g_{t+1,i} \\ &+ (1-\beta_{t+1})\alpha_{t+1}((q_{t,i}/\alpha_t) + g_{t+1,i} - \tilde{g}_{t,i}) \end{aligned}$$

Since in Algorithm 1 (as well as Algorithm 2) we choose $\alpha_t = t$ and $\beta_t = 1/\alpha_t$, we have $(1-\beta_{t+1})\alpha_{t+1} = \alpha_t$, therefore:

$$q_{t+1,i} = q_t + g_{t+1,i} + \alpha_t(g_{t+1,i} - \tilde{g}_{t,i}) \tag{8}$$

Now, let us define two sequences $\{s_{t,i}\}, \{\bar{s}_{t,i}\}$: $s_{1,i} := g_{1,i}$, and $\bar{s}_{1,i} := \bar{g}_{1,i}$, and $\forall t > 1$,

$$\begin{aligned} s_{t,i} &:= g_{t,i} + \alpha_{t-1}(g_{t,i} - \tilde{g}_{t-1,i}) \\ \bar{s}_{t,i} &:= \bar{g}_{t,i} + \alpha_{t-1}(\bar{g}_{t,i} - \bar{g}_{t-1,i}) \end{aligned} \tag{9}$$

Note that the definitions of $s_{t,i}, \bar{s}_{t,i}$ and $g_{t,i}, \tilde{g}_{t,i}, \bar{g}_{t,i}$ imply:

$$\mathbb{E}_{t-1}\left[s_{t,i} - \bar{s}_{t,i}\right] = \mathbb{E}_{t-1}\left[g_{t,i} - \bar{g}_{t,i}\right] = 0$$

where $\mathbb{E}_{t-1}[\cdot]$ denotes conditional expectation over all randomization prior to time t, i.e. $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}\left[\cdot|\{z_{\tau,i}\}_{\tau\in[t-1],i\in[M]}, \{Y_{\tau,i}\}_{\tau\in[t-1],i\in[M]}\right]$. Thus, that

above implies that the sequence $\{s_{t,i} - \bar{s}_{t,i}\}_t$ is a Martingale difference sequence w.r.t. the natural filtration induced by the data-samples and injected noises.

The next lemma shows that for our specific choices of $\{\alpha_t, \beta_t\}$, we can represent $q_{t,i}$ as a sum of $\{s_{\tau,i}\}_{\tau=1}^t$. Importantly, each such $s_{\tau,i}$ terms is related to an individual data sample $z_{\tau,i}$. Moreover, it shows that we can represent the error of the weighted gradient-estimate $\varepsilon_{t,i} := q_{t,i} - \alpha_t \nabla f_i(x_t)$ as sum of Martingale difference sequence $\{s_{\tau,i} - \bar{s}_{\tau,i}\}_{\tau=1}^t$.

Lemma 4.1. *The choices that we make in Algorithms 1 and 2, i.e. $\alpha_t = t$, and $\beta_t = 1/\alpha_t$; imply:*

$$q_{t,i} = \sum_{\tau=1}^{t} s_{\tau,i} \quad \& \quad \varepsilon_{t,i} = \sum_{\tau=1}^{t} (s_{\tau,i} - \bar{s}_{\tau,i})$$

We prove of the above lemma in Appendix C.1.

Our next lemma provides bound on the increments $s_{t,i}$:

Lemma 4.2. *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set of diameter D, and $\{f_i(\cdot; z^i)\}_{z^i \in \mathcal{Z}_i}$ be a family of convex G-Lipschitz and L-smooth functions. Also define $S := G + 2LD, \tilde{\sigma} = \sigma + 2\sigma_L D$, then:*

$$\|s_{t,i}\| \le S \quad \& \quad \mathbb{E}\|s_{t,i} - \bar{s}_{t,i}\|^2 \le \tilde{\sigma}^2$$

Where the in expectation bound further assumes that the samples in $\mathcal{S}_i$ arrive from i.i.d. $\mathcal{D}_i$.

We prove the above lemma in Appendix C.2.

Lemmas 4.1 and 4.2 enable us to bound $\varepsilon_{t,i}$:

Lemma 4.3. *Our algorithms ensure: $\mathbb{E}\|\varepsilon_{t,i}\|^2 \le \tilde{\sigma}^2 t$.*

We prove the above lemma in Appendix C.3.

Comparison to SGD Here we will make an informal discussion of the guarantees of the above lemma in comparison to its parallel noise-injected SGD version that appears. Concretely, in SGD the contribution of a single sample $z_{t,i}$ to the gradient estimate is encapsulated in $\nabla f_i(w_t; z_{t,i})$, and the magnitude of this contribution can be of order $O(G)$ (as per our bounded gradient assumption). Moreover, in SGD the expected error of the gradient estimate (without the injected noise $Y_{t,i}$) is $\mathbb{E}\left\|\mathbf{err}_t^{\text{SGD}}\right\|^2 = \mathbb{E}\|\nabla f_i(w_t; z_{t,i}) - \nabla f_i(w_t)\|^2 \le \sigma^2$. Now, to make a proper comparison with our algorithm, we will compare these bounds to our (unweighted) gradient estimate, namely $d_{t,i}$. Recalling that $d_{t,i} = q_{t,i}/\alpha_t$ and $\alpha_t = t$, Lemmas 4.1 and 4.2, imply that the contribution of a single sample $z_{\tau,i}$ to the gradient estimate $d_{t,i}$ is encapsulated in $s_{\tau,i}/\alpha_t$ (assuming $\tau \le t$), and the magnitude of this contribution can be of order $O(S/t)$. Moreover, $\mathbb{E}\|\mathbf{err}_t\|^2 = \mathbb{E}\|d_{t,i} - \nabla f_i(x_t)\|^2 \le \tilde{\sigma}^2 t/\alpha_t^2 = \tilde{\sigma}^2/t$.

Algorithm 1 DP-μ^2-FL for Untrusted Server

Inputs: #iterations T, #machines M, initial point x_1, learning rate $\eta > 0$, importance weights $\{\alpha_t = t\}$, corrected momentum weights $\{\beta_t = 1/\alpha_t\}$, noise distributions $\left\{P_{t,i} = \mathcal{N}(0, I\sigma_{t,i}^2)\right\}$, per-machine $i \in [M]$ a dataset of samples $\mathcal{S}_i = \{z_{1,i}, \ldots, z_{T,i}\}$
Initialize: set $w_1 = x_1$, and $x_0 = x_1$
for every Machine $i \in [M]$ **do**
  set $d_{0,i} = 0$
end for
for $t = 1, \ldots, T$ **do**
  for every Machine $i \in [M]$ **do**
    Actions of Machine i:
    Retrieve $z_{t,i}$ from $\mathcal{S}_i$, compute $g_{t,i} = \nabla f(x_t; z_{t,i})$, and $\tilde{g}_{t-1,i} = \nabla f(x_{t-1}; z_{t,i})$
    Update $d_{t,i} = g_{t,i} + (1 - \beta_t)(d_{t-1,i} - \tilde{g}_{t-1,i})$ and $q_{t,i} = \alpha_t d_{t,i}$
    Draw $Y_{t,i} \sim \mathcal{N}\left(0, I\sigma_{t,i}^2\right)$
    Update $\tilde{q}_{t,i} = q_{t,i} + Y_{t,i}$ {each machine sends private information}
  end for
  Actions of Server:
  Aggregate $\tilde{q}_t = \frac{1}{M}\sum_{i=1}^M \tilde{q}_{t,i}$ {average after adding the noise}
  Update $w_{t+1} = \Pi_{\mathcal{K}}(w_t - \eta \tilde{q}_t)$
  Update $x_{t+1} = (1 - \frac{\alpha_{t+1}}{\alpha_{1:t+1}})x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} w_{t+1}$
end for
Output: x_T

Thus, in comparison to SGD, the errors of our gradient estimates decay with t, and a single sample $z_{\tau,i}$ directly affects all estimates $\{d_{t,i}\}_{t \ge \tau}$, but its affect decays with t. Note that the property that $\mathbb{E}\|d_{t,i} - \nabla f_i(x_t)\|^2 = \tilde{\sigma}^2/t$ was already demonstrated in (Levy, 2023), and we provide its proof for completeness.

In the next sections we will see how Lemma 4.2 plays a key role in deriving privacy guarantees for our DP-μ^2approach appearing in Algorithms 1 and 2.

5. Untrusted Server

In the untrusted server case, the server is not allowed to access private information, thus each machine must apply a privacy mechanism before sending the gradient estimate to the server. Thus, each machine will have its own $q_{t,i}$, but send $\tilde{q}_{t,i} := q_{t,i} + Y_{t,i}$ instead. Since the server receives protected information, the server does not need to apply any privacy mechanism before sending the query point x_t to the machines. Algorithm 1 depicts our approach for the untrusted server case.

5.1. Privacy Guarantees

Here we establish the privacy guarantees of Algorithm 1. Concretely, the following theorem shows how does the privacy of our algorithm depends on the variances of injected noise $\{Y_{t,i}\}_t$.

Theorem 5.1 (Privacy Guarantees for Algorithm 1)**.** *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set of diameter D, and $\{f_i(\cdot; z^i)\}_{z^i \in \mathcal{Z}_i}$ be a family of convex G-Lipschitz and L-smooth functions. Then invoking Algorithm 1 with noise distributions $Y_{t,i} \sim P_{t,i} = \mathcal{N}\left(0, I\sigma_{t,i}^2\right)$, and any learning rate $\eta > 0$, ensures that for any machine $i \in [M]$, the resulting sequences $\{\tilde{q}_{t,i}\}_{t\in[T]}$ is $\left(\alpha, \frac{\alpha\rho_i^2}{2}\right)$-RDP for any $\alpha > 1$, where: $\rho_i = 2S\sqrt{\sum_{\tau=1}^{T} \frac{1}{\sigma_{\tau,i}^2}}$.*

The full proof is in Appendix D.1.

Proof Sketch. First assume that $\mathcal{S}_i$ and $\mathcal{S}_i'$ are neighboring datasets, meaning that there exists only a single index $\tau^* \in [T]$ where they differ, i.e. that $z_{\tau^*} \neq z'_{\tau^*}$.

Now, recalling that we obtain $\tilde{q}_{t,i}$ by adding a Gaussian noise $Y_{t,i}$ to $q_{t,i}$, we may use Lemma 2.2 and obtain:

$$\mathbb{D}_\alpha\left(\tilde{q}_{t,i}(\mathcal{S}_i) \| \tilde{q}_{t,i}(\mathcal{S}_i')\right) = \frac{\alpha\Delta_{t,i}^2}{2\sigma_{t,i}^2} \tag{10}$$

Using that fact that $q_{t,i} = \sum_{\tau=1}^{t} s_{\tau,i}$ together with the bound $\|s_{\tau,i}\| \leq S$, which holds for any $\tau \in [T]$ due to Lemma 4.2, we show that $\Delta_{t,i} \leq 2S \cdot \mathbb{I}\{t \geq \tau^*\}$, with τ^* being the time step where $z_{\tau^*} \neq z'_{\tau^*}$, and $S := G + 2LD$. We thus get:

$$\mathbb{D}_\alpha\left(\tilde{q}_{t,i}(\mathcal{S}_i) \| \tilde{q}_{t,i}(\mathcal{S}_i')\right) \leq \frac{2\alpha S^2}{\sigma_{t,i}^2} \cdot \mathbb{I}\{t \geq \tau^*\} \tag{11}$$

Using the above together with Lemma A.3, gives:

$$\mathbb{D}_\alpha\left(\{\tilde{q}_{\tau,i}(\mathcal{S}_i)\}_{\tau=1}^t \| \{\tilde{q}_{\tau,i}(\mathcal{S}_i')\}_{\tau=1}^t\right) \leq 2\alpha S^2 \sum_{\tau=1}^{t} \frac{1}{\sigma_{\tau,i}^2} \tag{12}$$

Using the definition of RDP concludes the proof. □

5.2. Convergence Guarantees

Next we state our main theorem for this setting.

Theorem 5.2. *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set of diameter D and $\{f_i(\cdot; z^i)\}_{i\in[M], z^i \in \mathcal{Z}_i}$ be a family of G-Lipschitz and L-smooth functions over $\mathcal{K}$, with $\sigma \in [0, G], \sigma_L \in [0, L]$ as defined in Equations (3) and (4), define $f(x; z) = \frac{1}{M}\sum_{i=1}^{M} f_i(x; z^i)$ and $G^* := \nabla f(x^*)$, where $x^* = \arg\min_{x\in\mathcal{K}} f(x)$, and $S := G + 2LD, \tilde{\sigma} := \sigma + 2\sigma_L D$, moreover let $T \in \mathbb{N}, \rho > 0$.*

Then upon invoking Algorithm 1 with $\alpha_t = t, \beta_t = 1/\alpha_t, \eta = \min\left\{\frac{\rho D\sqrt{M}}{2ST\sqrt{d}}, \frac{1}{4LT}\right\}$, and $\sigma_{t,i}^2 = 4S^2T/\rho^2$, and any starting point $x_1 \in \mathcal{K}$ and datasets $\{\mathcal{S}_i \in \mathcal{Z}_i^T\}_{i\in[M]}$, then for all $\alpha \geq 1$ Algorithm 1 satisfies $\left(\alpha, \frac{\alpha\rho^2}{2}\right)$-RDP w.r.t gradient estimate sequences that each machine produces, i.e. $\{\tilde{q}_{t,i}\}_{t\in[T], i\in[M]}$.

Furthermore, if $\mathcal{S}_i$ consists of i.i.d. samples from a distribution $\mathcal{D}_i$ for all $i \in [M]$, then Algorithm 1 guarantees:

$$\mathcal{R}(x_T^{untrust}) := \mathbb{E}\left[f\left(x_T^{untrust}\right)\right] - \min_{x\in\mathcal{K}} f(x)$$
$$\leq 4D\left(\frac{G^* + 2LD}{T} + \frac{2S\sqrt{d}}{\rho T\sqrt{M}} + \frac{\tilde{\sigma}}{\sqrt{TM}}\right)$$

The full proof is in Appendix E.1. Notably, the above bounds are optimal. Moreover, similarly to the standard minibatch-SGD analysis (Dekel et al., 2012), these bound do not depend on the level of heterogeneity between machines. Which is due to the full synchronicity and symmetry between machines, implying that the aggregated q_t satisfies $\mathbb{E}[q_t] = \alpha_t\mathbb{E}[\nabla f(x_t)]$, as well as $\mathbb{E}\|q_t - \alpha_t\nabla f(x_t)\|^2 \leq \tilde{\sigma}^2 t/M$.

Proof Sketch. The privacy guarantees follow directly from Theorem 5.1, and our choice of $\sigma_{t,i}^2$. Regrading convergence: in the spirit of μ^2-SGD analysis (Levy, 2023), we show that:

$$\alpha_{1:T}\mathcal{R}(x_T) \leq \frac{D^2}{\eta} + \eta\sum_{\tau=1}^{T}\mathbb{E}\|Y_\tau\|^2$$
$$+2\alpha_T DG^* + 2D\sum_{\tau=1}^{T}\sqrt{\mathbb{E}\|\varepsilon_\tau\|^2} \tag{13}$$

where we denoted $\varepsilon_\tau := \frac{1}{M}\sum_{i\in[M]}\varepsilon_{\tau,i}$, and $Y_\tau := \frac{1}{M}\sum_{i\in[M]} Y_{\tau,i}$. Then, by utilizing Lemma 4.3, we show that $\mathbb{E}\|\varepsilon_\tau\|^2 \leq \frac{1}{M}\mathbb{E}\|\varepsilon_{\tau,i}\|^2 \leq \tilde{\sigma}^2\tau/M$; and similarly that $\mathbb{E}\|Y_\tau\|^2 \leq \frac{1}{M}\mathbb{E}\|Y_{\tau,i}\|^2 \leq \sigma_\tau^2/M$. Plugging these into the bound above and using $\alpha_t = t$, and our choices of η and σ_t implies the bound. □

5.3. Experiments

We ran DP-μ^2on MNIST using a logistic regression model in the untrusted server case. The parameters are $G = \sqrt{2 \cdot 785} = 39.6, L = 785/2 = 392.5, D = 0.1$, which brings us $S = 118.1$. Our model has $d = 10 \cdot 785 = 7850$ parameters. We kept $M \cdot T = 60,000$, and checked $M = 1, 10, 100$ and $\rho = 4, 8, 16$. We show our results in Table 1.

We can see that by increasing ρ, we improve the loss and accuracy, as higher ρ means less privacy. By increasing

Table 1. DP-μ^2on MNIST with untrusted server

	M=1, T=60,000		M=10, T=6,000		M=100, T=600	
ρ	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
4	2.256	69.9%	2.267	69.4%	2.285	65.4%
8	2.253	70.2%	2.259	70.0%	2.274	69.8%
16	2.252	70.4%	2.255	70.1%	2.264	70.0%

M the loss and accuracy become worse, as the data is split between different machine and cannot be used optimally due to not trusting the server.

6. Trusted Server

In the trusted server case, the server is allowed to access private information, therefore machines do not apply a privacy mechanism before sending the gradient estimate to the server. And each machine maintains its own estimate $q_{t,i}$ (an estimate of $\alpha_t \nabla f_i(x_t)$), and sends it without adding noise. Since the server receives private information, it must apply a privacy mechanism before sending the query point x_t to the machines. Algorithm 2 depicts our approach in the trusted server case.

6.1. Privacy Guarantees

Here we establish the privacy guarantees of Algorithm 2. Concretely, the following theorem shows how does the privacy of our algorithm depends on the variances of injected noise $\{Y_t\}_t$.

Theorem 6.1 (Privacy Guarantees for Algorithm 2)**.** *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set of diameter D, and $\{f_i(\cdot; z^i)\}_{z^i \in \mathcal{Z}_i}$ be a family of convex G-Lipschitz and L-smooth functions. Then invoking Algorithm 2 with noise distributions $Y_t \sim P_t = \mathcal{N}\left(0, I\sigma_t^2\right)$, and any learning rate $\eta > 0$; ensures that for any machine $i \in [M]$, the query sequence $\{x_t\}_{t\in[T]}$ is $\left(\alpha, \frac{\alpha\rho^2}{2}\right)$-RDP for any $\alpha > 1$, where: $\rho = \frac{2S}{M}\sqrt{\sum_{\tau=1}^{T} \frac{1}{\sigma_\tau^2}}$.*

The full proof is in Appendix D.2.

Proof Sketch. Fix $i \in [M]$, and assume that $\mathcal{S}_i$ and $\mathcal{S}_i'$ are neighboring datasets, meaning that there exists only a single index $\tau^* \in [T]$ where they differ, i.e. that $z_{\tau^*,i} \neq z'_{\tau^*,i}$. For other machines $j \neq i$ assume that $\mathcal{S}_j$ are known and fixed.

Now, recalling that we obtain $\tilde{q}_t$ by adding a Gaussian noise Y_t to q_t, we may use Lemma 2.2 and obtain:

$$\mathbb{D}_\alpha\left(\tilde{q}_t(\mathcal{S}_i) \| \tilde{q}_t(\mathcal{S}_i')\right) = \frac{\alpha \Delta_t^2}{2\sigma_t^2} \tag{14}$$

Using that fact that $q_t = \frac{1}{M}\sum_{i=1}^{M}\sum_{\tau=1}^{t} s_{\tau,i}$ together with the bound $\|s_{\tau,i}\| \leq S$, which holds for any $\tau \in [T]$ due to Lemma 4.2, we show that $\Delta_t \leq (2S/M) \cdot \mathbb{I}\{t \geq \tau^*\}$, with τ^* being the time step where $z_{\tau^*,i} \neq z'_{\tau^*,i}$, and $S := G + 2LD$. We thus get:

$$\mathbb{D}_\alpha\left(\tilde{q}_t(\mathcal{S}_i) \| \tilde{q}_t(\mathcal{S}_i')\right) \leq \frac{2\alpha S^2}{M^2 \sigma_t^2} \cdot \mathbb{I}\{t \geq \tau^*\} \tag{15}$$

Combining the above with Lemma A.3:

$$\mathbb{D}_\alpha\left(\{\tilde{q}_\tau(\mathcal{S}_i)\}_{\tau=1}^t \| \{\tilde{q}_\tau(\mathcal{S}_i')\}_{\tau=1}^t\right) \leq \frac{2\alpha S^2}{M^2} \sum_{\tau=1}^{t} \frac{1}{\sigma_\tau^2} \tag{16}$$

Finally, we use the post processing lemma for RDPs (van Erven & Harremoës, 2014), to bound the privacy of $\{x_\tau\}_{\tau=1}^t$ with the privacy of $\{\tilde{q}_\tau\}_{\tau=1}^t$. □

Algorithm 2 DP-μ^2-FL for Trusted Server

Inputs: #iterations T, #machines M, initial point x_1, learning rate $\eta > 0$, importance weights $\{\alpha_t = t\}$, corrected momentum weights $\{\beta_t = 1/\alpha_t\}$, noise distributions $\left\{P_t = \mathcal{N}(0, I\sigma_t^2)\right\}$, per-machine $i \in [M]$ a dataset of samples $\mathcal{S}_i = \{z_{1,i}, \ldots, z_{T,i}\}$
Initialize: set $w_1 = x_1$, and $x_0 = x_1$
for every Machine $i \in [M]$ **do**
  set $d_{0,i} = 0$
end for
for $t = 1, \ldots, T$ **do**
  for every Machine $i \in [M]$ **do**
    Actions of Machine i:
    Retrieve $z_{t,i}$ from $\mathcal{S}_i$, compute $g_{t,i} = \nabla f(x_t; z_{t,i})$, and $\tilde{g}_{t-1,i} = \nabla f(x_{t-1}; z_{t,i})$
    Update $d_{t,i} = g_{t,i} + (1-\beta_t)(d_{t-1,i} - \tilde{g}_{t-1,i})$ and $q_{t,i} = \alpha_t d_{t,i}$
  end for
  Actions of Server:
  Aggregate $q_t = \frac{1}{M}\sum_{i=1}^{M} q_{t,i}$ {average before adding the noise}
  Draw $Y_t \sim \mathcal{N}\left(0, I\sigma_t^2\right)$
  Update $\tilde{q}_t = q_t + Y_t$ {the server is the one to induce privacy}
  Update $w_{t+1} = \Pi_{\mathcal{K}}(w_t - \eta\tilde{q}_t)$
  Update $x_{t+1} = (1 - \frac{\alpha_{t+1}}{\alpha_{1:t+1}})x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} w_{t+1}$
end for
Output: x_T

6.2. Convergence Guarantees

Next we state our main theorem for this setting.

Theorem 6.2. *Assume the same assumptions as in Theorem 5.2. Then upon invoking Algorithm 2 with $\alpha_t = t, \beta_t = 1/\alpha_t, \eta = \min\left\{\frac{\rho DM}{2ST\sqrt{d}}, \frac{1}{4LT}\right\}$, and $\sigma_t^2 = 4S^2T/\rho^2M^2$, and any starting point $x_1 \in \mathcal{K}$ and datasets $\{\mathcal{S}_i \in$*

Table 2. DP-μ^2on MNIST with trusted server

	M=1, T=60,000		M=10, T=6,000		M=100, T=600	
ρ	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
4	2.256	69.9%	2.256	69.7%	2.258	69.5%
8	2.253	70.2%	2.253	70.1%	2.257	69.6%
16	2.252	70.4%	2.252	70.3%	2.256	69.7%

$\mathcal{Z}_i^T\}_{i\in[M]}$, *then for all* $\alpha \geq 1$ *Algorithm 2 satisfies* $\left(\alpha, \frac{\alpha\rho^2}{2}\right)$*-RDP for the query and gardient estimate sequences that the server produces, i.e.* $\{x_t\}_{t\in[T]}, \{\tilde{q}_t\}_{t\in[T]}$. *Furthermore, if* $\mathcal{S}_i$ *consists of i.i.d. samples from a distribution* $\mathcal{D}_i$*, then Algorithm 2 guarantees:*

$$\mathcal{R}(x_T^{trust}) := \mathbb{E}\left[f\left(x_T^{trust}\right)\right] - \min_{x\in\mathcal{K}} f(x)$$
$$\leq 4D\left(\frac{G^* + 2LD}{T} + \frac{2S\sqrt{d}}{\rho TM} + \frac{\tilde{\sigma}}{\sqrt{TM}}\right)$$

The full proof is in Appendix E.2.

Proof Sketch. The privacy guarantees follow directly from Theorem 6.1, and our choice of σ_t^2. Regrading convergence: in the spirit of μ^2-SGD analysis (Levy, 2023), we show that:

$$\alpha_{1:T}\mathcal{R}(x_T) \leq \frac{D^2}{\eta} + \eta\sum_{\tau=1}^{T}\mathbb{E}\left\|Y_\tau\right\|^2$$
$$+2\alpha_T DG^* + 2D\sum_{\tau=1}^{T}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2} \qquad (17)$$

Where we denoted $\varepsilon_\tau := q_t - \alpha_t\nabla f(x_t)$. Then, by utilizing Lemma 4.3, we show that $\mathbb{E}\left\|\varepsilon_\tau\right\|^2 \leq \frac{1}{M}\mathbb{E}\left\|\varepsilon_{\tau,i}\right\|^2 \leq \tilde{\sigma}^2\tau/M$. Plugging these into the bound above and using $\alpha_t = t$, and our choices of η and σ_t^2 implies the bound. □

6.3. Experiments

We ran DP-μ^2on MNSIT using the same specification as the previous section, but for the trusted server case. We show our results in Table 2.

The First thing we can see, is that the results are the same for both cases when $M = 1$, as this case in not federated, and thus there is no difference between the trusted and untrusted server cases. When $M = 10$, the loss doesn't change in a noticeable way, and the accuracy decreases a little. When $M = 100$, the loss decreases a little, but less than the untrusted server case, and the accuracy also decreases. Another thing we can notice, is that with higher value of ρ, the trusted and untrusted server cases give more similar results. That is because with lower privacy, the privacy type (trusted/untrusted) matters less.

7. Conclusion

We built upon the original μ^2-SGD algorithm to accommodate differential privacy setting. We extended the algorithm to the federated learning setting, and shown it is optimal in both the untrusted server and trusted server setting. The optimal convergence is satisfied with linear computational complexity.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, and especially the aspect of Privacy. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This research was partially supported by Israel PBC-VATAT, by the Technion Artificial Intelligent Hub (Tech.AI) and by the Israel Science Foundation (grant No. 447/20).

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In Weippl, E. R., Katzenbeisser, S., Kruegel, C., Myers, A. C., and Halevi, S. (eds.), *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016.

Avella-Medina, M., Bradshaw, C., and Loh, P. Differentially private inference via noisy optimization. *CoRR*, abs/2103.11003, 2021.

Bassily, R., Smith, A. D., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS*. IEEE Computer Society, 2014.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. G. Private stochastic convex optimization with optimal rates. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 2011.

Cheu, A., Joseph, M., Mao, J., and Peng, B. Shuffle private stochastic convex optimization. In *The Tenth International Conference on Learning Representations, ICLR*. OpenReview.net, 2022.

Cutkosky, A. Anytime online-to-batch, optimism and acceleration. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2019.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 2012.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*. IEEE Computer Society, 2013.

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

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In Vaudenay, S. (ed.), *Advances in Cryptology - EUROCRYPT, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Lecture Notes in Computer Science. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T. (eds.), *Theory of Cryptography, Third Theory of Cryptography Conference, TCC*, Lecture Notes in Computer Science. Springer, 2006b.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: optimal rates in linear time. In Makarychev, K., Makarychev, Y., Tulsiani, M., Kamath, G., and Chuzhoy, J. (eds.), *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC*. ACM, 2020.

Ganesh, A., Haghifam, M., Steinke, T., and Thakurta, A. Faster differentially private convex optimization via second-order methods. *CoRR*, abs/2305.13209, 2023.

Girgis, A. M., Data, D., Diggavi, S. N., Kairouz, P., and Suresh, A. T. Shuffled model of federated learning: Privacy, accuracy and communication trade-offs. *IEEE J. Sel. Areas Inf. Theory*, 2021.

Huang, Z., Hu, R., Guo, Y., Chan-Tin, E., and Gong, Y. DP-ADMM: admm-based distributed learning with differential privacy. *IEEE Trans. Inf. Forensics Secur.*, 2020.

Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., and Wang, L. Towards practical differentially private convex optimization. In *IEEE Symposium on Security and Privacy, SP*. IEEE, 2019.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. Practical and private (deep) learning without sampling or shuffling. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2021a.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 2021b.

Kavis, A., Levy, K. Y., Bach, F. R., and Cevher, V. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

Kifer, D., Smith, A. D., and Thakurta, A. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In Mannor, S., Srebro, N., and Williamson, R. C. (eds.), *COLT 2012 - The 25th Annual Conference on Learning Theory*, JMLR Proceedings. JMLR.org, 2012.

Levy, K. Y. μ^2-sgd: Stable stochastic optimization via a double momentum mechanism. *CoRR*, abs/2304.04172, 2023.

Lowy, A. and Razaviyayn, M. Private federated learning without a trusted server: Optimal algorithms for convex losses. In *The Eleventh International Conference on Learning Representations, ICLR*. OpenReview.net, 2023.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. Private non-convex federated learning without a trusted server. In Ruiz, F. J. R., Dy, J. G., and van de Meent, J. (eds.), *International Conference on Artificial Intelligence and*

Statistics, Proceedings of Machine Learning Research. PMLR, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, Proceedings of Machine Learning Research. PMLR, 2017.

Mironov, I. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF*. IEEE Computer Society, 2017.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS*, Proceedings of Machine Learning Research. PMLR, 2022.

Rényi, A. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. The Regents of the University of California, 1961.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *COLT - The 22nd Conference on Learning Theory*, 2009.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP*. IEEE, 2013.

Talwar, K., Thakurta, A., and Zhang, L. Nearly optimal private LASSO. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, 2015.

Thakurta, A. and Smith, A. D. Differentially private feature selection via stability arguments, and the robustness of the lasso. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *COLT - The 26th Annual Conference on Learning Theory*, JMLR Workshop and Conference Proceedings. JMLR.org, 2013.

Ullman, J. R. Private multiplicative weights beyond linear queries. In Milo, T. and Calvanese, D. (eds.), *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS*. ACM, 2015.

van Erven, T. and Harremoës, P. Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 2014.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. *CoRR*, abs/1802.05251, 2018.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q. S., and Poor, H. V. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Secur.*, 2020.

Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. F. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In Salihoglu, S., Zhou, W., Chirkova, R., Yang, J., and Suciu, D. (eds.), *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference*. ACM, 2017.

A. Additional Theorems and Lemmas

Here we provide additional theorems and lemmas that are used for our proofs.

Theorem A.1 ((Cutkosky, 2019)). *Let $f : \mathcal{K} \to \mathbb{R}$ be a convex function. Also let $\{\alpha_t > 0\}$ and $\{w_t \in \mathcal{K}\}$. Let $\{x_t\}$ be the $\{\alpha_\tau\}_{\tau=1}^t$ weighted average of $\{w_\tau\}_{\tau=1}^t$, meaning: $x_t = \frac{1}{\alpha_{1:t}} \sum_{\tau=1}^t \alpha_\tau w_\tau$. Then the following holds for all $t \geq 1, x \in \mathcal{K}$:*

$$\alpha_{1:t}(f(x_t) - f(x)) \leq \sum_{\tau=1}^t \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle$$

Note that the above theorem holds generally for any sequences of iterates $\{w_t\}_t$ with weighted averages $\{x_t\}_t$, and as a private case it holds for the sequences generated by Anytime-SGD. Concretely, the theorem implies that the excess loss of the weighted average x_t can be related to the weighted regret $\sum_{\tau=1}^t \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle$.

Proof of Theorem A.1. Proof by induction.
Induction basis: $t = 1$

$$\alpha_1(f(x_1) - f(x)) \leq \alpha_1 \langle \nabla f(x_1), x_1 - x \rangle = \alpha_1 \langle \nabla f(x_1), w_1 - x \rangle$$

The inequality is from convexity of f, and the equality is because $x_1 = w_1$.
Induction assumption: for some $t \geq 1$:

$$\alpha_{1:t}(f(x_t) - f(x)) \leq \sum_{\tau=1}^t \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle$$

Induction step: proof for $t + 1$:

$$\begin{aligned}
\alpha_{1:t+1}(f(x_{t+1}) - f(x)) =& \alpha_{1:t}(f(x_{t+1}) - f(x_t) + f(x_t) - f(x)) + \alpha_{t+1}(f(x_{t+1}) - f(x)) \\
\leq& \sum_{\tau=1}^t \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle + \alpha_{1:t}(f(x_{t+1}) - f(x_t)) + \alpha_{t+1}(f(x_{t+1}) - f(x)) \\
\leq& \sum_{\tau=1}^t \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle + \langle \nabla f(x_{t+1}), \alpha_{1:t}(x_{t+1} - x_t) + \alpha_{t+1}(x_{t+1} - x) \rangle \\
=& \sum_{\tau=1}^t \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle + \alpha_{t+1} \langle \nabla f(x_{t+1}), w_{t+1} - x \rangle = \sum_{\tau=1}^{t+1} \alpha_\tau \langle \nabla f(x_\tau), w_\tau - x \rangle
\end{aligned}$$

In the first equality we rearranged the terms, and added and subtracted the same thing, we then used the induction assumption, then used convexity of f on both pairs and added them together, and finally we used the update rule of x_{t+1}: $\alpha_{1:t+1} x_{t+1} = \alpha_{1:t} x_t + \alpha_{t+1} w_{t+1}$, and added the last member of the sum to get our desired result. □

Lemma A.2. *Let $\{Z_t\}$ be a Martingale difference sequence w.r.t a Filtration $\{\mathcal{F}_t\}_t$, i.e. $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$, then:*

$$\mathbb{E} \left\| \sum_{\tau=1}^t Z_\tau \right\|^2 = \sum_{\tau=1}^t \mathbb{E} \|Z_\tau\|^2$$

Proof of Lemma A.2. Proof by induction.
Induction basis: $t = 1$

$$\mathbb{E} \left\| \sum_{\tau=1}^1 Z_\tau \right\|^2 = \mathbb{E} \|Z_1\|^2 = \sum_{\tau=1}^1 \mathbb{E} \|Z_\tau\|^2$$

Induction assumption: for some $t \geq 1$:

$$\mathbb{E}\left\|\sum_{\tau=1}^{t} Z_\tau\right\|^2 = \sum_{\tau=1}^{t} \mathbb{E}\left\|Z_\tau\right\|^2$$

Induction step: proof for $t+1$:

$$\mathbb{E}\left\|\sum_{\tau=1}^{t+1} Z_\tau\right\|^2 = \mathbb{E}\left\|\sum_{\tau=1}^{t+1} Z_\tau\right\|^2 + 2\mathbb{E}\left[\left\langle \sum_{\tau=1}^{t} Z_\tau, Z_{t+1}\right\rangle\right] + \mathbb{E}\left\|Z_{t+1}\right\|^2 = \sum_{\tau=1}^{t} \mathbb{E}\left\|Z_\tau\right\|^2 + \mathbb{E}\left\|Z_{t+1}\right\|^2 = \sum_{\tau=1}^{t+1} \mathbb{E}\left\|Z_\tau\right\|^2$$

The first equality is square rules, then we used the induction assumption and the fact that $\mathbb{E}\left[Z_{t+1}|Z_1,\ldots,Z_t\right] = 0$, and finally we added the last term into the sum. □

Lemma A.3 ((Mironov, 2017))**.** *If $\mathcal{A}_1,\ldots,\mathcal{A}_k$ are randomized algorithms satisfying (α,ϵ_1)-RDP, …, (α,ϵ_k)-RDP, respectively, then their composition $(\mathcal{A}_1(\mathcal{S})\ldots,\mathcal{A}_k(\mathcal{S}))$ is $(\alpha,\epsilon_1+\ldots,+\epsilon_k)$-RDP. Moreover, the i'th algorithm $\mathcal{A}_i$, can be chosen on the basis of the outputs of the previous algorithms $\mathcal{A}_1,\ldots,\mathcal{A}_{i-1}$.*

Proof of Lemma A.3. Proof by induction.
Induction basis: $k=1$

$$\mathbb{D}_\alpha\left(\mathcal{A}_1(\mathcal{S})\|\mathcal{A}_1(\mathcal{S}')\right) \leq \epsilon_1$$

Because $\mathcal{A}_1$ is (α,ϵ_1)-RDP.
Induction assumption: for some $k \geq 1$:

$$\mathbb{D}_\alpha\left(\{\mathcal{A}_i(\mathcal{S})\}_{i=1}^k\|\{\mathcal{A}_i(\mathcal{S}')\}_{i=1}^k\right) \leq \sum_{i=1}^{k}\epsilon_i$$

Induction step: proof for $k+1$

$$\begin{aligned}
\mathbb{D}_\alpha\left(\{\mathcal{A}_i(\mathcal{S})\}_{i=1}^{k+1}\|\{\mathcal{A}_i(\mathcal{S}')\}_{i=1}^{k+1}\right) =& \frac{1}{\alpha-1}\log\left(\mathbb{E}_{\mathcal{A}_i\sim\mathcal{A}_i(\mathcal{S})}\left[\left(\frac{\mathbb{P}\left\{\{\mathcal{A}_i(\mathcal{S})\}_{i=1}^{k+1}\right\}}{\mathbb{P}\left\{\{\mathcal{A}_i(\mathcal{S}')\}_{i=1}^{k+1}\right\}}\right)^{\alpha-1}\right]\right)\\
=& \frac{1}{\alpha-1}\log\left(\mathbb{E}_{\mathcal{A}_i\sim\mathcal{A}_i(\mathcal{S})}\left[\left(\frac{\mathbb{P}\left\{\mathcal{A}_{k+1}(\mathcal{S})|\{\mathcal{A}_i\}_{i=1}^{k}\right\}\mathbb{P}\left\{\{\mathcal{A}_i(\mathcal{S})\}_{i=1}^{k}\right\}}{\mathbb{P}\left\{\mathcal{A}_{k+1}(\mathcal{S}')|\{\mathcal{A}_i\}_{i=1}^{k}\right\}\mathbb{P}\left\{\{\mathcal{A}_i(\mathcal{S}')\}_{i=1}^{k}\right\}}\right)^{\alpha-1}\right]\right)\\
=& \frac{1}{\alpha-1}\log\left(\mathbb{E}_{\mathcal{A}_i\sim\mathcal{A}_i(\mathcal{S})}\left[\left(\frac{\mathbb{P}\left\{\mathcal{A}_{k+1}(\mathcal{S})|\{\mathcal{A}_i\}_{i=1}^{k}\right\}}{\mathbb{P}\left\{\mathcal{A}_{k+1}(\mathcal{S}')|\{\mathcal{A}_i\}_{i=1}^{k}\right\}}\right)^{\alpha-1}\right]\right)\\
&+\frac{1}{\alpha-1}\log\left(\mathbb{E}_{\mathcal{A}_i\sim\mathcal{A}_i(\mathcal{S})}\left[\left(\frac{\mathbb{P}\left\{\{\mathcal{A}_i(\mathcal{S})\}_{i=1}^{k}\right\}}{\mathbb{P}\left\{\{\mathcal{A}_i(\mathcal{S}')\}_{i=1}^{k}\right\}}\right)^{\alpha-1}\right]\right)\\
=& \mathbb{D}_\alpha\left(\mathcal{A}_{k+1}(\mathcal{S})\|\mathcal{A}_{k+1}(\mathcal{S}')|\{\mathcal{A}_i\}_{i=1}^{k}\right) + \mathbb{D}_\alpha\left(\{\mathcal{A}_i(\mathcal{S})\}_{i=1}^{k}\|\{\mathcal{A}_i(\mathcal{S}')\}_{i=1}^{k}\right)\\
\leq& \epsilon_{k+1} + \sum_{i=1}^{k}\epsilon_i = \sum_{i=1}^{k+1}\epsilon_i
\end{aligned}$$

The writing $\mathcal{A}_i \sim \mathcal{A}_i(\mathcal{S})$ means that the output of $\mathcal{A}_i$ is distributed in the case that the dataset is $\mathcal{S}$. The first equation is the definition of the Rényi divergence, next we use conditional probability rules, then we use the fact that the output of $\mathcal{A}_{k+1}$ conditioned on the outputs of the previous algorithms is independent of the outputs of these previous algorithms, and let the product out of the log as a summery. We then notice that each of the two members are divergences themselves, and finally, we invoke both the the fact that $\mathcal{A}_{k+1}$ is (α,ϵ_{k+1})-RDP, and the induction assumption, and add them together to the same sum. □

Lemma A.4 (Post Processing Lemma (van Erven & Harremoës, 2014))**.** *Let X, Y be random variables and $\mathcal{A}$ be a randomized or deterministic algorithm. Then for all $\alpha \geq 1$:*

$$\mathbb{D}_\alpha\left(\mathcal{A}(X) \| \mathcal{A}(Y)\right) \leq \mathbb{D}_\alpha\left(X \| Y\right)$$

We use this lemma to bound the privacy of $\{x_\tau\}_{\tau=1}^t$ with the privacy of $\{\tilde{q}_\tau\}_{\tau=1}^t$.

Lemma A.5. *Let $\eta > 0$, and $\mathcal{K} \subset \mathbb{R}^d$ be a convex domain of bounded diameter D, also let $\{\tilde{q}_t \in \mathbb{R}^d\}_{t=1}^T$ be a sequence of arbitrary vectors. Then for any starting point $w_1 \in \mathbb{R}^d$, and an update rule $w_{t+1} = \Pi_{\mathcal{K}}\left(w_t - \eta \tilde{q}_t\right)$, $\forall t \geq 1$, the following holds $\forall x \in \mathcal{K}$:*

$$\sum_{\tau=1}^t \langle \tilde{q}_\tau, w_{\tau+1} - x \rangle \leq \frac{D^2}{2\eta} - \frac{1}{2\eta} \sum_{\tau=1}^t \left\| w_\tau - w_{\tau+1} \right\|^2$$

Proof of Lemma A.5. The update rule $w_{t+1} = \Pi_{\mathcal{K}}\left(w_t - \eta \tilde{q}_t\right)$ can be re-written as a convex optimization problem over $\mathcal{K}$:

$$w_{t+1} = \Pi_{\mathcal{K}}\left(w_t - \eta \tilde{q}_t\right) = \arg\min_{x \in \mathcal{K}} \left\{ \left\| w_t - \eta \tilde{q}_t - x \right\|^2 \right\} = \arg\min_{x \in \mathcal{K}} \left\{ \langle \tilde{q}_t, x - w_t \rangle + \frac{1}{2\eta} \left\| x - w_t \right\|^2 \right\}$$

The first equality is our update definition, the second is by the definition of the projection operator, and then we rewrite it in a way that does not affect the minimum point.

Now, since w_{t+1} is the minimal point of the above convex problem, then from optimality conditions we obtain:

$$\left\langle \tilde{q}_t + \frac{1}{\eta}(w_{t+1} - w_t), x - w_{t+1} \right\rangle \geq 0, \quad \forall x \in \mathcal{K}$$

Re-arranging the above, we get that:

$$\langle \tilde{q}_t, w_{t+1} - x \rangle \leq \frac{1}{\eta} \langle w_t - w_{t+1}, w_{t+1} - x \rangle = \frac{1}{2\eta} \left\| w_t - x \right\|^2 - \frac{1}{2\eta} \left\| w_{t+1} - x \right\|^2 - \frac{1}{2\eta} \left\| w_t - w_{t+1} \right\|^2$$

Where the equality is an algebraic manipulation. After summing over t we get:

$$\begin{aligned}
\sum_{\tau=1}^t \langle \tilde{q}_\tau, w_{\tau+1} - x \rangle &\leq \frac{1}{2\eta} \sum_{\tau=1}^t \left(\left\| w_\tau - x \right\|^2 - \left\| w_{\tau+1} - x \right\|^2 - \left\| w_\tau - w_{\tau+1} \right\|^2 \right) \\
&= \frac{\left\| w_1 - x \right\|^2 - \left\| w_{t+1} - x \right\|^2}{2\eta} - \frac{1}{2\eta} \sum_{\tau=1}^t \left\| w_\tau - w_{\tau+1} \right\|^2 \leq \frac{D^2}{2\eta} - \frac{1}{2\eta} \sum_{\tau=1}^t \left\| w_\tau - w_{\tau+1} \right\|^2
\end{aligned}$$

Where the second line is due to splitting the sum into two sums, and using the fact that the first one is a telescopic sum, and lastly, we use the diameter of $\mathcal{K}$. This establishes the lemma. □

Lemma A.6. *If $f : \mathcal{K} \to \mathbb{R}$ is convex and L-smooth, and $x^* = \arg\min_{x \in \mathcal{K}} \{f(x)\}$, then $\forall x \in \mathbb{R}^d$:*

$$\left\| \nabla f(x) - \nabla f(x^*) \right\|^2 \leq 2L(f(x) - f(x^*))$$

Proof of Lemma A.6. Let us define a new function:

$$h(x) = f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle$$

Since f is convex and L-smooth, we know that:

$$0 \leq h(x) \leq \frac{L}{2} \left\| x - x^* \right\|^2$$

The gradient of this function is:

$$\nabla h(x) = \nabla f(x) - \nabla f(x^*)$$

We can see that $h(x^*) = 0, \nabla h(x^*) = 0$, and that x^* is the global minimum. The function h is also convex and L-smooth, since the gradient is the same as f up to a constant translation. We will add to the domain of h to include all $\mathbb{R}^d$, while still being convex and L-smooth. Since h is convex then:

$$h(y) \geq h(x^*) + \langle \nabla h(x^*), y - x^* \rangle = 0, \quad \forall y \in \mathbb{R}^d$$

It is true even for points outside of the original domain, meaning that x^* remains the global minimum even after this. For a smooth function, $\forall x, y \in \mathbb{R}^d$:

$$h(y) \leq h(x) + \langle \nabla h(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

By picking $y = x - \frac{1}{L} \nabla h(x)$, we get:

$$h(x) - h(y) \geq \frac{1}{2L} \|\nabla h(x)\|^2$$

Rearranging, we get:

$$\|\nabla h(x)\|^2 \leq 2L(h(x) - h(y)) \leq 2L \cdot h(x)$$

By using $x \in \mathcal{K}$ we get:

$$\|\nabla f(x) - \nabla f(x^*)\|^2 \leq 2L(f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle$$

Since x^* is the minimum point of f then:

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{K}$$

Thus we get that:

$$\|\nabla f(x) - \nabla f(x^*)\|^2 \leq 2L(f(x) - f(x^*))$$

□

Lemma A.7. *If $\mathcal{A}_t \leq \frac{1}{2T} \sum_{\tau=1}^{T} \mathcal{A}_\tau + \mathcal{B}, \forall t \in [T]$, then $\mathcal{A}_t \leq 2\mathcal{B}, \forall t \in [T]$.*

Proof of Lemma A.7. Let's sum the inequalities:

$$\sum_{t=1}^{T} \mathcal{A}_t \leq \sum_{t=1}^{T} \left(\frac{1}{2T} \sum_{\tau=1}^{T} \mathcal{A}_\tau + \mathcal{B} \right) = \frac{1}{2} \sum_{\tau=1}^{T} \mathcal{A}_\tau + T\mathcal{B}$$

The inequality is from the assumption, and the equality is because we sum constant values. If we rearrange this we get:

$$\sum_{\tau=1}^{T} \mathcal{A}_\tau \leq 2T\mathcal{B}$$

And then:

$$\mathcal{A}_t \leq \frac{1}{2T} \sum_{\tau=1}^{T} \mathcal{A}_\tau + \mathcal{B} \leq \mathcal{B} + \mathcal{B} = 2\mathcal{B}$$

□

B. Proofs of Section 2

B.1. Proof of Equations (3) and (4)

Both claims use the same principle:

$$\mathbb{E}\left\|X-\mathbb{E}\left[X\right]\right\|^2 \leq \mathbb{E}\left\|X\right\|^2$$

And so:

$$\begin{aligned}\mathbb{E}\left\|\nabla f(x;z)-\nabla f(x)\right\|^2 \leq & \mathbb{E}\left\|\nabla f(x;z)\right\|^2 \leq G^2\\ \mathbb{E}\left\|(\nabla f(x;z)-\nabla f(x))-(\nabla f(y;z)-\nabla f(y))\right\|^2 \leq & \mathbb{E}\left\|\nabla f(x;z)-\nabla f(y;z)\right\|^2 \leq L^2\left\|x-y\right\|^2\end{aligned}$$

Thus $\sigma \leq G$ and $\sigma_L \leq L$.

B.2. Proof of Lemma 2.2

Let us calculate $\mathbb{D}_\alpha\left(P\|Q\right)$ by definition.

$$\begin{aligned}\mathbb{D}_\alpha\left(P\|Q\right) =& \frac{1}{\alpha-1}\log\left(\mathbb{E}_{X\sim P}\left[\left(\frac{P(X)}{Q(X)}\right)^{\alpha-1}\right]\right) = \frac{1}{\alpha-1}\log\left(\mathbb{E}_{X\sim P}\left[e^{(\alpha-1)\frac{1}{2\sigma^2}\left(\|X-\mu-\Delta\|^2-\|X-\mu\|^2\right)}\right]\right)\\ =& \frac{1}{\alpha-1}\log\left(\mathbb{E}_{X\sim P}\left[e^{\frac{\alpha-1}{2\sigma^2}\left(\|\Delta\|^2-2\langle\Delta,X-\mu\rangle\right)}\right]\right) = \frac{1}{\alpha-1}\log\left(e^{\frac{\alpha-1}{2\sigma^2}\|\Delta\|^2+\frac{1}{2}\sigma^2\frac{(\alpha-1)^2}{4\sigma^4}4\|\Delta\|^2}\right)\\ =& \frac{\|\Delta\|^2}{2\sigma^2}+\frac{(\alpha-1)\|\Delta\|^2}{2\sigma^2} = \frac{\alpha\|\Delta\|^2}{2\sigma^2}\end{aligned}$$

The first equality is the definition of the Rényi divergence, in the second we input the values of P and Q, then we open the norm, and then we calculate the expectation using the moment generating function of a Gaussian random vector $\mathbb{E}\left[e^{\langle a,X\rangle}\right] = e^{\langle\mu,a\rangle+\frac{1}{2}\sigma^2\|a\|^2}$, and finally the log and the exponent cancel each other, and we fix things up.

B.3. Proof of Lemma 2.5

We will start with the first part. To do it, we will show another property of the Rényi divergence. Holder's inequality states that for any $p,q\geq 1$ such that $1/p+1/q=1$, we get $\|fq\|_1 \leq \|f\|_p\|g\|_q$. We will pick $p=\alpha, q=\frac{\alpha}{\alpha-1}, f=\frac{P}{Q^{1/q}}, g=Q^{1/q}$, and the norm to be on an arbitrary event A, and then:

$$P(A) = \int_A P(x)dx \leq \left(\int_A \frac{(P(x))^\alpha}{(Q(x))^{\alpha-1}}dx\right)^{\frac{1}{\alpha}}\left(\int_A Q(x)dx\right)^{\frac{\alpha-1}{\alpha}} \leq \left(e^{\mathbb{D}_\alpha(P\|Q)}Q(A)\right)^{\frac{\alpha-1}{\alpha}} \leq \left(e^\epsilon Q(A)\right)^{\frac{\alpha-1}{\alpha}}$$

The equality is the definition of the probability of the event, the inequality is from Holder's, next we expand the integral over everything to get the Rényi divergence, and finally we bound the Rényi divergence by ϵ. We pick the event A to be $A=\mathcal{O}$, and then:

$$\mathbb{P}\left\{\mathcal{A}(\mathcal{S})=\mathcal{O}\right\} \leq \left(e^\epsilon\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\}\right)^{1-\frac{1}{\alpha}}$$

If $\left(e^\epsilon\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\}\right)^{1-\frac{1}{\alpha}} \leq \delta$ then $\mathbb{P}\left\{\mathcal{A}(\mathcal{S})=\mathcal{O}\right\} \leq \delta$. If $\left(e^\epsilon\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\}\right)^{1-\frac{1}{\alpha}} > \delta$ then $\left(e^\epsilon\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\}\right)^{-\frac{1}{\alpha}} < \delta^{-\frac{1}{\alpha-1}} = e^{\frac{\log(1/\delta)}{\alpha-1}}$, and then:

$$\mathbb{P}\left\{\mathcal{A}(\mathcal{S})=\mathcal{O}\right\} \leq e^{\epsilon+\frac{\log(1/\delta)}{\alpha-1}}\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\}$$

If we combine both cases, we showed that:

$$\mathbb{P}\left\{\mathcal{A}(\mathcal{S})=\mathcal{O}\right\} \leq \max\left\{e^{\epsilon+\frac{\log(1/\delta)}{\alpha-1}}\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\},\delta\right\} \leq e^{\epsilon+\frac{\log(1/\delta)}{\alpha-1}}\mathbb{P}\left\{\mathcal{A}(\mathcal{S}')=\mathcal{O}\right\}+\delta$$

Now for the second part. Since $\mathcal{A}$ is $\left(\alpha,\frac{\alpha\rho^2}{2}\right)$-RDP for every $\alpha\geq 1$, then it is also $\left(\frac{\alpha\rho^2}{2}+\frac{\log(1/\delta)}{\alpha-1},\delta\right)$-DP, for every $\alpha\geq 1,\delta\in(0,1)$. We will pick the value that minimize this expression $\alpha = 1+\frac{1}{\rho}\sqrt{2\log(1/\delta)}$, and get that for every $\delta\in(0,1)$, $\mathcal{A}$ is $\left(\frac{\rho^2}{2}+\rho\sqrt{2\log(1/\delta)},\delta\right)$-DP.

C. Proofs of Section 4

C.1. Proof of Lemma 4.1

First note that for $t = 1$ we have $\alpha_t = 1$, and therefore we have, $s_{1,i} := g_{1,i} = d_{1,i} = q_{1,i}$, which implies that the lemma holds for $t = 1$. For that case of $t > 1$, our choices for $\{\alpha_t, \beta_t\}$ imply that $\beta_{t+1}\alpha_{t+1} = 1, (1 - \beta_{t+1})\alpha_{t+1} = \alpha_t$, which prove Equation (8). Unrolling the above equation yields $q_{t+1,i} = \sum_{\tau=1}^{t+1} s_{\tau,i}$, which establishes the first part of the lemma. For the second part regarding $\varepsilon_{t,i}$, note that our choice $\alpha_t = t$, together with the definition of $\bar{s}_t$ (Equation (9)) implies that we can write $\bar{s}_{t,i} = \alpha_t \bar{g}_{t,i} - \alpha_{t-1}\bar{g}_{t-1,i}$ (where for consistency we denote $\alpha_0 := 0$). Therefore, for any t:

$$\sum_{\tau=1}^{t} \bar{s}_{\tau,i} = \alpha_t \bar{g}_{t,i}$$

Using the above together with $q_t = \sum_{\tau}^{t} s_{\tau,i}$, immediately gives:

$$\varepsilon_{t,i} := q_{t,i} - \alpha_t \bar{g}_{t,i} = \sum_{\tau=1}^{t} (s_{\tau,i} - \bar{s}_{\tau,i})$$

Which establishes the proof.

C.2. Proof of Lemma 4.2

Before we begin, we shall bound the difference between consecutive query points:

$$x_t - x_{t-1} = x_t - \frac{\alpha_{1:t}x_t - \alpha_t w_t}{\alpha_{1:t-1}} = \frac{\alpha_t}{\alpha_{1:t-1}}(w_t - x_t)$$

Where the first equality is due to $\alpha_{1:t}x_t = \alpha_{1:t-1}x_{t-1} + \alpha_t w_t$. Using the above enables to bound the following scaled difference:

$$\alpha_{t-1} \|x_t - x_{t-1}\| = \left(\frac{\alpha_{t-1}\alpha_t}{\alpha_{1:t-1}}\right) \|w_t - x_t\| \leq 2D \tag{18}$$

Where we use that fact that $\alpha_t = t$, which implies $\alpha_{t-1}\alpha_t = 2\alpha_{1:t-1}$, as well as the bounded diameter assumption.

First part:

$$\begin{aligned}
\|s_{t,i}\| &= \|g_{t,i} + \alpha_{t-1}(g_{t,i} - \tilde{g}_{t-1,i})\| \leq \|g_{t,i}\| + \alpha_{t-1} \|g_{t,i} - \tilde{g}_{t-1,i}\| \\
&= \|\nabla f(x_t; z_{t,i})\| + \alpha_{t-1} \|\nabla f(x_t; z_{t,i}) - \nabla f(x_{t-1}; z_{t,i})\| \\
&\leq G + \alpha_{t-1} L \|x_t - x_{t-1}\| \leq G + 2LD := S
\end{aligned}$$

The first equality is from the definition of $s_{t,i}$, then we use the triangle inequality, then explicitly employ the definitions of $g_{t,i}, \tilde{g}_{t-1,i}$, next we use Lipschitz and smoothness, and finally, we employ Equation (18), and the definition of S.

Second part:

$$\begin{aligned}
\mathbb{E}\|s_{t,i} - \bar{s}_{t,i}\|^2 &= \mathbb{E}\|(g_{t,i} - \bar{g}_{t,i}) + \alpha_{t-1}((g_{t,i} - \bar{g}_{t,i}) - (\tilde{g}_{t-1,i} - \bar{g}_{t-1,i}))\|^2 \\
&\leq \left(\sqrt{\mathbb{E}\|g_{t,i} - \bar{g}_{t,i}\|^2} + \alpha_{t-1}\sqrt{\mathbb{E}\|(g_{t,i} - \bar{g}_{t,i}) - (\tilde{g}_{t-1,i} - \bar{g}_{t-1,i})\|^2}\right)^2 \\
&\leq (\sigma + \alpha_{t-1}\sigma_L \|x_t - x_{t-1}\|)^2 \leq (\sigma + 2\sigma_L D)^2 := \tilde{\sigma}^2
\end{aligned}$$

The first equality is by the definition of $s_{t,i} - \bar{s}_{t,i}$, then we use the inequality: $\mathbb{E}\|X + Y\|^2 \leq \left(\sqrt{\mathbb{E}\|X\|^2} + \sqrt{\mathbb{E}\|Y\|^2}\right)^2$, which holds since:

$$\begin{aligned}
\mathbb{E}\|X + Y\|^2 &= \mathbb{E}\|X\|^2 + 2\langle X, Y\rangle + \|Y\|^2 \\
&\leq \mathbb{E}\|X\|^2 + \sqrt{\mathbb{E}\|X\|^2 \cdot \|Y\|^2} + \mathbb{E}\|Y\|^2 = \left(\sqrt{\mathbb{E}\|X\|^2} + \sqrt{\mathbb{E}\|Y\|^2}\right)^2
\end{aligned}$$

Then, we use the bounded variance assumption as well as the bounded smoothness variance assumption (Equations (3) and (4)), and finally, we employ Equation (18), and the definition of $\tilde{\sigma}$.

C.3. Proof of Lemma 4.3

Using the previous lemmas will naturally allow us to bound $\varepsilon_{t,i}$:

$$\mathbb{E}\left\|\varepsilon_{t,i}\right\|^2 = \mathbb{E}\left\|\sum_{\tau=1}^{t}(s_{\tau,i} - \bar{s}_{\tau,i})\right\|^2 = \sum_{\tau=1}^{t}\mathbb{E}\left\|s_{\tau,i} - \bar{s}_{\tau,i}\right\|^2 \leq \sum_{\tau=1}^{t}\tilde{\sigma}^2 = \tilde{\sigma}^2 t$$

The first equality is Lemma 4.1, then we use Lemma A.2, and finally Lemma 4.2.

D. Proof of Privacy

Here we will provide the proof of the privacy of Algorithms 1 and 2.

D.1. Proof of Theorem 5.1

Let us look at a single machine i. Let $\mathcal{S}_i, \mathcal{S}_i'$ be *neighbouring datasets* of T samples which differ on a single datapoint; i.e. assume there exists $\tau^* \in [T]$ such that $\mathcal{S}_i := \{z_{1,i}, z_{2,i}, \ldots, z_{\tau^*,i}, \ldots, z_{T,i}\}$ and $\mathcal{S}_i' := \{z_{1,i}, z_{2,i}, \ldots, z'_{\tau^*,i}, \ldots, z_{T,i}\}$, and $z_{\tau^*} \neq z'_{\tau^*}$.

Notation: We will employ $q_{t,i}(\mathcal{S}_i), \tilde{q}_{t,i}(\mathcal{S}_i), x_t(\mathcal{S}), s_{t,i}(\mathcal{S}_i)$ to denote the resulting values of these quantities when we invoke Algorithm 1 with the datasets $\mathcal{S}$, we will similarly denote $q_{t,i}(\mathcal{S}_i'), \tilde{q}_{t,i}(\mathcal{S}_i'), x_t(\mathcal{S}'), s_{t,i}(\mathcal{S}_i')$.

First Part: Here we will bound the Rényi Divergence between $\tilde{q}_{t,i}(\mathcal{S}_i)$ and $\tilde{q}_{t,i}(\mathcal{S}_i')$ conditioned on the values of $\{x_1, \{\tilde{q}_\tau\}_{\tau=1}^{t-1}\}$. Note that given $\{x_1, \{\tilde{q}_\tau\}_{\tau=1}^{t-1}\}$ we can directly compute the values of $\{w_\tau\}_{\tau=1}^{t}$ and therefore also $\{x_\tau\}_{\tau=1}^{t}$ (see Algorithm 1). Thus, we may assume that prior to the computation of $q_{t,i}$, we are given $\{x_\tau\}_{\tau=1}^{t}$.

Note that given the query history $x_1, \ldots, x_t$, and the dataset $\mathcal{S}_i$, $q_{t,i}(\mathcal{S}_i)$ is known (respectively $q_{t,i}(\mathcal{S}_i')$ is known given the dataset $\mathcal{S}_i'$ and the query history). Now, since we obtain $\tilde{q}_{t,i}$ by adding a Gaussian noise $Y_{t,i}$ to $q_{t,i}$, we may use Lemma 2.2 and obtain:

$$\mathbb{D}_\alpha\left(\tilde{q}_{t,i}(\mathcal{S}_i)\|\tilde{q}_{t,i}(\mathcal{S}_i')\right) = \frac{\alpha\Delta_{t,i}^2}{2\sigma_{t,i}^2}$$

Where $\sigma_{t,i}^2$ is the variance of the noise $Y_{t,i}$, and $\Delta_{t,i} := \|q_{t,i}(\mathcal{S}_i) - q_{t,i}(\mathcal{S}_i')\|$.
Next, we will bound $\Delta_{t,i}$: First note that given $x_1, \ldots, x_t$ then by the definition of $s_{t,i}$ (Equation (9)) it only depends on $z_{t,i}$, i.e.:

$$s_{t,i} = g_{t,i} + \alpha_{t-1,i}(g_{t,i} - \tilde{g}_{t-1,i}) = \nabla f(x_t; z_{t,i}) + \alpha_{t-1}(\nabla f(x_t; z_{t,i}) - \nabla f(x_{t-1}; z_{t,i}))$$

Thus, for any $t \in [T]$, then given $x_1, \ldots x_t$ we have:

$$s_{t,i}(\mathcal{S}_i) - s_{t,i}(\mathcal{S}_i') = 0, \quad \forall t \neq \tau^*$$

where we used the fact that the datasets $\mathcal{S}_i, \mathcal{S}_i'$ only differ on the τ^*'th sample. Using the above together with the fact that $q_{t,i} = \sum_{\tau=1}^{t} s_{\tau,i}$ (see Lemma 4.1), implies the following $\forall t$:

$$\begin{aligned}\Delta_{t,i} := \|q_{t,i}(\mathcal{S}_i) - q_{t,i}(\mathcal{S}_i')\| &= \mathbb{I}\{t \geq \tau\}\,\|s_{\tau^*,i}(\mathcal{S}_i) - s_{\tau^*,i}(\mathcal{S}_i')\| \\ &\leq \mathbb{I}\{t \geq \tau^*\}(\|s_{\tau^*,i}(\mathcal{S}_i)\| + \|s_{\tau^*,i}(\mathcal{S}_i')\|) \leq 2S \cdot \mathbb{I}\{t \geq \tau^*\}\end{aligned}$$

Where $\mathbb{I}\{t \geq \tau^*\}$ denotes the indicator function (i.e. it equals zero if $t < \tau^*$, and is equal to 1 otherwise). Moreover, we also used Lemma 4.2 to bound the norms of $s_{\tau^*}(\mathcal{S}), s_{\tau^*}(\mathcal{S}')$. Using the above together with Equation (10) we can bound for any $t \in [T]$:

$$\mathbb{D}_\alpha\left(\tilde{q}_{t,i}(\mathcal{S}_i)\|\tilde{q}_{t,i}(\mathcal{S}_i')\right) = \frac{\alpha\Delta_{t,i}^2}{2\sigma_{t,i}^2} \leq \frac{2\alpha S^2}{\sigma_{t,i}^2} \cdot \mathbb{I}\{t \geq \tau^*\}$$

Second Part: Here we will bound the Rényi Divergence of the sequence $\{\tilde{q}_{t,i}\}_{t\in[T]}$. In Equation (11) we bound the Rényi Divergence of the t'th step of our algorithm, which produces $\tilde{q}_{t,i}$ based on the previously computed $\{x_1, \{\tilde{q}_{\tau,i}\}_{\tau=1}^{t-1}\}$ as well

as based on the dataset, and the injected noise $Y_{t,i}$. We shall relate this t'th step update as a procedure, and denote by $\mathcal{A}_t$. Thus, the procedure $\mathcal{A}_t$ computes its output $\tilde{q}_{t,i}$ based on the outputs of the previous procedures $\mathcal{A}_1, \dots \mathcal{A}_{t-1}$, as well as on the dataset and the initialization point x_1 (which does not depend on the dataset). This allows us to directly apply the RDP Composition rule in Lemma A.3, together with Equation (11), yielding $\forall t \in [T]$:

$$\mathbb{D}_\alpha\left(\{\tilde{q}_{\tau,i}(\mathcal{S}_i)\}_{\tau=1}^t \| \{\tilde{q}_{\tau,i}(\mathcal{S}_i')\}_{\tau=1}^t\right) \leq \sum_{\tau=1}^t \frac{\alpha\Delta_{\tau,i}^2}{2\sigma_{\tau,i}^2} \leq 2\alpha S^2 \sum_{\tau=\tau^*}^t \frac{1}{\sigma_{\tau,i}^2} \leq 2\alpha S^2 \sum_{\tau=1}^t \frac{1}{\sigma_{\tau,i}^2} \tag{19}$$

Thus, the theorem follows as per Definition 2.4 of RDP, and by Equation (19).

D.2. Proof of Theorem 6.1

The proof follows the same steps as the proof of Appendix D.1. Since we obtain $\tilde{q}_t$ by adding a Gaussian noise Y_t to q_t, we may use Lemma 2.2 and obtain:

$$\mathbb{D}_\alpha\left(\tilde{q}_t(\mathcal{S}) \| \tilde{q}_t(\mathcal{S}')\right) = \frac{\alpha\Delta_t^2}{2\sigma_t^2}$$

Where σ_t^2 is the variance of the noise Y_t, and $\Delta_t := \|q_t(\mathcal{S}) - q_t(\mathcal{S}')\|$. Since $q_t = \frac{1}{M}\sum_{i=1}^M \sum_{\tau=1}^t s_{\tau,i}$ and only $z_{\tau^*,i}$ is different between $\mathcal{S}, \mathcal{S}'$, we get:

$$\begin{aligned}\Delta_t := \|q_t(\mathcal{S}) - q_t(\mathcal{S}')\| &= \frac{1}{M}\mathbb{I}\{t \geq \tau\} \|s_{\tau^*}(\mathcal{S}) - s_{\tau^*}(\mathcal{S}')\| \\ &\leq \frac{1}{M}\mathbb{I}\{t \geq \tau^*\}(\|s_{\tau^*}(\mathcal{S})\| + \|s_{\tau^*}(\mathcal{S}')\|) \leq \frac{2S}{M} \cdot \mathbb{I}\{t \geq \tau^*\}\end{aligned}$$

Using the above together with Equation (14) we can bound for any $t \in [T]$:

$$\mathbb{D}_\alpha\left(\tilde{q}_t(\mathcal{S}) \| \tilde{q}_t(\mathcal{S}')\right) = \frac{\alpha\Delta_t^2}{2\sigma_t^2} \leq \frac{2\alpha S^2}{M^2\sigma_t^2} \cdot \mathbb{I}\{t \geq \tau^*\}$$

As before, we apply the RDP Composition rule in Lemma A.3, together with Equation (15), yielding $\forall t \in [T]$:

$$\mathbb{D}_\alpha\left(\{\tilde{q}_\tau(\mathcal{S})\}_{\tau=1}^t \| \{\tilde{q}_\tau(\mathcal{S}')\}_{\tau=1}^t\right) \leq \sum_{\tau=1}^t \frac{\alpha\Delta_\tau^2}{2\sigma_\tau^2} \leq \frac{2\alpha S^2}{M^2} \sum_{\tau=\tau^*}^t \frac{1}{\sigma_\tau^2} \leq \frac{2\alpha S^2}{M^2} \sum_{\tau=1}^t \frac{1}{\sigma_\tau^2}$$

To bound the RDP of the $\{x_t\}$ sequence, recall from Algorithm 2 that given $\{x_1, \{\tilde{q}_\tau\}_{\tau=1}^{t-1}\}$ we can directly compute the values of $\{w_\tau\}_{\tau=1}^t$ and therefore also $\{x_\tau\}_{\tau=1}^t$ (see Algorithm 2). This allows us to bound privacy of the queries $\{x_\tau\}_{\tau=1}^t$ as follows for any $t \in [T]$:

$$\mathbb{D}_\alpha\left(\{x_\tau(\mathcal{S})\}_{\tau=1}^t \| \{x_\tau(\mathcal{S}')\}_{\tau=1}^t\right) \leq \mathbb{D}_\alpha\left(\{\tilde{q}_\tau(\mathcal{S})\}_{\tau=1}^t \| \{\tilde{q}_\tau(\mathcal{S}')\}_{\tau=1}^t\right) \leq \frac{2\alpha S^2}{M^2} \sum_{\tau=1}^t \frac{1}{\sigma_\tau^2} \tag{20}$$

Where we have used the Post-Processing property of RDP appearing in Lemma A.4.

Thus, the theorem follows as per Definition 2.4 of RDP, and by Equation (20).

E. Proof of Convergence

Here we will provide the proof of the convergence of Algorithms 1 and 2.

Our starting point is the following lemma regarding the update rule $w_{t+1} = \Pi_{\mathcal{K}}\left(w_t - \eta\tilde{q}_t\right)$ that we employ in Algorithms 1 and 2. Employing Lemma A.5 together with the Anytime guarantees, i.e. Theorem A.1, and recalling that x_t is a weighted average of the $\{w_t\}_t$ sequence, implies that we may bound the excess loss $\mathcal{R}(x_t) := \mathbb{E}\left[f(x_t)\right] - \min_{x\in\mathcal{K}}\{f(x)\}$ of Algorithms 1

and 2 as follows:

$$
\begin{aligned}
\alpha_{1:t}\mathcal{R}(x_t) =& \alpha_{1:t}(\mathbb{E}\left[f(x_t)\right] - f(x^*)) \leq \sum_{\tau=1}^{t} \mathbb{E}\left[\alpha_\tau \left\langle \nabla f(x_\tau), w_\tau - x^* \right\rangle\right] = \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \tilde{q}_\tau - \varepsilon_\tau - Y_\tau, w_\tau - x^* \rangle\right] \\
=& \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \tilde{q}_\tau, w_\tau - x^* \rangle\right] - \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \varepsilon_\tau, w_\tau - x^* \rangle\right] \\
=& \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \tilde{q}_\tau, w_{\tau+1} - x^* \rangle\right] + \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \tilde{q}_\tau, w_\tau - w_{\tau+1} \rangle\right] - \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \varepsilon_\tau, w_\tau - x^* \rangle\right] \\
\leq& \frac{D^2}{2\eta} - \frac{1}{2\eta} \sum_{\tau=1}^{t} \mathbb{E}\left\|w_\tau - w_{\tau+1}\right\|^2 + \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \tilde{q}_\tau, w_\tau - w_{\tau+1} \rangle\right] - \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \varepsilon_\tau, w_\tau - x^* \rangle\right] \\
\leq& \frac{D^2}{2\eta} + \underbrace{\sum_{\tau=1}^{t} \mathbb{E}\left[\langle \alpha_\tau \nabla f(x_\tau) - \alpha_\tau \nabla f(x^*) + Y_\tau, w_\tau - w_{\tau+1} \rangle - \frac{1}{2\eta}\left\|w_\tau - w_{\tau+1}\right\|^2\right]}_{\text{(A)}} \\
&+ \underbrace{\sum_{\tau=1}^{t} \mathbb{E}\left[\langle \alpha_\tau \nabla f(x^*), w_\tau - w_{\tau+1} \rangle\right]}_{\text{(B)}} + \underbrace{\sum_{\tau=1}^{t} \mathbb{E}\left[\langle -\varepsilon_\tau, w_{\tau+1} - x^* \rangle\right]}_{\text{(C)}}
\end{aligned}
$$

The first inequality is due to Theorem A.1, the next one is due to $\tilde{q}_\tau = \alpha_\tau \nabla f(x_\tau) + \varepsilon_\tau + Y_\tau$, then we use the fact that Y_τ is zero mean and is independent of w_τ, and after that we add and subtract $\pm \langle \tilde{q}_\tau, w_{\tau+1} \rangle$, and then use Lemma A.5. Next, we reuse the definition of $\tilde{q}_\tau$ and move terms together plus adding and subtracting $\pm \alpha_\tau \nabla f(x^*)$. Next we separately bound the above terms (A), (B), (C).

Bounding (A) This term can be bounded as follows:

$$
\begin{aligned}
\text{(A)} :=& \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \alpha_\tau \nabla f(x_\tau) - \alpha_\tau \nabla f(x^*) + Y_\tau, w_\tau - w_{\tau+1} \rangle - \frac{1}{2\eta} \sum_{\tau=1}^{t} \left\|w_\tau - w_{\tau+1}\right\|^2\right] \\
\leq& \frac{\eta}{2} \sum_{\tau=1}^{t} \mathbb{E}\left\|\alpha_\tau(\nabla f(x_\tau) - \nabla f(x^*)) + Y_\tau\right\|^2 = \frac{\eta}{2} \sum_{\tau=1}^{t} \alpha_\tau^2 \mathbb{E}\left\|\nabla f(x_\tau) - \nabla f(x^*)\right\|^2 + \frac{\eta}{2} \sum_{\tau=1}^{t} \mathbb{E}\left\|Y_\tau\right\|^2
\end{aligned}
$$

First, we use Young's inequality, and equality holds since Y_τ is independent of x_τ, and has zero mean.

Bounding (B) This term can be written as follows:

$$
\begin{aligned}
\text{(B)} :=& \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \alpha_\tau \nabla f(x^*), w_\tau - w_{\tau+1} \rangle\right] = \sum_{\tau=1}^{t} (\alpha_\tau - \alpha_{\tau-1}) \mathbb{E}\left[\langle \nabla f(x^*), w_\tau \rangle\right] - \alpha_t \mathbb{E}\left[\langle \nabla f(x^*), w_{t+1} \rangle\right] \\
=& \sum_{\tau=1}^{t} (\alpha_\tau - \alpha_{\tau-1}) \mathbb{E}\left[\langle \nabla f(x^*), w_\tau - w_{t+1} \rangle\right] \leq \sum_{\tau=1}^{t} (\alpha_\tau - \alpha_{\tau-1}) \left\|\nabla f(x^*)\right\| \mathbb{E}\left[\left\|w_\tau - w_{t+1}\right\|\right] \\
\leq& D \left\|\nabla f(x^*)\right\| \sum_{\tau=1}^{t} (\alpha_\tau - \alpha_{\tau-1}) = \alpha_t D \left\|\nabla f(x^*)\right\| = \alpha_t D G^*
\end{aligned}
$$

The first equality is rearrangement of the sum while defining $\alpha_0 = 0$, then we put the last term into the sum, and use Cauchy-Schwartz. Finally we use the diameter bound and telescope the sum. Note that we use the definition $G^* := \|\nabla f(x^*)\|$, and that $G^* \in [0, G]$.

Bounding (C) This term is bounded as follows:

$$
\text{(C)} := \sum_{\tau=1}^{t} \mathbb{E}\left[\langle \varepsilon_\tau, w_{\tau+1} - x^* \rangle\right] \leq \sum_{\tau=1}^{t} \mathbb{E}\left[\|\varepsilon_\tau\| \cdot \|w_{\tau+1} - x^*\|\right] \leq D \sum_{\tau=1}^{t} \sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2}
$$

The first inequality is Cauchy-Schwartz, and the second is using the bounded diameter and Jensen's inequality w.r.t. concave function $\sqrt{\cdot}$. In total, we get that:

$$\alpha_{1:t}\mathcal{R}(x_t) \leq \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{\tau=1}^{t}\alpha_\tau^2\mathbb{E}\left\|\nabla f(x_\tau) - \nabla f(x^*)\right\|^2 + \frac{\eta}{2}\sum_{\tau=1}^{t}\mathbb{E}\left\|Y_\tau\right\|^2 + \alpha_t DG^* + D\sum_{\tau=1}^{t}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2} \tag{21}$$

Using Lemma A.6, together with $\alpha_\tau^2 = \tau^2 \leq 2\frac{\tau(\tau+1)}{2} = 2\alpha_{1:\tau}$:

$$\alpha_\tau^2\mathbb{E}\left\|\nabla f(x_\tau) - \nabla f(x^*)\right\|^2 \leq 4\alpha_{1:\tau}L\mathcal{R}(x_\tau)$$

Using the above and plugging it back into Equation (21) yields:

$$\alpha_{1:t}\mathcal{R}(x_t) \leq \frac{D^2}{2\eta} + 2\eta L\sum_{\tau=1}^{t}\alpha_{1:\tau}\mathcal{R}(x_\tau) + \frac{\eta}{2}\sum_{\tau=1}^{t}\mathbb{E}\left\|Y_\tau\right\|^2 + \alpha_t DG^* + D\sum_{\tau=1}^{t}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2}$$

We will now increase t to T in the right-hand-side (since the RHS is monotonically non-decreasing with t) and use our choice of the learning rate which implies $\eta \leq \frac{1}{4LT}$, to get that for any $t \leq T$,

$$\alpha_{1:t}\mathcal{R}(x_t) \leq \frac{1}{2T}\sum_{\tau=1}^{T}\alpha_{1:\tau}\mathcal{R}(x_\tau) + \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{\tau=1}^{T}\mathbb{E}\left\|Y_\tau\right\|^2 + \alpha_T DG^* + D\sum_{\tau=1}^{T}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2}$$

Now, using the Lemma A.7 with:

$$\begin{aligned}\mathcal{A}_t =& \alpha_{1:t}\mathcal{R}(x_t)\\ \mathcal{B} =& \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{\tau=1}^{T}\mathbb{E}\left\|Y_\tau\right\|^2 + \alpha_T DG^* + D\sum_{\tau=1}^{T}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2}\end{aligned}$$

We finally get:

$$\alpha_{1:T}\mathcal{R}(x_T) \leq \frac{D^2}{\eta} + \eta\sum_{\tau=1}^{T}\mathbb{E}\left\|Y_\tau\right\|^2 + 2\alpha_T DG^* + 2D\sum_{\tau=1}^{T}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2}$$

We can see that the bound depends on our learning rate η, the estimation error ε_τ and the injected noise Y_τ.

Using Lemma 4.3, and the fact the $\{\varepsilon_{t,i}\}_{i\in[M]}$ are independent given the queries $\{x_\tau\}_{\tau\in[t]}$, since each of them depends on a different dataset, we get that: $\mathbb{E}\left\|\varepsilon_t\right\|^2 \leq \frac{\tilde{\sigma}^2 t}{M}$. Now we can bound:

$$\sum_{\tau=1}^{T}\sqrt{\mathbb{E}\left\|\varepsilon_\tau\right\|^2} \leq \sum_{\tau=1}^{T}\sqrt{\frac{\tilde{\sigma}^2\tau}{M}} \leq \frac{\tilde{\sigma}T^{1.5}}{\sqrt{M}}$$

The bound on Y_τ will be different between our two versions.

E.1. Proof of Theorem 5.2

Since we pick $\sigma_{t,i}^2 = \frac{4S^2T}{\rho^2}$ and $Y_\tau = \frac{1}{M}\sum_{i=1}^{M}Y_{\tau,i}$, then:

$$\sum_{\tau=1}^{T}\mathbb{E}\left\|Y_\tau\right\|^2 = \frac{1}{M^2}\sum_{\tau=1}^{T}\sum_{i=1}^{M}\mathbb{E}\left\|Y_{\tau,i}\right\|^2 = \frac{1}{M^2}\sum_{\tau=1}^{T}\sum_{i=1}^{M}d\sigma_{t,i}^2 = \frac{4S^2T^2d}{\rho^2M}$$

When we put our bounds together in Equation (13) we get:

$$\alpha_{1:T}\mathcal{R}(x_T) \leq \frac{D^2}{\eta} + \frac{4\eta S^2T^2d}{\rho^2M} + 2\alpha_T DG^* + \frac{2D\tilde{\sigma}T^{1.5}}{\sqrt{M}}$$

Finally, dividing by $\alpha_{1:T} = T(T+1)/2 \geq T^2/2$ and recalling our choice $\eta = \min\left\{\frac{\rho D\sqrt{M}}{2ST\sqrt{d}}, \frac{1}{4LT}\right\}$, using $\eta \leq \frac{\rho D\sqrt{M}}{2ST\sqrt{d}}$, $\frac{1}{\eta} \leq \frac{2ST\sqrt{d}}{\rho D\sqrt{M}} + 4LT$, and that $\alpha_T = T$, yields the final bound:

$$\mathcal{R}(x_T) \leq 4D\left(\frac{G^* + 2LD}{T} + \frac{2S\sqrt{d}}{\rho T\sqrt{M}} + \frac{\tilde{\sigma}}{\sqrt{TM}}\right)$$

Thus concluding the proof.

E.2. Proof of Theorem 6.2

Since we pick $\sigma_t^2 = \frac{4S^2T}{\rho^2M^2}$, then:

$$\sum_{\tau=1}^{T}\mathbb{E}\left\|Y_\tau\right\|^2 = \sum_{\tau=1}^{T}\sum_{i=1}^{M} d\sigma_t^2 = \frac{4S^2T^2d}{\rho^2M^2}$$

When we put our bounds together in Equation (17) we get:

$$\alpha_{1:T}\mathcal{R}(x_T) \leq \frac{D^2}{\eta} + \frac{4\eta S^2T^2d}{\rho^2M^2} + 2\alpha_T DG^* + \frac{2D\tilde{\sigma}T^{1.5}}{\sqrt{M}}$$

Finally, dividing by $\alpha_{1:T} = T(T+1)/2 \geq T^2/2$ and recalling our choice $\eta = \min\left\{\frac{\rho DM}{2ST\sqrt{d}}, \frac{1}{4LT}\right\}$, using $\eta \leq \frac{\rho DM}{2ST\sqrt{d}}$, $\frac{1}{\eta} \leq \frac{2ST\sqrt{d}}{\rho DM} + 4LT$, and that $\alpha_T = T$, yields the final bound:

$$\mathcal{R}(x_T) \leq 4D\left(\frac{G^* + 2LD}{T} + \frac{2S\sqrt{d}}{\rho TM} + \frac{\tilde{\sigma}}{\sqrt{TM}}\right)$$

Thus concluding the proof.

F. Review of (Lowy & Razaviyayn, 2023)

In (Lowy & Razaviyayn, 2023) they propose an algorithm for private federated learning. In this paper they assume to have M machines, and n is the size of the dataset used by each machine, and therefore the total dataset is of size $|\mathcal{S}| = nM$. Moreover, they denote by R to total number of update rounds, and by K the minibatch size used by each machine in every round, totalling RKM gradient computations overall. In theorem D.2 of this paper we can see that for the convex case of ISRL-DP (untrusted server) they use $R = \frac{\beta D\sqrt{M}}{L}\min\left\{\sqrt{n}, \frac{\epsilon_0 n}{\sqrt{d}}\right\} + \min\left\{nM, \frac{\epsilon_0^2 n^2 M}{d}\right\}$ time steps, and batch size of $K \geq \frac{\epsilon_0 n}{4\sqrt{2R\ln(2/\delta_0)}}$. The total number of gradient computations each silo use is RK, and the total number of computations across all silos is RKM:

$$\begin{aligned} RKM \geq & \frac{\epsilon_0 nM\sqrt{R}}{4\sqrt{2\ln(2/\delta_0)}} = \frac{\epsilon_0 nM}{4\sqrt{2\ln(2/\delta_0)}}\sqrt{\frac{\beta D}{L}\sqrt{nM}\min\left\{1, \frac{\epsilon_0\sqrt{n}}{\sqrt{d}}\right\} + nM\min\left\{1, \frac{\epsilon_0^2 n}{d}\right\}} \\ \geq & \frac{\epsilon_0 (nM)^{3/2}}{4\sqrt{2\ln(2/\delta_0)}}\min\left\{1, \frac{\epsilon_0\sqrt{n}}{\sqrt{d}}\right\} \end{aligned}$$

The first inequality is inputting K, then we input R, and then remove the first term under the square root. Thus in the prevalent regime where $\epsilon_0 \in \Theta(1)$ and $d \in O(n)$, then we can say that the total number of gradient computations is:

$$RKM \geq \Omega\left((nM)^{3/2}\right) = \Omega\left(|\mathcal{S}|^{3/2}\right)$$

Thus (Lowy & Razaviyayn, 2023) has computational complexity that is proportional to $|\mathcal{S}|^{3/2}$.