Private Heterogeneous Federated Learning Without a Trusted Server Revisited: Error-Optimal and Communication-Efficient Algorithms for Convex Losses

Changyu Gao [*1] **Andrew Lowy** [*1] **Xingyu Zhou** [*2] **Stephen J. Wright** [1]

Abstract

We revisit the problem of federated learning (FL) with private data from people who do not trust the server or other silos/clients. In this context, every silo (e.g. hospital) has data from several people (e.g. patients) and needs to protect the privacy of each person's data (e.g. health records), even if the server and/or other silos try to uncover this data. Inter-Silo Record-Level Differential Privacy (ISRL-DP) prevents each silo's data from being leaked, by requiring that silo i's *communications* satisfy item-level differential privacy. Prior work (Lowy & Razaviyayn, 2023a) characterized the optimal excess risk bounds for ISRL-DP algorithms with *homogeneous* (i.i.d.) silo data and convex loss functions. However, two important questions were left open: 1) Can the same excess risk bounds be achieved with *heterogeneous* (non-i.i.d.) silo data? 2) Can the optimal risk bounds be achieved with *fewer communication rounds*? In this paper, we give positive answers to both questions. We provide novel ISRL-DP FL algorithms that achieve the optimal excess risk bounds in the presence of heterogeneous silo data. Moreover, our algorithms are more *communication-efficient* than the prior state-of-the-art. For smooth loss functions, our algorithm achieves the *optimal* excess risk bound and has *communication complexity that matches the non-private lower bound*. Additionally, our algorithms are more *computationally efficient* than the previous state-of-the-art.

1. Introduction

Federated learning (FL) is a distributed machine learning paradigm in which multiple *silos* (a.k.a. clients), such as hospitals or cell-phone users, collaborate to train a global model. In FL, silos store their data locally and exchange focused updates (e.g. stochastic gradients), sometimes making use of a central server (Kairouz et al., 2021). FL has been applied in myriad domains, from consumer digital products (e.g. Google's mobile keyboard (Hard et al., 2018) and Apple's iOS (Apple, 2019)) to healthcare (Courtiol et al., 2019), finance (FedAI, 2019), and large language models (LLMs) (Hilmkil et al., 2021).

One of the primary reasons for the introduction of FL was to enhance protection of the privacy of people's data (McMahan et al., 2017). Unfortunately, local storage is not sufficient to prevent data from being leaked, because the model parameters and updates communicated between the silos and the central server can reveal sensitive information (Zhu & Han, 2020; Gupta et al., 2022). For example, Gupta et al. (2022) attacked a FL model for training LLMs to uncover private text data.

Differential privacy (DP) (Dwork et al., 2006) guarantees that private data cannot be leaked. Different variations of DP have been considered for FL. *Central DP* prevents the *final trained FL model* from leaking data to an *external adversary* (Jayaraman & Wang, 2018; Noble et al., 2022). However, central DP has two major drawbacks: 1) it does not provide a privacy guarantee for each individual silo; and 2) it does not ensure privacy when an attacker/eavesdropper has access to the server or to another silo.

Another notion of DP for FL is *user-level DP* (McMahan et al., 2018; Geyer et al., 2017; Levy et al., 2021). User-level DP mitigates the drawback 1) of central DP, by providing privacy for the *complete local data set* of each silo. User-level DP is practical in the *cross-device FL* setting, where each silo is a single person (e.g. cell-phone user) with a large number of records (e.g. text messages). However, user-level DP still permits privacy breaches if an adversary has access to the server or eavesdrops on the communications between silos. Moreover, user-level DP is not well-suited for *cross-silo* federated learning, where silos represent organizations

[*]Equal contribution [1]University of Wisconsin-Madison, Madison, WI, USA [2]Wayne State University, Detroit, MI, USA. Correspondence to: Changyu Gao <changyu.gao@wisc.edu>, Andrew Lowy <alowy@wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

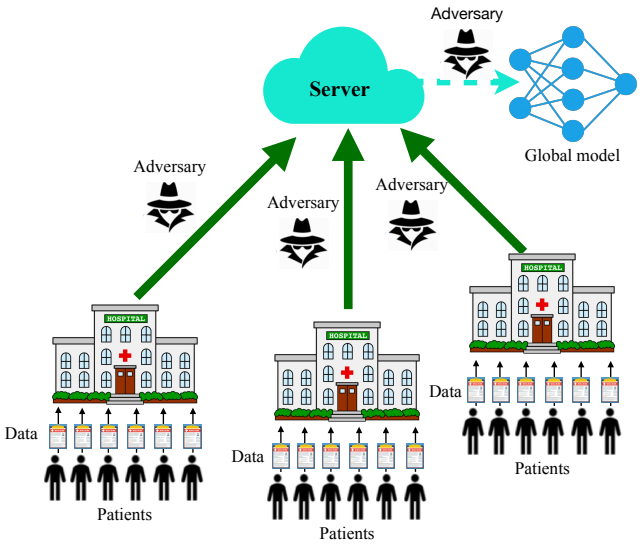

Figure 1. ISRL-DP maintains the privacy of each patient's record, provided the patient's *own hospital* is trusted. Silo i's messages are item-level DP, preventing data leakage, even if the server/other silos collude to decode the data of silo i.

like hospitals, banks, or schools that house data from many individuals (e.g. patients, customers, or students). In the cross-silo FL context, each person possesses a record, referred to as an "item," which may include sensitive data. Therefore, an appropriate notion of DP for cross-silo FL should safeguard the privacy of each individual record (i.e. "item-level differential privacy") within silo i, rather than the complete aggregated data of silo i.

Following Lowy & Razaviyayn (2023a); Lowy et al. (2023a); Liu et al. (2022) (among others), this work considers *inter-silo record-level DP* (ISRL-DP). ISRL-DP requires that the full transcript of *messages* sent by silo i satisfy item-level DP, for all silos i. Thus, ISRL-DP guarantees the privacy of each silo's local data, even in the presence of an adversary with access to the server, other silos, or the communication channels between silos and the server. See Figure 1 for an illustration. If each silo only has one record, then ISRL-DP reduces to *local DP* (Kasiviswanathan et al., 2011; Duchi et al., 2013). However, in the FL setting where each silo has many records, ISRL-DP FL is more appropriate and permits much higher accuracy than local DP (Lowy & Razaviyayn, 2023a). Moreover, ISRL-DP implies user-level DP if the ISRL-DP parameters are sufficiently small (Lowy & Razaviyayn, 2023a). Thus, ISRL-DP is a practical privacy notion for cross-silo and cross-device FL when individuals do not trust the server or other silos/clients/devices with their sensitive data.

Problem Setup. Consider a FL environment with N silos, where each silo has a local data set with n samples: $X_i = (x_{i,1}, \cdots, x_{i,n})$ for $i \in [N] := \{1, \dots, N\}$. At the beginning of every communication round r, silos receive the global model w_r. Silos then utilize their local data to enhance the model, and send local updates to the server (or to other silos for peer-to-peer FL). The server (or other silos) uses the local silo updates to update the global model.

For each silo i, let $\mathcal{D}_i$ be an unknown probability distribution on a data domain $\mathcal{X}$. Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be a loss function (e.g. cross-entropy loss for logistic regression), where $\mathcal{W} \subset \mathbb{R}^d$ is a parameter domain. Assume $f(\cdot, x)$ is convex. Silo i's local objective is to minimize its expected population/test loss, which is

$$F_i(w) := \mathbb{E}_{x_i \sim \mathcal{D}_i}[f(w, x_i)]. \tag{1}$$

Our (global) objective is to find a model that achieves small error for all silos, by solving the FL problem

$$\min_{w \in \mathcal{W}} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} F_i(w) \right\}, \tag{FL}$$

while maintaining the privacy of each silo's local data under ISRL-DP.

Assume that the samples $\{x_{i,j}\}_{i \in [N], j \in [n]}$ are independent. The FL problem is *homogeneous* if the data is i.i.d.: i.e. $\mathcal{D}_i = \mathcal{D}_j$ for all $i, j \in [N]$. In this work, we focus on the more challenging *heterogeneous* (non-i.i.d.) case: i.e. the data distributions $\{\mathcal{D}_i\}_{i=1}^N$ may be arbitrary. This case arises in many practical FL settings (Kairouz et al., 2021).

Contributions. The quality of a private algorithm $\mathcal{A}$ for solving Problem FL is measured by its *excess risk* (a.k.a. excess population/test loss), $\mathbb{E}[F(\mathcal{A}(\mathbf{X}))] - F^*$, where $F^* = \inf_{w \in \mathcal{W}} F(w)$ and the expectation is over the random draw of the data $\mathbf{X} = (X_1, \dots, X_N)$ as well as the randomness of $\mathcal{A}$. Lowy & Razaviyayn (2023a) characterized the minimax optimal excess risk (up to logarithms) for convex ISRL-DP FL in the *homogeneous* (i.i.d.) case:

$$\widetilde{\Theta}\left(\frac{1}{\sqrt{N}} \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \log(1/\delta)}}{\varepsilon n} \right) \right), \tag{2}$$

where d is the dimension of the parameter space and ε, δ are the privacy parameters. For *heterogeneous* FL, the state-of-the-art excess risk bound is $O\left(1/\sqrt{Nn} + \left(\sqrt{d \log(1/\delta)}/(\varepsilon n \sqrt{N})\right)^{4/5}\right)$, assuming smoothness of the loss function (Lowy & Razaviyayn, 2023a). Lowy & Razaviyayn (2023a) left the following question open:

Question 1. Can the optimal ISRL-DP excess risk in (2) for solving Problem FL be attained in the presence of *heterogeneous* silo data?

Contribution 1. We give a positive answer to **Question 1**. Moreover, we do not require smoothness of the loss function to attain the optimal rate: see Theorem 3.1.

In practical FL settings (especially cross-device settings), it can be expensive or slow for silos to communicate with the server or with each other (Kairouz et al., 2021). Thus, another important desiderata for FL algorithms is communication efficiency, which is measured by *communication complexity*: the number of communication rounds R that are required to achieve excess risk α. In the *homogeneous* setting, the state-of-the-art communication complexity for an algorithm achieving optimal excess risk is due to Lowy & Razaviyayn (2023a, Theorem D.1):

$$R_{\text{SOTA}} = \widetilde{O}\left(\min\left\{Nn, \frac{Nn^2\varepsilon^2}{d}\right\}\right). \tag{3}$$

Question 2. Can the optimal ISRL-DP excess risk in (2) for solving Problem FL be attained in *fewer communication rounds* R, for $R \ll R_{SOTA}$?

Contribution 2. We answer **Question 2** positively. For smooth losses, our algorithm achieves optimal excess risk with significantly improved communication complexity:

$$R_{\text{smooth}} = \widetilde{O}\left(\min\left\{(Nn)^{1/4}, \frac{N^{1/4}n^{1/2}\varepsilon^{1/2}}{d^{1/4}}\right\}\right). \tag{4}$$

Our communication complexity in (4) *matches the non-private lower bound* of Woodworth et al. (2020) in the high heterogeneity regime (Theorem 2.4), hinting at the communication-optimality of our algorithm.

For non-smooth loss functions, our communication complexity is

$$R_{\text{non-smooth}} = \widetilde{O}\left(\min\left\{(Nn)^{1/2}, \frac{N^{1/2}n\varepsilon}{d^{1/2}}\right\}\right), \tag{5}$$

a major improvement over the prior state-of-the-art bound in (3).

Moreover, we achieve these improved communication complexity and optimal excess risk bounds *without assuming homogeneity* of silo data.

When $N = 1$ in (FL), ISRL-DP FL reduces to central DP stochastic convex optimization (SCO), which has been studied extensively (Bassily et al., 2019; Feldman et al., 2020; Asi et al., 2021; Zhang et al., 2022). In this centralized setting, communication complexity is usually referred to as *iteration complexity*. Even in the special case of $N = 1$, *our iteration complexity for smooth losses improves over the prior state-of-the-art* result (Zhang et al., 2022).

Another important property of FL algorithms is *computational efficiency*, which we measure by *(sub)gradient complexity*: the number of stochastic (sub)gradients T that an algorithm must compute to achieve excess risk α.

In the *homogeneous* case, the state-of-the-art gradient complexity for an ISRL-DP FL algorithm that attains optimal excess risk is due to Lowy & Razaviyayn (2023a, Theorem D.1). For smooth losses and $\varepsilon = \Theta(1)$, Lowy et al. (2023a) obtain gradient complexity

$$T_{\text{SOTA}} = \widetilde{O}\left(N^2 \min\left\{n, \frac{n^2}{d}\right\} + N^{3/2} \min\left\{n^{3/2}, \frac{n^2}{d^{1/2}}\right\}\right). \tag{6}$$

Question 3. Can the optimal ISRL-DP excess risk in (2) for solving Problem FL be attained with *smaller gradient complexity* $T \ll T_{\text{SOTA}}$?

Contribution 3. We give a positive answer to **Question 3**. We provide an ISRL-DP FL algorithm with optimal excess risk and improved gradient complexity (see Theorem 2.1). When $d = \Theta(n)$ and $\varepsilon = \Theta(1)$, our gradient complexity bound simplifies to

$$T_{\text{smooth}} = \widetilde{O}\left(N^{5/4}n^{1/4} + (Nn)^{9/8}\right). \tag{7}$$

In Theorem 3.4, we also improve over the state-of-the-art subgradient complexity bound of Lowy & Razaviyayn (2023a, Theorem D.2) for *non-smooth* losses. Moreover, in contrast to these earlier results, we do not assume homogeneous data.

We summarize our main results in Table 1 and Table 2.

Our Algorithms and Techniques. Our algorithms combine and extend various private optimization techniques in novel ways.

Our ISRL-DP Algorithm 1 for smooth FL builds on the *ISRL-DP Accelerated MB-SGD* approach used by Lowy & Razaviyayn (2023a) for federated empirical risk minimization (ERM). We call this algorithm repeatedly to iteratively solve a carefully chosen sequence of regularized ERM subproblems, using the *localization* technique of Feldman et al. (2020). We obtain our novel optimal excess risk bounds for heterogeneous FL with an algorithmic *stability* (Bousquet & Elisseeff, 2002) argument. We make the key observation that the stability and generalization of regularized ERM in Shalev-Shwartz et al. (2009) does not require the data to be identically distributed. By combining this observation with a bound for ISRL-DP Accelerated MB-SGD, we obtain our excess risk bound.

It is not immediately clear that the approach just described should work, because the iterates of Accelerated MB-SGD are not necessarily stable. The instability of acceleration may explain why Lowy & Razaviyayn (2023a) used ISRL-DP Accelerated MB-SGD only for ERM and not for minimizing the population risk. We overcome this obstacle

Table 1. Comparison vs. SOTA for *Smooth* Loss Functions. In Tables 1-2, [LR'23] refers to Lowy & Razaviyayn (2023a); we omit logs and fix $M = N$, $d = \Theta(n)$, $\varepsilon = \Theta(1)$, $L = \beta = D = 1$. See theorems and the appendix for more general cases.

Algorithm & Setting	Excess Risk	Communication Complexity	Gradient Complexity
[LR'23] Alg. 2 (i.i.d.)	optimal	Nn	N^2n^2
[LR'23] Alg. 1 (non-i.i.d.)	suboptimal	$N^{1/5}n^{1/5}$	Nn
Alg. 1 (non-i.i.d.)	optimal	$N^{1/4}n^{1/4}$	$N^{5/4}n^{1/4} + (Nn)^{9/8}$

Table 2. Comparison vs. SOTA for *Nonsmooth* Loss Functions. First 3 algorithms use Nesterov smoothing.

Algorithm & Setting	Excess Risk	Communication Complexity	Gradient Complexity
[LR'23] Alg. 2 (i.i.d.)	optimal	Nn	N^2n^2
[LR'23] Alg. 1 (non-i.i.d.)	suboptimal	$N^{1/3}n^{1/3}$	Nn
Alg. 1 (non-i.i.d.)	optimal	$N^{1/2}n^{1/2}$	Omitted (see Section 3.1)
Alg. 4 (non-i.i.d.)	optimal	Nn	$N^2n + N^{3/2}n^{3/2}$
Algorithm 6 (non-i.i.d.)	optimal	$Nn^{5/4}$	$N^{3/2}n^{3/4} + N^{5/4}n^{11/8}$

with an alternative analysis that leverages the stability of regularized ERM, the convergence of ISRL-DP Accelerated MB-SGD to an approximate minimizer of the regularized empirical loss, and localization. This argument enables us to show that our algorithm has optimal excess risk. By carefully choosing algorithmic parameters, we also obtain state-of-the-art communication and gradient complexities.

To extend our algorithm to the non-smooth case, we use two different techniques. One is Nesterov smoothing (Nesterov, 2005), which results in an algorithm with favorable communication complexity. The other approach is to replace the ISRL-DP Accelerated MB-SGD ERM subsolver by an *ISRL-DP Subgradient Method*. This technique yields an algorithm with favorable (sub)gradient complexity. Third, we use randomized convolution smoothing, as in Kulkarni et al. (2021), to obtain another favorable gradient complexity bound.

1.1. Preliminaries

Differential Privacy. Let $\mathbb{X} = \mathcal{X}^{n\times N}$ and $\rho : \mathbb{X}^2 \to [0,\infty)$ be a distance between distributed data sets. Two distributed data sets $\mathbf{X}, \mathbf{X}' \in \mathbb{X}$ are *ρ-adjacent* if $\rho(\mathbf{X}, \mathbf{X}') \leq 1$. Differential privacy (DP) prevents an adversary from distinguishing between the outputs of algorithm $\mathcal{A}$ when it is run on adjacent databases:

Definition 1.1 (Differential Privacy (Dwork et al., 2006))**.** Let $\varepsilon \geq 0$, $\delta \in [0,1)$. A randomized algorithm $\mathcal{A} : \mathbb{X} \to \mathcal{W}$ is *(ε,δ)-differentially private* (DP) for all ρ-adjacent data sets $\mathbf{X}, \mathbf{X}' \in \mathbb{X}$ and all measurable subsets $S \subseteq \mathcal{W}$, we have

$$\mathbb{P}(\mathcal{A}(\mathbf{X}) \in S) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(\mathbf{X}') \in S) + \delta. \tag{8}$$

Definition 1.2 (Inter-Silo Record-Level Differential Privacy)**.** Let $\rho : \mathcal{X}^{2n} \to [0,\infty)$, $\rho(X_i, X_i') := \sum_{j=1}^{n} \mathbb{1}_{\{x_{i,j} \neq x'_{i,j}\}}$, $i \in [N]$. A randomized algorithm $\mathcal{A}$ is (ε,δ)-ISRL-DP if for all $i \in [N]$ and all ρ-adjacent silo data sets X_i, X_i', the full transcript of silo i's sent messages satisfies (8) for any fixed settings of other silos' data.

Notation and Assumptions. Let $\|\cdot\|$ be the ℓ_2 norm and $\Pi_{\mathcal{W}}(z) := \arg\min_{w\in\mathcal{W}} \|w - z\|^2$ denote the projection operator. Function $h : \mathcal{W} \to \mathbb{R}^m$ is L-Lipschitz if $\|h(w) - h(w')\| \leq L\|w - w'\|$, $\forall w, w' \in \mathcal{W}$. A differentiable function $h(\cdot)$ is *β-smooth* if its derivative ∇h is β-Lipschitz. For differentiable (w.r.t. w) $f(w,x)$, we denote its gradient w.r.t. w by $\nabla f(w,x)$.

We write $a \lesssim b$ if $\exists C > 0$ such that $a \leq Cb$. We use $a = \widetilde{O}(b)$ to hide poly-logarithmic factors.

We assume the following throughout:

Assumption 1.3.

1. $\mathcal{W} \subset \mathbb{R}^d$ is closed, convex. We assume $\|w - w'\| \leq D$, $\forall w, w' \in \mathcal{W}$.
2. $f(\cdot, x)$ is L-Lipschitz and convex for all $x \in \mathcal{X}$. In some places, we assume that $f(\cdot, x)$ is β-smooth.
3. In each round r, a uniformly random subset $S_r \subset [N]$ of M silos is available to communicate with the server.

For simplicity, in the main body, we often assume $M = N$. The Appendix contains the general statements of all results, complete proofs, and a further discussion of related work.

2. Localized ISRL-DP Accelerated MB-SGD for Smooth Losses

We start with the smooth case. Combining iterative localization techniques of Feldman et al. (2020); Asi et al. (2021) with the multi-stage ISRL-DP Accelerated MB-SGD of Lowy & Razaviyayn (2023a), our proposed Algorithm 1

Algorithm 1 Localized ISRL-DP Accelerated MB-SGD

Require: Datasets $X_l \in \mathcal{X}^n$ for $l \in [N]$, loss function f, constraint set $\mathcal{W}$, initial point w_0, subroutine parameters $\{R_i\}_{i=1}^{\lfloor \log_2 n \rfloor} \subset \mathbb{N}$, $\{K_i\}_{i=1}^{\lfloor \log_2 n \rfloor} \subset [n]$.

1: Set $\tau = \lfloor \log_2 n \rfloor$.
2: Set $p = \max(\frac{1}{2}\log_n(M) + 1, 3)$
3: **for** $i = 1$ **to** τ **do**
4: Set $\lambda_i = \lambda \cdot 2^{(i-1)p}$, $n_i = \lfloor n/2^i \rfloor$, $D_i = 2L/\lambda_i$.
5: Each silo $l \in [N]$ draws disjoint batch $B_{i,l}$ of n_i samples from X_l.
6: Let $\hat{F}_i(w) = \frac{1}{n_i N}\sum_{l=1}^{N}\sum_{x_{l,j}\in B_{i,l}} f(w; x_{l,j}) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$.
7: Call the multi-stage (ε, δ)-ISRL-DP implementation of Algorithm 2 with loss function $\hat{F}_i(w)$, data $X_l = B_{i,l}$, $R = R_i$, $K = K_i$, initialization w_{i-1}, constraint set $\mathcal{W}_i = \{w \in \mathcal{W} : \|w - w_{i-1}\| \le D_i\}$, and $\mu = \lambda_i$. Denote the output by w_i.
8: **end for**
9: **return** the last iterate w_τ

achieves optimal excess risk and state-of-the-art communication complexity for heterogeneous FL under ISRL-DP.

We recall ISRL-DP Accelerated MB-SGD in Algorithm 2. It is a distributed, ISRL-DP version of the AC-SA algorithm (Ghadimi & Lan, 2012). For strongly convex losses, the multi-stage implementation of ISRL-DP Accelerated MB-SGD, given in Algorithm 5 in Appendix B, offers improved excess risk (Lowy & Razaviyayn, 2023a). Algorithm 5 is a distributed, ISRL-DP version of the multi-stage AC-SA (Ghadimi & Lan, 2013).

Building on Algorithm 5, we describe our algorithm as follows (see Algorithm 1 for pseudocode). The distributed learning process is divided into $\tau = \lfloor \log_2 n \rfloor$ phases. In each phase $i \in [\tau]$, all silos work together (via communication with the central server) to iteratively solve a regularized ERM problem with ISRL-DP. The regularized ERM problem in phase i is defined over N local batches of data, each containing n_i *disjoint* samples (cf. $\hat{F}_i(w)$ in Line 6). To find the approximate constrained minimizer of $\hat{F}_i(w)$ privately, we apply Algorithm 5 for a careful choice of the number of rounds R_i and the batch size $K_i \in [n_i]$ (cf. Line 7). The output w_i of phase i affects phase $i+1$ in three ways: (i) regularization center used to define $\hat{F}_{i+1}$; (ii) initialization for the next call of Algorithm 5; and (iii) constraint set $\mathcal{W}_{i+1}$. We enforce *stability* (hence generalization) of our algorithm via regularization and localization: e.g., as i increases, we increase the regularization parameter λ_i to prevent w_{i+1} from moving too far away from w_i and w^*.

The following theorem captures our main results of this section: Algorithm 1 can achieve the optimal excess risk,

Algorithm 2 Accelerated ISRL-DP MB-SGD (Lowy & Razaviyayn, 2023a)

Require: Datasets $X_l \in \mathcal{X}^n$ for $l \in [N]$, loss function $\hat{F}(w) = \frac{1}{nN}\sum_{l=1}^{N}\sum_{x\in X_l} f(w, x)$, constraint set $\mathcal{W}$, initial point w_0, strong convexity modulus $\mu \ge 0$, privacy parameteres (ε, δ), iteration count $R \in \mathbb{N}$, batch size $K \in [n]$, step size parameters $\{\eta_r\}_{r\in[R]}, \{\alpha_r\}_{r\in[R]}$ specified in Appendix B.

Initialize $w_0^{ag} = w_0 \in \mathcal{W}$ and $r = 1$.
for $r \in [R]$ **do**
Server updates and broadcasts $w_r^{md} = \frac{(1-\alpha_r)(\mu+\eta_r)}{\eta_r + (1-\alpha_r^2)\mu} w_{r-1}^{ag} + \frac{\alpha_r[(1-\alpha_r)\mu + \eta_r]}{\eta_r + (1-\alpha_r^2)\mu} w_{r-1}$
for $i \in S_r$ **in parallel do**
Silo i draws $\{x_{i,j}^r\}_{j=1}^K$ from X_i (with replacement) and privacy noise $u_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ for proper σ^2.
Silo i computes $\widetilde{g}_r^i := \frac{1}{K}\sum_{j=1}^{K} \nabla f(w_r^{md}, x_{i,j}^r) + u_i$.
end for
Server aggregates $\widetilde{g}_r := \frac{1}{M}\sum_{i\in S_r} \widetilde{g}_r^i$ and updates:
$w_r := \arg\min_{w\in\mathcal{W}} \{\alpha_r [\langle \widetilde{g}_r, w\rangle + \frac{\mu}{2}\|w_r^{md} - w\|^2] + [(1-\alpha_r)\frac{\mu}{2} + \frac{\eta_r}{2}] \|w_{r-1} - w\|^2\}$.
Server updates and broadcasts $w_r^{ag} = \alpha_r w_r + (1-\alpha_r) w_{r-1}^{ag}$.
end for
return: w_R^{ag}.

regardless of the heterogeneity, in a communication-efficient and gradient-efficient manner.

Theorem 2.1 (Upper Bound for Smooth Losses)**.** *Let $f(\cdot, x)$ be β-smooth and $M = N$. Assume $\varepsilon \le 2\ln(2/\delta), \delta \in (0, 1)$. Then, there exist parameter choices such that Algorithm 1 is (ϵ, δ)-ISRL-DP and has the following excess risk*

$$\mathbb{E}F(w_\tau) - F(w^*) = \widetilde{O}\left(\frac{LD}{\sqrt{N}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right). \quad (9)$$

Moreover, the communication complexity of Algorithm 1 is

$$\widetilde{O}\left(\frac{\sqrt{\beta D} N^{1/4}}{\sqrt{L}}\left(\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}\right)^{1/2} + 1\right).$$

Assuming $d = \Theta(n)$ and $\varepsilon = \Theta(1)$, the gradient complexity of Algorithm 1 is

$$\widetilde{O}\left(N^{5/4} n^{1/4} (\beta D/L)^{1/2} + Nn + (Nn)^{9/8} (\beta D/L)^{1/4}\right).$$

For general d, n, ε, the gradient complexity expression is complicated, and is given in the Appendix C.1.

Remark 2.2 (Optimal risk in non-i.i.d private FL). The excess risk bound in (9) matches the optimal *i.i.d* risk bound

up to log factors (cf. Theorem 2.2 in Lowy & Razaviyayn (2023a)), *even when silo data is arbitrarily heterogeneous across silos*. To the best of our knowledge, our algorithm is the first to have this property, resolving an open question of Lowy & Razaviyayn (2023a). The prior state-of-the-art bound in Lowy & Razaviyayn (2023a, Theorem 3.1) is suboptimal by a factor of $\widetilde{O}((\sqrt{d}/(\epsilon n\sqrt{N}))^{1/5})$.

Remark 2.3 (Improved communication and gradient complexity). The communication and gradient complexities of our Algorithm 1 significantly improve over the previous state-of-the-art for ISRL-DP FL: recall (3), (4), (6), and (7).

Our algorithm has state-of-the-art communication complexity, *even in the simple case of $N = 1$, where ISRL-DP FL reduces to central DP SCO*. In fact, the prior state-of-the-art iteration complexity bound for DP SCO was $O((\beta D/L)\min\{\sqrt{n}, \varepsilon n/\sqrt{d\ln(1/\delta)}\})$ (Zhang et al., 2022). By comparison, when $N = 1$, our algorithm's communication complexity is the square root of this bound. Note that when $N = 1$, our algorithm is essentially the same as the algorithm of Kulkarni et al. (2021), but we do not incorporate convolution smoothing here since we are assuming smoothness of the loss.

We can also compare our *gradient complexity* results against the state-of-the-art central DP SCO algorithms when $N = 1$. As an illustration, consider the interesting regime $\varepsilon = \Theta(1)$ and $d = \Theta(n)$. For smooth and convex losses, when (i) $\beta \leq O(\sqrt{n}L/D)$, algorithms in both Feldman et al. (2020) and Zhang et al. (2022) achieve optimal risk using $\widetilde{O}(n)$ gradient complexity. For (ii) $\beta \geq \Omega(\sqrt{n}L/D)$, the two algorithms proposed by Feldman et al. (2020) fail to guarantee optimal risk, whereas (Zhang et al., 2022, Algorithm 2) continues to attain optimal risk with gradient complexity $\widetilde{O}(n^{3/4}\sqrt{\beta D/L})$. By comparison with these results, for case (i), our Algorithm 1 achieves optimal risk with gradient complexity $\widetilde{O}(n^{5/4})$. For case (ii), as in (Zhang et al., 2022), our Algorithm 1 continues to achieve optimal risk with gradient complexity $\widetilde{O}(n^{9/8}(\beta D/L)^{1/4} + n^{1/4}(\beta D/L)^{1/2} + n)$. Thus, our algorithm is faster than the state-of-the-art result of Zhang et al. (2022) when $\beta D/L \geq n^{3/2}$. In the complementary parameter regime, however, the algorithm of Zhang et al. (2022) (which is not ISRL-DP) is faster. We discuss possible ways to close this gap in Section 5.

Comparison with Non-Private Communication Complexity Lower Bound. Theorem 2.1 trivially extends to the unconstrained optimization setting in which $D = \|w_0 - w^*\|$ for $w^* \in \arg\min_{w\in\mathbb{R}^d} F(w)$. Moreover, our excess risk bound is still optimal for the unconstrained case: a matching lower bound is obtained by combining the technique for unconstrained DP lower bounds in Liu & Lu (2021) with the constrained ISRL-DP FL lower bounds of Lowy & Razaviyayn (2023a).

Let us compare our communication complexity upper bound against the non-private communication complexity lower bound of Woodworth et al. (2020). Define the following parameter which describes the heterogeneity of the silos at the optimum $w^* = \arg\min_{w\in\mathbb{R}^d} F(w)$:

$$\zeta_*^2 = \frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(w^*)\|^2.$$

The lower bound holds for the class of *distributed zero-respecting* algorithms (defined in Appendix F), which includes most *non-private* FL algorithms, such as MB-SGD, Accelerated MB-SGD, local SGD/FedAvg, and so on.

Theorem 2.4 (Communication Lower Bound (Woodworth et al., 2020)). *Fix $M = N$ and suppose $\mathcal{A}$ is a distributed zero-respecting algorithm with excess risk*

$$\mathbb{E}F(\mathcal{A}(X)) - F^* \lesssim \frac{LD}{\sqrt{N}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)$$

in $\leq R$ rounds of communications on any β-smooth FL problem with heterogeneity ζ_* and $\|w_0 - w^*\| \leq D$. Then,

$$R \gtrsim N^{1/4}\left(\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}\right)^{1/2} \times \min\left(\frac{\sqrt{\beta D}}{\sqrt{L}}, \frac{\zeta_*}{\sqrt{\beta L D}}\right).$$

Remark 2.5. Our communication complexity in Theorem 2.1 matches the lower bound on the communication cost in Theorem 2.4 when $\zeta_* \gtrsim \beta D$ (i.e., high heterogeneity).

There are several reasons why we cannot quite assert that Theorem 2.4 implies that Algorithm 1 is *communication-optimal*. First, the lower bound does not hold for randomized algorithms that are not zero-respecting (e.g. our ISRL-DP algorithms). However, Woodworth et al. (2020) note that the lower bound should be extendable to all randomized algorithms by using the random rotations techniques of Woodworth & Srebro (2016); Carmon et al. (2020). Second, the lower bound construction of Woodworth et al. (2020) is not L-Lipschitz. However, we believe that their quadratic construction can be approximated by a Lipschitz function (e.g. by using an appropriate Huber function). Third, as is standard in non-private complexity lower bounds, the construction requires the dimension d to grow with R. This third issue seems challenging to overcome, and may require a fundamentally different proof approach. Thus, a rigorous proof of a communication complexity lower bound for ISRL-DP algorithms and Lipschitz functions is an interesting topic for future work.

Sketch of the Proof of Theorem 2.1. We end this section with a high-level overview of the key steps needed to establish Theorem 2.1.

Proof sketch. **Privacy:** We choose parameters so that each call to Algorithm 5 is (ε, δ)-ISRL-DP. Then, the full Algorithm 1 is (ε, δ)-ISRL-DP by parallel composition, since silo l samples *disjoint* local batches across phases ($\forall l \in [N]$).

Excess risk: Following Feldman et al. (2020); Asi et al. (2021), we start with the following error decomposition: Let $\hat{w}_0 = w^*$ for analysis only, and write

$$\mathbb{E}[F(w_\tau)] - F(w^*) = \mathbb{E}[F(w_\tau) - F(\hat{w}_\tau)] \tag{10}$$

$$+ \sum_{i=1}^{\tau} \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})], \tag{11}$$

where $\hat{w}_i = \arg\min_{w \in \mathcal{W}} \hat{F}_i(w)$ for $i \in [\tau]$. Then, it remains to bound (10) and (11), respectively. To this end, we first show that w_i in Algorithm 1 is close to $\hat{w}_i$ for all $i \in [\tau]$. Here, we mainly leverage the excess empirical risk bound of Algorithm 5 for ERM with strongly convex and smooth loss functions. In particular, by the empirical risk bound in Lowy & Razaviyayn (2023a, Theorem F.1) and λ_i-strongly convex of $\hat{F}_i$, we can show that the following bound holds for all $i \in [\tau]$:

$$\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \leq \widetilde{O}\left(\frac{L^2}{\lambda_i^2 N} \cdot \frac{d \log(1/\delta)}{n_i^2 \varepsilon^2}\right). \tag{12}$$

We bound (10) via (12), Lipschitzness, and Jensen's Inequality.

To bound (11), we first leverage a key observation that *stability and generalization of the empirical minimizer of strongly-convex loss does not require homogeneous data*. This observation that allows us to handle the non-i.i.d. case optimally. Specifically, by the stability result, we have

$$\mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] \lesssim \frac{\lambda_i \mathbb{E}[\|\hat{w}_{i-1} - w_{i-1}\|^2]}{2} + \frac{L^2}{\lambda_i n_i M}.$$

We obtain a bounds on (11) by combining the above inequality with (12). Finally, by putting everything together and leveraging the geometrical schedule of λ_i, n_i, we obtain the final excess population risk bound. □

An alternative proof approach would be to try to show that Algorithm 5 is stable *directly*, e.g. by using the tools that Hardt et al. (2016) use for proving stability of SGD. However, this approach seems unlikely to yield the same tight bounds that we obtain, since acceleration may impede stability of the iterates. Instead, we establish our stability and generalization guarantee *indirectly*: We use the facts that reguralized ERM is stable and that Algorithm 5 approximates regularized ERM.

3. Error-Optimal Heterogeneous ISRL-DP FL for Non-Smooth Losses

In this section, we turn to the case of non-smooth loss functions. We modify Algorithm 1 to obtain two algorithms that can handle the non-smooth case. Our algorithms are the first to achieve the optimal excess risk for heterogeneous ISRL-DP FL with non-smooth loss functions.

3.1. Error-Optimal and Communication-Efficient ISRL-DP FL for Non-Smooth Losses

Our first algorithm is based on the technique of Nesterov smoothing (Nesterov, 2005): For a non-smooth function f, Moreau-Yosida regularization is used to approximate f by the β-smooth β-Moreau envelope

$$f_\beta(w) := \min_{v \in \mathcal{W}} \left(f(v) + \frac{\beta}{2}\|w - v\|^2 \right).$$

We then optimize this smooth function using our Algorithm 1. See Appendix E for details.

Theorem 3.1 (Non-smooth FL via smoothing)**.** *Let $M = N$. Then the combination of Algorithm 1 with Nesterov smoothing yields an (ε, δ)-ISRL-DP algorithm with optimal excess population risk as in (9). The communication complexity is*

$$\widetilde{O}\left(\sqrt{N} \min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d \ln(1/\delta)}}\right\} + 1\right).$$

Remark 3.2 (Optimal excess risk in non-i.i.d. private FL). Theorem 3.1 gives the first optimal rate for non-smooth heterogeneous FL under ISRL-DP. Note that the current best result for heterogeneous FL in Lowy & Razaviyayn (2023a) is suboptimal and holds only for the *smooth* case. One can combine the same smoothing technique above with the algorithm in Lowy & Razaviyayn (2023a) to obtain a risk bound of $O(1/\sqrt{nN} + (\sqrt{d \ln(1/\delta)}/\varepsilon n\sqrt{N})^{2/3})$ for the non-smooth case, which is again suboptimal.

Remark 3.3 (Improved communication complexity). The communication complexity bound in Theorem 3.1 improves over the previous state-of-the-art result for an algorithm achieving optimal excess risk (Lowy & Razaviyayn, 2023a); recall (3). Moreover, the result of Lowy & Razaviyayn (2023a) assumed i.i.d. silo data, whereas our result holds for the non-i.i.d. case.

We have omitted the gradient complexity for this smoothing approach. This is mainly because the gradient computation of f_β needs additional computation in the form of the prox operator. One may consider using an *approximate* prox operator to get a handle on the total gradient complexity as in Bassily et al. (2019). However, this approach would still introduce an additional $\Theta(n^3)$ gradient complexity per step. Thus, a natural question is whether we can design a more *computation-efficient* algorithm for the non-smooth case — providing a motivation for our next algorithm.

Algorithm 3 Noisy ISRL-DP MB-Subgradient Method

Require: Datasets $X_l \in \mathcal{X}^n$ for $l \in [N]$, loss function $\hat{F}(w) = \frac{1}{nN}\sum_{l=1}^N \sum_{x\in X_l} f(w,x)$, constraint set $\mathcal{W}$, initial point w_0, privacy parameters (ε, δ), iteration count $R \in \mathbb{N}$, batch size $K \in [n]$, step sizes $\{\gamma_r\}_{r=0}^{R-1}$, initial point $w_0 \in \mathcal{W}$.

1: **for** $r \in \{0, 1, \ldots, R-1\}$ **do**
2: **for** $l \in S_r$ **in parallel do**
3: Server sends global model w_r to silo l.
4: Silo l draws K samples $x_{l,j}^r$ uniformly from X_i (for $j \in [K]$) and noise $u_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ for proper σ^2.
5: Silo l computes $\widetilde{g}_r^l := \frac{1}{K}\sum_{j=1}^K g_{r,j}^l + u_i$ and sends to server, where $g_{r,j}^l \in \partial f(w_r, x_{l,j}^r)$ (subgradient).
6: **end for**
7: Server aggregates $\widetilde{g}_r := \frac{1}{M_r}\sum_{l\in S_r} \widetilde{g}_r^l$.
8: Server updates $w_{r+1} := \Pi_{\mathcal{W}}[w_r - \gamma_r \widetilde{g}_r]$.
9: **end for**
10: **Output:** $\bar{w}_R = \frac{2}{R(R+1)}\sum_{r=1}^R r w_r$.

Algorithm 4 Localized ISRL-DP MB-Subgradient Method

Require: Dataset $X_l \in \mathcal{X}^n$, $l \in [N]$, constraint set $\mathcal{W}$, $\eta > 0$, subroutine parameters (specified in Appendix) including batch size K_i, number of rounds R_i, noise parameters σ_i.

1: Choose any $w_0 \in \mathcal{W}$.
2: Set $\tau = \lfloor \log_2 n \rfloor$, $p = \max(\frac{1}{2}\log_n(M) + 1, 3)$.
3: **for** $i = 1$ **to** τ **do**
4: Set $\eta_i = \eta/2^{ip}$, $n_i = n/2^i$, $\lambda_i = 1/(\eta_i n_i)$, $D_i = 2L/\lambda_i$.
5: Each silo $l \in [N]$ draws disjoint batch $B_{i,l}$ of n_i samples from X_l.
6: Let $\hat{F}_i(w) = \frac{1}{n_i N}\sum_{l=1}^N \sum_{x_{l,j}\in B_{i,l}} f(w; x_{l,j}) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$.
7: Call the (ε, δ)-ISRL-DP Algorithm 3 with loss function $\hat{F}_i(w)$, data $X_l = B_{i,l}$, $R = R_i$, $K = K_i$, step sizes $\gamma_r = \frac{2}{\lambda_i(r+1)}$ for $r = 0, 1, \ldots, R_i - 1$, initialization w_{i-1}, and constraint set $\mathcal{W}_i = \{w \in \mathcal{W} : \|w - w_{i-1}\| \le D_i\}$. Let w_i denote the output.
8: **end for**
9: **return** the last iterate w_τ.

3.2. Error-Optimal and Computationally Efficient ISRL-DP FL for Non-Smooth Losses

We propose another variation of Algorithm 1 that uses *subgradients* to handle the non-smooth case in a computationally efficient way: see Algorithm 4. Algorithm 4 follows the same structure as Algorithm 1: we iteratively solve a carefully chosen sequence of regularized ERM problems with a ISRL-DP solver and use localization. Compared to Algorithm 1 for the smooth case, Algorithm 4 does not use an accelerated solver (due to non-smoothness). Instead, we use *ISRL-DP Minibatch Subgradient* method (Algorithm 3) to solve the non-smooth strongly convex ERM problem in each phase of Algorithm 4. There are two key differences between our subroutine Algorithm 3 and the ISRL-DP MB-SGD of Lowy & Razaviyayn (2023a): (i) Instead of the gradient, a subgradient of the non-smooth objective is used in Line 5; (ii) A different and simpler averaging step in Line 10 is used for strongly convex non-smooth losses.

Theorem 3.4 (Non-smooth FL via subgradient)**.** *Let $M = N$. Then, there exist parameter choices such that Algorithm 4 is (ε, δ)-ISRL-DP and achieves the optimal excess population risk in (9). The communication complexity is*

$$\widetilde{O}\left(\min\left(nN, \frac{N\varepsilon^2 n^2}{d}\right) + 1\right).$$

Assuming $\varepsilon = \Theta(1)$, the subgradient complexity is

$$\widetilde{O}\left(Nn + N^2 \min\left(n, \frac{n^2}{d}\right) + N^{3/2}\min\left(n^{3/2}, \frac{n^2}{\sqrt{d}}\right)\right).$$

Remark 3.5 (Improved gradient complexity). The above subgradient complexity improves over the state-of-the-art gradient complexity (Lowy & Razaviyayn, 2023a) for an ISRL-DP FL algorithm with optimal excess risk. Lowy & Razaviyayn (2023a) apply Nesterov smoothing to ISRL-DP MB-SGD. As discussed earlier, implementing the smoothing approach is computationally costly. Moreover, the results in Lowy & Razaviyayn (2023a) assume i.i.d. silo data. The subgradient complexity result in Theorem 3.4 also improves over the gradient complexity of Algorithm 1 with smoothing.

An alternative approach is to use the *convolutional smoothing* technique (Kulkarni et al., 2021) to optimize non-smooth functions. For certain regimes, this approach improves the gradient complexity over the subgradient method, with a worse communication complexity compared to the Nesterov smoothing approach. See Appendix D for details.

Depending on the FL application, communication or computation may be more of a bottleneck. If communication efficiency is more of a priority, the smoothed version of Algorithm 1 should be used. On the other hand, if computational efficiency is more pressing, Algorithm 4 or Algorithm 6 is recommended.

4. Numerical Experiments

We validate our theoretical findings with numerical experiments on MNIST data. As shown in Figures 2 and 3, *our algorithm consistently outperforms Lowy et al. (2023a)*. We use a similar experimental setup to Lowy et al. (2023a), as outlined below.

Task/model/data set: We run binary logistic regression with heterogeneous MNIST data: each of the $N = 25$ silos contains data corresponding to one odd digit class and one even digit class (e.g. (1, 0), (3, 2), etc.) and the goal is to classify each digit as odd or even. We randomly sample roughly $1/5$ of MNIST data to expedite the experiments. We borrow the code from Woodworth et al. (2020) to transform and preprocess MNIST data.

Preprocessing: We preprocess MNIST data, flatten the images, and reduce the dimension to $d = 50$ using PCA. We use an 80/20 train/test split, yielding a total of $n = 1734$ training samples.

Our algorithm: Localized ISRL-DP MB-SGD, which is a practical (non-accelerated) variant of our Algorithm 1: For simplicity and to expedite parameter search, we use vanilla MB-SGD in place of accelerated MB-SGD as our regularized ERM subsolver in our implementation of Algorithm 1.

Baseline: We compare our algorithm against the *One-pass ISRL-DP MB-SGD* of Lowy et al. (2023a). Recall that Lowy et al. (2023a) did not provide theoretical guarantees for their multi-pass ISRL-DP MB-SGD with heterogeneous silos.

Hyperparameter tuning and evaluation: We evaluate the algorithms across a range of privacy parameters ε and fix $\delta = 1/n^2$. For each algorithm and each setting of ε, we search a range of step sizes η.

Simulating unreliable communication: In addition to the reliable communication setting where all silos communicate in each round $M = N = 25$, we simulate unreliable communication by randomly selecting a subset of silos to communicate in each round. In each communication round, $M = 18$ of the $N = 25$ silos are chosen uniformly at random to communicate with the server.

Evaluation: Each evaluation consists of 5 trials, each with different data due to random sampling. In each trial, for each parameter setting, we repeat 3 runs and choose the hyperparameters with the lowest average loss, and record the average test error as the test error. We plot the average test error and the standard deviation across trials.

As shown in the plots, *in both reliable and unreliable communication settings, our localized ISRL-DP MB-SGD algorithm outperforms the baseline one-pass ISRL-DP MB-SGD algorithm across all privacy parameters*. Despite being an algorithm designed to achieve theoretical guarantees, our algorithm evidently performs well in practice.

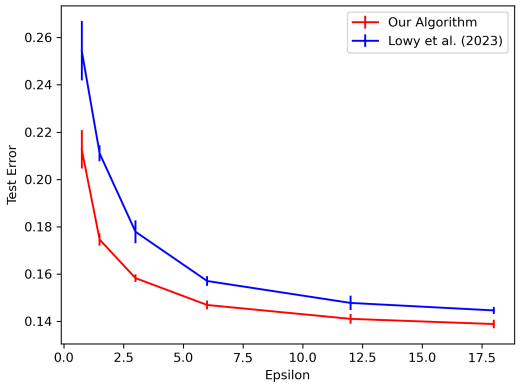

Figure 2. Reliable Communication

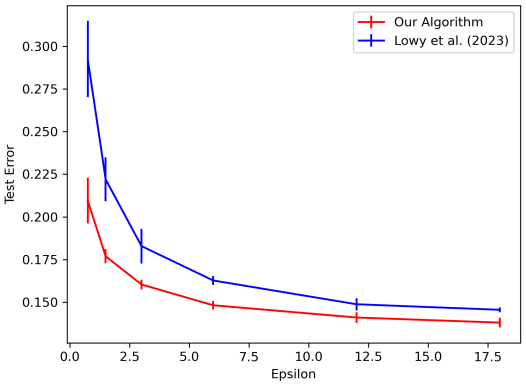

Figure 3. Unreliable Communication

5. Concluding Remarks and Open Questions

We have studied private federated learning in the absence of a trusted server. We characterized the minimax optimal excess risk bounds for heterogeneous ISRL-DP FL, answering an open question posed by Lowy & Razaviyayn (2023a). Further, our algorithms advanced the state-of-the-art in terms of communication and computational efficiency. For smooth losses, the communication complexity of our optimal algorithm matches the non-private lower bound.

To conclude, we discuss some open problems that arise from our work. (1) A rigorous proof of a ISRL-DP communication complexity lower bound. (2) Is there an optimal ISRL-DP algorithm with $O(nN)$ gradient complexity? A promising approach may be to combine Algorithm 1 with ISRL-DP variance-reduction. (Note that the gradient-efficient variance-reduced central DP algorithm of Zhang et al. (2022) uses output perturbation, which requires a trusted server.) (3) Is it possible to achieve both optimal communication complexity and optimal gradient complexity simultaneously with a single algorithm? Even in the simpler centralized setting ($N = 1$), this question is open. (4) What are the optimal rates for ISRL-DP FL problems beyond convex and uniformly Lipschitz loss functions?

Acknowledgements

XZ is supported in part by NSF CNS-2153220 and CNS-2312835. CG, AL, and SW supported in part by NSF CCF-2023239 and CCF-2224213 and AFOSR FA9550-21-1-0084.

Impact Statement

In this research, we advance the field of federated learning (FL) by introducing algorithms that enhance privacy and communication efficiency in settings where trust in a central server is not assumed. Our work has significant implications for industries handling sensitive data, such as healthcare and finance, by offering an improved method for leveraging collective data while safeguarding individual privacy. However, the increased algorithmic complexity of these algorithms could limit their accessibility, especially for organizations with limited resources. Additionally, while our approach reduces the risk of data leakage, it does not entirely eliminate the possibility of misuse or unintended consequences, such as reinforcing existing biases in data. Future research should focus on refining these algorithms to make them more accessible and addressing potential ethical implications. By doing so, we aim to contribute positively to the development of safe and equitable data handling practices in various sectors of society and the economy.

References

Apple. Private federated learning. *NeurIPS 2019 Expo Talk Abstract*, 2019. URL https://nips.cc/ExpoConferences/2019/schedule?talk_id=40.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Bassily, R., Guzmán, C., and Nandi, A. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pp. 474–499. PMLR, 2021.

Boob, D. and Guzmán, C. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming*, pp. 1–43, 2023.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, 2013. doi: 10.1109/FOCS.2013.53.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

FedAI. Webank and swiss re signed cooperation mou. *Fed AI Ecosystem*, 2019. URL https://www.fedai.org/news/webank-and-swiss-re-signed-cooperation-mou/.

Feldman, V., Koren, T., and Talwar, K. Private Stochastic Convex Optimization: Optimal Rates in Linear Time. *arXiv:2005.04763 [cs, math, stat]*, May 2020.

Gao, C. and Wright, S. J. Differentially private optimization for smooth nonconvex erm. *arXiv preprint arXiv:2302.04972*, 2023.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *CoRR*, abs/1712.07557, 2017. URL http://arxiv.org/abs/1712.07557.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012. doi: 10.1137/110848864. URL https://doi.org/10.1137/110848864.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23

(4):2061–2089, 2013. doi: 10.1137/110848876. URL https://doi.org/10.1137/110848876.

Gupta, S., Huang, Y., Zhong, Z., Gao, T., Li, K., and Chen, D. Recovering private text in federated learning of language models. *Advances in Neural Information Processing Systems*, 35:8130–8143, 2022.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Heikkilä, M. A., Koskela, A., Shimizu, K., Kaski, S., and Honkela, A. Differentially private cross-silo federated learning. *arXiv preprint arXiv:2007.05553*, 2020.

Hilmkil, A., Callh, S., Barbieri, M., Sütfeld, L. R., Zec, E. L., and Mogren, O. Scaling federated learning for fine-tuning of large language models. In *International Conference on Applications of Natural Language to Information Systems*, pp. 15–23. Springer, 2021.

Jayaraman, B. and Wang, L. Distributed learning without distress: Privacy-preserving empirical risk minimization. *Advances in Neural Information Processing Systems*, 2018.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Kulkarni, J., Lee, Y. T., and Liu, D. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

Levy, D., Sun, Z., Amin, K., Kale, S., Kulesza, A., Mohri, M., and Suresh, A. T. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34: 12466–12479, 2021.

Liu, D. and Lu, Z. Lower bounds for differentially private erm: Unconstrained and non-euclidean. *arXiv preprint arXiv:2105.13637*, 2021.

Liu, Z., Hu, S., Wu, Z. S., and Smith, V. On privacy and personalization in cross-silo federated learning. *arXiv preprint arXiv:2206.07902*, 2022.

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*, 2023a.

Lowy, A. and Razaviyayn, M. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *International Conference on Algorithmic Learning Theory*, pp. 986–1054. PMLR, 2023b.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, pp. 5749–5786. PMLR, 2023a.

Lowy, A., Gupta, D., and Razaviyayn, M. Stochastic differentially private and fair learning. In *The Eleventh International Conference on Learning Representations*, 2023b.

Lowy, A., Li, Z., Huang, T., and Razaviyayn, M. Optimal differentially private learning with public data. *arXiv preprint arXiv:2306.15056*, 2023c.

Lowy, A., Ullman, J., and Wright, S. J. How to make the gradients small privately: Improved rates for differentially private non-convex optimization. *arXiv preprint arXiv:2402.11173*, 2024.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

McMahan, B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/pdf?id=BJ0hF1Z0b.

McSherry, F. D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30, 2009.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10110–10145. PMLR, 2022.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *COLT*, pp. 5, 2009.

Woodworth, B. E. and Srebro, N. Tight complexity bounds for optimizing composite objectives. *Advances in neural information processing systems*, 29, 2016.

Woodworth, B. E., Patel, K. K., and Srebro, N. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020.

Wu, Y., Zhou, X., Tao, Y., and Wang, D. On private and robust bandits. *arXiv preprint arXiv:2302.02526*, 2023.

Zhang, L., Thekumparampil, K. K., Oh, S., and He, N. Bring Your Own Algorithm for Optimal Differentially Private Stochastic Minimax Optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187, December 2022.

Zhou, X. and Chowdhury, S. R. On differentially private federated linear contextual bandits. *arXiv preprint arXiv:2302.13945*, 2023.

Zhu, L. and Han, S. Deep leakage from gradients. In *Federated learning*, pp. 17–31. Springer, 2020.

A. Further Discussion of Related Work

DP Optimization. There is a large and growing body of work on DP optimization. Most of this work focuses on the centralized setting, with Lipschitz convex loss functions in ℓ_2 geometry (Bassily et al., 2014; 2019; Feldman et al., 2020; Zhang et al., 2022) and ℓ_p geometry (Asi et al., 2021; Bassily et al., 2021). Recently, we have started to learn more about other central DP optimization settings, such as DP optimization with non-uniformly Lipschitz loss functions/heavy-tailed data (Lowy & Razaviyayn, 2023b), non-convex loss functions (Gao & Wright, 2023; Lowy et al., 2024), and min-max games (Boob & Guzmán, 2023). There has also been work on the interactions between DP and other ethical desiderata, like fairness (Lowy et al., 2023b) and robustness (Wu et al., 2023), as well as DP optimization with side access to public data (Lowy et al., 2023c). Despite this progress, much less is known about DP distributed optimization/federated learning, particularly in the absence of a trustworthy server.

DP Federated Learning. There have been many works attempting to ensure privacy of people's data during the federated learning (FL) process. Some of these works have utilized *user-level differential privacy* (McMahan et al., 2018; Geyer et al., 2017; Levy et al., 2021), which can be practical for cross-device FL with a trusted server. Several works have also considered inter-silo record-level DP (ISRL-DP) or similar notions to ensure privacy without a trusted server (Heikkilä et al., 2020; Liu et al., 2022; Lowy & Razaviyayn, 2023a; Lowy et al., 2023a; Zhou & Chowdhury, 2023).

The state-of-the-art theoretical bounds for convex ISRL-DP FL are due to Lowy & Razaviyayn (2023a). Lowy & Razaviyayn (2023a) gave minimax error-optimal algorithms and lower bounds for the i.i.d. setting, and suboptimal algorithms for the heterogeneous setting. We close this gap by providing optimal algorithms for the heterogeneous setting. Additionally, we improve over the communication complexity and gradient complexity bounds in Lowy & Razaviyayn (2023a).

B. Multi-Stage Implementation of ISRL-DP Accelerated MB-SGD

Algorithm 5 Multi-stage Accelerated Noisy MB-SGD (Lowy & Razaviyayn, 2023a)

Require: Inputs: Constraint set $\mathcal{W}$, L-Lipschitz and μ-strongly convex loss function $\widehat{F}$, $U \in [R]$ such that $\sum_{k=1}^{U} R^{(k)} \leq R$ for $R^{(k)}$ defined below; $w_0 \in \mathcal{W}$, $\Delta \geq \widehat{F}(w_0) - \widehat{F}^*$, and $q_0 = 0$.

1: **for** $k \in [U]$ **do**

2: $\quad R^{(k)} = \left\lceil \max\left\{ 4\sqrt{\frac{2\beta}{\mu}}, \frac{128L^2}{3\mu\Delta 2^{-(k+1)}} \right\} \right\rceil$

3: $\quad \upsilon_k = \max\left\{ 2\beta, \left[\frac{\mu V^2}{3\Delta 2^{-(k-1)} R^{(k)}(R^{(k)}+1)(R^{(k)}+2)} \right]^{1/2} \right\}$

4: $\quad \alpha_r = \frac{2}{r+1}$, $\eta_r = \frac{4\upsilon_k}{r(r+1)}$, for $r \in [R^{(k)}]$.

5: $\quad$ Call Algorithm 2 with $R = R^{(k)}$, using $w_0 = q_{k-1}$, and $\{\alpha_r\}_{r\in[R^{(k)}]}$ and $\{\eta_r\}_{r\in[R^{(k)}]}$ defined above.

6: $\quad$ Set q_k to be the output of stage k.

7: **end for**

8: **return** q_U.

We will need the following result for the excess risk bound, which is due to (Lowy & Razaviyayn, 2023a) (Theorem F.1).

Lemma B.1 (Smooth ERM Upper Bound for Algorithm 5)**.** *Assume $f(\cdot, x)$ is β-smooth and λ-strongly convex for all x. Let $\varepsilon \leq 2\ln(2/\delta), \delta \in (0,1)$. Then, there exist algorithmic parameters such that Algorithm 5 is (ε, δ)-ISRL-DP. Moreover, Algorithm 5 has the following excess empirical risk bound*

$$\mathbb{E}[\widehat{F}(q_U) - \widehat{F}^*] = \widetilde{O}\left(\frac{L^2}{\lambda} \frac{d\ln(1/\delta)}{\varepsilon^2 n^2 M} \right), \tag{13}$$

and the communication complexity is

$$R = \max\left\{ 1, \sqrt{\frac{\beta}{\lambda}} \ln\left(\frac{\Delta\lambda M\varepsilon^2 n^2}{L^2 d} \right), \mathbb{1}_{\{MK<Nn\}} \frac{\varepsilon^2 n^2}{Kd\ln(1/\delta)} \right\}. \tag{14}$$

C. Precise Statement and Proof of Theorem 2.1

Theorem C.1 (Precise Statement of Theorem 2.1)**.** *Assume $f(\cdot, x)$ is β-smooth for all x. Let $\varepsilon \leq 2\ln(2/\delta)$, $\delta \in (0,1)$. Choose $R_i \approx \max\left(\sqrt{\frac{\beta+\lambda_i}{\lambda_i}}\ln\left(\frac{\Delta_i \lambda_i M \varepsilon^2 n_i^2}{L^2 d}\right), \mathbb{1}_{\{MK_i<Nn_i\}}\frac{\varepsilon^2 n_i^2}{K_i d \ln(1/\delta)}\right)$, where $LD \geq \Delta_i \geq \hat{F}_i(w_{i-1}) - \hat{F}_i(\hat{w}_i)$, $K_i \geq \frac{\varepsilon n_i}{4\sqrt{2R_i\ln(2/\delta)}}$, $\sigma_i^2 = \frac{256L^2R_i\ln(\frac{2.5R_i}{\delta})\ln(2/\delta)}{n_i^2\varepsilon^2}$, and*

$$\lambda = \frac{L}{Dn\sqrt{M}}\max\left\{\sqrt{n}, \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon}\right\}. \tag{15}$$

Then, the output of Algorithm 1 is (ε,δ)-ISRL-DP and achieves the following excess risk bound:

$$\mathbb{E}F(w_\tau) - F(w^*) = \widetilde{O}\left(\frac{LD}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right). \tag{16}$$

The communication complexity is

$$\widetilde{O}\left(\max\left\{1, \frac{\sqrt{\beta D}M^{1/4}}{\sqrt{L}}\left(\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}\right)^{1/2}, \mathbb{1}_{\{M<N\}}\frac{\varepsilon^2 n}{d\ln(1/\delta)}\right\}\right), \tag{17}$$

when $K_i = n_i$. If $d = \Theta(n)$, $M = N$, and $\varepsilon = \Theta(1)$, then the gradient complexity is

$$\widetilde{O}\left(N^{5/4}n^{1/4}(\beta D/L)^{1/2} + Nn + (Nn)^{9/8}(\beta D/L)^{1/4}\right).$$

Remark C.2. In Theorem 2.1, we state the results only for the case $M = N$. Here, we do not assume $M = N$, and present the complete results. The complete analysis of gradient complexity for other regimes can be found in Appendix C.1.

Before proving the theorem, we need several lemmas.

1. We first show the following result.

Lemma C.3. *Let $\hat{w}_i = \arg\min_{w\in\mathcal{W}}\hat{F}_i(w)$. We have $\hat{w}_i \in \mathcal{W}_i$ and $\hat{F}_i$ is $3L$-Lipschitz, $(\beta+\lambda_i)$-smooth.*

Proof. The optimality of $\hat{w}_i$ implies that

$$\frac{1}{n_i M}\sum_{l=1}^{M}\sum_{j=1}^{n_i} f(\hat{w}_i; x_{l,j}) + \frac{\lambda_i}{2}\left\|\hat{w}_i - w_{i-1}\right\|^2 \leq \frac{1}{n_i M}\sum_{l=1}^{M}\sum_{j=1}^{n_i} f(w_{i-1}; x_{l,j}) + 0.$$

By rearranging and using the L-Lipschitzness of $f(\cdot, x)$, We obtain

$$\frac{\lambda_i}{2}\left\|\hat{w}_i - w_{i-1}\right\|^2 \leq L\left\|\hat{w}_i - w_{i-1}\right\|.$$

It follows that $\hat{w}_i \in \mathcal{W}_i = \left\{w : \|w - w_{i-1}\| \leq \frac{2L}{\lambda_i}\right\}$.

For Lipschitzness, the norm of the derivative of the regularizer $r_i(w) = \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$ is $\lambda_i\|w - w_{i-1}\| \leq \lambda_i D_i = 2L$. The hessian of the regularizer is $\lambda_i I$.

Therefore $r_i(w)$ is $2L$-Lipschitz and λ_i-smooth. It follows that $\hat{F}_i$ is $3L$-Lipschitz and $(\beta+\lambda_i)$-smooth. □

2. We have the following bounds that relate the private solution w_i and the true solution $\hat{w}_i$ of $\hat{F}_i$.

Lemma C.4. *In each phase i, the following bounds hold:*

$$\mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)] = \widetilde{O}\left(\frac{L^2}{\lambda_i M}\cdot\frac{d\ln(1/\delta)}{\varepsilon^2 n_i^2}\right), \tag{18}$$

$$\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \leq \widetilde{O}\left(\frac{L^2}{\lambda_i^2 M}\cdot\frac{d\log(1/\delta)}{n_i^2\varepsilon^2}\right). \tag{19}$$

Proof. Applying Lemma B.1 to $\hat{F}_i$, we have

$$\mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)] = \widetilde{O}\left(\frac{L^2}{\lambda_i M} \cdot \frac{d\ln(1/\delta)}{\varepsilon^2 n_i^2}\right).$$

By using the λ_i-strong convexity, we have

$$\frac{\lambda_i}{2}\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \leq \mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)].$$

The bound (19) follows. □

3. As a consequence, we have the following bound.

Lemma C.5. *Let $w \in \mathcal{W}$. We have*

$$\mathbb{E}[F(\hat{w}_i)] - F(w) \leq \frac{\lambda_i \mathbb{E}[\|w - w_{i-1}\|^2]}{2} + \frac{4 \cdot (3L)^2}{\lambda_i n_i M}.$$

Proof. Applying the stability result in Lemma H.1 to $\hat{F}_i$ with $m = Mn_i$, which is λ_i-strongly convex and $3L$-Lipschitz, we have

$$\mathbb{E}[\hat{F}_i(\hat{w}_i) - \hat{F}_i(w)] \leq \frac{4 \cdot (3L)^2}{\lambda_i n_i M}.$$

It follows from the definition of $\hat{F}_i$ that

$$\begin{aligned}
\mathbb{E}[F(\hat{w}_i)] - F(w) &= \mathbb{E}[\hat{F}_i(\hat{w}_i)] - \frac{\lambda_i \mathbb{E}[\|\hat{w}_i - w_{i-1}\|^2]}{2} - \left(\hat{F}_i(w) - \frac{\lambda_i \mathbb{E}[\|w - w_{i-1}\|^2]}{2}\right) \\
&\leq \frac{\lambda_i \mathbb{E}[\|w - w_{i-1}\|^2]}{2} + \mathbb{E}[\hat{F}_i(\hat{w}_i) - \hat{F}_i(w)] \\
&\leq \frac{\lambda_i \mathbb{E}[\|w - w_{i-1}\|^2]}{2} + \frac{4 \cdot (3L)^2}{\lambda_i n_i M}.
\end{aligned} \tag{20}$$

□

By putting these results together, prove the theorem.

Proof of Theorem C.1. **Privacy.** By the privacy guarantee of Algorithm 5 given in Lemma B.1, each phase of the algorithm is (ε, δ)-ISRL-DP. Since the batches $\{B_{i,l}\}_{i=1}^{\tau}$ are disjoint for all $l \in [N]$, the privacy guarantee of the entire algorithm follows by parallel composition of differential privacy (McSherry, 2009).

Excess risk. Recall that we define $\hat{w}_0 = w^*$. Write

$$\mathbb{E}F(w_\tau) - F(w^*) = \mathbb{E}[F(w_\tau) - F(\hat{w}_\tau)] + \sum_{i=1}^{\tau} \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})].$$

Since $\tau = \lfloor \log_2 n \rfloor$, we have $n_\tau = \Theta(1)$ and $\lambda_\tau = \Theta(\lambda n^p)$. By (19) and Jensen's Inequality $\mathbb{E}Z \leq \sqrt{\mathbb{E}Z^2}$, we bound the first term as follows:

$$\begin{aligned}
\mathbb{E}[F(w_\tau) - F(\hat{w}_\tau)] \leq L\mathbb{E}\left[\|w_\tau - \hat{w}_\tau\|\right] &\leq \widetilde{O}\left(\frac{L^2}{\sqrt{M}\lambda_\tau n_\tau} \cdot \frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\right) \\
&\leq \widetilde{O}\left(\frac{L^2}{\lambda n^p \sqrt{M}} \cdot \frac{\sqrt{d\log(1/\delta)}}{\varepsilon}\right) \\
&\leq \widetilde{O}\left(\frac{L^2}{\frac{L}{D\sqrt{nM}} \cdot n^p \sqrt{M}}\right) \leq \widetilde{O}\left(\frac{DL}{n^{p-\frac{1}{2}}}\right),
\end{aligned}$$

where the last step is due to the choice of λ, per (15). Now recall $p = \max(\frac{1}{2}\log_n(M) + 1, 3)$. We have

$$n^{p-\frac{1}{2}} \geq \sqrt{n \cdot n^{\log_n(M)}} = \sqrt{nM}. \tag{21}$$

It follows that $\mathbb{E}[F(w_\tau) - F(\hat{w}_\tau)] \leq \widetilde{O}\left(\frac{DL}{\sqrt{nM}}\right)$.

Note that $\lambda_i n_i^2 = \Theta(\lambda n^2 \cdot 2^{(p-2)i})$ and $p \geq 3$. We know that $\lambda_i n_i^2$ and $\lambda_i n_i$ increase geometrically. By combining Lemma C.5 and (18), we obtain

$$\begin{aligned}
\sum_{i=1}^{\tau} \mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] &\leq \sum_{i=1}^{\tau} \left(\frac{\lambda_i \mathbb{E}[\|w_{i-1} - \hat{w}_{i-1}\|^2]}{2} + \frac{4 \cdot (3L)^2}{\lambda_i n_i M} \right) \\
&\leq \widetilde{O}\left(\lambda D^2 + \sum_{i=2}^{\tau} \frac{L^2}{\lambda_i n_i^2 M} \cdot \frac{d \log(1/\delta)}{\varepsilon^2} + \sum_{i=1}^{\tau} \frac{L^2}{\lambda_i n_i M} \right) \\
&\leq \widetilde{O}\left(\lambda D^2 + \frac{L^2}{\lambda n^2 M} \cdot \frac{d \log(1/\delta)}{\varepsilon^2} + \frac{L^2}{\lambda n M} \right),
\end{aligned}$$

Setting λ as per (15) gives the result.

Communication complexity. When we use the full batch in each round, that is, $K_i = n_i$, hiding logarithmic factors, communication complexity is

$$\begin{aligned}
\sum_{i=1}^{\tau} R_i &= \sum_{i=1}^{\tau} \max\left\{ 1, \sqrt{\frac{\beta + \lambda_i}{\lambda_i}} \ln\left(\frac{\Delta \lambda_i M \varepsilon^2 n_i^2}{L^2 d} \right), \mathbb{1}_{\{M<N\}} \frac{\varepsilon^2 n_i^2}{n_i d \ln(1/\delta)} \right\} \\
&= \widetilde{O}\left(\max\left\{ \lfloor \log_2 n \rfloor, \sqrt{\frac{\beta}{\lambda}}, \mathbb{1}_{\{M<N\}} \frac{\varepsilon^2 n}{d \ln(1/\delta)} \right\} \right) \\
&= \widetilde{O}\left(\max\left\{ 1, \frac{\sqrt{\beta D} M^{1/4}}{\sqrt{L}} \left(\min\left\{ \sqrt{n}, \frac{\varepsilon n}{\sqrt{d \ln(1/\delta)}} \right\} \right)^{1/2}, \mathbb{1}_{\{M<N\}} \frac{\varepsilon^2 n}{d \ln(1/\delta)} \right\} \right).
\end{aligned} \tag{22}$$

For gradient complexity, we defer the analysis to the following section. □

C.1. Gradient Complexity of Algorithm 1 under Different Parameter Regimes

Theorem C.6. *When $N = M$ and the full batch $K_i = n_i$ is used, the gradient complexity of Algorithm 1 is*

$$\begin{aligned}
\sum_{i=1}^{\tau} N n_i R_i &= \widetilde{O}\left(Nn \max\left\{ 1, \sqrt{\frac{\beta}{\lambda}} \right\} \right) \\
&= \widetilde{O}\left(Nn + \frac{n \cdot \sqrt{\beta D} N^{5/4}}{\sqrt{L}} \min\left\{ \sqrt{n}, \frac{\varepsilon n}{\sqrt{d \ln(1/\delta)}} \right\}^{1/2} \right).
\end{aligned} \tag{23}$$

When we choose $K_i < n_i$, the communication cost can be worse due to the second term of $R_i = \max\left\{1, \sqrt{\frac{\beta+\lambda_i}{\lambda_i}} \ln\left(\frac{\Delta_i \lambda_i M \varepsilon^2 n_i^2}{L^2 d}\right), \mathbb{1}_{\{MK_i < Nn_i\}} \frac{\varepsilon^2 n_i^2}{K d \ln(1/\delta)}\right\}$. However, under some regimes, the gradient complexity can be better than that of the full batch. In the case where $M < N$, the second case will always be present. Now we relax the assumption $M = N$, and discuss general results.

For simplicity, let us assume $\varepsilon = \Theta(1)$, keep terms involving ε, M, n, and d only and omit $\widetilde{O}$. We summarize the results as follows:

- *When $d \lesssim n$, there are two subcases to consider: if $d \gtrsim \frac{n^{3/4}}{M^{1/4}}$, then gradient complexity is $\min\left\{(M + M^{5/4} n^{1/4}) \max\left(\frac{n^{7/4}}{M^{1/4} d}, \frac{n^{7/8}}{M^{1/8}}, 1\right), \frac{Mn^2}{d}\right\}$. If $d \lesssim \frac{n^{3/4}}{M^{1/4}}$, the complexity is $\frac{Mn^2}{d}$.*

- *When $d \gtrsim n$, there are two subcases to consider: if $d \gtrsim \frac{n^{2/3}}{M^{1/3}}$, then gradient complexity is $\min\left(\left(M + \frac{M^{5/4}n^{1/2}}{d^{1/4}}\right)\max\left(\frac{n^{3/2}}{M^{1/4}d^{3/4}}, \frac{n^{3/4}d^{1/8}}{M^{1/8}}, 1\right), Mn\right)$. If $d \lesssim \frac{n^{2/3}}{M^{1/3}}$, the complexity is Mn.*

In particular, when $d = \Theta(n)$, the gradient complexity is $M^{5/4}n^{1/4} + (Mn)^{9/8}$. More precisely, the complexity is

$$\widetilde{O}\left(N^{5/4}n^{1/4}(\beta D/L)^{1/2} + Nn + (Nn)^{9/8}(\beta D/L)^{1/4}\right) \tag{24}$$

Proof. The bound in (23) is an immediate consequence of Theorem 2.1.

Since terms in the sum decrease geometrically, we only need to consider the first term where $i = 1$. For simplicity we drop the index i.

We write $R = \max(R', R'') + 1$, where we let $R' = Q \cdot \left(\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}\right)^{1/2}$, $R'' = \frac{\varepsilon^2 n^2}{Kd\ln(1/\delta)}$, and $Q := \sqrt{\beta D}M^{1/4}/\sqrt{L}$. Recall the batch size constraint $K \geq \frac{\varepsilon n}{4\sqrt{2R\ln(2/\delta)}}$. We analyze the complexity for (1) $d \lesssim n$, (2) $d \gtrsim n$ separately below. We omit $\widetilde{O}$ for simplicity.

1. When $d \lesssim n$, we have $\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\} = O\left(\sqrt{n}\right)$.

1a. If $K \gtrsim \frac{\varepsilon^2 n^{7/4}}{Qd\ln(1/\delta)} = K_a$, we have $R' \gtrsim R''$ and $R = R' + 1$. The constraint on the batch size simplifies to $K \gtrsim \frac{\varepsilon n}{\sqrt{Qn^{1/4}\ln(1/\delta)}} =: K_b$ when $K_a \lesssim n$. Together, the gradient complexity is $M(R'+1)\max(K_a, K_b, 1))$.

1b. Otherwise if $K \lesssim K_a$, we have $R = R'' + 1$. The constraint on the batch size simplifies to $K \gtrsim \sqrt{Kd}$. Therefore, we need $d \lesssim K < n$, which is already assumed. In this case, $K = d$, and the gradient complexity is $M(R''+1)\cdot K = Md + M \cdot \frac{\varepsilon^2 n^2}{d\ln(1/\delta)}$.

2. When $d \gtrsim n$, we have $\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\} = O\left(\frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right)$.

2a. If $K \gtrsim \frac{(\varepsilon n)^{3/2}}{Q(d\ln(1/\delta))^{3/4}} =: K_c$, we have $R' \gtrsim R''$ and $R = R' + 1$. The constraint on the batch size reduces to $K \gtrsim \frac{(\varepsilon n)^{3/4}d^{1/8}}{Q^{1/2}(\ln(1/\delta))^{3/8}} =: K_d$. In the case when $K_d \geq n$, we will use the full batch. Therefore, the gradient complexity is $M(R'+1)\max(K_c, \min(K_d, n), 1)$ if $K_c \lesssim n$.

2b. If $K \lesssim K_c$, we have $R = R''+1$. As in the case 1b., we need $d \lesssim n$, which contradicts with our assumption $d = \Omega(n)$. In this case, we need to use the full batch. We have $K = n$ and the gradient complexity is $M(R''+1)\cdot K = Mn + M\cdot\frac{\varepsilon^2 n^2}{d\ln(1/\delta)}$.

We summarize as follows.

1. When $d \lesssim n$, the gradient complexity is $\min\left\{M(R'+1)\max(K_a, K_b, 1), Md + M\cdot\frac{\varepsilon^2 n^2}{d\ln(1/\delta)}\right\}$ if $K_a \lesssim n$, and $Md + M\cdot\frac{\varepsilon^2 n^2}{d\ln(1/\delta)}$ if $K_a \gtrsim n$.

2. When $d \gtrsim n$, the gradient complexity is $\min\left\{M(R'+1)\max(K_c, \min(K_d, n), 1)), Mn + M\cdot\frac{\varepsilon^2 n^2}{d\ln(1/\delta)}\right\}$ if $K_c \lesssim n$, and $Mn + M\cdot\frac{\varepsilon^2 n^2}{d\ln(1/\delta)}$ if $K_c \gtrsim n$.

Keeping terms involving ε, M, n, and d only, we have $Q = M^{1/4}$, $R' = M^{1/4}\min\{n^{1/4}, \frac{n^{1/2}}{d^{1/4}}\}$, $K_a = \frac{n^{7/4}}{M^{1/4}d}$, $K_b = \frac{n^{7/8}}{M^{1/8}}$, $K_c = \frac{n^{3/2}}{M^{1/4}d^{3/4}}$, $K_d = \frac{n^{3/4}d^{1/8}}{M^{1/8}}$. The results follow by plugging in the expressions. It is straightforward to verify the complexity for the special case $d = \Theta(n)$. □

D. Optimize Non-Smooth Losses via Convolutional Smoothing

D.1. Preliminaries for Convolutional Smoothing

We first provide a brief overview of convolutional smoothing and its key properties. For simplicity, let $\mathcal{U}_s$ denote the uniform distribution over the ℓ_2 ball of radius s.

Definition D.1 (Convolutional Smoothing)**.** For ERM with convex loss f, that is $\hat{F}(w) = \frac{1}{n}\sum_{i=1}^n f(w, x_i)$. We define the convolutional smoother of f with radius s as $\tilde{f}_s := \mathbb{E}_{v\sim\mathcal{U}_s} f(w+v, x_i)$. Then the ERM smoother is defined accordingly, as follows, $\hat{F}_s(w) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{v\sim\mathcal{U}_s} f(w+v, x_i)$.

We have the following properties of the smoother $\hat{F}_{n_s}$ (Kulkarni et al., 2021):

Lemma D.2. *Suppose $\{f(\cdot, x)\}_{x\in\Xi}$ is convex and L-Lipschitz over $\mathcal{K} + B_2(0, s)$. For $w \in \mathcal{K}$, $\tilde{F}_s(w)$ has following properties:*

1. $\widehat{F}(w) \le \tilde{F}_s(w) \le \widehat{F}(w) + Ls$;
2. $\tilde{F}_s(w)$ *is L-Lipschitz;*
3. $\tilde{F}_s(w)$ *is* $\frac{L\sqrt{d}}{r}$*-Smooth;*
4. *For random variables $v \sim \mathcal{U}_s$ and x uniformly from $\{1, 2, \ldots, n\}$, one has*

$$\mathbb{E}[\nabla f(w+v, x)] = \nabla \tilde{F}_s(w)$$

and

$$\mathbb{E}\left[\left\|\nabla \tilde{F}_s(w) - \nabla f(w+v, x)\right\|_2^2\right] \le L^2.$$

Definition D.3 (Poisson Sampling for Convolutional Smoothing)**.** Since the smoother takes the form of an expectation, we can (independently)sample $v_i \overset{\text{iid}}{\sim} \mathcal{U}_s$, and compute $\nabla f(w+v_i, x_i)$ for an estimate of the gradient of $\tilde{f}_s(w, x_i)$. Similar calculation can be done for stochastic gradient. Let K denote the batch size (in expectation). With Poisson sampling of a rate $p = K/n$, we compute an estimate of the stochastic gradient of the ERM smoother

$$\hat{g} = \frac{1}{K}\sum_{i=1}^n Z_i \nabla f(w+v_i, x_i), \tag{25}$$

where $Z_i \overset{\text{iid}}{\sim} Bernoulli(p)$ and $v_i \overset{\text{iid}}{\sim} \mathcal{U}_s$. Each sample x_i is included in the sum independently with probability p.

Similar to the case for the regular stochstic gradient, we can obtain a $O(L^2/m)$ bound for the variance of the estimate of the smoother, proved as follows.

Theorem D.4. *With Poisson sampling of a rate $p = K/n$, let $\hat{g}$ denote the estimate of the above ERM smoother (25). Then $\hat{g}$ is an unbiased estimator of $\nabla \hat{F}_{n_s}(w)$, and the variance of the estimate is $4L^2/K$.*

Proof. By boundedness of ∇f, we can interchange the gradient and the expectation. We have $\mathbb{E}[\nabla f(w+v_i, x_i)] = \nabla \mathbb{E}_{v_i\sim\mathcal{U}_s} f(w+v_i, x_i) = \tilde{f}_s$. Therefore, by linearity of expectation, we have

$$\mathbb{E}[\hat{g}] = \frac{1}{K}\sum_{i=1}^n \mathbb{E}\left[Z_i \nabla f(w+v_i, x_i)\right] = \nabla \tilde{F}_s(w).$$

The variance of the estimate is

$$\begin{aligned} V &:= \mathbb{E}\|g - \nabla \tilde{F}_s(w)\|^2 \\ &= \frac{1}{K^2}\mathbb{E}\left\|\sum_{i=1}^n \left(Z_i \nabla f(w+v_i, x_i) - p\nabla \tilde{F}_s(w)\right)\right\|^2. \end{aligned} \tag{26}$$

Algorithm 6 Convolutional Smoothing-Based Localized ISRL-DP Accelerated MB-SGD

Require: Datasets $X_l \in \mathcal{X}^n$ for $l \in [N]$, loss function f, constraint set $\mathcal{W}$, initial point w_0, subroutine parameters $\{R_i\}_{i=1}^{\lfloor \log_2 n \rfloor} \subset \mathbb{N}$, $\{K_i\}_{i=1}^{\lfloor \log_2 n \rfloor} \subset [n]$, smoothing parameter s.

1: Set $\tau = \lfloor \log_2 n \rfloor$.
2: Set $p = \max(\frac{1}{2}\log_n(M) + 1, 3)$
3: **for** $i = 1$ **to** τ **do**
4: Set $\lambda_i = \lambda \cdot 2^{(i-1)p}$, $n_i = \lfloor n/2^i \rfloor$, $D_i = 2L/\lambda_i$.
5: Each silo $l \in [N]$ draws disjoint batch $B_{i,l}$ of n_i samples from X_l.
6: Let $\hat{F}_i(w) = \frac{1}{n_i N}\sum_{l=1}^{N}\sum_{x_{l,j}\in B_{i,l}} \tilde{f}_s(w; x_{l,j}) + \frac{\lambda_i}{2}\|w - w_{i-1}\|^2$, where $\tilde{f}_s$ is the convolutional smoother of f with radius s.
7: Call the multi-stage (ε, δ)-ISRL-DP implementation of Algorithm 7 with loss function $\hat{F}_i(w)$, data $X_l = B_{i,l}$, $R = R_i$, $K = K_i$, initialization w_{i-1}, constraint set $\mathcal{W}_i = \{w \in \mathcal{W} : \|w - w_{i-1}\| \le D_i\}$, and $\mu = \lambda_i$. Denote the output by w_i.
8: **end for**
9: **return** the last iterate w_τ

Since $\mathbb{E}\left[Z_i \nabla f(w + v_i, x_i) - p\nabla \tilde{F}_s(w)\right] = 0$ and the samples (Z_i, v_i) are independent, the cross terms in the expectation vanish. We have

$$\begin{aligned} V &= \frac{1}{K^2}\sum_{i=1}^{n} \mathbb{E}\|Z_i \nabla f(w + v_i, x_i) - p\nabla \tilde{F}_s(w)\|^2 \\ &= \frac{1}{K^2}\sum_{i=1}^{n} \left(p\mathbb{E}\|\nabla f(w + v_i, x_i) - p\nabla \tilde{F}_s(w)\|^2 + (1-p)\mathbb{E}\|p\nabla \tilde{F}_s(w)\|^2\right) \\ &\le \frac{1}{K^2}\sum_{i=1}^{n} \left(p(1+p)^2 L^2 + (1-p)p^2 L^2\right) \\ &= \frac{1+3p}{K} \cdot L^2 \le \frac{4L^2}{K}, \end{aligned} \tag{27}$$

where the second line follows by conditioning on Z_i, and the third line is due to the Lipschitz property of f and $\tilde{F}_s$. □

D.2. Algorithm and Analysis

For nonsmooth loss, we apply Algorithm 1 to the convolutional smoother with radius s, where s is to be determined. We describe the algorithm in Algorithm 6 and the subroutine in Algorithm 7. The changes compared to Algorithm 1 are listed below.

In the main algorithm, f is replaced by the smoother f_s. For the subroutine call in Line 7, we modify Algorithm 2 as follows. Let $\frac{\lambda}{2}\|w - w_0\|^2$ be the regularizer applied for the subroutine call.

- We change the subsampling regime to Poisson sampling with rate $p = K/n$ in line 5. For silo i, we sample $Z_{i,j} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ for $j \in [n]$ and include $x_{i,j}$ in the sum with probability p. DP noise is sampled $u_i \sim \mathcal{N}(0, \sigma^2 I_d)$ as before.
- We change the gradient computation in Line 6. We sample $v_{i,j} \overset{\text{iid}}{\sim} \mathcal{U}_s$ and compute the estimate of the gradient of the smoother $\frac{1}{pn}\sum_{j=1}^{n} Z_{i,j}\nabla \tilde{f}_s(w + v_{i,j}, x_j) + \lambda(w - w_0) + u_i$.

For the analysis of the algorithm, we note that using Poisson sampling does not change the privacy analysis (up to some constant factors). We can apply the same proof in Appendix C to the smoother f_s except for one minor change in (18), since Lemma B.1 does not directly apply as we have changed the gradient estimator computation. However, by Theorem D.4, the variance of the estimate of the smoother is still $O\left(\frac{L^2}{m}\right)$ without considering the DP noise. Therefore, the conclusion of Lemma B.1 still holds for our smoother.

Algorithm 7 Accelerated ISRL-DP MB-SGD for Convolutional Smoother

Require: Datasets $X_l \in \mathcal{X}^n$ for $l \in [N]$, initial point w_0, loss function $\hat{F}(w) = \frac{1}{nN}\sum_{l=1}^{N}\sum_{x\in X_l}\tilde{f}_s(w,x) + \frac{\lambda}{2}\|w-w_0\|^2$, constraint set $\mathcal{W}$, strong convexity modulus $\mu \geq 0$, privacy parameteres (ε, δ), iteration count $R \in \mathbb{N}$, (expected) batch size $K \in [n]$, step size parameters $\{\eta_r\}_{r\in[R]}, \{\alpha_r\}_{r\in[R]}$ specified in Appendix B.

Initialize $w_0^{ag} = w_0 \in \mathcal{W}$ and $r = 1$.
for $r \in [R]$ **do**
  Server updates and broadcasts
  $w_r^{md} = \frac{(1-\alpha_r)(\mu+\eta_r)}{\eta_r+(1-\alpha_r^2)\mu} w_{r-1}^{ag} + \frac{\alpha_r[(1-\alpha_r)\mu+\eta_r]}{\eta_r+(1-\alpha_r^2)\mu} w_{r-1}$
  for $i \in S_r$ **in parallel do**
    Let the Poisson sampling rate be $p = K/n$.
    Silo i draws $Z_{i,j} \overset{\text{iid}}{\sim} Bernoulli(p)$, $j \in [n]$ and $v_{i,j} \overset{\text{iid}}{\sim} \mathcal{U}_s$, $j \in [n]$
    Sample privacy noise $u_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ for proper σ^2.
    Silo i computes $\widetilde{g}_r^i := \frac{1}{K}\sum_{j=1}^{n} Z_{i,j}\nabla f(w_r^{md} + v_{i,j}, x_{i,j}^r) + \lambda(w_r^{md} - w_0) + u_i$.
  end for
  Server aggregates $\widetilde{g}_r := \frac{1}{M}\sum_{i\in S_r}\widetilde{g}_r^i$ and updates:
  $w_r := \arg\min_{w\in\mathcal{W}} \left\{\alpha_r\left[\langle \widetilde{g}_r, w\rangle + \frac{\mu}{2}\|w_r^{md} - w\|^2\right] + \left[(1-\alpha_r)\frac{\mu}{2} + \frac{\eta_r}{2}\right]\|w_{r-1} - w\|^2\right\}$.
  Server updates and broadcasts
  $w_r^{ag} = \alpha_r w_r + (1-\alpha_r) w_{r-1}^{ag}$.
end for
return: w_R^{ag}.

By the first property in Lemma D.2 that relates $\tilde{F}_s$ to $\hat{F}$. It suffices to choose s to match the difference Ls with the excess risk bound. We have the following result for nonsmooth loss via convolutional smoothing.

Theorem D.5 (Non-smooth FL via convolutional smoothing)**.** *Assume only that $f(\cdot, x)$ is L-Lipschitz and convex for all $x \in \mathcal{X}$. Let $\varepsilon \leq 2\ln(2/\delta)$, $\delta \in (0,1)$. Choose the following convolutional smoothing parameter:*

$$s = \frac{D}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right).$$

Let $\beta = L\sqrt{d}/s$ and choose $R_i \approx \max\left(\sqrt{\frac{\beta+\lambda_i}{\lambda_i}}\ln\left(\frac{\Delta_i\lambda_i M\varepsilon^2 n_i^2}{L^2 d}\right), \frac{\varepsilon^2 n_i^2}{K_i d\ln(1/\delta)}\right)$, where $LD \geq \Delta_i \geq \hat{F}_i(w_{i-1}) - \hat{F}_i(\hat{w}_i)$, $K_i \geq \frac{\varepsilon n_i}{4\sqrt{2R_i\ln(2/\delta)}}$, $\sigma_i^2 = \frac{256L^2R_i\ln(\frac{2.5R_i}{\delta})\ln(2/\delta)}{n_i^2\varepsilon^2}$, and

$$\lambda = \frac{L}{Dn\sqrt{M}}\max\left\{\sqrt{n}, \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon}\right\}. \tag{28}$$

Then, the output of Algorithm 6 is (ε, δ)-ISRL-DP and achieves the following excess risk bound:

$$\mathbb{E}F(w_\tau) - F(w^*) = \widetilde{O}\left(\frac{LD}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right). \tag{29}$$

The communication complexity is

$$\widetilde{O}\left(\max\left\{1, d^{1/4}\sqrt{M}\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}, \frac{\varepsilon^2 n}{d\ln(1/\delta)}\right\}\right), \tag{30}$$

when $K_i = n_i$. When $d = \Theta(n)$, and $\varepsilon = \Theta(1)$, the gradient complexity is

$$\widetilde{O}\left(M^{3/2}n^{3/4} + M^{5/4}n^{11/8}\right).$$

We present the complete proof below. Let $\beta := L\sqrt{d}/r$. By properties of the smoother (D.2), we know that $\tilde{f}_s$ is convex, L-Lipschitz and β-smooth. We can thus reuse Lemma C.3, restated below.

Lemma D.6. *Let $\hat{w}_i = \arg\min_{w\in\mathcal{W}} \hat{F}_i(w)$. We have $\hat{w}_i \in \mathcal{W}_i$ and $\hat{F}_i$ is $3L$-Lipschitz and $(\beta+\lambda_i)$-smooth.*

2. We have the following bounds (same as Lemma C.4) that relate the private solution w_i and the true solution $\hat{w}_i$ of $\hat{F}_i$.

Lemma D.7. *In each phase i, the following bounds hold:*

$$\mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)] = \widetilde{O}\left(\frac{L^2}{\lambda_i M}\cdot\frac{d\ln(1/\delta)}{\varepsilon^2 n_i^2}\right), \tag{31}$$

$$\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \le \widetilde{O}\left(\frac{L^2}{\lambda_i^2 M}\cdot\frac{d\log(1/\delta)}{n_i^2\varepsilon^2}\right). \tag{32}$$

To prove this, it suffices to show that the results Lemma B.1 hold for the multi-stage implementation of Algorithm 7. The proof is essentially the same as that in (Lowy et al., 2023a) by noticing that the variance of the estimate is still $O(L^2/m)$ by Theorem D.4.

For simplicity and consistency with the proof in Appendix C, we will redefine F as the smoother $\tilde{F}_s$ in the following and use F_0 to denote the original loss function. We will apply the following bound for the multi-stage implementation.

Lemma D.8. *Let $f : \mathcal{W} \to \mathbb{R}^d$ be μ-strongly convex and β-smooth, and suppose that the unbiased stochastic gradients $\widetilde{g}(w_r)$ at each iteration r have bounded variance $\mathbb{E}\|\widetilde{g}(w_r) - \nabla F(w_t)\|^2 \le V^2$. If $\widehat{w}_R^{ag}$ is computed by R steps of the Multi-Stage Accelerated MB-SGD, then*

$$\mathbb{E}F(\widehat{w}_R^{ag}) - F^* \lesssim \Delta\exp\left(-\sqrt{\frac{\mu}{\beta}}R\right) + \frac{V^2}{\mu R},$$

where $\Delta = F(w_0) - F^*$.

Proof. By Definition D.3, we know that our gradient estimator in Algorithm 7

$$\tilde{g}_r = \lambda(w_r^{md} - w_0) + \frac{1}{KM}\sum_{i\in S_r}\sum_{j=1}^{n} Z_{i,j}\nabla f(w_r^{md} + v_{i,j}, x_{i,j}^r) + \frac{1}{M}\sum_{i\in S_r} u_i,$$

is an unbiased estimator of $\nabla\hat{F}(w_r^{md})$, and the variance of the estimate has the bound $V := \frac{4L^2}{KM} + \frac{d\sigma^2}{M}$, where the first term is the variance bound in Theorem D.4, and the second term is due to the DP noise.

Applying Lemma D.8 to $\hat{F}_i$, with β replaced by $\beta+\lambda_i$, μ set to λ_i, we get the following the excess risk bound

$$\mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)] \lesssim \Delta_i\exp\left(-\sqrt{\frac{\lambda_i}{\beta+\lambda_i}}R_i\right) + \frac{L^2}{\lambda_i}\left(\frac{1}{MKR} + \frac{d\ln^2(R_i\delta)}{M\varepsilon^2 n^2}\right).$$

The excess risk bound (31) follows by plugging in our choice of R_i. The bound (32) then follows from the inequality below due to λ_i-strong convexity,

$$\frac{\lambda_i}{2}\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \le \mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)].$$

□

3. As a consequence, we can reuse Lemma C.5, restated below.

Lemma D.9. *Let $w \in \mathcal{W}$. We have*

$$\mathbb{E}[F(\hat{w}_i)] - F(w) \le \frac{\lambda_i\mathbb{E}[\|w - w_{i-1}\|^2]}{2} + \frac{4\cdot(3L)^2}{\lambda_i n_i M}.$$

Putting these results together, we now prove the theorem.

Proof. We note that the proof is essentially the same as that in Appendix C.

Privacy. Note that using Poisson sampling does not change the privacy analysis (up to some constant factors). By our choice of σ_i, each phase of the algorithm is (ε, δ)-ISRL-DP. Since the batches $\{B_{i,l}\}_{i=1}^{\tau}$ are disjoint for all $l \in [M]$, the privacy guarantee of the entire algorithm follows from parallel composition.

Excess risk. Given the lemmas introduced above, by the identical reasoning as in Appendix C, we have the following excess risk bound for the smoother,

$$\mathbb{E}F(w_\tau) - F(w^*) \leq \widetilde{O}\left(\frac{LD}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right).$$

By Lemma D.2, we can relate this to the true loss F_0 as follows

$$\mathbb{E}F_0(w_\tau) - F_0(w^*) \leq \mathbb{E}F(w_\tau) - F(w^*) + Ls.$$

The bound then follows from our choice of s.

Complexity. When we use the full batch in each round, that is, when $K_i = n_i$, hiding logarithmic factors, communication complexity is

$$\begin{aligned}\sum_{i=1}^{\tau} R_i &= \sum_{i=1}^{\tau} \max\left\{1, \sqrt{\frac{\beta+\lambda_i}{\lambda_i}}\ln\left(\frac{\Delta\lambda_i M\varepsilon^2 n_i^2}{L^2 d}\right), \frac{\varepsilon^2 n_i^2}{n_i d\ln(1/\delta)}\right\} \\ &= \widetilde{O}\left(\max\left\{\lfloor\log_2 n\rfloor, \sqrt{\frac{\beta}{\lambda}}, \frac{\varepsilon^2 n}{d\ln(1/\delta)}\right\}\right) \\ &= \widetilde{O}\left(\max\left\{1, d^{1/4}\sqrt{M}\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}, \frac{\varepsilon^2 n}{d\ln(1/\delta)}\right\}\right),\end{aligned} \tag{33}$$

where in the last step is due to our choice of $\beta = L\sqrt{d}/s$ and λ.

For gradient complexity, we can follow the same analysis in Appendix C.1 except that due to Poisson sampling our gradient complexity will be in expectation, instead of deterministic. We only consider the case when $d = \Theta(n)$ and $\varepsilon = \Theta(1)$ here to make it simple. Keep terms involving M, n only and drop $\widetilde{O}$ for simplicity. We have $s = \frac{D}{\sqrt{nM}}$, $\beta = L\sqrt{d}/s = \frac{nL\sqrt{M}}{D}$.

Plugging β into Equation (24), we obtain the gradient complexity of $\widetilde{O}\left(M^{3/2}n^{3/4} + M^{5/4}n^{11/8}\right)$.

□

E. Precise Statement and Proof of Theorem 3.1

Lemma E.1. *(Nesterov, 2005) Let $f : \mathcal{W} \to \mathbb{R}^d$ be convex and L-Lipschitz and $\beta > 0$. Then the β-Moreau envelope $f_\beta(w) := \min_{v\in\mathcal{W}}\left(f(v) + \frac{\beta}{2}\|w - v\|^2\right)$ satisfies:*
(1) f_β is convex, $2L$-Lipschitz, and β-smooth;
(2) $\forall w \in \mathcal{W}$, $f_\beta(w) \leq f(w) \leq f_\beta(w) + \frac{L^2}{2\beta}$;
(3) $\forall w \in \mathcal{W}$, $\nabla f_\beta(w) = \beta(w - \mathrm{prox}_{f/\beta}(w))$;
where the prox operator of $f : \mathcal{W} \to \mathbb{R}^d$ is defined as $\mathrm{prox}_f(w) := \arg\min_{v\in\mathcal{W}}\left(f(v) + \frac{1}{2}\|w - v\|^2\right)$.

Our algorithm for non-smooth functions uses Lemma E.1 and Algorithm 1 as follows: First, property (1) above allows us to apply Algorithm 1 to optimize f_β and obtain an excess population risk bound for f_β, via Theorem 2.1. Inside Algorithm 1, we will use property (3) to compute the gradient of f_β. Then, property (2) enables us to extend the excess risk guarantee to the true function f, for a proper choice of β.

Theorem E.2 (Precise Statement of Theorem 3.1)**.** *Let $\varepsilon \leq 2\ln(2/\delta)$, $\delta \in (0,1)$. Choose $R_i \approx \max\left(\sqrt{\frac{\beta+\lambda_i}{\lambda_i}}\ln\left(\frac{\Delta_i\lambda_i M\varepsilon^2 n_i^2}{L^2 d}\right), \mathbb{1}_{\{MK_i<Nn_i\}}\frac{\varepsilon^2 n_i^2}{K_i d\ln(1/\delta)}\right)$, where $LD \geq \Delta_i \geq \hat{F}_i(w_{i-1}) - \hat{F}_i(\hat{w}_i)$, $K_i \geq \frac{\varepsilon n_i}{4\sqrt{2R_i\ln(2/\delta)}}$,*

$\sigma_i^2 = \frac{256L^2R_i\ln(\frac{2.5R_i}{\delta})\ln(2/\delta)}{n_i^2\varepsilon^2}$, *and*

$$\lambda = \frac{L}{Dn\sqrt{M}}\max\left\{\sqrt{n}, \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon}\right\}.$$

There exist choices of β *such that running Algorithm 1 with* $f_\beta(w,x) := \min_{v\in\mathcal{W}}\left(f(v,x) + \frac{\beta}{2}\|w-v\|^2\right)$ *yields an* (ε,δ)*-ISRL-DP algorithm with optimal excess population risk as in* (16)*. The communication complexity is*

$$\widetilde{O}\left(\max\left\{1, \sqrt{M}\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}, \mathbb{1}_{\{M<N\}}\frac{\varepsilon^2 n}{d\ln(1/\delta)}\right\}\right).$$

Proof. **Privacy.** For any given β, the privacy guarantee is immediate since we have already showed that Algorithm 1 is (ε,δ)-ISRL-DP.

Excess risk. By Lemma E.1 part 2, we have

$$\mathbb{E}F(w_\tau) - F(w^*) \le \mathbb{E}F(w_\tau) - F(w^*) + \frac{L^2}{2\beta}.$$

It suffices to choose $\beta = \frac{L\sqrt{M}}{D}\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\log(1/\delta)}}\right\}$.

Communication complexity. The result follows by plugging in our choice of β into (17).

□

F. More Details on the Communication Complexity Lower Bound (Theorem 2.4)

We begin with a definition (Woodworth et al., 2020):

Definition F.1 (Distributed Zero-Respecting Algorithm)**.** For $v\in\mathbb{R}^d$, let $\text{supp}(v) := \{j\in[d] : v_j \neq 0\}$. Denote the random seed of silo m's gradient oracle in round t by z_t^m. An optimization algorithm is *distributed zero-respecting with respect to* f if for all $t\ge 1$ and $m\in[N]$, the t-th query on silo m, w_t^m satisfies

$$\text{supp}(w_t^m) \subseteq \bigcup_{s<t}\text{supp}(\nabla f(w_s^m, z_s^m) \cup \bigcup_{m'\neq m}\bigcup_{s\le\pi_m(t,m')}\text{supp}(\nabla f(w_s^{m'}, z_s^{m'}),$$

where $\pi_m(t,m')$ is the most recent time before t when silos m and m' communicated with each other.

Theorem F.2 (Re-statement of Theorem 2.4)**.** *Fix* $M = N$ *and suppose* $\mathcal{A}$ *is a distributed zero-respecting algorithm with excess risk*

$$\mathbb{E}F(\mathcal{A}(X)) - F^* \lesssim \frac{LD}{\sqrt{N}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)$$

in $\le R$ *rounds of communications on any* β*-smooth FL problem with heterogeneity* ζ_* *and* $\|w_0 - w^*\| \le D$*. Then,*

$$R \gtrsim N^{1/4}\left(\min\left\{\sqrt{n}, \frac{\varepsilon n}{\sqrt{d\ln(1/\delta)}}\right\}\right)^{1/2} \times \min\left(\frac{\sqrt{\beta D}}{\sqrt{L}}, \frac{\zeta_*}{\sqrt{\beta LD}}\right).$$

Proof. Denote the worst-case excess risk of $\mathcal{A}$ by

$$\alpha := \mathbb{E}F(\mathcal{A}(X)) - F^*.$$

Now, (Woodworth et al., 2020, Theorem 4) implies that any zero-respecting algorithm has

$$R \gtrsim \min\left\{\frac{\zeta_*}{\sqrt{\beta\alpha}}, \frac{D\sqrt{\beta}}{\sqrt{\alpha}}\right\}.$$

Plugging in $\alpha = \frac{LD}{\sqrt{N}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)$ proves the lower bound. □

G. Precise Statement and Proof of Theorem 3.4

Theorem G.1 (Precise statement of Theorem 3.4)**.** *Choose*

$$\eta = \frac{D\sqrt{M}}{L} \min\left\{ \frac{1}{\sqrt{n}}, \frac{\varepsilon}{\sqrt{d\ln(1/\delta)}} \right\}, \tag{34}$$

$K_i \geq \max\left(1, \frac{\varepsilon n}{4\sqrt{2R\ln(2/\delta)}}\right)$, $R_i = \min\left(Mn_i, \frac{M\varepsilon^2 n_i^2}{d}\right) + 1$, *and* $\sigma_i^2 = \frac{256L^2 R_i \ln(\frac{2.5R}{\delta})\ln(2/\delta)}{n_i^2\varepsilon^2}$ *for* $i \in [k]$*in Algorithm 4. Then, Algorithm 4 is* (ε, δ)*-ISRL-DP and achieves the following excess risk bound:*

$$\mathbb{E}F(w_\tau) - F(w^*) = \widetilde{O}\left(\frac{LD}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)\right).$$

Further, the communication complexity is

$$\sum_{i=1}^{\tau} R_i = \widetilde{O}\left(\min\left(nM, \frac{M\varepsilon^2 n^2}{d}\right) + 1\right),$$

and the gradient complexity is

$$\sum_{i=1}^{\tau} R_i K_i \cdot M = \widetilde{O}\left(M + M^2 \min\left(n, \frac{\varepsilon^2 n^2}{d}\right) + \varepsilon Mn + M^{3/2}\min\left(\varepsilon n^{3/2}, \frac{\varepsilon^2 n^2}{\sqrt{d}}\right)\right).$$

We need the following result in convex optimization to bound the excess risk in each call of ISRL-DP MB-Subgradient Method.

Lemma G.2. *(Bubeck et al., 2015, Theorem 6.2) Let* g *be* λ*-strongly convex, and assume that the stochastic subgradient oracle returns a stochastic subgradient* $\widetilde{g}(w)$ *such that* $\mathbb{E}\widetilde{g}(w) \in \partial g(w)$ *and* $\mathbb{E}\|\widetilde{g}(w)\|_2^2 \leq B^2$*. Then, the stochastic subgradient method* $w_{r+1} = w_r - \gamma_r \widetilde{g}(w_r)$ *with* $\gamma_r = \frac{2}{\lambda(r+1)}$ *satisfies*

$$\mathbb{E}g\left(\sum_{r=1}^{R} \frac{2r}{R(R+1)} w_r\right) - g\left(w^*\right) \leq \frac{2B^2}{\lambda(R+1)}. \tag{35}$$

As a consequence, we have the following result:

Lemma G.3. *In each phase* i *of Algorithm 4, the following bounds hold:*

$$\mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)] = \widetilde{O}\left(\frac{L^2\eta_i}{M} + \frac{dL^2\eta_i}{Mn_i\varepsilon^2}\right), \tag{36}$$

$$\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \leq \widetilde{O}\left(\frac{L^2\eta_i^2 n_i}{M} + \frac{dL^2\eta_i^2}{M\varepsilon^2}\right). \tag{37}$$

Proof. Note that the second moment of each noisy aggregated subgradient in round r of Algorithm 3 is bounded by

$$\mathbb{E}\left\|\tilde{g}_r\right\|^2 = \left\| u_r + \frac{1}{M_r}\sum_{i\in S_r}\frac{1}{K}\sum_{j=1}^{K}\nabla f(w_r, x_{i,j}^r)\right\|^2 \leq 2L^2 + \frac{2d\sigma^2}{M}.$$

Given our choice of step sizes in Algorithm 4, we can verify the assumptions in Lemma G.2 are met for $\hat{F}_i$. Recall w_i is the output of each phase and $\hat{w}_i = \arg\min_{w\in\mathcal{W}} \hat{F}_i(w)$. It follows from Lemma G.2 that,

$$\begin{aligned}
\mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)] &\leq \frac{2}{\lambda_i(R_i+1)} \cdot \left(2L^2 + \frac{2d\sigma_i^2}{M}\right) \\
&= \widetilde{O}\left(\frac{L^2\eta_i n_i}{R_i} + \frac{dL^2\eta_i}{Mn_i\varepsilon^2}\right) \\
&= \widetilde{O}\left(\frac{L^2\eta_i}{M} + \frac{dL^2\eta_i}{Mn_i\varepsilon^2}\right),
\end{aligned}$$

as desired (36), where the last step is due to our choice of $R_i = \min\left(Mn_i, \frac{M\varepsilon^2 n_i^2}{d}\right) + 1$.

Using the λ_i-strong convexity, we have

$$\frac{\lambda_i}{2}\mathbb{E}\left[\|w_i - \hat{w}_i\|^2\right] \leq \mathbb{E}[\hat{F}_i(w_i) - \hat{F}_i(\hat{w}_i)].$$

The bound (37) follows. □

Now we are ready to prove the theorem.

Proof of Theorem G.1. **Privacy.** The privacy analysis of (Lowy & Razaviyayn, 2023a, Theorem D.1) for ISRL-DP MB-SGD holds verbatim in the nonsmooth case when we replace subgradients by gradients: the only property of $f(\cdot, x)$ that is used in the proof is Lipschitz continuity. Thus, Algorithm 3 is (ε, δ)-ISRL-DP if the noise variance is $\sigma^2 \geq \frac{256L^2R\ln(2.5R/\delta)\ln(2/\delta)}{\varepsilon^2 n^2}$. By our choice of σ_i^2, we see that phase i of Algorithm 4 is (ε, δ)-ISRL-DP on data $\{B_{i,l}\}_{l=1}^N$. Since the batches $\{B_{i,l}\}_{i=1}^\tau$ are disjoint for all $l \in [N]$, the full Algorithm 4 is (ε, δ)-ISRL-DP by parallel composition (McSherry, 2009).

Excess Risk. Define $\hat{w}_0 = w^*$ and write

$$\mathbb{E}F(w_\tau) - F(w^*) = \mathbb{E}[F(w_\tau) - F(\hat{w}_\tau)] + \sum_{i=1}^{\tau}\mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})].$$

Since $\tau = \lfloor \log_2 n \rfloor$, we have $n_\tau = \Theta(1)$ and $\eta_\tau = \eta\Theta(n^{-p})$. By (37) and Jensen's Inequality $\mathbb{E}Z \leq \sqrt{\mathbb{E}Z^2}$, we bound the first term as follows

$$\begin{aligned}
\mathbb{E}[F(w_\tau) - F(\hat{w}_\tau)] \leq L\mathbb{E}\left[\|w_\tau - \hat{w}_\tau\|\right] &\leq \widetilde{O}\left(\frac{L^2\eta_\tau\sqrt{n_\tau}}{\sqrt{M}} + \frac{\sqrt{d}L^2\eta_\tau}{\sqrt{M}\varepsilon}\right) \\
&\leq \widetilde{O}\left(\frac{L^2\eta}{n^p\sqrt{M}} + \frac{\sqrt{d}L^2\eta}{\sqrt{M}n^p\varepsilon}\right) \\
&\leq \widetilde{O}\left(\frac{LD}{n^{p-\frac{1}{2}}}\right) \leq \widetilde{O}\left(\frac{LD}{\sqrt{nM}}\right),
\end{aligned}$$

where the last two steps are due to the choice of η per (34) and (21).

By (36) and using the fact that $1/n_i$ and η_i decrease geometrically, we have

$$\begin{aligned}
\sum_{i=1}^{\tau}\mathbb{E}[F(\hat{w}_i) - F(\hat{w}_{i-1})] &\leq \sum_{i=1}^{\tau}\left(\frac{\lambda_i\mathbb{E}[\|w_{i-1} - \hat{w}_{i-1}\|^2]}{2} + \frac{4\cdot(3L)^2}{\lambda_i n_i M}\right) \\
&\leq \widetilde{O}\left(\frac{D^2}{\eta n} + \sum_{i=2}^{\tau}\frac{L^2\eta_i}{M} + \frac{dL^2\eta_i}{Mn_i\varepsilon^2} + \sum_{i=1}^{\tau}\frac{L^2\eta_i}{M}\right) \\
&\leq \widetilde{O}\left(\frac{D^2}{\eta n} + \frac{L^2\eta}{M} + \frac{L^2\eta d}{Mn\varepsilon^2}\right).
\end{aligned}$$

Plugging in the choice of η per (34) gives the desired excess risk.

Communication complexity. Summing geometric series, we obtain the communication complexity as follows

$$\sum_{i=1}^{\tau} R_i = \widetilde{O}\left(\min\left(nM, \frac{M\varepsilon^2 n^2}{d}\right)\right) + 1.$$

Gradient complexity. Recall $K_i \geq \max\left(1, \frac{\varepsilon n_i}{4\sqrt{2R_i\ln(2/\delta)}}\right)$. Choosing the minimum K_i, we have

$$\begin{aligned}
\sum_{i=1}^{\tau} R_i K_i \cdot M &= \widetilde{O}\left(\max\left(RM, \varepsilon nM\sqrt{R}\right)\right) \\
&= \widetilde{O}\left(M + M^2\min\left(n, \frac{\varepsilon^2 n^2}{d}\right) + \varepsilon nM + M^{3/2}\min\left(\varepsilon n^{3/2}, \frac{\varepsilon^2 n^2}{\sqrt{d}}\right)\right).
\end{aligned}$$

□

H. A Stability Result

In order to bound the excess risk of the function F_i, we require the following generalization of Theorem 6 from (Shalev-Shwartz et al., 2009) which provides a stability result.

Lemma H.1. *Let $g(w, x)$, $w \in \mathcal{W}$ be λ-strongly convex and L-Lipschitz in w for all $x \in \mathcal{X}$. Let $X = (x_1, x_2, \ldots, x_m)$ be a set of m independent samples such that x_i is sampled from their corresponding distribution $\mathcal{D}_i$ for $i \in [m]$. We write $X \sim \mathcal{D}$ for short.*

Let $\hat{G}(w) = \frac{1}{m}\sum_{i=1}^m g(w, x_i)$ and let $\hat{w} = \arg\min_{w\in\mathcal{W}} \hat{G}(w)$ be the empirical minimizer. Let $G(w) = \mathbb{E}_{X\sim\mathcal{D}}\left[\frac{1}{m}\sum_{i=1}^m g(w, x_i)\right]$ and let $w^* = \arg\min G(w)$ be the population minimizer. Then, for any $w \in \mathcal{W}$, we have

$$\mathbb{E}[G(\hat{w})] - G(w) \leq \frac{4L^2}{\lambda m}.$$

Proof. We will use a stability argument. Let $X' = (x'_1, x'_2, \ldots, x'_n)$ be a set of m independent samples from $\mathcal{D}$, and let $X^{(i)} = (x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_m)$ be the dataset with the i-th element replaced by x'_i.

Let $\hat{G}^{(i)}(w) = \frac{1}{m}\left(f(w, x'_i) + \sum_{j\neq i} f(w, x_j)\right)$ and $\hat{w}^{(i)} = \min \hat{G}^{(i)}(w)$ be its minimizer. We have

$$\begin{aligned}
\hat{G}(w^{(i)}) - \hat{G}(\hat{w}) &= \frac{g(\hat{w}^{(i)}, z_i) - g(\hat{w}, z_i)}{m} + \frac{\sum_{j\neq i}\left(g(\hat{w}^{(i)}, z_i) - g(\hat{w}, z_i)\right)}{m} \\
&= \frac{g(\hat{w}^{(i)}, z_i) - g(\hat{w}, z_i)}{m} + \frac{g(\hat{w}, z'_i) - g(\hat{w}^{(i)}, z'_i)}{m} \\
&+ \left(\hat{G}^{(i)}(\hat{w}^{(i)}) - \hat{G}^{(i)}(\hat{w})\right) \\
&\leq \frac{\left|g(\hat{w}^{(i)}, z_i) - g(\hat{w}, z_i)\right|}{m} + \frac{\left|g(\hat{w}, z'_i) - g(\hat{w}^{(i)}, z'_i)\right|}{m} \\
&\leq \frac{2L}{m}\left\|\hat{w}^{(i)} - \hat{w}\right\|,
\end{aligned}$$

where the first inequality follows from the definition of the minimizer $w^{(i)}$, and the second inequality follows from the Lipschitzness of g.

By strong convexity of $\hat{G}$, we have

$$\hat{G}(\hat{w}^{(i)}) \geq \hat{G}(\hat{w}) + \frac{\lambda}{2}\left\|\hat{w}^{(i)} - \hat{w}\right\|^2.$$

Therefore, we have $\left\|\hat{w}^{(i)} - \hat{w}\right\| \leq 4L/(\lambda m)$. Thus, ERM satisfies $4L/\lambda m$ uniform argument stability, and (by Lipschitz continuity of $g(\cdot, x)$) $4L^2/\lambda m$ uniform stability.

Next, we show that this stability bound implies the desired generalization error bound (even if the samples are not drawn from an identical underlying distribution). Note that X and X' are independently sampled from $\mathcal{D}$. By symmetry (renaming x_i to x'_i), we know that $\mathbb{E}_{X,X'}[g(\hat{w}, x_i)] = \mathbb{E}_{X,X'}[g(\hat{w}^{(i)}, x'_i)]$. Therefore, we have

$$\begin{aligned}
\mathbb{E}[\hat{G}(\hat{w})] &= \mathbb{E}_X\left[\frac{1}{m}\sum_{i=1}^n g(\hat{w}, x_i)\right] \\
&= \mathbb{E}_{X,X'}\left[\frac{1}{m}\sum_{i=1}^n g(\hat{w}^{(i)}, x'_i)\right] \\
&= \mathbb{E}_{X,X'}\left[\frac{1}{m}\sum_{i=1}^n g(\hat{w}, x'_i) + \delta\right] \\
&= \mathbb{E}[G(\hat{w})] + \delta,
\end{aligned}$$

where by Lipschitz continuity of f, we have

$$\begin{aligned}\delta &= \mathbb{E}_{X,X'} \frac{1}{m} \sum_{i=1}^{n} \left(g(\hat{w}^{(i)}, x_i') - g(\hat{w}, x_i') \right) \\ &\leq \mathbb{E}_{X,X'} \frac{1}{m} \sum_{i=1}^{n} L \cdot \left\| w^{(i)} - w \right\| \\ &\leq \frac{4L^2}{\lambda m}.\end{aligned}$$

Now for given $w \in \mathcal{W}$ we have $G(w) = \mathbb{E}[\hat{G}(w)] \geq \mathbb{E}[\hat{G}(\hat{w})]$, by definition of $\hat{w}$. Therefore, we conclude that

$$\mathbb{E}[G(\hat{w})] - G(w) \leq \frac{4L^2}{\lambda m}.$$

□