Private Federated Learning Without a Trusted Server: Optimal Algorithms for Convex Losses

Andrew Lowy
University of Southern California
lowya@usc.edu

Meisam Razaviyayn
University of Southern California
razaviya@usc.edu

Abstract

This paper studies the problem of federated learning (FL) in the absence of a trustworthy server/clients. In this setting, each client needs to ensure the privacy of its own data without relying on the server or other clients. We study local differential privacy (LDP) at the client level and provide *tight upper and lower bounds that establish the minimax optimal rates* (up to logarithms) for LDP convex/strongly convex federated stochastic optimization. Our rates match the optimal statistical rates in certain practical parameter regimes ("privacy for free"). Second, we develop a novel time-varying noisy SGD algorithm, leading to the first non-trivial LDP risk bounds for FL with *non-i.i.d.* clients. Third, we consider the special case where each client's loss function is empirical and develop an *accelerated* LDP FL algorithm to improve communication complexity compared to existing works. We also provide matching lower bounds, establishing the optimality of our algorithm for convex/strongly convex settings. Fourth, with a secure shuffler to anonymize client reports (but without a trusted server), our algorithm attains the optimal central DP rates for stochastic convex/strongly convex optimization, thereby achieving optimality in the local and central models simultaneously. Our upper bounds quantify the role of network communication reliability in performance. We illustrate the practical utility of our algorithm with numerical experiments.

1 Introduction

In federated learning, each "client" (e.g. cell-phone users or organizations such as hospitals) stores its data locally and a central server coordinates updates to achieve the global learning objective [40]. Federated learning (FL) has been deployed across application domains such as the internet of things [61], wireless communication [55], medicine [16], finance [1], and by companies such as Google [63] and Apple [4]. One of the primary reasons for the introduction of FL was to offer greater privacy for sensitive user data [54]. Unfortunately, merely storing data locally is not sufficient to prevent data leakage, as model parameters or updates can still reveal sensitive information [27, 35, 66, 80]. These leaks can occur when clients send updates to the server, which an adversary may be able to access, or (in decentralized/peer-to-peer FL) directly to other clients. Therefore, it is important to develop privacy-preserving mechanisms for FL that do not rely on the server or other clients to implement.

Consider a FL setting with N clients, each containing a local data set with n_i samples: $X_i = (x_{i,1}, \cdots, x_{i,n_i})$ for $i \in [N]$. In each round of communication r, a uniformly random subset S_r of $M_r \in [N]$ clients is able to participate, where $\{M_r\}_{r=1}^R$ are i.i.d. random variables. For all i, let $\mathcal{D}_i$ be a probability distribution on a set $\mathcal{X}_i$ containing data and denote $\mathcal{X} := \bigcup_{i=1}^N \mathcal{X}_i$. Given a convex (in w) loss function $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, define client i's local objective as

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

where $\mathcal{W} \subset \mathbb{R}^d$ is closed, convex, and $\|w\|_2 \leqslant D$, $\forall w \in \mathcal{W}$. At times, we may focus on empirical risk minimization (ERM) framework where $\widehat{F}_i(w) := \frac{1}{n_i} \sum_{j=1}^{n_i} f(w, x_{i,j})$. Our goal is to solve

1st NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021), Virtual Meeting.

$$\min_{w\in\mathcal{W}}\left\{F(w) := \sum_{i=1}^{N} p_i F_i(w)\right\}, \tag{2}$$

or, in the ERM case, $\min_{w\in\mathcal{W}}\{\widehat{F}(w) := \sum_{i=1}^N p_i \widehat{F}_i(w)\}$, while maintaining the privacy of each client. Here $p_i \geqslant 0$ with $\sum_i p_i = 1$. We assume WLOG (see Appendix B) that $p_i = \frac{1}{N}, \forall i \in [N]$. We say problem (2) is "i.i.d." if $\mathcal{X}_i = \mathcal{X}$ and $\mathcal{D}_i = \mathcal{D} \;\; \forall i$. When F_i takes the general form (1) (not necessarily ERM), we refer to the problem as "SCO" (stochastic convex optimization) for emphasis.

Popular algorithms for FL include Local SGD/Federated Averaging [55] and Minibatch SGD. Both of these are *fully interactive* algorithms, meaning they can adaptively query each client multiple times. A subset of fully interactive algorithms is the set of *sequentially interactive* algorithms, which can query clients adaptively in sequence, but cannot query any one client more than once. *Non-interactive* algorithms are non-adaptive sequentially interactive algorithms: they query each client once, independently of other clients' reports. See [39] for further discussion.

Notions of Privacy for FL: Given the practical importance of maintaining the privacy of user data during the FL process, numerous different definitions of private FL have been proposed. Some of these have used secure multi-party computation (MPC) [12, 53], but this approach leaves users vulnerable to inference attacks on the trained model. This is in contrast to differential privacy (DP), which by now has been widely accepted as the gold standard of rigorous data privacy notions. DP is defined with respect to a database space $\mathbb{X}$ and a measure of distance $\rho : \mathbb{X}^2 \to [0, \infty)$ between databases. We say two databases $\mathbf{X}, \mathbf{X}' \in \mathbb{X}$ are ρ-adjacent if $\rho(\mathbf{X}, \mathbf{X}') \leqslant 1$.

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

$$\mathbb{P}(\mathcal{A}(\mathbf{X}) \in S) \leqslant e^{\epsilon}\mathbb{P}(\mathcal{A}(\mathbf{X}') \in S) + \delta. \tag{3}$$

If (3) holds for all measurable subsets S, then we denote this property by $\mathcal{A}(\mathbf{X}) \underset{(\epsilon,\delta)}{\simeq} \mathcal{A}(\mathbf{X}')$. For FL, we will fix $\mathbb{X} := \mathcal{X}_1^{n_1} \times \cdots \times \mathcal{X}_N^{n_N}$ so that databases $\mathbf{X}$ in our FL setting consist of N client datasets $\mathbf{X} = (X_1, \cdots X_N)$. Definition 1 says that an algorithm is DP if with high probability, an adversary cannot distinguish between the outputs of the algorithm when it is run on adjacent databases. Depending on the choice of ρ, we can get different variations of DP. For example, in the classical notion of central differential privacy (CDP) (often simply referred to as "differential privacy") [19], $\rho(\mathbf{X}, \mathbf{X}') := \sum_{i=1}^N \sum_{j=1}^{n_i} \mathbb{1}_{x_{i,j} \neq x'_{i,j}}$ is the hamming distance and adjacent databases are those differing in a single sample. In the context of FL, a major problem with CDP is that it does not preclude the untrusted server from accessing non-private updates (which may leak clients' data).

Client-level DP (also called user-level DP) has been proposed as an alternative to CDP where a single person may contribute many samples to the database (e.g. language modeling) [55, 29, 38, 28, 72, 79, 47]. It is defined by taking $\rho(\mathbf{X}, \mathbf{X}') = \sum_{i=1}^N \mathbb{1}_{X_i \neq X'_i}$. Client-level DP shares the same drawback as CDP: it allows sensitive data to be leaked to the untrusted server. In contrast to the centralized models of CDP and client-level DP, this work imposes the stronger requirement that *client updates be private before they are sent to the untrusted server (or other clients)* for aggregation.

First, we consider local differential privacy (LDP), a generalization of the the classic notion with the same name [44] to FL. An R-round fully interactive randomized algorithm $\mathcal{A} : \mathbb{X} \to \mathcal{Z}^{R\times N}$ for FL is characterized in every round $r \in [R]$ by N local client functions called *randomizers* $\mathcal{R}_r^{(i)} : \mathcal{Z}^{(r-1)\times N} \times \mathcal{X}_i^{n_i} \to \mathcal{Z}$ $(i \in [N])$ and an aggregation mechanism. The randomizers send messages $Z_r^{(i)} := \mathcal{R}_r^{(i)}(\mathbf{Z}_{1:r-1}, X_i)$ (which may depend on client data X_i and the outputs $\mathbf{Z}_{1:r-1} := \{Z_t^{(j)}\}_{j\in[N], t\in[r-1]}$ of clients' randomizers in prior rounds) to the server or (in peer-to-peer FL) other clients. [1] Then, the server (or clients, for peer-to-peer FL) updates the global model. Algorithm $\mathcal{A}$ is $\{(\epsilon_i, \delta_i)\}_{i=1}^N$-LDP if for all $i \in [N]$, the full transcript of client i's communications, i.e. the collection

[1] We assume that $\mathcal{R}_r^{(i)}(\mathbf{Z}_{1:r-1}, X_i)$ is conditionally independent of X_j $(j \neq i)$ given $\mathbf{Z}_{1:r-1}$ and X_i. That is, the randomizers of i cannot "eavesdrop" on another client's data (consistent with the local data principle of FL). We allow for $Z_t^{(i)}$ to be empty/zero if client i does not output anything to the server in round t.

of all R messages $\{Z_r^{(i)}\}_{r\in[R]}$, is (ϵ_i, δ_i)-DP, conditional on the messages and data of all other clients. See Fig. 1. [2] More precisely:

Definition 2. *(Local Differential Privacy) A randomized algorithm $\mathcal{A} : \mathcal{X}_1^{n_1} \times \cdots \times \mathcal{X}_N^{n_N} \to \mathcal{Z}^{R\times N}$ is $\{(\epsilon_i, \delta_i)\}_{i=1}^N$-LDP if for all $i \in [N]$ and all ρ_i-adjacent $X_i, X_i' \in \mathcal{X}_i^{n_i}$, we have*

$$(\mathcal{R}_1^{(i)}(X_i), \mathcal{R}_2^{(i)}(\mathbf{Z}_1, X_i), \cdots, \mathcal{R}_R^{(i)}(\mathbf{Z}_{1:R-1}, X_i)) \underset{(\epsilon_i,\delta_i)}{\simeq} (\mathcal{R}_1^{(i)}(X_i'), \mathcal{R}_2^{(i)}(\mathbf{Z}_1', X_i'), \cdots, \mathcal{R}_R^{(i)}(\mathbf{Z}_{1:R-1}', X_i')),$$

where for all r the distribution of $\mathcal{R}_r^{(i)}(\mathbf{Z}_{1:r-1}, X_i)$ is conditional on the transcripts $\mathbf{Z}_{1:r-1}^{(j\neq i)}$ of all other clients in all previous rounds. Here $\rho_i : \mathcal{X}_i^2 \to [0, \infty)$ is given by $\rho_i(X_i, X_i') = \sum_{j=1}^{n_i} \mathbb{1}_{x_{i,j}\neq x_{i,j}'}$.

We sometimes assume for simplicity that privacy parameters are the same across clients and denote these common parameters by (ϵ_0, δ_0). Note that *LDP is stronger than CDP*: (ϵ_0, δ_0)-LDP implies (ϵ_0, δ_0)-CDP but the converse is false. Moreover, *LDP is stronger than client-level DP* in the following sense: (ϵ_0, δ_0)-LDP implies $(n\epsilon, ne^{(n-1)\epsilon}\delta)$ client-level DP, but (ϵ, δ)-client-level DP does not imply (ϵ', δ')-LDP for any ϵ', δ'. See Appendix C for proofs.

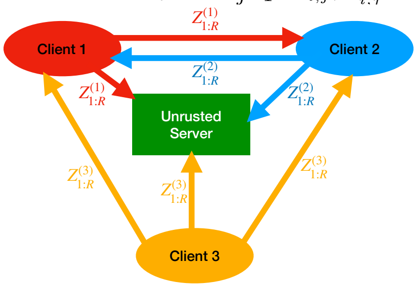

Figure 1: LDP protects the privacy of each client's data regardless of whether the server or other clients are trustworthy and regardless of the network topology (e.g. peer-to-peer or server-orchestrated). The messages $Z_{1:R}^{(i)}$ of client i are DP, ensuring that *client i's data cannot be leaked, even if the server or other clients are curious or leak data themselves.*

It is also illuminating to compare Definition 2 with classical item-level LDP [44, 18], which requires each individual *person* (rather than client) to randomize their own data. When $n = 1$, so that each client has just one person's data, classical LDP is equivalent to Definition 2. But if $n > 1$, then classical LDP would require each person (e.g. patient) to randomize her reports (e.g. medical test results) before sending them to the data collector (doctors/researchers) within the client (hospital). Since we assume in FL that clients can be trusted with their own data, this intra-client randomization is practically unnecessary. Thus, *Definition 2 is more practically relevant for FL than classical item-level LDP*, and it results in higher accuracy models.

An intermediate trust model between the low-trust local model and the high-trust central/client-level model is the shuffle model [10, 14, 21, 22, 25, 50, 32] where clients have access to a secure shuffler (also known as a mixnet) that receives randomized reports from the clients and randomly permutes them (effectively anonymizing them), before the reports are sent to the untrusted server/other clients.[3]

Definition 3. *(Shuffle Differential Privacy) A randomized algorithm $\mathcal{A} : \mathcal{X}_1^{n_1} \times \cdots \times \mathcal{X}_N^{n_N} \to \mathcal{Z}^{N\times R}$ is (ϵ, δ)-shuffle DP (SDP) if for all ρ-adjacent databases $\mathbf{X}, \mathbf{X}' \in \mathcal{X}_1^{n_1} \times \cdots \times \mathcal{X}_N^{n_N}$ and all measurable subsets S, the collection of all uniformly randomly permuted messages that are sent by the shuffler satisfies (3), with $\rho(\mathbf{X}, \mathbf{X}') := \sum_{i=1}^N \sum_{j=1}^{n_i} \mathbb{1}_{x_{i,j}\neq x_{i,j}'}$ (same ρ as CDP).*

That is, SDP prohibits the server from viewing non-private functions of clients' data, instead restricting clients to randomize their own data and use shuffling. Since (ϵ_0, δ_0)-LDP implies (ϵ_0, δ_0)-CDP and shuffling can be seen as post-processing, it follows that (ϵ_0, δ_0)*-LDP implies* (ϵ_0, δ_0)*-SDP*.

Notation and Assumptions: Denote by $\|\cdot\|$ the Euclidean norm. A differentiable function $g : \mathcal{W} \to \mathbb{R}$ ($\mathcal{W} \subseteq \mathbb{R}^d$) is μ*-strongly convex* ($\mu \geqslant 0$) if $g(w) \geqslant g(w') + \langle \nabla g(w'), w - w' \rangle + \frac{\mu}{2}\|w - w'\|^2 \ \forall \ w, w' \in \mathcal{W}$. If $\mu = 0$, we say g is *convex*. A function $h : \mathcal{W} \to \mathbb{R}^m$ is L*-Lipschitz* if $\|h(w) - h(w')\| \leqslant L\|w - w'\|$ for all $w, w' \in \mathcal{W}$. h is β*-smooth* if its derivative ∇h is β-Lipschitz. Denote $w^* \in \operatorname{argmin}_{w\in\mathcal{W}} F(w)$. We write $a \lesssim b$ if $\exists C > 0$ such that $a \leqslant Cb$. We write $a = \widetilde{O}(b)$ if $a \lesssim \log(\theta)b$ for some parameters θ. We assume the following throughout this work:

Assumption 1. *$f(\cdot, x)$ is L_i-Lipschitz and μ-strongly convex (with $\mu = 0$ for convex) $\forall x \in \mathcal{X}_i$.*

Assumption 2. *In each round r, a uniformly random subset S_r of $M_r \in [N]$ distinct clients can communicate with the server, where $\{M_r\}_{r\geqslant 0}$ are i.i.d. random variables with $\frac{1}{M} := \mathbb{E}(\frac{1}{M_r})$.*

Assumption 2 is more realistic and general than existing FL works, which usually assume $M_r = M$ is deterministic [40]. M_r is determined by the network and is not a design parameter: M_r is the number

[2] Ultimately, the algorithm $\mathcal{A}$ may output some $\widehat{w} \in \mathcal{W}$ that is a function of the client transcripts $(\mathbf{Z}_1, \cdots, \mathbf{Z}_R)$. By the post-processing property of DP [20, Proposition 2.1], the privacy of $\widehat{w}$ will be guaranteed if the client transcripts are DP. Thus, here we simply consider the output of $\mathcal{A}$ to be the client transcripts.

[3] Assume that client reports can be decrypted by the server, but not by the shuffler [21, 25].

of clients that are able to contribute to global updates in each round, which is not the same as the number of clients that a given algorithm queries in each round. For example, the one-pass sequential algorithms of [18, 65] query just $O(1)$ clients in each round, but require $M_r = N$ to implement since they dictate exactly which client(s) must communicate with the server in every round and do not allow any client to contribute more than one report.

Related Work and Our Contributions: Below we discuss the most relevant related work and describe our main contributions. See Appendix A for additional discussion of related work.

1. Tight minimax risk bounds for LDP Federated SCO with i.i.d. clients and reliable communication (Theorem D.1, Theorem 2.1, and Theorem 3.1): [18] studied a special case of the i.i.d. ($\mathcal{D}_i = \mathcal{D}$) FL problem with $n = 1$ and $M = N$, establishing minimax risk bounds for the class of *sequentially interactive* ϵ_0-LDP algorithms and convex, Lipschitz loss functions. We consider the (ϵ_0, δ_0)-LDP ($\delta_0 > 0$) i.i.d. FL problem in a more general setting where *each client has an arbitrary number of samples ($n \geqslant 1$)* and establish tight minimax risk bounds for two Lipschitz function classes–*strongly convex and convex*. Crucially, our minmax risk bounds hold with respect to a wider class of algorithms than that considered by [18]; in addition to sequentially interactive algorithms, our algorithm class also includes a broad subset of *fully interactive algorithms*. These risk bounds match (up to logarithms) the respective non-private rates if $\frac{d}{\epsilon_0^2} \lesssim n$ ("privacy for free").

2. The first non-trivial upper bound for LDP SCO with non-i.i.d. clients (Theorem 2.2): For the challenging problem of LDP FL (SCO) with *non-i.i.d.* client distributions, we develop an *accelerated* distributed noisy SGD algorithm based on [30], which runs in linear time and obtains the first non-trivial risk bound for smooth convex/strongly convex loss. Unsurprisingly, our non-i.i.d. bound does not match the i.i.d. minimax rate, leading to the open question of what the minimax rate is for non-i.i.d. LDP FL.

3. Improved communication complexity for LDP Federated ERM, plus matching lower bounds for a subset of LDP algorithms (Theorem D.2, Theorem 2.3, Section 3 and Theorem E.4): The special case of problem (2), which we refer to as LDP Federated ERM, has been extensively studied in recent years [70, 37, 36, 76, 73, 17, 78, 5, 64, 32], but only one of these works, [32], provides an algorithm that achieves a tight upper bound. [32] focuses primarily on the shuffled model of DP, but we observe that their algorithm also yields an (ϵ_0, δ_0)-LDP empirical risk bound for convex loss. We employ a variation of our accelerated LDP algorithm–combined with Nesterov smoothing [60] in the non-smooth case–to achieve the same risk bound as [32] in fewer rounds of server communication. We also address strongly convex ERM. Further, we provide matching lower bounds, implying that our algorithm is optimal among a subclass of fully interactive LDP algorithms.

4. Achieving the optimal CDP i.i.d. SCO rates without a trusted curator (Theorem 4.1): [32, 21] showed that the optimal CDP convex federated ERM rate [9] can be attained in the lower trust (relative to the central model) shuffle model of DP. The concurrent work [13, Theorem 4.9] shows that when all $M = N$ clients can communicate in every round and each client has just $n = 1$ sample, the optimal CDP SCO rate can be attained with an SDP algorithm. We show that our algorithm (with shuffling) attains the optimal CDP SCO rates for clients with any number $n \geqslant 1$ of samples. In particular, with shuffling, *our algorithm is simultaneously optimal in both the local and central models of DP for i.i.d. FL.* We also provide upper bounds for $M < N$.

In non-private distributed optimization, [51, 68, 58] provide convergence results with random connectivity graphs. Our upper bounds describe the effect of the mean/variance of $1/M_r$ on DP FL. Tight lower bounds for DP FL when $M < N$ is an open problem stemming from our work.

Finally, numerical experiments demonstrate the practical performance of our algorithm.

2 Upper Bounds for LDP FL

To simplify the presentation of our results, we will assume that $n_i = n$, $\epsilon_i = \epsilon_0$, $\delta_i = \delta_0$, and $L_i = L$ for all i. Appendix D contains the general versions of these upper bounds and their proofs.

Noisy Minibatch SGD for i.i.d. Clients: Consider the case of i.i.d. clients: $\mathcal{X}_i = \mathcal{X}$, $\mathcal{D}_i = \mathcal{D}$ for all i. We derive tight loss bounds via Algorithm 1 (Appendix D.2), an LDP version of distributed minibatch SGD (MB-SGD). In each round r, all M_r available clients send noisy stochastic gradients to the server: $\widetilde{g}_r^i := \frac{1}{K_i}\sum_{j=1}^{K_i} \nabla f(w_r, x_{i,j}^r) + u_i$, where $u_i \sim N(0, \sigma_i^2 \mathbf{I}_d)$ and $\{x_{i,j}\}_{j \in [K_i]}$ are drawn uniformly from X_i (and then replaced). The server averages these M_r reports, updates

$w_{r+1} := \Pi_{\mathcal{W}}[w_r - \frac{\eta_r}{M_r}\sum_{i\in S_r}\widetilde{g}_r^i]$ and reports w_{r+1} to all N clients. Algorithm 1 can also be seen as a distributed version of [8, Algorithm 1]. We now give privacy and excess population loss guarantees for Algorithm 1, run with $\sigma_i^2 := \frac{256L^2R\ln(2.5R/\delta_0)\ln(2/\delta_0)}{n^2\epsilon_0^2}$:

Theorem 2.1. [Informal] *Assume* $\epsilon_0 \leqslant \ln(2/\delta_0)$ *and choose* $K \geqslant \frac{\epsilon_0 n}{4\sqrt{2R\ln(2/\delta_0)}}$*. Then Algorithm 1 is* (ϵ_0, δ_0)*-LDP. Further, there exists* $\beta > 0$ *such that running Algorithm 1 on* $f_\beta(w) := \min_{v\in\mathcal{W}}\left(f(v) + \frac{\beta}{2}\|w-v\|^2\right)$ *yields:*

1. (Convex) Setting $R = M\min\left\{n, \frac{\epsilon_0^2 n^2}{d}\right\}$ *yields*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{LD}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(2/\delta_0)}}{\epsilon_0 n}\right)\right). \tag{4}$$

2. (μ*-strongly convex) Setting* $R = M\min\left\{n, \frac{\epsilon_0^2 n^2}{d}\right\}\ln\left(\frac{D^2\mu^2M^2\epsilon_0^2n^3}{dL^2}\right)$ *yields*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{L^2}{\mu M}\left(\frac{1}{n} + \frac{d\ln(2/\delta_0)}{\epsilon_0^2n^2}\right)\right). \tag{5}$$

We will see (Section 3) that these upper bounds are tight up to logarithmic factors when $M = N$, implying that Algorithm 1 is optimal among a large class of fully interactive LDP algorithms. Further, if $M = N$ and $\frac{d}{\epsilon_0^2} \lesssim n$, then both of these upper bounds match (up to logarithms) the respective non-private lower bounds for SCO ("privacy for free") [59, 3, 34]. The proof of Theorem 2.1 uses the corresponding smooth result Theorem D.1 in the Appendix and Nesterov smoothing [60]. Theorem D.1 is proved by bounding the empirical loss and using a *uniform stability* [11] argument, similar to how [8] proceeded for the case $N = 1$.

One-pass Accelerated Noisy Distributed SGD for Non-i.i.d. Clients: Consider the general **non-i.i.d.** FL problem, where $F_i(w)$ takes the general form (1) for some unknown distributions $\mathcal{D}_i$ on $\mathcal{X}_i$ ($i \in [N]$). The uniform stability approach that we used to obtain our i.i.d. upper bounds does not work in this setting. [4] Instead, we directly minimize F by modifying Algorithm 1 as follows:
1. We draw $K := 1$ local samples *without replacement* and set $R := n$. Thus, each sample is used at most one time during the algorithm, so that the bounds we obtain apply to F.
2. We use *acceleration* to increase the convergence rate.
3. To provide LDP, we use Gaussian *noise with larger variance* $\sigma^2 = \frac{8L^2\ln(1.25/\delta_0)}{\epsilon_0^2}$.
We call this algorithm **One-pass Accelerated Noisy Distributed SGD**. It is an instantiation of Accelerated Noisy MB-SGD, described in Algorithm 2 (Appendix D.4). Algorithm 2 is a noisy LDP version of the accelerated MB-SGD of [30], which was analyzed in the distributed setting by [75].

Theorem 2.2. [Informal] *Let* $f(\cdot, x)$ *be* β*-smooth for all* $x \in \mathcal{X}$*. Assume* $\epsilon_0 \leqslant 1$*. Then One-pass Accelerated Noisy Distributed SGD is* (ϵ_0, δ_0)*-LDP. Moreover:*
1. If $f(\cdot, x)$ *is convex, then*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = O\left(\frac{\beta D^2}{n^2} + \frac{LD\sqrt{d\ln(1/\delta_0)}}{\epsilon_0\sqrt{Mn}}\right). \tag{6}$$

2. If $f(\cdot, x)$ *is* μ*-strongly convex, then*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = O\left(LD\exp\left(-\sqrt{\frac{\mu}{\beta}}n\right) + \frac{L^2}{\mu}\frac{d\ln(1/\delta_0)}{\epsilon_0^2Mn}\right). \tag{7}$$

This upper bound is looser than the optimal bounds of Theorem D.1 (yet the tightest known bound for non-i.i.d. LDP FL), leaving open the question of what the optimal rate is for non-i.i.d. LDP FL.

Accelerated Noisy MB-SGD for Federated ERM: With non-random $M_r = M$, [32] provides an upper bound for convex LDP ERM that nearly matches the one we provide below.[5] We use an accelerated LDP algorithm, Algorithm 2 (Appendix D.4), which achieves the upper bounds for convex and strongly convex loss in fewer rounds of communications than [32]. Unlike the one-pass version of Algorithm 2 used for non-i.i.d. FL, here we sample local minibatches from each client *with replacement* to get tighter (in fact, optimal) bounds. Here we present just the non-smooth result:

[4] Specifically, Lemma D.1 in the Appendix does not apply without the i.i.d. assumption.

[5] The bound in [32] is looser than the bound in (35) by logarithmic factor.

Theorem 2.3. **[Informal]** *Assume $\epsilon_0 \leqslant \ln(2/\delta_0)$ and choose $K \geqslant \frac{\epsilon_0 n}{4\sqrt{2R\ln(2/\delta_0)}}$. Then Algorithm 2 is (ϵ_0, δ_0)-LDP. Further, there is $\beta > 0$ such that running Algorithm 2 on $f_\beta(w) := \min_{v \in \mathcal{W}} \left(f(v) + \frac{\beta}{2}\|w - v\|^2\right)$ yields the following bounds on the excess empirical loss (w.r.t. f):*

1. (Convex) Setting $R = \max\left(\frac{\sqrt{M}\epsilon_0 n}{\sqrt{d}}, \frac{\epsilon_0^2 n^2}{d}\begin{cases}\frac{1}{K} & \textit{if } M = N \\ 1 & \textit{otherwise}\end{cases}\right)$ yields

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{O}\left(LD\left(\frac{\sqrt{d\ln(2/\delta_0)}}{\epsilon_0 n\sqrt{M}}\right)\right). \tag{8}$$

2. (μ-strongly convex) Setting $R = \max\left(\frac{\sqrt{M}\epsilon_0 n}{\sqrt{d}}\ln\left(\frac{D\mu M\epsilon_0^2 n^2}{Ld}\right), \frac{\epsilon_0^2 n^2}{d}\begin{cases}\frac{1}{K} & \textit{if } M = N \\ 1 & \textit{otherwise}\end{cases}\right)$ yields

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{O}\left(\frac{L^2}{\mu}\left(\frac{d\ln(2/\delta_0)}{\epsilon_0^2 n^2 M}\right)\right). \tag{9}$$

The upper bounds in Theorem 2.3 match (up to logarithms) the lower bounds in Theorem E.4 when $M = N$, establishing the optimality of Algorithm 2 for convex/strongly convex Federated ERM among a wide subclass of fully interactive algorithms. The algorithm in [32] was not analyzed for random M_r. For fixed $M_r = M$, their algorithm requires $R = \widetilde{\Omega}(\epsilon_0^2 n^2 M/d)$ communication rounds to attain (8), making Algorithm 2 faster by a factor of $\min(\sqrt{M}\epsilon_0 n/\sqrt{d}, M)$.

3 Lower Bounds for LDP FL

We provide tight lower bounds on the excess population/empirical loss of LDP algorithms for the federated SCO/ERM problems when $M = N$. As a consequence, Algorithm 1 and Algorithm 2 are minimax optimal for i.i.d. SCO and ERM respectively, for two function classes: $\mathcal{F}_{L,D} := \{f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^d \mid \forall x \in \mathcal{X}\ f(\cdot, x)$ is convex, L-Lipschitz, and $\mathcal{W} \subseteq B_2(0, D)\}$; and $\mathcal{G}_{\mu,L,D} := \{f \in \mathcal{F}_{L,D} \mid \forall x \in \mathcal{X}\ f(\cdot, x)$ is μ-strongly convex$\}$. For SCO, the (ϵ_0, δ_0)-LDP algorithm class $\mathbb{A}_{(\epsilon_0, \delta_0)} = \mathbb{A}$ that we consider contains all sequentially interactive algorithms, as well as fully interactive algorithms that are compositional (c.f. [39]):

Definition 4. *Let $\mathcal{A}$ be an R-round (ϵ_0, δ_0)-LDP FL algorithm with data domain $\mathcal{X}$. Let $\{(\epsilon_r^0, \delta_r^0)\}_{r=1}^R$ denote the minimal (non-negative) parameters of the local randomizers $\mathcal{R}_r^{(i)}$ selected at round r ($r \in [R]$) such that $\mathcal{R}_r^{(i)}(\mathbf{Z}_{(1:r-1)}, \cdot) : \mathcal{X}^n \to \mathcal{Z}$ is $(\epsilon_r^0, \delta_r^0)$-DP for all $i \in [N]$ and all $\mathbf{Z}_{(1:r-1)} \in \mathcal{Z}^{(r-1)N}$. For an absolute constant $C > 0$, we say that $\mathcal{A}$ is C-compositional if $\sqrt{\sum_{r \in [R]}(\epsilon_0^r)^2} \leqslant C\epsilon_0$. If such a C exists, we simply say $\mathcal{A}$ is compositional.*

Any (ϵ_0, δ_0)-LDP $\mathcal{A}$ has $\epsilon_0^r \leqslant \epsilon_0$. The vast majority of (ϵ_0, δ_0)-LDP algorithms studied in the literature satisfy Definition 4. For example, any algorithm that uses the strong composition theorems of [20, Thm. 3.20] or [41] for its privacy analysis is 1-compositional. In particular, the three algorithms presented in Section 2 are 1-compositional, hence they are in $\mathbb{A}$. See also Appendix E.1.

Theorem 3.1. *Let $n, d, N, R \in \mathbb{N}$, $\epsilon_0 \in (0, \sqrt{N}]$, $\delta_0 \in (0, 1)$ and $\mathcal{A} \in \mathbb{A}_{(\epsilon_0, \delta_0)}$ such that in every round $r \in [R]$, the local randomizers $\mathcal{R}_r^{(i)}(\mathbf{Z}_{(1:r-1)}, \cdot) : \mathcal{X}^n \to \mathcal{Z}$ are $(\epsilon_0^r, \delta_0^r)$-DP for all $i \in [N]$, $\mathbf{Z}_{(1:r-1)} \in \mathcal{Z}^{r-1 \times N}$, with $\epsilon_0^r \leqslant \frac{1}{n}$, and $N \geqslant 16\ln(2/\delta_0^{\min} n)$, where $\delta_0^{\min} := \min_r \delta_0^r$. If $\mathcal{A}$ is if $\mathcal{A}$ is sequentially interactive, assume $\delta_0 = o(1/n^2N^2)$; if $\mathcal{A}$ is compositional, assume $\sum_r \delta_0^r = o(1/n^2N^2)$ instead. Then there exists a β-smooth ($\forall \beta \geqslant 0$) loss $f \in \mathcal{F}_{L,D}$ and a distribution $\mathcal{D}$ on a set $\mathcal{X}$ such that if the local data sets are drawn i.i.d. $X_i \sim \mathcal{D}^n$, then:*

$$\mathbb{E}F(\mathcal{A}(\mathbf{X})) - F(w^*) = \widetilde{\Omega}\left(LD\left(\frac{1}{\sqrt{Nn}} + \min\left\{1, \frac{\sqrt{d}}{\epsilon_0 n\sqrt{N}}\right\}\right)\right). \tag{10}$$

Furthermore, there exists another (μ-smooth) $f \in \mathcal{G}_{\mu,L,D}$ and distribution $\mathcal{D}$ such that

$$\mathbb{E}F(\mathcal{A}(\mathbf{X})) - F(w^*) = \widetilde{\Omega}\left(\frac{L^2}{\mu nN} + LD\min\left\{1, \frac{d}{\epsilon_0^2 n^2 N}\right\}\right). \tag{11}$$

These lower bounds are essentially tight [6] by Theorem 2.1. The first term in each of the lower bounds is the optimal non-private rate; the second parts of the bounds are what we prove in Appendix E.2.

[6] Up to logarithms, and for strongly convex case–a factor of $\mu D/L$. If $d > \epsilon_0^2 n^2 N$, then the trivial algorithm attains the matching upper bound $O(LD)$.

Theorem 3.1 is more generally applicable than the lower bound in [18, Proposition 3] (for the L_2 setting), which only applies to sequentially interactive algorithms and databases with $n = 1$ sample per client. As in [18], our lower bounds hold for sufficiently private LDP algorithms–those for which the privacy loss on each client in each round $\epsilon_0^r \lesssim 1/n$ [7]. For $n = 1$, $\epsilon_0^r \lesssim 1/n = 1$ is also required in [18, Proposition 3] since they consider sequential algorithms, where $\epsilon_0^r = \epsilon_0$. Also, the assumptions on δ_0, δ_0^r are not very restrictive in practice; see Remark E.1. For federated ERM, we prove a tight lower bound in Theorem E.4, establishing the optimality of Algorithm 2 for a class $\mathbb{B} \subset \mathbb{A}$.

4 Optimal Algorithm for Shuffle DP FL

Assume access to a secure shuffler and fix $M_r = M \in [N]$. In each round r, the shuffler receives the reports $(Z_r^{(1)}, \cdots Z_r^{(M)})$ from active clients (we assume $S_r = [M]$ here for concreteness), draws a uniformly random permutation of $[M]$, π, and then sends $(Z_r^{(\pi(1))}, \cdots, Z_r^{(\pi(M))})$ to the server for aggregation. We show that this "shuffled version" of Algorithm 1 achieves the optimal convex/strongly convex CDP bounds for i.i.d. SCO when $M = N$, even with the weaker trust assumptions of the shuffle model:

Theorem 4.1. *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^d$ be β-smooth, $\epsilon \leqslant \ln(2/\delta)$, $\delta \in (0,1)$, and $M \geqslant 16\ln(18RM^2/N\delta)$ for R specified below. Then $\exists C > 0$ such that $\sigma_i^2 := \frac{CL^2RM\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}{n^2N^2\epsilon^2}$ ensures that the shuffled Algorithm 1 is (ϵ, δ)-CDP $\forall K \in [n]$. Further:*

1. (Convex) Setting $R := \max\left(\frac{n^2N^2\epsilon^2}{M}, \frac{N}{M}, \min\left\{n, \frac{\epsilon^2 n^2 N^2}{dM}\right\}, \frac{\beta D}{L}\min\left\{\sqrt{nM}, \frac{\epsilon n N}{\sqrt{d}}\right\}\right)$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = O\left(LD\left(\frac{1}{\sqrt{nM}} + \frac{\sqrt{d\ln(1/\delta)}}{\epsilon n N}\right)\right). \tag{12}$$

2. (Strongly convex) $R := \max\left(\frac{n^2N^2\epsilon^2}{M}, \frac{N}{M}, \frac{8\beta}{\mu}\ln\left(\frac{\beta D^2 \mu \epsilon^2 n^2 N^2}{dL^2}\right), \min\left\{n, \frac{\epsilon^2 n^2 N^2}{dM}\right\}\right)$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{L^2}{\mu}\left(\frac{1}{nM} + \frac{d\ln(1/\delta)}{\epsilon^2 n^2 N^2}\right)\right). \tag{13}$$

Together with our LDP results, Theorem 4.1 implies that, with shuffling, *Algorithm 1 is simultaneously optimal for i.i.d. FL in the local and central models of DP* if $M = N$. Nesterov smoothing can be used to obtain the same bounds without the β-smoothness assumption, similar to how we proceeded for Theorem 2.1. We omit the details here.

5 Numerical Experiments

Linear Regression with Health Insurance Data: We divide the data set [15] into N heterogeneous groups based on the level of the target (medical charges). See Appendix G.1 for additional details. Fig. 2 shows that *LDP MB-SGD outperforms LDP Local SGD across all privacy levels*. Also, as $\epsilon \uparrow 10$, the price of LDP shrinks towards 0. Here, υ_*^2 describes the amount of heterogeneity between clients' data (increasing with heterogeneity); see Appendix D.1 for the precise definition.

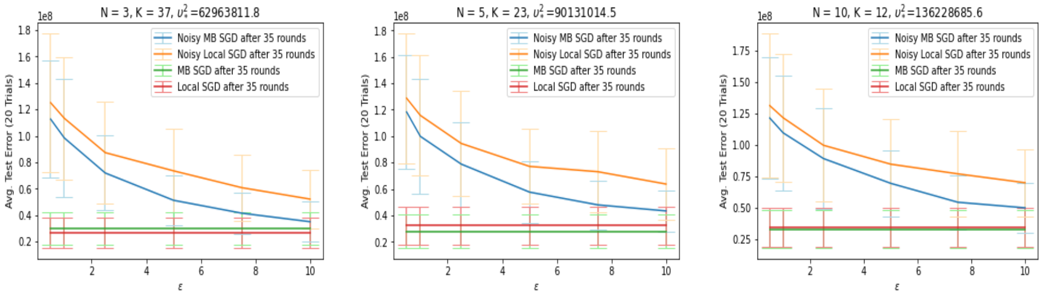

Figure 2: Test error vs. ϵ for linear regression on heterogeneous health insurance data. We display 90% error bars over the 20 trials (train/test splits). $\delta = 1/n^2$; $n = 1070/N$ is the number of training examples per client.

Logistic Regression with MNIST: See Appendix G for experiments with $M < N$. Algorithm 1 still uniformly outperform LDP Local SGD and even outperforms *non-private* Local SGD for $\epsilon \geqslant 12.5$.

[7] If there are $N = O(1)$ clients, then existing CDP lower bounds [8] apply and match our LDP upper bounds as long as $\epsilon_0 \lesssim 1$; thus, this restriction on ϵ_0^r disappears if $N = O(1)$.

Acknowledgements

We would like to thank Tianjian Huang, Dmitrii Ostrovskii, and Adam Smith for helpful comments and conversations.

References

[1] Webank and swiss re signed cooperation mou. *Fed AI Ecosystem*, Nov 2019.

[2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.

[3] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[4] Apple. Private federated learning. *NeurIPS 2019 Expo Talk Abstract*, 2019.

[5] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842, 2019.

[6] B. Balle, J. Bell, A. Gascón, and K. Nissim. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, pages 638–667. Springer, 2019.

[7] B. Balle, P. Kairouz, B. McMahan, O. D. Thakkar, and A. Thakurta. Privacy amplification via random check-ins. 33, 2020.

[8] R. Bassily, V. Feldman, K. Talwar, and A. Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, 2019.

[9] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[10] A. Bittau, U. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the Symposium on Operating Systems Principles (SOSP)*, pages 441–459, 2017.

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

[12] Y.-R. Chen, A. Rezapour, and W.-G. Tzeng. Privacy-preserving ridge regression on distributed data. *Information Sciences*, 451:34–49, 2018.

[13] A. Cheu, M. Joseph, J. Mao, and B. Peng. Shuffle private stochastic convex optimization. *arXiv preprint arXiv:2106.09805*, 2021.

[14] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 375–403. Springer, 2019.

[15] M. Choi. Medical insurance charges data. 2018. `https://www.kaggle.com/mirichoi0218/insurance`.

[16] P. Courtiol, C. Maussion, M. Moarii, E. Pronier, S. Pilcer, M. Sefta, P. Manceron, S. Toldo, M. Zaslavskiy, and N. Le Stang. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, page 1–7, 2019.

[17] R. Dobbe, Y. Pu, J. Zhu, K. Ramchandran, and C. Tomlin. Customized local differential privacy for multi-agent distributed optimization, 2020.

[18] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[19] C. Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

[20] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. 2014.

[21] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, S. Song, K. Talwar, and A. Thakurta. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. *arXiv preprint arXiv:2001.03618*, 2020.

[22] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity, 2020.

[23] M. S. Exchange. Total variation distance of two random vectors whose components are independent. `https://math.stackexchange.com/questions/1558845/total-variation-distance-of-two-random-vectors-whose-components-are-independent.`

[24] V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[25] V. Feldman, A. McMillan, and K. Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling, 2020.

[26] V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279, Phoenix, USA, 25–28 Jun 2019. PMLR.

[27] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[28] S. Gade and N. H. Vaidya. Privacy-preserving distributed learning via obfuscated stochastic gradients. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 184–191. IEEE, 2018.

[29] R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. *CoRR*, abs/1712.07557, 2017.

[30] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[31] S. Ghadimi and G. Lan. 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.

[32] A. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. Theertha Suresh. Shuffled model of differential privacy in federated learning. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2521–2529. PMLR, 13–15 Apr 2021.

[33] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR.

[34] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[35] Z. He, T. Zhang, and R. B. Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 148–162, 2019.

[36] Z. Huang and Y. Gong. Differentially private ADMM for convex distributed learning: Improved accuracy via multi-step approximation. *arXiv preprint:2005.07890*, 2020.

[37] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong. DP-ADMM: ADMM-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012, 2020.

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

[39] M. Joseph, J. Mao, S. Neel, and A. Roth. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 94–105. IEEE, 2019.

[40] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *arXiv preprint:1912.04977*, 2019.

[41] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy, 2015.

[42] G. Kamath. Cs 860: Algorithms for private data analysis, 2020. `http://www.gautamkamath.com/CS860notes/lec5.pdf`.

[43] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.

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

[45] A. Khaled, K. Mishchenko, and P. Richtárik. Better communication complexity for local SGD. *arXiv preprint*, 2019.

[46] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized SGD with changing topology and local updates. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5381–5393. PMLR, 13–18 Jul 2020.

[47] D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh. Learning with user-level privacy. *arXiv preprint arXiv:2102.11845*, 2021.

[48] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of FedAvg on non-iid data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[49] J. Liu and K. Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

[50] R. Liu, Y. Cao, H. Chen, R. Guo, and M. Yoshikawa. Flame: Differentially private federated learning in the shuffle model. In *AAAI*, 2020.

[51] I. Lobel and A. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306, 2010.

[52] A. Lowy and M. Razaviyayn. Output perturbation for differentially private convex optimization with improved population loss bounds, runtimes and applications to private adversarial training. *arXiv preprint:2102.04704*, 2021.

[53] X. Ma, F. Zhang, X. Chen, and J. Shen. Privacy preserving multi-party computation delegation for deep learning in cloud computing. *Information Sciences*, 459:103–116, 2018.

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

[55] B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018.

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

[57] J. Murtagh and S. Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2016.

[58] A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[59] A. S. Nemirovskii and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. 1983.

[60] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[61] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys and Tutorials*, page 1–1, 2021.

[62] N. Papernot and T. Steinke. Hyperparameter tuning with renyi differential privacy, 2021.

[63] S. Pichai. Google's Sundar Pichai: Privacy should not be a luxury good. *The New York Times*, May 2019.

[64] M. Seif, R. Tandon, and M. Li. Wireless federated learning with local differential privacy. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2604–2609, 2020.

[65] A. Smith, A. Thakurta, and J. Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77, 2017.

[66] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi. Analyzing user-level privacy attack against federated learning. *IEEE Journal on Selected Areas in Communications*, 38(10):2430–2444, 2020.

[67] S. U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint:1907.04232*, 2019.

[68] B. Touri and B. Gharesifard. Continuous-time distributed convex optimization on time-varying directed networks. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 724–729, 2015.

[69] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 1–11, 2019.

[70] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei. LDP-Fed: federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, page 61–66. Association for Computing Machinery, 2020.

[71] J. Ullman. CS7880: rigorous approaches to data privacy, 2017. `http://www.ccs.neu.edu/home/jullman/cs7880s17/HW1sol.pdf`.

[72] K. Wei, J. Li, M. Ding, C. Ma, H. Su, B. Zhang, and H. V. Poor. User-level privacy-preserving federated learning: Analysis and performance optimization. *arXiv preprint:2003.00229*, 2020.

[73] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[74] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

[75] B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6281–6292. Curran Associates, Inc., 2020.

[76] N. Wu, F. Farokhi, D. Smith, and M. A. Kaafar. The value of collaboration in convex machine learning with differential privacy, 2019.

[77] H. Yuan and T. Ma. Federated accelerated stochastic gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5332–5344. Curran Associates, Inc., 2020.

[78] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam. Local differential privacy based federated learning for internet of things. *IEEE Internet of Things Journal*, 2020.

[79] Y. Zhou and S. Tang. Differentially private distributed learning. *INFORMS Journal on Computing*, 32(3):779–789, 2020.

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

Appendix

A Further Discussion of Related Work

In the absence of differential privacy constraints, federated learning has received a lot of attention from researchers in recent years. Among these, the most relevant works to us are [46, 48, 43, 74, 75, 77], which have proved bounds on the convergence rate of federated learning algorithms. From an algorithmic standpoint, all of these works propose and analyze either Minibatch SGD (MB-SGD), FedAvg/Local SGD [54], or an extension or accelerated/variance-reduced variation of one of these. Notably, [75] proves tight upper and lower bounds that establish the near optimality of accelerated MB-SGD for the heterogeneous SCO problem with non-random $M_r = M = N$ in a fairly wide parameter regime.

More recently, there have been many proposed attempts to ensure the privacy of individuals' data during and after the federated learning process. Some of these have used secure multi-party computation (MPC) [12, 53], but this approach leaves users vulnerable to inference attacks on the trained model and does not provide the rigorous guarantee of DP. Others [55, 29, 38, 28, 72, 79] have used client-level DP or global DP (CDP), which rely on a trusted third party, or hybrid DP/MPC approaches [38, 69]. The work of [38] is particularly relevant in that they prove CDP empirical risk bounds and high probability guarantees on the population loss when the data is i.i.d. across clients. However they do not address LDP, non-i.i.d. FL, or provide expected excess loss bounds for i.i.d. FL. It is also worth mentioning that [29] considers random M_r but does not prove any bounds.

Despite this progress, prior to our present work, far less was known about the convergence rate and excess risk potential of LDP FL algorithms. The only exceptions are in the two extreme corner cases of $N = 1$ and $n = 1$. When $N = 1$, LDP and CDP are essentially equivalent; tight ERM [9] and i.i.d. SCO [8, 24] bounds are known for this case. In addition, for the special case of pure LDP i.i.d. SCO when $n = 1$ and $M_r = N$ is fixed, [18] establishes the minimax optimal rate for the class of sequentially interactive algorithms and convex loss functions. To the best of our knowledge, all works examining the general LDP FL problem with arbitrary $n, M, N \geqslant 1$ either focus on ERM and/or do not provide excess risk bounds that scale with both M and n_i. Furthermore, none provide communication/gradient complexity guarantees, lower bounds, or bounds for random M_r. We discuss each of these works in turn below:

[70] gives an LDP FL algorithm but no risk bounds.

[37] and [36] use LDP ADMM algorithms for smooth convex Federated ERM. However, their utility bounds are stated in terms of an average of the client functions evaluated at different points, so it is not clear how to relate their result to the standard performance measure for learning (which we consider in this paper): expected excess risk at the point $\widehat{w}$ output by the algorithm.

[76, Theorem 2] provides an $\{(\epsilon_i, 0)\}_{i=1}^N$-LDP ERM bound for fixed $M_r = M = N$ of $O\left(\kappa \frac{L^2}{\mu} \frac{d}{\widetilde{N}\epsilon^2} + \epsilon\right)$ for μ-strongly convex, β-smooth f with condition number $\kappa = \beta/\mu$, $\widetilde{N} = \sum_{i=1}^N n_i$, and $1/\epsilon^2$ is an average of $1/\epsilon_i^2$. The additive ϵ term is clearly problematic: e.g. if $\epsilon = 1$, then the bound becomes trivial. Ignoring this term, the first term in their bound is still looser than the bound that we provide in Theorem 2.3. Namely, our bound in part 2 of Theorem D.2 is tighter by a factor of $O\left(\frac{\ln(1/\delta)}{\kappa n}\right)$. Additionally, the bounds in [76] require R "large enough" and do not come with communication complexity guarantees. In the convex case, the LDP ERM bound reported in [76, Theorem 3] is not interpretable because the unspecified "constants" in the upper bound on $\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*)$ are said to be allowed to depend on R.

[73, Theorems 2 and 3] provide convergence rates for smooth PL convex LDP ERM, which are complicated non-monotonic functions of R. Since they do not prescribe a choice of R, it is unclear what excess risk and communication complexity bounds are attainable with their algorithm. In particular, they do not prove any excess risk bounds.

[17] proposes LDP Inexact Alternating Minimization Algorithm (IAMA) with Laplace noise and [17, Theorem 3.11] gives a convergence rate for smooth, strongly convex LDP FL of order $O\left(\frac{\Theta \sum_{i \in [M]} \sigma_i^2}{R}\right)$ ignoring smoothness and strong convexity factors, where Θ is a parameter that they

only provide an upper bound for in special cases (e.g. quadratic objective). Thus, their bounds are not complete for general strongly convex loss functions. Even in the special cases where they do provide a bound for Θ, our bounds are much tighter. Assuming that $\Theta = 1$ and (for simplicity of exposition) that parameters are the same across clients, their Theorem 3.1 implies taking $\sigma^2 = 1/\epsilon^2$ to ensure $(\epsilon, 0)$-LDP. The resulting convergence rate is then $O(M/\epsilon^2)$, which does not scale with n_i and is *increasing* with M. Also, the dependence of their rate on the dimension d is unclear, as it does not appear explicitly in their theorem. [8] Ignoring this issue, their bound (particularly the dependence on M and n_i) is clearly much looser than all of the excess risk bounds we give in the present work.

[78] and [5] apply the LDP FL framework to Internet of Things, and [64] uses noisy (deterministic) GD for LDP wireless channels in the FL (smooth strongly convex) ERM setting. Their bounds do not scale with the number of data points n_i, however (only with the number of clients N). Therefore, our bounds are much tighter, and apply to general convex FL problems besides wireless channels.

B Assumption that $p_i = 1/N$ for all $i \in [N]$

The assumption that $p_i = \frac{1}{N}$ for all $i \in [N]$ in (2) is without loss of generality by considering the transformation $\widetilde{F}_i(w) := p_i N F_i(w)$. Then $F(w) = \sum_{i=1}^N p_i F_i(w) = \frac{1}{N}\sum_{i=1}^N \widetilde{F}_i(w)$, so our results for $p_i = 1/N$ apply for general p_i with L_i replaced by $\widetilde{L_i} := p_i N L_i$, μ_i replaced by $\widetilde{\mu_i} := p_i N \mu_i$, and β_i gets replaced by $\widetilde{\beta_i} := p_i N \beta_i$. In particular, if $\mathcal{X}_i = \mathcal{X}$ for all i (as we assume for our non-ERM results), then L gets replaced with $\widetilde{L} = \max_{i \in [N]} p_i N L$, μ gets replaced with $\widetilde{\mu} = \max_{i \in [N]} p_i N \mu$, β gets replaced with $\widetilde{\beta} = \max_{i \in [N]} p_i N \beta$.

C Relationships between notions of DP

C.1 LDP is stronger than CDP

Assume $\mathcal{A}$ is (ϵ_0, δ_0)-LDP. Let $\mathbf{X}, \mathbf{X}'$ be adjacent databases in the CDP sense; i.e. there exists a unique $i \in [N]$, $j \in [n_i]$ such that $x_{i,j} \neq x'_{i,j}$. Then for all $r \in [R]$, $l \neq i$, $X_l = X'_l$, so the conditional distributions of $\mathcal{R}_r^{(l)}(\mathbf{Z}_{1:r-1}, X_l)$ and $\mathcal{R}_r^{(l)}(\mathbf{Z}'_{1:r-1}, X'_l)$ given $Z_{1:r-1}^{(l' \neq l)} = z_{1:r-1}^{(l' \neq l)}$ are identical for all $z_{1:r-1}^{(l' \neq l)} \in \mathcal{Z}^{r-1 \times N-1}$. Integrating both sides of this equality with respect to the joint density of $Z_{1:r-1}^{(l' \neq l)}$ shows that $\mathcal{R}_r^{(l)}(\mathbf{Z}_{1:r-1}, X_l) = \mathcal{R}_r^{(l)}(\mathbf{Z}'_{1:r-1}, X'_l)$ (unconditional equality of distributions). Hence the full transcript of client l is (unconditionally) $(0,0)$-CDP for all $l \neq i$. A similar argument (using the inequality (3) instead of equality) shows that client i's full transcript is unconditionally (ϵ_0, δ_0)-CDP. Therefore, by the basic composition theorem for DP [20], the full combined transcript of all N clients is (ϵ_0, δ_0)-CDP, which implies that $\mathcal{A}$ is (ϵ_0, δ_0)-CDP.

Conversely, (ϵ, δ)-CDP does not imply (ϵ', δ')-LDP for any ϵ', δ'. This is because a CDP algorithm may send non-private updates to the server and rely on the server to randomize, completely violating the requirement of LDP.

C.2 LDP is "stronger" than client-level DP

Precisely, we claim that if $\mathcal{A}$ is (ϵ_0, δ_0)-LDP then $\mathcal{A}$ is $(n\epsilon, ne^{(n-1)\epsilon}\delta)$ client-level DP; but conversely (ϵ, δ)-client-level DP does not imply (ϵ', δ')-LDP for any ϵ', δ'. The first part of the claim is due to group privacy [42, Theorem 10] (and the argument used above in Appendix C.1 to get rid of the "conditional"). The second part of the claim is true because a client-level DP algorithm may send non-private updates to the server and rely on the server to randomize, completely violating the requirement of LDP.

[8] Note that in order for their result to be correct, by [9, Theorem 5.4] when $N = M = 1$, their bound must scale at least as $d^2/\epsilon^2 n^2$, unless their bound is trivial ($\geqslant LD$).

D Complete Versions and Proofs of Upper Bounds for LDP FL

D.1 Notation and assumptions for stating the complete versions of our upper bounds

We will require the following additional notations and assumptions to state the complete, general forms of our upper bound theorems.

Assumption 3. *In each round r, a uniformly random subset S_r of $M_r \in [N]$ distinct clients can communicate with the server, where $\{M_r\}_{r \geqslant 0}$ are i.i.d. random variables with $\frac{1}{M} := \mathbb{E}(\frac{1}{M_r})$ and $\frac{1}{M'} := \sqrt{\mathbb{E}\left(\frac{1}{M_r^2}\right)}$.*

Assumption 4. *For all $i \in [N]$:*

1. *$\mathbb{E}_{x_i \sim \mathcal{D}_i} \|\nabla f(w, x_i) - \nabla F_i(w)\|^2 \leqslant (\phi_i)^2$ for all $w \in \mathcal{W}$, or:*
2. *$\mathbb{E}_{x_i \sim \mathcal{D}_i} \|\nabla f(w^*, x_i) - \nabla F_i(w^*)\|^2 \leqslant (\phi_i^*)^2$ for any $w^* \in \operatorname{argmin}_{w \in \mathcal{W}} F(w)$.*

Assumption 4 is standard in FL (e.g. [48, 43, 75]).

For $M \in [N]$, denote $\overline{\phi_M^2} := \frac{1}{M}\sum_{i=1}^{M} \phi_{(i)}^2$, where $\phi_{(1)} := \phi_{\max} := \max_{i \in [N]} \phi_i \geqslant \phi_{(2)} \geqslant \cdots \geqslant \phi_{(N)} := \phi_{\min} := \min_{i \in [N]} \phi_i$, and define $(\overline{\phi_M^*})^2$ similarly. More generally, whenever a bar and M subscript are appended to a parameter in this paper, it denotes the average of the M largest values. Also, define $\Phi^2 := \sqrt{\mathbb{E}[(\overline{\phi_{M_1}^2})^2]}$ and $\Sigma^2 := \sqrt{\mathbb{E}(\overline{\sigma_{M_1}^2})^2}$ for any $\{\sigma_i^2\}_{i=1}^N \subseteq [0, \infty)$.

Next, define the heterogeneity parameters $\upsilon_*^2 := \frac{1}{N}\sum_{i=1}^N \|\nabla F_i(w^*)\|^2$ and $\upsilon^2 := \sup_{w \in \mathcal{W}} \frac{1}{N}\sum_{i=1}^N \|\nabla F_i(w) - \nabla F(w)\|^2$, which have appeared in [43, 45, 46, 75]. If the data is homogeneous, then all F_i share the same minimizers, so $\upsilon_*^2 = 0$, but the converse is false. Also, $\upsilon^2 = 0$ iff $F_i = F + a_i$ for constants $a_i \in \mathbb{R}$, $i \in [N]$ (homogeneous up to translation). Denote $\Upsilon_*^2 := \begin{cases} \left(1 - \frac{M-1}{N-1}\right)\upsilon_*^2 & \text{if } N > 1 \\ 0 & \text{otherwise} \end{cases}$ and its counterpart Υ^2, defined similarly but with υ^2.

In the unbalanced data case, we will assume that each client uses proportionally sized local minibatches, i.e. that $K_i/n_i = K_j/n_j$ for all i, j. Also, we denote $K := \min_{i \in [N]} K_i$.

Lastly, for given parameters, denote $\psi_i := \left(\frac{L_i}{n_i \epsilon_i}\right)^2 \ln(2.5R/\delta_i)\ln(2/\delta_i)$ for $i \in [N]$, $\Psi := \sqrt{\mathbb{E}_{M_1}\left(\frac{1}{M_1}\sum_{i=1}^{M_1} \psi_{(i)}\right)^2}$, $\xi_i := \psi_i / L_i^2$, and $\Xi := \sqrt{\mathbb{E}_{M_1}\left(\frac{1}{M_1}\sum_{i=1}^{M_1} \xi_{(i)}\right)^2}$.

D.2 Statement and proof of upper bounds for smooth i.i.d. SCO

We first describe the noisy minibatch SGD algorithm that we use:

Algorithm 1 Noisy MB-SGD

```
Require: Number of clients $N \in \mathbb{N}$, dimension $d \in \mathbb{N}$ of data, noise parameters $\{\sigma_i\}_{i \in [N]}$, data sets $X_i \in \mathcal{X}_i^{n_i}$ for $i \in [N]$, convex loss function $f(w, x)$, number of communication rounds $R \in \mathbb{N}$, number $\{K\}_{i=1}^N \subset \mathbb{N}$ of local samples drawn per round, step sizes $\{\eta_r\}_{r \in [R]}$ and weights $\{\gamma_r\}_{r \in [R]}$.
Initialize $w_0 = 0$.
for $r \in \{0, 1, \cdots, R-1\}$ do
    for $i \in S_r$ do in parallel
        Server sends global model $w_r$ to client $i$.
        Client $i$ draws $K_i$ samples $x_{i,j}^r$ (uniformly with replacement) from $X_i$ (for $j \in [K_i]$) and noise $u_i \sim N(0, \sigma_i^2 \mathbf{I}_d)$.
        Client $i$ computes $\widetilde{g}_r^i := \frac{1}{K_i} \sum_{j=1}^{K_i} \nabla f(w_r, x_{i,j}^r) + u_i$ and sends to server.
    end for
    Server aggregates $\widetilde{g}_r := \frac{1}{M_r} \sum_{i \in S_r} \widetilde{g}_r^i$.
    Server updates $w_{r+1} := \Pi_{\mathcal{W}}[w_r - \eta_r \widetilde{g}_r]$.
end for
return $\widehat{w}_R = \frac{1}{\Gamma_R} \sum_{r=0}^{R-1} \gamma_r w_r$, where $\Gamma_R := \sum_{r=0}^{R-1} \gamma_r$.
```

Here is the informal guarantee of Algorithm 1 for smooth losses:

Theorem D.1. [Informal] *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^d$ be L-Lipschitz, β-smooth, and μ-strongly convex (with $\mu = 0$ for convex case). Assume $\epsilon_0 \leqslant \ln(2/\delta_0)$ and choose $K \geqslant \frac{\epsilon_0 n}{4\sqrt{2R\ln(2/\delta_0)}}$. Then Algorithm 1 is (ϵ_0, δ_0)-LDP. Moreover, there exist choices of $\eta_r = \eta$ and $\{\gamma_r\}_{r=0}^{R-1}$ such that the output $\widehat{w}_R = \sum_{r=0}^{R-1} \gamma_r w_r$ of Algorithm 1 achieves the following upper bounds on excess loss:*
1. (Convex) Setting $R = \max\left(\frac{\beta D\sqrt{M}}{L} \min\left\{\sqrt{n}, \frac{\epsilon_0 n}{\sqrt{d}}\right\}, \min\left\{n, \frac{\epsilon_0^2 n^2}{d}\right\}/K\right)$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = O\left(\frac{LD}{\sqrt{M}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(2.5R/\delta_0)\ln(2/\delta_0)}}{\epsilon_0 n}\right)\right). \tag{14}$$

2. (Strongly convex) Setting $R = \max\left(\frac{8\beta}{\mu}\ln\left(\frac{\beta D^2 \mu M \epsilon_0^2 n^2}{dL^2}\right), \min\left\{n, \frac{\epsilon_0^2 n^2}{d}\right\}/K\right)$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{L^2}{\mu M}\left(\frac{1}{n} + \frac{d\ln(2.5R/\delta_0)\ln(2/\delta_0)}{\epsilon_0^2 n^2}\right)\right). \tag{15}$$

We get rid of the restriction on β that appears in [8, Theorem 3.2] for $N = 1$, non-strongly convex loss by using a different (smaller, when $N = 1$) step size. This allows us to extend our convex upper bound to non-smooth functions in the distributed setting (Theorem 2.1) via Nesterov smoothing [60], like [8] did for $N = 1$.

We now state the fully general version of Theorem D.1 for arbitrary $n_i, \epsilon_i, \delta_i$.

Theorem D.1 [Complete Version] *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be convex, L-Lipschitz, and β-smooth in w for all $x \in \mathcal{X}$, where $\mathcal{W}$ is a closed convex set in $\mathbb{R}^d$ s.t. $\|w\| \leqslant D$ for all $w \in \mathcal{W}$. Let $\mathcal{D}$ be an arbitrary probability distribution on $\mathcal{X}$. For each client $i \in [N]$, draw a local i.i.d. data set $X_i \sim \mathcal{D}^{n_i}$. Run Algorithm 1 and $\sigma_i^2 = \frac{256 L^2 R \ln(\frac{2.5R}{\delta_i})\ln(2/\delta_i)}{n_i^2 \epsilon_i^2}$. Then the algorithm is $\{(\epsilon_i, \delta_i)\}_{i \in [N]}$-LDP for any $\epsilon_i \in (0, \ln(\frac{2}{\delta_i})]$ and $\delta_i \in (0, 1)$. Moreover, we get the following excess population loss bounds:*
1. (Convex): Choose $\gamma_r = \frac{1}{R}$ and constant step-size $\eta = \min\{1/4\beta, \widetilde{\eta}\}$, where $\widetilde{\eta} = \frac{D\sqrt{\widetilde{M}}}{LR}\min\left\{\sqrt{n_{\min}}, \frac{L}{\sqrt{d\widetilde{\Psi}}}\right\}$, where $\widetilde{M} := \begin{cases}\sqrt{M'} & \text{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \sqrt{M} & \text{otherwise}\end{cases}$ and $\widetilde{\Psi} := \begin{cases}\Psi & \text{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \psi_{\max} & \text{otherwise}\end{cases}$. Similarly, denote $\widetilde{\Xi} := \widetilde{\Psi}/L^2$. Then,

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) \leqslant 272\frac{LD}{\sqrt{\widetilde{M}}}\max\left\{\frac{1}{\sqrt{n_{\min}}}, \sqrt{d\widetilde{\Xi}}\right\}$$

in $R := \left\lceil \max\left(\frac{\beta D \sqrt{\widetilde{M}}}{L} \min\left\{ \sqrt{n_{\min}}, \frac{1}{\sqrt{d\widetilde{\Xi}}} \right\}, \min\left\{ n_{\min}, \frac{1}{d\widetilde{\Xi}} \right\} / K \right) \right\rceil$ *rounds of communication.*

2. (Strongly Convex) Assume additionally that $f(\cdot, x)$ is μ-strongly convex for all $x \in \mathcal{X}$. Choose the following constant step-size and averaging weights:

$$\eta_r = \eta = \min\left\{ \frac{1}{4\beta}, \frac{\ln\left(\max\left\{ 2, \frac{\mu^2 D^2 R^2}{2\left(\frac{4L^2}{MK} + d \min\left\{ \frac{\Sigma^2}{M'}, \frac{\sigma^2_{\max}}{M} \right\} \right)} \right\} \right)}{\mu R} \right\}, \quad \gamma_r = (1 - \mu\eta)^{-(r+1)}.$$

Then choosing $R = \left\lceil \begin{cases} \max\left\{ \frac{8\beta}{\mu} \ln\left(\frac{\beta D^2 \mu M'}{d\Psi} \right), \frac{L^2}{d\Psi} \right\} & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \max\left\{ \frac{8\beta}{\mu} \ln\left(\frac{\beta D^2 \mu M}{d\psi_{\max}} \right), \frac{L^2}{d\psi_{\max}} \right\} & \textit{otherwise} \end{cases} \right\rceil$ *implies*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{L^2}{\mu} \left(\frac{1}{Mn_{\min}} + d \min\left\{ \frac{\Xi}{M'}, \frac{\xi_{\max}}{M} \right\} \right) \right). \tag{16}$$

Remark D.1. *Note that $1/M' \geqslant 1/M$ by Cauchy-Schwartz. Both of the upper bounds in the complete version of Theorem D.1 involve minima of the terms Ξ/M' and $\xi_{\max}/M$, which trade off the unbalancedness of client data and privacy needs with the variance of $1/M_r$. In particular, if the variance of $1/M_r$ is small enough that $\frac{\Xi}{M'} \leqslant \frac{\xi_{\max}}{M}$, then the communication complexity and excess risk bounds in the complete version of Theorem D.1 depend on averages of the parameters across clients, rather than maximums. In FL problems with unbalanced/heterogeneous data and disparate privacy needs across a large number of clients, the difference between "average" and "max" can be substantial. On the other hand, if data is balanced and privacy needs are the same across clients, then $\xi_i = \xi_{\max} = \Xi = \ln(2.5R/\delta_0)\ln(2/\delta_0)/n^2\epsilon_0^2$ for all i and $\frac{\Xi}{M'} \geqslant \frac{\xi_{\max}}{M}$, so we recover the informal version of Theorem D.1 stated in the main body, with dependence only on the mean of $1/M_r$ and not the second moment.*

Remark D.2. *Increasing K allows for a smaller choice of R needed to attain the excess risk bounds in Theorem D.1. On the other hand, as K increases, the overall gradient complexity ($= KR$ per client) may still increase. Thus, there is a trade-off between computational complexity and communication complexity. Depending on the particular problem and whether computation or communication is more of a bottleneck, K can be tuned to minimize runtime. Also, in Theorem D.1, the dependence of our communication complexity bounds on the number of active clients M is favorable: in the convex case, we have a $\sqrt{M}$ dependence and for strongly convex, it is logarithmic (yet excess risk decreases linearly in M). This is an attractive feature of Algorithm 1 for large-scale FL problems.*

Remark D.3. *The alternative form of noise in [2] with $\sigma^2 = \frac{8L^2 R \ln(1/\delta_0)}{n^2\epsilon_0^2}$ can be used to eliminate the $\ln(R/\delta_0)$ factor in the above risk bounds. However, this choice of noise would require $K \approx \max\{1, \sqrt{\frac{\epsilon_0}{4R}} n\}$ to ensure LDP, providing less flexibility. This remark applies verbatim to the rest of the LDP upper bounds in this paper with the exception of Theorem 2.2.*

We will require some preliminaries before we move to the proof of Theorem D.1. We begin with the following definition from [11]:

Definition 5. *(Uniform Stability) A randomized algorithm $\mathcal{A} : \mathcal{W} \times \mathcal{X}^{\widetilde{N}}$ is said to be α-uniformly stable (w.r.t. loss function $f : \mathcal{W} \times \mathcal{X}$) if for any pair of adjacent data sets $X, X' \in \mathcal{X}^{\widetilde{N}}$, $|X \Delta X'| \leqslant 2$, we have*

$$\sup_{x \in \mathcal{X}} \mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(X), x) - f(\mathcal{A}(X'), x)] \leqslant \alpha.$$

The following lemma, which is well-known, allows us to easily pass from empirical risk to population loss when the algorithm in question is uniformly stable:

Lemma D.1. *Let $\mathcal{A} : \mathcal{X}^{\widetilde{N}} \to \mathcal{W}$ be α-uniformly stable w.r.t. convex loss function $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{D}$ be any distribution over $\mathcal{X}$ and let $X \sim \mathcal{D}^{\widetilde{N}}$. Then the excess population loss is upper bounded* *by the excess expected empirical loss plus α:*

$$\mathbb{E}[F(\mathcal{A}(X),\mathcal{D}) - F^*] \leqslant \alpha + \mathbb{E}[\widehat{F}(\mathcal{A}(X),X) - \min_{w\in\mathcal{W}} \widehat{F}(w,X)],$$

where the expectations are over both the randomness in $\mathcal{A}$ and the sampling of $X \sim \mathcal{D}^{\widetilde{N}}$. Here we denote the empirical loss by $\widehat{F}(w,X)$ and the population loss by $F(w,\mathcal{D})$ for additional clarity, and $F^* := \min_{w\in\mathcal{W}} F(w,\mathcal{D}) = \min_{w\in\mathcal{W}} F(w)$.

Proof. By [33, Theorem 2.2],

$$\mathbb{E}[F(\mathcal{A}(X),\mathcal{D}) - \widehat{F}(\mathcal{A}(X),X)] \leqslant \alpha.$$

Hence

$$\begin{aligned}\mathbb{E}[F(\mathcal{A}(X),\mathcal{D}) - F^*] &= \mathbb{E}[F(\mathcal{A}(X),\mathcal{D}) - \widehat{F}(\mathcal{A}(X),X) + \widehat{F}(\mathcal{A}(X),X) \\ &\quad - \min_{w\in\mathcal{W}} \widehat{F}(w,X) + \min_{w\in\mathcal{W}} \widehat{F}(w,X) - F^*] \\ &\leqslant \alpha + \mathbb{E}[\widehat{F}(\mathcal{A}(X),X) - \min_{w\in\mathcal{W}} \widehat{F}(w,X)],\end{aligned}$$

since $\mathbb{E}_{X\sim\mathcal{D}^N} \min_{w\in\mathcal{W}} \widehat{F}(w,X) \leqslant \min_{w\in\mathcal{W}} \mathbb{E}\left[\widehat{F}(w,X)\right] = \min_{w\in\mathcal{W}} F(w,\mathcal{D}) = F^*$. □

The next step is to bound the uniform stability of Algorithm 1:

Lemma D.2. *Let $f(\cdot,x)$ be convex, L-Lipschitz, and β-smooth loss for all $x \in \mathcal{X}$. Then under Assumption 3, Algorithm 1 with constant stepsize $\eta \leqslant \frac{1}{\beta}$ is α-uniformly stable with respect to f for $\alpha = \frac{2L^2 R\eta}{n_{\min} M}$, where $n_{\min} = \min_{i\in[N]} n_i$. If, in addition $f(\cdot,x)$ is μ-strongly convex, for all $x \in \mathcal{X}$, then under Assumption 3, Algorithm 1 with constant step size $\eta_r = \eta \leqslant \frac{1}{\beta}$ and any averaging weights γ_r is α-uniformly stable with respect to f for $\alpha = \frac{4L^2}{\mu(Mn_{\min}-1)}$ (assuming $\min\{M, n_{\min}\} > 1$).*

Proof of Lemma D.2. The proof of the convex case is similar to proofs of [33, Theorem 3.8], [26, Lemma 4.3], and [8, Lemma 3.4]. For simplicity, assume $K_i = K$ for all i: it will be clear from the proof that K_i does not affect the result. For now, fix the randomness of $\{M_r\}_{r\geqslant 0}$. Let $X, X' \in \mathcal{X}^{\widetilde{N}}$ be two data sets, denoted $X = (X_1, \cdots, X_N)$ for $X_i \in \mathcal{X}^{n_i}$ for all $i \in [N]$ and similarly for X', and assume $|X\Delta X'| = 2$. Then there is a unique $a \in [N]$ and $b \in [n_i]$ such that $x_{a,b} \neq x'_{a,b}$. For $t \in \{0, 1, \cdots, R\}$, denote the t-th iterates of Algorithm 1 on these two data sets by $w_t = w_t(X)$ and $w'_t = w_t(X')$ respectively. We claim that

$$\mathbb{E}\left[\|w_t - w'_t\| \,|\{M_r\}_{0\leqslant r\leqslant t}\right] \leqslant \frac{2L\eta}{n_{\min}} \sum_{r=0}^{t} \frac{1}{M_r} \tag{17}$$

for all t. We prove the claim by induction. It is trivially true when $t = 0$. Suppose (17) holds for all $t \leqslant \tau$. Denote the samples in each local mini-batch at iteration τ by $\{x_{i,j}\}_{i\in[N],j\in[K]}$ (dropping the τ for brevity). First condition on the randomness due to minibatch sampling and due to the Gaussian noise. That is, assume that the averages of the Gaussian vectors u_i, u'_i added to the stochastic gradients at iteration τ of the algorithm run on X and X' respectively are fixed (non-random) vectors: $\bar{u} := \frac{1}{M_\tau}\sum_{i\in S_\tau} u_i$ and $\bar{u}' := \frac{1}{M_\tau}\sum_{i\in S_\tau} u'_i$. Assume WLOG that $S_\tau = [M_\tau]$. Observe that the function $\widetilde{G}(w) = g(w) + \frac{\|\bar{u}\|^2}{2}$ whose gradient equals $\nabla g(w) + \bar{u}$ is still convex and β-smooth if g is. Apply this observation to the convex β-smooth function $g(w) = \frac{1}{M_\tau K}\sum_{i\in[M_\tau]} f(w_\tau, x_{i,j})$. Denote $\widetilde{g}_\tau = \left(\frac{1}{M_\tau K}\sum_{i\in[M_\tau],j\in[K]} \nabla f(w_\tau, x_{i,j})\right) + \bar{u}$ and $\widetilde{g}'_\tau = \left(\frac{1}{M_\tau K}\sum_{i\in[M_\tau],j\in[K]} \nabla f(w_\tau, x'_{i,j})\right) + \bar{u}'$.

Then by non-expansiveness of projection, we have

$$\begin{aligned}
&\|w_{\tau+1} - w'_{\tau+1}\| \\
&= \left\|\Pi_{\mathcal{W}}\left(w_\tau - \eta_\tau \widetilde{g}_\tau\right) - \Pi_{\mathcal{W}}\left(w'_\tau - \eta_\tau \widetilde{g}'_\tau\right)\right\| \\
&\leqslant \left\|\left(w_\tau - \eta_\tau \widetilde{g}_\tau\right) - \left(w'_\tau - \eta_\tau \widetilde{g}'_\tau\right)\right\| \\
&\leqslant \Bigg\| w_r - \eta_r\left(\left(\frac{1}{M_r K}\sum_{(i,j)\neq(a,b)} \nabla f(w_r, x_{i,j})\right) + \bar{u}\right) \\
&\quad - \left(w'_r - \eta_r\left(\left(\frac{1}{M_r K}\sum_{(i,j)\neq(a,b)} \nabla f(w'_r, x_{i,j})\right) + \bar{u}'\right)\right)\Bigg\| \\
&\quad + \frac{q_\tau \eta_\tau}{M_\tau K}\left\|\nabla f(w_\tau, x_{a,b}) - \nabla f(w'_\tau, x'_{a,b})\right\| \\
&\leqslant \|w_\tau - w'_\tau\| + \frac{q_\tau \eta_\tau}{M_\tau K}\left\|\nabla f(w_\tau, x_{a,b}) - \nabla f(w'_\tau, x'_{a,b})\right\|,
\end{aligned}$$

where $q_\tau \in \{0, 1, \cdots, K\}$ is a realization of the random variable Q_τ that counts the number of times index b occurs in worker a's local minibatch at iteration τ, and we used non-expansiveness of the gradient descent step [33, Lemma 3.7] for $\eta \leqslant \frac{2}{\beta}$ (and the observation above about translating a smooth convex function by $\bar{u}$) in the last inequality. (Recall that we sample uniformly with replacement.) Now Q_τ is a sum of K independent Bernoulli($\frac{1}{n_a}$) random variables, hence $\mathbb{E}Q_\tau = \frac{K}{n_a}$. Then using the inductive hypothesis and taking expected value over the randomness of the Gaussian noise and the minibatch sampling proves the claim. (Note that in the worst case, we would have $a \in \operatorname{argmin}_{i\in[N]} n_i$.) Next, taking expectation with respect to the randomness of $\{M_r\}_{r\in[t]}$ implies

$$\mathbb{E}\|w_t - w'_t\| \leqslant \frac{2Lt}{n_{\min} M},$$

since the M_r are i.i.d. with $\mathbb{E}(\frac{1}{M_1}) = \frac{1}{M}$. Then Jensen's inequality and Lipschitz continuity of $f(\cdot, x)$ imply that for any $x \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}[f(\overline{w_R}, x) - f(\overline{w'_R}, x')] &\leqslant L\mathbb{E}\|\overline{w_R} - \overline{w'_R}\| \\
&\leqslant \frac{L}{R}\sum_{t=0}^{R-1}\mathbb{E}\|w_t - w'_t\| \\
&\leqslant \frac{2L^2\eta}{RMn_{\min}}\frac{R(R+1)}{2} = \frac{L^2\eta(R+1)}{Mn_{\min}},
\end{aligned}$$

completing the proof of the convex case.

Next suppose f is μ-strongly convex. The proof begins identically to the convex case. We condition on M_r, u_i, and S_r as before and (keeping the same notation used there) get for any $r \geqslant 0$

$$\begin{aligned}
&\|w_{r+1} - w'_{r+1}\| \\
&\leqslant \Bigg\| w_r - \eta_r\left(\left(\frac{1}{M_r K}\sum_{(i,j)\neq(a,b)} \nabla f(w_r, x_{i,j})\right) + \bar{u}\right) \\
&\quad - \left(w'_r - \eta_r\left(\left(\frac{1}{M_r K}\sum_{(i,j)\neq(a,b)} \nabla f(w'_r, x_{i,j})\right) + \bar{u}'\right)\right)\Bigg\| \\
&\quad + \frac{q_r \eta_r}{M_r K}\left\|\nabla f(w_r, x_{a,b}) - \nabla f(w'_r, x'_{a,b})\right\|.
\end{aligned}$$

We will need the following tighter estimate of the non-expansiveness of the gradient updates to bound the first term on the right-hand side of the inequality above:

Lemma D.3. *[33, Lemma 3.7.3] Let $G : \mathcal{W} \to \mathbb{R}^d$ be μ-strongly convex and β-smooth. Assume $\eta \leqslant \frac{2}{\beta+\mu}$ Then for any $w, v \in \mathcal{W}$, we have*

$$\|(w - \eta\nabla G(w)) - (v - \eta\nabla G(v))\| \leqslant \left(1 - \frac{\eta\beta\mu}{\beta+\mu}\right)\|v - w\| \leqslant \left(1 - \frac{\eta\mu}{2}\right)\|v - w\|.$$

Note that $G_r(w_r) := \frac{1}{M_r K}\sum_{(i,j)\neq(a,b),(i,j)\in[M_r]\times[K]} f(w_r, x^r_{i,j})$ is $(1-\frac{q_r}{M_r K})\beta$-smooth and $(1-\frac{q_r}{M_r K})\mu$-strongly convex and hence so is $G_r(w_r) + \bar{u}$. Therefore, invoking Lemma D.3 and the assumption $\eta_r = \eta \leqslant \frac{1}{\beta}$, as well as Lipschitzness of $f(\cdot, x) \forall x \in \mathcal{X}$, yields

$$\|w_{r+1} - w'_{r+1}\| \leqslant \left(1 - \frac{\eta\mu(1-\frac{q_r}{M_r K})}{2}\right)\|w_r - w'_r\| + \frac{2q_r\eta L}{M_r K}.$$

Next, taking expectations over the M_r (with mean $\mathbb{E}(\frac{1}{M_r}) = \frac{1}{M}$), the minibatch sampling (recall $\mathbb{E}q_r = \frac{K}{n_a}$), and the Gaussian noise implies

$$\mathbb{E}\|w_{r+1} - w'_{r+1}\| \leqslant \left(1 - \frac{\eta\mu(1-\frac{1}{n_a M})}{2}\right)\mathbb{E}\|w_r - w'_r\| + \frac{2\eta L}{n_a M}.$$

One can then prove the following claim by an inductive argument very similar to the one used in the proof of the convex part of Lemma D.2: for all $t \geqslant 0$,

$$\mathbb{E}\|w_t - w'_t\| \leqslant \frac{2\eta L}{n_a M}\sum_{r=0}^{t}(1-b)^r,$$

where $b := \frac{\mu\eta}{2}\left(\frac{n_a M - 1}{n_a M}\right) < 1$. The above claim implies that

$$\begin{aligned}\mathbb{E}\|w_t - w'_t\| &\leqslant \frac{2\eta L}{n_a M}\left(\frac{1-(1-b)^{t+1}}{b}\right)\\ &\leqslant \frac{4L}{\mu(n_a M - 1)}\\ &\leqslant \frac{4L}{\mu(n_{\min} M - 1)}.\end{aligned}$$

Finally, using the above bound together with Lipschitz continuity of f and Jensen's inequality, we obtain that for any $x \in \mathcal{X}$,

$$\begin{aligned}\mathbb{E}[f(\widehat{w}_R, x) - f(\widehat{w}'_R, x)] &\leqslant L\mathbb{E}\|\widehat{w}_R - \widehat{w}'_R\|\\ &= L\mathbb{E}\left\|\frac{1}{\Gamma_R}\sum_{r=0}^{R-1}\gamma_r(w_r - w'_r)\right\|\\ &\leqslant L\mathbb{E}\left[\frac{1}{\Gamma_R}\sum_{r=0}^{R-1}\gamma_r\|w_r - w'_r\|\right]\\ &\leqslant L\left[\frac{1}{\Gamma_R}\sum_{r=0}^{R-1}\gamma_r\left(\frac{4L}{\mu(n_{\min} M - 1)}\right)\right]\\ &= \frac{4L^2}{\mu(n_{\min} M - 1)},\end{aligned}$$

which completes the proof of Lemma D.2. □

Finally, we bound the empirical loss of Algorithm 1:

Lemma D.4. *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be μ-strongly convex (with $\mu = 0$ for convex case), L-Lipschitz, and β-smooth in w for all $x \in \mathcal{X}$, where $\mathcal{W}$ is a closed convex set in $\mathbb{R}^d$ s.t. $\|w\| \leqslant D$ for all $w \in \mathcal{W}$. Let $\mathbf{X} \in \mathcal{X}^{n_1} \times \cdots \mathcal{X}^{n_N}$. Then Algorithm 1 with $\sigma_i^2 = \frac{256L^2R\ln(\frac{2.5R}{\delta_i})\ln(2/\delta_i)}{n_i^2\epsilon_i^2}$ attains the following empirical loss bounds:*

1. (Convex) For any $\eta \leqslant 1/4\beta$ and $R \in \mathbb{N}$, $\gamma_r := 1/R$, we have

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) \leqslant \frac{D^2}{\eta R} + 2\eta\left[\min\left\{\frac{\Phi_*^2}{M'K}, \frac{(\phi_{\max}^*)^2}{MK}\right\} + \frac{\Upsilon_*^2}{M} + \frac{d}{2}\min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M}\right\}\right]$$

.

2. (Strongly Convex) Choose the following constant step-size and averaging weights:

$$\eta_r = \eta = \min \left\{ \frac{1}{4\beta}, \frac{\ln\left(\max\left\{ 2, \frac{\mu^2 D^2 R^2}{2\left(\min\left\{ \frac{\bar{\mathcal{L}}^2}{M'}, \frac{\mathcal{L}_{\max}^2}{M} \right\} + d \min\left\{ \frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M} \right\} \right)} \right\} \right)}{\mu R} \right\}, \quad \gamma_r = (1-\mu\eta)^{-(r+1)}.$$

Then choosing $R = \begin{cases} \max\left\{ \frac{8\beta}{\mu} \ln\left(\frac{\beta D^2 \mu M'}{d\Psi} \right), \frac{L^2}{d\Psi} \right\} & \text{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \max\left\{ \frac{8\beta}{\mu} \ln\left(\frac{\beta D^2 \mu M}{d\psi_{\max}} \right), \frac{L^2}{d\psi_{\max}} \right\} & \text{otherwise} \end{cases}$ *(and assuming* $R \geqslant 1$*) implies*

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{O}\left(\frac{L^2}{\mu} d \min\left\{ \frac{\Xi}{M'}, \frac{\xi_{\max}}{M} \right\} \right). \tag{18}$$

The proof of Lemma D.4 will require some additional lemmas (some of which will come in handy for later results too): First, observe that if we assume that the subset S_r of $M_r \in [N]$ active clients is drawn uniformly (i.e. Assumption 3), then the stochastic gradients $\widetilde{g}_r$ in line 7 of Algorithm 1 are unbiased estimates of $\nabla F(w_r)$. Furthermore, their variance is upper bounded as follows:

Lemma D.5. *Suppose Assumption 3 holds. Let* $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ *be a convex loss function and let* $\widetilde{g}_r := \frac{1}{M_r} \sum_{i \in S_r} \frac{1}{K_i} \sum_{j \in [K_i]} (\nabla f(w_r, x_{i,j}^r) + u_i)$*, where* $(x_{i,j}^r)_{j \in [K]}$ *are sampled (with replacement) from* X_i *and* $u_i \sim N(0, \sigma_i^2 \mathbf{I}_d)$ *is independent of* $\nabla f(w_r, x_{i,j}^r)$ *for all* $i \in [N], j \in [K_i]$*. Denote* $K := \min_{i \in [N]} K_i$*. Assume* $N > 1$*. If* f *satisfies the first part of Assumption 4, then*

$$\mathbb{E}\|\widetilde{g}_r - \nabla F(w_r)\|^2 \leqslant \min\left\{ \frac{\Phi^2}{M'K}, \frac{\phi_{\max}^2}{MK} \right\} + \left(1 - \frac{M-1}{N-1} \right) \frac{\upsilon^2}{M} + d \min\left\{ \frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M} \right\},$$

where $\Phi^2 := \sqrt{\mathbb{E}(\overline{\phi_{M_1}^2})^2}$*, and* $\Sigma^2 = \sqrt{\mathbb{E}(\overline{\sigma_{M_1}^2})^2}$*. If instead* f *satisfies the second part of Assumption 4, then for* $\widetilde{g}_r$ *evaluated at* $w_r = w^* \in \operatorname{argmin}_{w \in \mathcal{W}} F(w)$*, we have*

$$\mathbb{E}\|\widetilde{g}_r\|^2 \leqslant \min\left\{ \frac{\Phi_*^2}{M'K}, \frac{(\phi_{\max}^*)^2}{MK} \right\} + \left(1 - \frac{M-1}{N-1} \right) \frac{\upsilon_*^2}{M} + d \min\left\{ \frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M} \right\},$$

where $\Phi_*^2 := \sqrt{\mathbb{E}[(\overline{\phi_{M_1}^*)^2}]^2}$*. If* $N = 1$*, then the second (middle) term on the right-hand side of each inequality above vanishes.*

The three terms on the right-hand side of each inequality correspond (from left to right) to the variances of: local minibatch sampling within each client, the draw of the client set S_r of size M_r under Assumption 3, and the Gaussian noise. Also, note that $M \geqslant M'$ by Cauchy-Schwartz. Each minimum on the right-hand side is attained by the first term if $1/M_r$ has small variance (so that $1/M' \approx 1/M$) and/or if the clients are fairly heterogeneous, so that parameters measuring averages (e.g. Σ^2) are much smaller than the corresponding maxima (e.g. $\sigma_{\max}^2$). In the complementary case, the minima are attained by the second terms. We now turn to the proof of Lemma D.5.

Proof of Lemma D.5. Assume f satisfies the first part of Assumption 4. First, fix the randomness due to the size of the client set M_r. Now $\widetilde{g}_r = g_r + \bar{u}_r$, where $\bar{u}_r = \frac{1}{M_r} \sum_{i=1}^{M_r} u_i \sim N(0, \frac{\sigma^2}{M_r} \mathbf{I}_d)$ for some $\sigma^2 \leqslant \overline{\sigma_{M_r}^2} = \frac{1}{M_r} \sum_{i=1}^{M_r} \sigma_{(i)}^2$ and $\bar{u}_r$ is independent of $g_r := \frac{1}{M_r} \sum_{i \in S_r} \frac{1}{K_i} \sum_{j \in [K_i]} \nabla f(w_r, x_{i,j}^r)$. Hence,

$$\begin{aligned} \mathbb{E}[\|\widetilde{g}_r - \nabla F(w_r)\|^2 | M_r] &= \mathbb{E}[\|g_r - \nabla F(w_r)\|^2 | M_r] + \mathbb{E}[\|\bar{u}\|^2 | M_r] \\ &\leqslant \mathbb{E}[\|g_r - \nabla F(w_r)\|^2 | M_r] + d \frac{\overline{\sigma_{M_r}^2}}{M_r}. \end{aligned}$$

Let us drop the r subscripts for brevity (denoting $g = g_r$, $w = w_r$, $S = S_r$, and $M_r = M_1$ since they have the same distribution) and denote $h_i := \frac{1}{K_i} \sum_{j=1}^{K_i} \nabla f(w, x_{i,j})$. Now, we have (conditionally on

M_1)

$$
\begin{aligned}
\mathbb{E}[\|g - \nabla F(w)\|^2 | M_1] &= \mathbb{E}\left[\left\| \frac{1}{M_1}\sum_{i\in S}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w, x_{i,j}) - \nabla F(w)\right\|^2 \,\middle|\, M_1\right] \\
&= \mathbb{E}\left[\left\| \frac{1}{M_1}\sum_{i\in S}(\nabla h_i - \nabla F_i(w)) + \frac{1}{M_1}\sum_{i\in S}\nabla F_i(w) - \nabla F(w)\right\|^2 \,\middle|\, M_1\right] \\
&= \frac{1}{M_1^2}\underbrace{\mathbb{E}\left[\|\sum_{i\in S} h_i(w) - \nabla F_i(w)\|^2 \middle| M_1\right]}_{\textcircled{a}} \\
&\quad + \frac{1}{M_1^2}\underbrace{\mathbb{E}\left[\|\sum_{i\in S} \nabla F_i(w) - \nabla F(w)\|^2 \middle| M_1\right]}_{\textcircled{b}},
\end{aligned}
$$

since, conditional on S, the cross-terms vanish by (conditional) independence of h_i and the non-random $\sum_{i'\in S}\nabla F_{i'}(w) - \nabla F(w)$ for all $i \in S$. Now we bound $\textcircled{a}$:

$$
\begin{aligned}
\textcircled{a} &= \mathbb{E}_S\left[\mathbb{E}_{h_i}\|\sum_{i\in S} h_i(w) - \nabla F_i(w)\|^2 \middle| S, M_1\right] \\
&= \mathbb{E}_S\left[\sum_{i\in S}\mathbb{E}_{h_i}\|h_i(w) - \nabla F_i(w)\|^2 \middle| S, M_1\right] \\
&\leqslant \mathbb{E}_S\left[\sum_{i\in S}\frac{\phi_i^2}{K_i}\right] \\
&\leqslant \mathbb{E}_S\left[\frac{M_1\overline{\phi^2_{M_1}}}{K}\right],
\end{aligned}
$$

by conditional independence of $h_i - \nabla F_i$ and $h_{i'} - \nabla F_{i'}$ given S. Hence

$$
\frac{1}{M_1^2}\mathbb{E}\left[\|\sum_{i\in S} h_i(w) - \nabla F_i(w)\|^2 \middle| M_1\right] \leqslant \frac{\overline{\phi^2_{M_1}}}{M_1 K}.
$$

Next we bound $\textcircled{b}$. Fix any $w \in \mathcal{W}$ and denote $y_i := \nabla F_i(w)$ and $\bar{y} := \frac{1}{N}\sum_{i=1}^N y_i = \nabla F(w)$. We claim $\textcircled{b} = \mathbb{E}\left[\|\sum_{i\in S} y_i - \bar{y}\|^2 \middle| M_1\right] \leqslant M_1\left(\frac{N-M_1}{N-1}\right)\upsilon^2$. Assume WLOG that $\bar{y} = 0$ (otherwise, consider $y_i' = y_i - \bar{y}$, which has mean 0). Also, we omit the "conditional on M_1" notation (but continue to condition on M_1) in the below and denote by Ω the collection of all $\binom{N}{M_1}$ subsets of $[N]$

of size M_1. Now,

$$
\begin{aligned}
\textcircled{b} &= \frac{1}{\binom{N}{M_1}} \sum_{S \in \Omega} \left\| \sum_{i \in S} y_i \right\|^2 \\
&= \frac{1}{\binom{N}{M_1}} \sum_{S \in \Omega} \left(\sum_{i \in S} \|y_i\|^2 + 2 \sum_{i, i' \in S, i < i'} \langle y_i, y_{i'} \rangle \right) \\
&= \frac{1}{\binom{N}{M_1}} \left(\binom{N-1}{M_1 - 1} \sum_{i=1}^{N} \|y_i\|^2 + 2 \binom{N-2}{M_1-2} \sum_{1 \leqslant i < i' \leqslant N} \langle y_i, y_{i'} \rangle \right) \\
&= \frac{M_1}{N} \sum_{i=1}^{N} \|y_i\|^2 + 2 \frac{M_1(M_1-1)}{N(N-1)} \sum_{1 \leqslant i < i' \leqslant N} \langle y_i, y_{i'} \rangle \\
&= \frac{M_1}{N} \left(\frac{M_1 - 1}{N-1} + \frac{N - M_1}{N-1} \right) \sum_{i=1}^{N} \|y_i\|^2 + \frac{2M_1(M_1-1)}{N(N-1)} \sum_{1 \leqslant i < i' \leqslant N} \langle y_i, y_{i'} \rangle \\
&= \frac{M_1}{N} \frac{M_1 - 1}{N-1} \left\| \sum_{i=1}^{N} y_i \right\|^2 + \frac{M_1}{N} \frac{N - M_1}{N-1} \sum_{i=1}^{N} \|y_i\|^2 \\
&= \frac{M_1}{N} \frac{N - M_1}{N-1} \sum_{i=1}^{N} \|y_i\|^2 \\
&\leqslant M_1 \left(\frac{N - M_1}{N-1} \right).
\end{aligned}
$$

Hence $\frac{1}{M_1^2} \mathbb{E}\left[\| \sum_{i \in S} \nabla F_i(w) - \nabla F(w) \|^2 \middle| M_1 \right] \leqslant \frac{N - M_1}{N-1} \frac{\upsilon^2}{M_1}$. Finally, we take expectation over the randomness in M_1 to get

$$
\begin{aligned}
\mathbb{E}\|\widetilde{g}_r - \nabla F(w_r)\|^2 &\leqslant \mathbb{E}\left[\frac{\overline{\phi^2_{M_1}}}{M_1 K} + \left(1 - \frac{M_1 - 1}{N-1}\right) \frac{\upsilon^2}{M_1} \right] + d\mathbb{E} \frac{\overline{\sigma^2_{M_1}}}{M_1} \\
&\leqslant \min\left\{ \frac{\Phi^2}{M' K}, \frac{\phi^2_{\max}}{MK} \right\} + \left(1 - \frac{M-1}{N-1}\right) \frac{\upsilon^2}{M} + d \min\left\{ \frac{\Sigma^2}{M'}, \frac{\sigma^2_{\max}}{M} \right\},
\end{aligned}
$$

where, in each minima, the first respective term was obtained using Cauchy-Schwartz and the second term by bounding the numerator by the deterministic "max," which can be pulled outside the expectation. The second statement in the lemma is proved in a nearly identical manner. The statement when $N = 1$ follows from the first part of the proof alone, since the $\textcircled{b}$ term is zero when there is no variance in client sampling (which is the case when $N = 1$). □

We will also need the following lemmas for the proof of Lemma D.4 (and hence Theorem D.1):

Lemma D.6. *(Projection lemma) Let $\mathcal{W} \subset \mathbb{R}^d$ be a closed convex set. Then $\|\Pi_{\mathcal{W}}(a) - b\|^2 \leqslant \|a - b\|^2$ for any $a \in \mathbb{R}^d, b \in \mathcal{W}$.*

Lemma D.6 is well-known.[9] We also recall a standard property of smooth convex functions for convenience:

Lemma D.7. *(Co-coercivity of the gradient) For any convex, β-smooth function $F : \mathcal{W} \to \mathbb{R}$ and any $w, w' \in \mathcal{W}$, we have*

$$
\|\nabla F(w) - \nabla F(w')\|^2 \leqslant \beta \langle \nabla F(w) - \nabla F(w'), w - w' \rangle,
$$

and

$$
\|\nabla F(w) - \nabla F(w')\|^2 \leqslant 2\beta (F(w) - F(w') - \langle \nabla F(w'), w - w' \rangle).
$$

Lastly, we need the following lemmas to optimize the step-sizes for our strongly convex excess risk bounds:

[9] It can be proved by expanding both sides and applying [52, Lemma B.4] multiple times.

Lemma D.8. *[67, Lemma 2] Let $b > 0$, let $a, c \geqslant 0$, and $\{\eta_t\}_{t \geqslant 0}$ be non-negative step-sizes such that $\eta_t \leqslant \frac{1}{g}$ for all $t \geqslant 0$ for some parameter $g \geqslant a$. Let $\{r_t\}_{t \geqslant 0}$ and $\{s_t\}_{t \geqslant 0}$ be two non-negative sequences of real numbers which satisfy*

$$r_{t+1} \leqslant (1 - a\eta_t)r_t - b\eta_t s_t + c\eta_t^2$$

for all $t \geqslant 0$. Then there exist particular choices of step-sizes $\eta_t \leqslant \frac{1}{g}$ and averaging weights $\gamma_t \geqslant 0$ such that

$$\frac{b}{\Gamma_T}\sum_{t=0}^{T} s_t\gamma_t + ar_{T+1} = \widetilde{O}\left(gr_0 \exp\left(\frac{-aT}{g}\right) + \frac{c}{aT}\right),$$

where $\Gamma_T := \sum_{t=0}^{T}\gamma_t$. In fact, we can choose η_t and γ_t as follows:

$$\eta_t = \eta = \min\left\{\frac{1}{g}, \frac{\ln\left(\max\left\{2, a^2 r_0 T^2/c\right\}\right)}{aT}\right\}, \ \gamma_t = (1 - a\eta)^{-(t+1)}.$$

Finally, we are ready to prove Lemma D.4:

Proof of Lemma D.4. We essentially follow the proof of the excess risk bound for the non-private version of Algorithm 1 given in [75, Theorem 1] (adapted for possibly constrained $\mathcal{W} \neq \mathbb{R}^d$ using projection), but accounting for the added Gaussian noise and random M_r. First, condition on the random M_r and consider M_r as fixed. Let $w^* \in \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}(w)$ be any minimizer of $\widehat{F}$ with norm less than or equal to D, and denote the average of the i.i.d. Gaussian noises across all clients in one round by $\bar{u}_r := \frac{1}{M_r}\sum_{i \in S_r} u_i$. Note that $\bar{u}_r \sim N\left(0, \frac{\overline{\sigma_{M_r}^2}}{M_r}\mathbf{I}_d\right)$ by independence of the $\{u_i\}_{i \in [N]}$ and hence $\mathbb{E}\|\bar{u}_r\|^2 = \frac{d\overline{\sigma_{M_r}^2}}{M_r}$. Then for any $r \geqslant 0$, conditional on M_r, we have that

$$\begin{aligned}
&\mathbb{E}\left[\|w_{r+1} - w^*\|^2 \middle| M_r\right] && (19)\\
=&\mathbb{E}\left[\left\|\Pi_{\mathcal{W}}\left[w_r - \eta_r\left(\frac{1}{M_r}\sum_{i \in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r) - u_i\right)\right] - w^*\right\|^2 \middle| M_r\right]\\
\leqslant&\mathbb{E}\left[\left\|w_r - \eta_r\left(\frac{1}{M_r}\sum_{i \in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r) - u_i\right) - w^*\right\|^2 \middle| M_r\right]\\
=&\mathbb{E}\left[\|w_r - w^*\|^2 \middle| M_r\right] - 2\eta_r\mathbb{E}\left[\langle\nabla\widehat{F}(w_r) + \bar{u}_r, w_r - w^*\rangle \middle| M_r\right] && (20)\\
&\quad + \eta_r^2\mathbb{E}\left[\left\|\bar{u}_r + \frac{1}{M_r}\sum_{i \in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r)\right\|^2 \middle| M_r\right]\\
\leqslant&(1 - \mu\eta_r)\mathbb{E}\left[\|w_r - w^*\|^2 \middle| M_r\right] - 2\eta_r\mathbb{E}[\widehat{F}(w_r) - \widehat{F}^*|M_r] && (21)\\
&\quad + \eta_r^2\mathbb{E}\left[\left\|\bar{u}_r + \frac{1}{M_r}\sum_{i \in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r)\right\|^2 \middle| M_r\right], && (22)
\end{aligned}$$

where we used Lemma D.6 in the first inequality, and μ-strong convexity of $\widehat{F}$ (for $\mu \geqslant 0$) and the fact that $\bar{u}_r$ is independent of the gradient estimate and mean zero in the last inequality. Now, omitting the "conditional on M_r" notation for brevity (but still conditioning on M_r), we can bound the last term by

$$
\begin{aligned}
&\mathbb{E}\left\|\bar{u}_r + \frac{1}{M_r}\sum_{i\in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r)\right\|^2 \\
=&\mathbb{E}\left\|\frac{1}{M_r}\sum_{i\in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r) - \nabla f(w^*, x_{i,j}^r) + \nabla f(w^*, x_{i,j}^r)\right\|^2 + d\frac{\overline{\sigma_{M_r}^2}}{M_r} \\
\leqslant& 2\mathbb{E}\left\|\frac{1}{M_r}\sum_{i\in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w_r, x_{i,j}^r) - \nabla f(w^*, x_{i,j}^r)\right\|^2 \\
&\quad + 2\mathbb{E}\left\|\frac{1}{M_r}\sum_{i\in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\nabla f(w^*, x_{i,j}^r)\right\|^2 + d\frac{\overline{\sigma_{M_r}^2}}{M_r} \\
&\leqslant \frac{2}{M_r}\sum_{i\in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\mathbb{E}\left\|\nabla f(w_r, x_{i,j}^r) - \nabla f(w^*, x_{i,j}^r)\right\|^2 \\
&\quad + \frac{2}{M_r}\left(\frac{\overline{(\phi_{M_r}^*)^2}}{K} + \Upsilon_r^2\right) + d\frac{\overline{\sigma_{M_r}^2}}{M_r} \\
&\leqslant \frac{4\beta}{M_r}\sum_{i\in S_r}\frac{1}{K_i}\sum_{j=1}^{K_i}\mathbb{E}\left[f(w_r, x_{i,j}^r) - f(w^*, x_{i,j}^r) - \langle\nabla f(w^*, x_{i,j}^r), w_r - w^*\rangle\right] \\
&\quad + \frac{2}{M_r}\left(\frac{\overline{(\phi_{M_r}^*)^2}}{K} + \Upsilon_r^2\right) + d\frac{\overline{\sigma_{M_r}^2}}{M_r} \\
&\leqslant 4\beta\mathbb{E}[\widehat{F}(w_r) - \widehat{F}^*] + \frac{2}{M_r}\left(\frac{\overline{(\phi_{M_r}^*)^2}}{K} + \Upsilon_r^2\right) + d\frac{\overline{\sigma_{M_r}^2}}{M_r}
\end{aligned}
$$

where we denote $\Upsilon_r^2 := \begin{cases}\left(1 - \frac{M_r-1}{N-1}\right)\upsilon_*^2 & \text{if } N > 1 \\ 0 & \text{otherwise.}\end{cases}$ Above, we used the "relaxed triangle inequality" (see e.g. [43, Lemma 3]) in the first inequality, Lemma D.5 (with M_r considered fixed/non-random due to conditioning and replacing $\widetilde{g}_r$ by the noiseless minibatch gradient) in the second inequality, Lemma D.7 in the third inequality, and the first-order optimality conditions for constrained optimization in the final inequality. The first equality is due to independence of the Gaussian noise and the stochastic gradients. Next, plugging this estimate back into Equation 22 and noting that $\eta_r \leqslant \frac{1}{4\beta}$ for all $r \geqslant 0$, we obtain

$$
\mathbb{E}\left[\|w_{r+1} - w^*\|^2 \middle| M_r\right] \leqslant (1 - \mu\eta_r)\mathbb{E}\left[\|w_r - w^*\|^2 \middle| M_r\right] - 2\eta_r(1 - 2\beta\eta_r)\mathbb{E}[\widehat{F}(w_r) - \widehat{F}^*|M_r] \tag{23}
$$

$$
\begin{aligned}
&+ \frac{2\eta_r^2}{M_r}\left(\frac{\overline{(\phi_{M_r}^*)^2}}{K} + \Upsilon_r^2\right) + \eta_r^2 d\frac{\overline{\sigma_{M_r}^2}}{M_r} \\
&\leqslant (1 - \mu\eta_r)\mathbb{E}[\|w_r - w^*\|^2|M_r] - \eta_r\mathbb{E}[\widehat{F}(w_r) - \widehat{F}^*|M_r] \\
&+ \frac{2\eta_r^2}{M_r}\left(\frac{\overline{(\phi_{M_r}^*)^2}}{K} + \Upsilon_r^2\right) + \eta_r^2 d\frac{\overline{\sigma_{M_r}^2}}{M_r},
\end{aligned} \tag{24}
$$

which implies

$$
\begin{aligned}
\mathbb{E}[\widehat{F}(w_r) - \widehat{F}^*|M_r] \leqslant& \left(\frac{1}{\eta_r} - \mu\right)\mathbb{E}[\|w_r - w^*\|^2|M_r] - \frac{1}{\eta_r}\mathbb{E}[\|w_{r+1} - w^*\|^2|M_r] \\
&+ \frac{2\eta_r}{M_r}\left(\frac{\overline{(\phi_{M_r}^*)^2}}{K} + \Upsilon_r^2\right) + \eta_r d\frac{\overline{\sigma_{M_r}^2}}{M_r}.
\end{aligned} \tag{25}
$$

Now we consider the convex ($\mu = 0$) and strongly convex ($\mu > 0$) cases separately.

Convex ($\mu = 0$) **case:** By our choice of $\eta_r = \eta$ and (25), the average iterate $\widehat{w_R}$ satisfies:

$$\begin{aligned}\mathbb{E}[\widehat{F}(\widehat{w_R}) - \widehat{F}^*|\{M_r\}_{r \leqslant R}] &\leqslant \frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}[\widehat{F}(w_r) - \widehat{F}^*|M_r] \\ &\leqslant \frac{1}{R}\sum_{r=0}^{R-1} \frac{1}{\eta}(\mathbb{E}[\|w_r - w^*\|^2 - \|w_{r+1} - w^*\||M_r]) \\ &+ \frac{1}{R}\sum_{r=0}^{R-1} \frac{2\eta}{M_r}\left(\frac{\overline{(\phi^*_{M_r})^2}}{K} + \Upsilon_r^2 + d\overline{\sigma^2_{M_r}}/2\right) \\ &\leqslant \frac{\|w_0 - w^*\|^2}{\eta R} + \frac{1}{R}\sum_{r=0}^{R-1} \frac{2\eta}{M_r}\left(\frac{\overline{(\phi^*_{M_r})^2}}{K} + \Upsilon_r^2 + d\overline{\sigma^2_{M_r}}/2\right).\end{aligned}$$

Then taking expectation over the randomness in M_r, we get by Assumption 3 that

$$\mathbb{E}\widehat{F}(\hat{w}_R) - \widehat{F}(w^*) \leqslant \frac{D^2}{\eta R} + 2\eta\left[\min\left\{\frac{\Phi_*^2}{M'K}, \frac{(\phi^*_{\max})^2}{MK}\right\} + \left(1 - \frac{M-1}{N-1}\right)\frac{\upsilon_*^2}{M} + \frac{d}{2}\min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma^2_{\max}}{M}\right\}\right] \tag{26}$$

if $N > 1$; and if $N = 1$, then the term involving υ_*^2 vanishes. Above we used Cauchy-Schwartz to get the first term in each of the minima and upper bounded the numerator by the non-random "max" quantity (and invoked linearity of expectation) for the second.

Strongly convex ($\mu > 0$) **case:** Recall from (24) that

$$\begin{aligned}\mathbb{E}[\|w_{t+1} - w^*\|^2|M_t] \leqslant (1 - \mu\eta_t)\mathbb{E}[\|w_t - w^*\|^2|M_t] - \eta_r\mathbb{E}[\widehat{F}(w_t) - \widehat{F}^*|M_t] \\ + \frac{2\eta_t^2}{M_t}\left(\frac{\overline{(\phi^*_{M_t})^2}}{K} + \Upsilon_t^2 + d\overline{\sigma^2_{M_t}}/2\right)\end{aligned} \tag{27}$$

for all $t \geqslant 0$. Taking expectation over M_t gives

$$\mathbb{E}\|w_{t+1} - w^*\|^2 \leqslant (1 - \mu\eta_t)\mathbb{E}\|w_t - w^*\|^2 - \eta_t\mathbb{E}[\widehat{F}(w_t) - \widehat{F}^*] \tag{28}$$

$$+ 2\eta_t^2\left[\frac{4L^2}{MK} + \frac{\Upsilon^2}{M} + \frac{d}{2}\min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma^2_{\max}}{M}\right\}\right], \tag{29}$$

which satisfies the conditions for Lemma D.8, with sequences

$$r_t = \mathbb{E}\|w_t - w^*\|^2, s_t = \mathbb{E}[\widehat{F}(w_t) - \widehat{F}^*]$$

and parameters

$$a = \mu,\ b = 1,\ c = 2\left(\frac{4L^2}{MK} + \frac{\Upsilon_*^2}{M}\right) + d\min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma^2_{\max}}{M}\right\},\ g = 4\beta,\ T = R.$$

Then applying Lemma D.8 and Jensen's inequality completes the proof. □

At last, we are prepared to prove Theorem D.1.

Proof of Theorem D.1. **Privacy:** By independence of the Gaussian noise across clients, it suffices to show that transcript of client i's interactions with the server is DP for all $i \in [N]$ (conditional on the transcripts of all other clients). WLOG consider $i = 1$ and denote client 1's privacy parameters by ϵ and δ. Then the proof begins along similar lines as the proof of [9, Theorem 2.1]. By the advanced composition theorem [20, Theorem 3.20], it suffices to show that each of the R rounds of the algorithm is $(\widetilde{\epsilon}, \widetilde{\delta})$-LDP, where $\widetilde{\epsilon} = \frac{\epsilon}{2\sqrt{2R\ln(2/\delta)}}$ (we used the assumption $\epsilon \leqslant \ln(2/\delta)$ here) and $\widetilde{\delta} = \frac{\delta}{2R}$. First, condition on the randomness due to local sampling of the local data point $x^r_{1,1}$ (line 4 of Algorithm 1). Now, the L_2 sensitivity of each local step of SGD is bounded by

$\Delta := \sup_{|X_1 \Delta X_1'| \leqslant 2, w \in \mathcal{W}} \| \frac{1}{K} \sum_{j=1}^{K} \nabla f(w, x_{1,j}) - \nabla f(w, x'_{1,j}) \| \leqslant 2L/K$, by L-Lipschitzness of f. Thus, since $\widetilde{\epsilon} \leqslant 1$ by assumption, the standard privacy guarantee of the Gaussian mechanism [20, Theorem A.1] implies that (conditional on the randomness due to sampling) taking $\sigma_1^2 \geqslant \frac{8L^2 \ln(1.25/\widetilde{\delta})}{\widetilde{\epsilon}^2 K^2}$ suffices to ensure that round r (in isolation) is $(\widetilde{\epsilon}, \widetilde{\delta})$-LDP. Now we invoke the randomness due to sampling: [71, Problem 1b] implies that round r (in isolation) is $(\frac{2\widetilde{\epsilon}K}{n_1}, \widetilde{\delta})$-LDP (where we used the fact that $\widetilde{\epsilon} \leqslant 1$ implies $e^{\widetilde{\epsilon}} - 1 \leqslant 2$). Therefore, with sampling, it suffices to take $\sigma_1^2 \geqslant \frac{32L^2 \ln(1.25/\widetilde{\delta})}{n_1^2 \widetilde{\epsilon}^2} = \frac{256L^2 R \ln(2.5R/\delta) \ln(2/\delta)}{n_1^2 \epsilon^2}$ to ensure that round r (in isolation) is $(\widetilde{\epsilon}, \widetilde{\delta})$-LDP for all r and hence that the full algorithm (R rounds) is (ϵ, δ)-LDP.

Excess loss: First suppose f is merely convex ($\mu = 0$). By Lemma D.2, Lemma D.1, and Lemma D.4 (and noting $\phi_{\max}^* \leqslant 4L^2$ and $\upsilon_*^2 = 0$ since client data is i.i.d.), we have:

$$\begin{aligned}
\mathbb{E}F(\widehat{w}_R) - F(w^*) &\leqslant \alpha + \mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) && (30)\\
&\leqslant \frac{2L^2 R\eta}{n_{\min} M} + \frac{D^2}{\eta R} && (31)\\
&+ 2\eta \left[\min\left\{ \frac{\Phi_*^2}{M'K}, \frac{(\phi_{\max}^*)^2}{MK} \right\} + \left(1 - \frac{M-1}{N-1}\right) \frac{\upsilon_*^2}{M} + \frac{d}{2} \min\left\{ \frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M} \right\} \right] && (32)\\
&\leqslant \frac{2L^2 R\eta}{n_{\min} M} + \frac{D^2}{\eta R} + 2\eta \left[4Rd \min\left\{ \frac{\Psi}{M'}, \frac{\psi_{\max}}{M} \right\} + \frac{4L^2}{MK} \right] && (33)\\
&\leqslant \eta L^2 \left[R \left(\frac{2}{M n_{\min}} + 256 d \min\left\{ \frac{\Xi}{M'}, \frac{\xi_{\max}}{M} \right\} \right) + \frac{8}{MK} \right] + \frac{D^2}{\eta R} && (34)
\end{aligned}$$

for any $\eta \leqslant \frac{1}{4\beta}$. Then one can verify that the prescribed choice of η and R yields the desired bound.

Now suppose f is μ-strongly convex. The prescribed choice of η and R imply that

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{O}\left(\frac{d}{\mu} \min\left\{ \frac{\Psi}{M'}, \frac{\psi_{\max}}{M} \right\} \right)$$

by Lemma D.4. Then the result follows from Lemma D.1 and Lemma D.2. □

D.3 Complete statement and proof of Theorem 2.1

Using the same notation as in the complete version of Theorem D.1, we have:

Theorem 2.1 [Complete Version] *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ be μ-strongly convex (with $\mu = 0$ for convex case) and L-Lipschitz, in w for all $x \in \mathcal{X}$, where $\mathcal{W}$ is a closed convex set in $\mathbb{R}^d$ s.t. $\|w\| \leqslant D$ for all $w \in \mathcal{W}$. Let $\mathcal{D}$ be an arbitrary probability distribution on $\mathcal{X}$. For each client $i \in [N]$, draw a local i.i.d. data set $X_i \sim \mathcal{D}^{n_i}$. Running Algorithm 1 on $f_\beta(w, x) := \min_{v \in \mathcal{W}} \left(f(v, x) + \frac{\beta}{2}\|w - v\|^2 \right)$ with β as prescribed below and the same σ_i^2, $\eta_r = \eta$ and $\{\gamma_r\}_{r=0}^{R-1}$ in Theorem D.1[Complete Version] results in the following upper bounds on the excess loss (w.r.t. f):*

1. *(Convex) Setting $\beta := \frac{L\sqrt{\widetilde{M}}}{D} \min\left(\sqrt{n_{\min}}, \frac{1}{\sqrt{d\widetilde{\Xi}}} \right)$ and $R := \left\lceil \widetilde{M} \min\left\{ n_{\min}, \frac{1}{d\widetilde{\Xi}} \right\} \right\rceil$ yields*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) \leqslant 545 \frac{LD}{\sqrt{\widetilde{M}}} \max\left\{ \frac{1}{\sqrt{n_{\min}}}, \sqrt{d\widetilde{\Xi}} \right\}.$$

2. *(Strongly convex) Setting $\beta := \mu\widetilde{M} \min\left\{ n_{\min}, \frac{1}{d\widetilde{\Xi}} \right\}$ and $R := \left\lceil \max\left\{ 8\widetilde{M} \min\left\{ n_{\min}, \frac{1}{d\widetilde{\Xi}} \right\} \ln\left(\frac{\beta D^2 \mu M'}{dL^2 \widetilde{\Xi}} \right), \frac{1}{d\widetilde{\Xi}} \right\} \right\rceil$ yields*

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{L^2}{\mu} \left(\frac{1}{M n_{\min}} + d \min\left\{ \frac{\Xi}{M'}, \frac{\xi_{\max}}{M} \right\} \right) \right).$$

The proof follows from applying Theorem D.1 to the β-smooth, Lipschitz, convex loss F_β to upper bound $\mathbb{E}F_\beta(\widehat{w}_R) - F_\beta^*$ and then relating this quantity to $\mathbb{E}F(\widehat{w}_R) - F(w^*)$ by using:

Lemma D.9. *(see [60], [8, Lemma 4.3]) Let $f : \mathcal{W} \to \mathbb{R}^d$ be convex and L-Lipschitz and let $\beta > 0$. 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. $\forall w$, $f_\beta(w) \leqslant f(w) \leqslant f_\beta(w) + \frac{L^2}{2\beta}$.

Proof of Theorem 2.1. We have $\mathbb{E}F(\widehat{w}_R) - F(w^*) \leqslant \mathbb{E}F_\beta(\widehat{w}_R) - F^*_\beta + \frac{L^2}{2\beta}$, by part 2 of Lemma D.9. Then plugging in β and combining part 1 of Lemma D.9 with Theorem D.1 completes the proof. □

D.4 Noisy Accelerated MB-SGD Algorithm and the complete statement and proof of smooth upper bounds

We first state the informal guarantee of Algorithm 2 for smooth losses:

Theorem D.2. [Informal] *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^d$ be L-Lipschitz, β-smooth, and μ-strongly convex (with $\mu = 0$ for convex case). Then, Algorithm 2 with $\sigma_i^2 := \frac{256 L^2 R \ln(2.5R/\delta_0)\ln(2/\delta_0)}{n^2 \epsilon_0^2}$ is (ϵ_0, δ_0)-LDP. Moreover, for any $\mathbf{X} \in \mathcal{X}^{n \times N}$, its output $\widehat{w}_R$ satisfies:*
1. (Convex) Setting $R = \max\left(\left(\frac{\beta D \sqrt{M} \epsilon_0 n}{L\sqrt{d}}\right)^{1/2}, \frac{\epsilon_0^2 n^2}{d} \begin{cases} \frac{1}{K} & \text{if } M = N \\ 1 & \text{otherwise} \end{cases} \right)$ yields

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = O\left(LD \left(\frac{\sqrt{d \ln(2.5R/\delta_0)\ln(2/\delta_0)}}{\epsilon_0 n \sqrt{M}} \right) \right). \tag{35}$$

2. (Strongly convex) Setting $R = \max\left(\sqrt{\frac{\beta}{\mu}} \ln\left(\frac{D\mu M \epsilon_0^2 n^2}{Ld}\right), \frac{\epsilon_0^2 n^2}{d} \begin{cases} \frac{1}{K} & \text{if } M = N \\ 1 & \text{otherwise} \end{cases} \right)$ yields

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = O\left(\frac{L^2}{\mu} \left(\frac{d \ln(2.5R/\delta_0)\ln(2/\delta_0)}{\epsilon_0^2 n^2 M} \right) \right). \tag{36}$$

The Noisy Accelerated MB-SGD algorithm is formally described in Algorithm 2.

Algorithm 2 Noisy Accelerated MB-SGD

```
Require: Number of clients $N \in \mathbb{N}$, dimension $d \in \mathbb{N}$ of data, noise parameters $\{\sigma_i\}_{i \in [N]}$, closed convex set $\mathcal{W} \subset \mathbb{R}^d$, data sets $X_i \in \mathcal{X}_i^{n_i}$ for $i \in [N]$, loss function $f(w, x)$, number of communication rounds $R \in \mathbb{N}$, number $K_i \in \mathbb{N}$ of local samples drawn per round, step size parameters $\{\eta_r\}_{r \in [R]}, \{\alpha_r\}_{r \in [R]}$ such that $\alpha_1 = 1, \alpha_r \in (0,1)$ for all $r \geqslant 2$ and $\eta_r > 0$ for all $r \geqslant 1$, norm $D$ of some optimum $w^*$ of $F$.
Set initial point $w_0^{ag} = w_0 \in \mathcal{W}$ 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 $i \in S_r$ do in parallel
        Client draws $K_i$ samples $x_{i,j}^r$ (uniformly with replacement) from $X_i$ (for $j \in [K_i]$) and noise $u_i \sim N(0, \sigma_i^2 \mathbf{I}_d)$.
        Client computes $\widetilde{g}_r^i := \frac{1}{K_i} \sum_{j=1}^{K_i} \nabla f(w_r^{md}, x_{i,j}^r) + u_i$.
    end for
    Server aggregates $\widetilde{g}_r := \frac{1}{M} \sum_{i=1}^{M} \widetilde{g}_r^i$.
    Server updates and broadcasts:
$w_r := \operatorname{argmin}_{w \in \mathcal{W}} \left\{ \alpha_r \left[ \langle \widetilde{g}_r, w \rangle + \frac{\mu}{2}\|w_r^{md} - w\|^2 \right] + \left[ (1-\alpha_r)\frac{\mu}{2} + \frac{\eta_r}{2} \right] \|w_{r-1} - w\|^2 \right\}$.
    Server updates and broadcasts $w_r^{ag} = \alpha_r w_r + (1-\alpha_r) w_{r-1}^{ag}$.
end for
return $w_R^{ag}$.
```

Next, we state the complete version of Theorem D.2:

Theorem D.2 [Complete Version] *Let $\mathbf{X} \in \mathcal{X}_1^{n_1} \times \cdots \mathcal{X}_N^{n_N}$. Suppose $f(\cdot, x_i)$ is L_i-Lipschitz and* *convex on $\mathcal{W}$ for all $x_i \in \mathcal{X}_i, i \in [N]$, $\widehat{F}(\cdot, \mathbf{X}) := \widehat{F}(\cdot)$ is $\bar{\beta}$-smooth, Assumption 4 part 1 and Assumption 3 hold. Set $\sigma_i^2 = \frac{256 L_i^2 R \ln(\frac{2.5R}{\delta_i}) \ln(2/\delta_i)}{n_i^2 \epsilon_i^2}$. Denote*

$$\Upsilon^2 = \begin{cases} \left(1 - \frac{M-1}{N-1}\right) \upsilon^2 & \textit{if } N > 1 \\ 0 & \textit{otherwise} \end{cases}$$

and

$$V^2 := \min\left\{\frac{\Phi^2}{M'K}, \frac{\phi_{\max}^2}{MK}\right\} + d \min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M}\right\} + \frac{\Upsilon^2}{M},$$

where $K := \min_{i \in [N]} K_i$. Then Algorithm 2 is $\{(\epsilon_i, \delta_i)\}_{i \in [N]}$-LDP for any $\epsilon_i \in (0, \ln(\frac{2}{\delta_i})]$ and $\delta_i \in (0,1)$. Moreover:
1. Running Algorithm 2 on $\widetilde{F}(w) := \widehat{F}(w) + \frac{\lambda}{2}\|w\|^2$ with $\lambda := \frac{V}{2D\sqrt{R}}$ for R rounds yields (for some choice of stepsizes)

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim \frac{\bar{\beta} D^2}{R^2} + D\left[\min\left\{\sqrt{\frac{(\Phi^2/KR) + d\Psi}{M'}}, \sqrt{\frac{(\phi_{\max}^2/KR) + d\psi_{\max}}{M}}\right\} + \sqrt{\frac{\Upsilon^2}{MR}}\right]. \tag{37}$$

In particular, setting $R = \begin{cases} \max\left\{\left(\frac{\bar{\beta} D \sqrt{M'}}{\sqrt{d\Psi}}\right)^{1/2}, \frac{\Phi^2/K + \Upsilon^2}{d\Psi}\right\} & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \max\left\{\left(\frac{\bar{\beta} D \sqrt{M}}{\sqrt{d\psi_{\max}}}\right)^{1/2}, \frac{\phi_{\max}^2/K + \Upsilon^2}{d\psi_{\max}}\right\} & \textit{otherwise} \end{cases}$ *implies*

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim D\sqrt{d} \min\left\{\frac{\sqrt{\Psi}}{\sqrt{M'}}, \frac{\sqrt{\psi_{\max}}}{\sqrt{M}}\right\}, \tag{38}$$

assuming $R \geqslant 1$.
2. If, in addition, $\widehat{F}$ is $\bar{\mu}$-strongly convex, then running a multi-stage implementation of Algorithm 2 directly on $\widehat{F}$ with the same choices of σ_i^2 and K yields

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim \Delta \exp\left(-\sqrt{\frac{\bar{\mu}}{\bar{\beta}}} R\right) + \frac{1}{\bar{\mu}}\left[\min\left\{\frac{(\Phi^2/KR) + d\Psi}{M'}, \frac{(\phi_{\max}^2/KR) + d\psi_{\max}}{M}\right\} + \frac{\Upsilon^2}{MR}\right], \tag{39}$$

where $F(w_0) - F^* \leqslant \Delta$.
In particular, choosing $R = \begin{cases} \max\left\{\sqrt{\frac{\bar{\beta}}{\bar{\mu}}} \ln\left(\frac{\Delta \bar{\mu} M'}{d\Psi}\right), \frac{\Phi^2/K + \Upsilon^2}{d\Psi}\right\} & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \max\left\{\sqrt{\frac{\bar{\beta}}{\bar{\mu}}} \ln\left(\frac{\Delta \bar{\mu} M}{d\psi_{\max}}\right), \frac{\phi_{\max}^2/K + \Upsilon^2}{d\psi_{\max}}\right\} & \textit{otherwise} \end{cases}$ *implies*

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim \frac{d}{\bar{\mu}} \min\left\{\frac{\Psi}{M'}, \frac{\psi_{\max}}{M}\right\}, \tag{40}$$

provided $R \geqslant 1$.

Proof of Theorem D.2. LDP of Algorithm 1 follows from LDP of Algorithm 1 (see Theorem D.1) and the post-processing property of DP [20, Proposition 2.1]. Namely, since the choice of noise given in Theorem D.2 ensures that clients' local stochastic minibatch gradients are LDP and the iterates in Algorithm 2 are functions of these private noisy gradients, it follows that the iterates themselves are LDP.
For convergence, we begin with the convex ($\bar{\mu} = 0$) part. We will need the following lemma:

Lemma D.10. *[75, Lemma 4] Let $F : \mathcal{W} \to \mathbb{R}^d$ be convex and β-smooth, and suppose that the unbiased stochastic gradients $\widetilde{g}(w_t)$ at each iteration have bounded variance $\mathbb{E}\|\widetilde{g}(w) - \nabla F(w)\|^2 \leqslant V^2$. If $\widehat{w}^{ag}$ is computed by T steps of AC-SA on the regularized objective $\widetilde{F}(w) = F(w) + \frac{V}{2\|w_0 - w^*\|\sqrt{T}}\|w - w_0\|^2$, then*

$$\mathbb{E}F(\widehat{w}^{ag}) - F^* \lesssim \frac{\beta\|w_0 - w^*\|^2}{T^2} + \frac{V\|w_0 - w^*\|}{\sqrt{T}}.$$

Combining Lemma D.10 with the estimates of the variance of the stochastic gradients from Lemma D.5 (and replacing β by $\bar{\beta}$) proves the first (convex) part of Theorem D.2.

For the second (strongly convex) part, we follow [31, 75] and use the following **multi-stage implementation of Algorithm 2** to further accelerate convergence: Let $F(0) - F^* \leqslant \Delta$ and $q_0 = 0$. Then for $k \in [U]$, do the following:

1. Run R_k rounds of Algorithm 2 using $w_0 = q_{k-1}$, $\{\alpha_r\}_{r\geqslant 1}$ and $\{\eta_r\}_{r\geqslant 1}$, where
$$R_k = \left\lceil \max\left\{ 4\sqrt{\frac{2\beta}{\mu}}, \frac{128V^2}{3\mu\Delta 2^{-(k+1)}} \right\} \right\rceil,$$
$$\alpha_r = \frac{2}{r+1}, \; \eta_r = \frac{4\upsilon_k}{r(r+1)},$$
$$\upsilon_k = \max\left\{ 2\beta, \left[\frac{\mu V^2}{3\Delta 2^{-(k-1)} R_k(R_k+1)(R_k+2)} \right]^{1/2} \right\}$$
2. Set $q_k = w_{R_k}^{ag}$, where $w_{R_k}^{ag}$ is the output of Step 1 above. Then update $k \leftarrow k+1$ and return to Step 1.

We then have the following risk bound for the multi-stage protocol:

Lemma D.11. *[31, Proposition 7] Let $F : \mathcal{W} \to \mathbb{R}^d$ be μ-strongly convex and β-smooth, and suppose that the unbiased stochastic gradients $\widetilde{g}(w_t)$ at each iteration have bounded variance $\mathbb{E}\|\widetilde{g}(w) - \nabla F(w)\|^2 \leqslant V^2$. If $\widehat{w}^{ag}$ is computed by T steps of the multi-stage AC-SA, then*
$$\mathbb{E}F(\widehat{w}^{ag}) - F^* \lesssim \Delta \exp\left(-\sqrt{\frac{\mu}{\beta}}T\right) + \frac{V^2}{\mu T},$$
where $\Delta = F(w_0) - F^*$.

In our notation, $T = R$ and F is replaced by $\widehat{F}$, which is $\bar{\mu}$-strongly convex and $\bar{\beta}$-smooth by assumption. If U is chosen so that $\sum_{k=1}^{U} R_k \leqslant R$ total rounds of Algorithm 2, then the full algorithm, run with noise σ_i^2 specified in Theorem D.2 [Complete Version], is LDP. Applying Lemma D.11 with
$$V^2 = \min\left\{\frac{\Phi^2}{M'K}, \frac{\phi_{\max}^2}{MK}\right\} + \frac{\Upsilon^2}{M} + d\min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M}\right\}$$
from Lemma D.5 proves the strongly convex portion of Theorem D.2. □

D.5 Proof of Theorem 2.2

Proof of Theorem 2.2. **1. Privacy:** By post-processing, it suffices to show that the $R = n$ noisy gradients computed in line 6 of Algorithm 2 are (ϵ_0, δ_0)-LDP. Further, since the points sampled locally are disjoint/distinct (because we sample locally *without replacement*), parallel composition [56] implies that if each update in line 6 is (ϵ_0, δ_0)-LDP, then the full algorithm is (ϵ_0, δ_0)-LDP. Now recall that the Gaussian mechanism [20, Appendix A.1] provides (ϵ_0, δ_0)-DP if $\sigma^2 \geqslant \frac{2\Delta_2^2 \ln(1.25/\delta_0)}{\epsilon_0^2}$, where $\Delta_2 = \sup_{w,x}\|\nabla f(w,x) - \nabla f(w,x')\| \leqslant 2L$ is the L_2 sensitivity of the non-private gradient update in line 6 of Algorithm 2. Therefore, conditional on the private transcript of all other clients, we see that client i's transcript is (ϵ_0, δ_0)-DP for all $i \in [N]$, which means that One-Pas Noisy Accelerated Distributed SGD is (ϵ_0, δ_0)-LDP.
2. Excess loss: For the convex case, we plug the estimate for the variance of the noisy stochastic gradients from Lemma D.5 for V^2 in Lemma D.10 and set $T = n$. Note that L-Lipschitzness implies that $V^2 \leqslant \frac{\phi^2}{M} + \frac{\upsilon^2}{M} + \frac{d\sigma^2}{M} \leqslant \frac{5L^2 + d\sigma^2}{M}$. Similarly, for strongly convex loss, we plug the same estimate for V^2 into Lemma D.11 with $T = n$ (using the multi-stage implementation of Algorithm 2, described in the previous subsection of this appendix). This completes the proof. □

D.6 Complete version and proof of Theorem 2.3

Theorem 2.3 [Complete Version] *Suppose $f(\cdot, x_i)$ is L_i-Lipschitz and convex on $\mathcal{W}$ for all $x_i \in \mathcal{X}_i, i \in [N]$, Assumption 4 part 1 and Assumption 3 hold. Assume $L_i = L$ for all i. Set $\sigma_i^2 = \frac{256 L^2 R \ln(\frac{2.5R}{\delta_i}) \ln(2/\delta_i)}{n_i^2 \epsilon_i^2}$. Denote*

$$\Upsilon^2 = \begin{cases} \left(1 - \frac{M-1}{N-1}\right) \upsilon^2 & \textit{if } N > 1 \\ 0 & \textit{otherwise} \end{cases}$$

and

$$V^2 := \min\left\{\frac{\Phi^2}{M'K}, \frac{\phi_{\max}^2}{MK}\right\} + d \min\left\{\frac{\Sigma^2}{M'}, \frac{\sigma_{\max}^2}{M}\right\} + \frac{\Upsilon^2}{M},$$

where $K := \min_{i \in [N]} K_i$. Also denote $\widetilde{M} := \begin{cases} \sqrt{M'} & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \sqrt{M} & \textit{otherwise} \end{cases}$ and $\widetilde{\Psi} := \begin{cases} \Psi & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \psi_{\max} & \textit{otherwise} \end{cases}$. Fix any $\mathbf{X} \in \mathcal{X}_1^{n_1} \times \cdots \times \mathcal{X}_N^{n_N}$ and denote $\widehat{F}_\beta(w) := \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^{n_i} \frac{1}{n_i} f_\beta(w, x_{i,j})$, with f_β defined earlier (see complete version of Theorem 2.1). Then:

1. Running Algorithm 2 on $\widetilde{F}_\beta(w) := \widehat{F}_\beta(w) + \frac{\lambda}{2}\|w\|^2$ with $\lambda := \frac{V}{2D\sqrt{R}}$ and $\beta := \frac{L^2\sqrt{\widetilde{M}}}{D\sqrt{d\widetilde{\Psi}}}$ for R rounds yields (for some choice of stepsizes)

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim \frac{L^2 D\sqrt{\widetilde{M}}}{R^2\sqrt{d\widetilde{\Psi}}} + D\left[\min\left\{\sqrt{\frac{(\Phi^2/KR) + d\Psi}{M'}}, \sqrt{\frac{(\phi_{\max}^2/KR) + d\psi_{\max}}{M}}\right\} + \sqrt{\frac{\Upsilon^2}{MR}}\right]$$
$$+ D\sqrt{d}\min\left\{\frac{\sqrt{\Psi}}{\sqrt{M'}}, \frac{\sqrt{\psi_{\max}}}{\sqrt{M}}\right\}.$$

In particular, setting $R = \begin{cases} \max\left\{\left(\frac{L\sqrt{M'}}{\sqrt{d\Psi}}\right), \frac{\Phi^2/K + \Upsilon^2}{d\Psi}\right\} & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \max\left\{\left(\frac{L\sqrt{M}}{\sqrt{d\psi_{\max}}}\right), \frac{\phi_{\max}^2/K + \Upsilon^2}{d\psi_{\max}}\right\} & \textit{otherwise} \end{cases}$ implies

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim D\sqrt{d}\min\left\{\frac{\sqrt{\Psi}}{\sqrt{M'}}, \frac{\sqrt{\psi_{\max}}}{\sqrt{M}}\right\}, \tag{41}$$

assuming $R \geqslant 1$.

2. If, in addition, $\widehat{F}$ is $\bar{\mu}$-strongly convex, then running a multi-stage implementation of Algorithm 2 on $\widehat{F}_\beta$ with $\beta := \frac{\bar{\mu}\widetilde{M}L^2}{d\widetilde{\Psi}}$ yields

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim \Delta \exp\left(-R\sqrt{\frac{d\widetilde{\Psi}}{\widetilde{M}L^2}}\right)$$
$$+ \frac{1}{\bar{\mu}}\left[\min\left\{\frac{(\Phi^2/KR) + d\Psi}{M'}, \frac{(\phi_{\max}^2/KR) + d\psi_{\max}}{M}\right\} + \frac{\Upsilon^2}{MR}\right], \tag{42}$$

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

In particular, choosing $R = \begin{cases} \max\left\{\sqrt{\frac{M'L^2}{d\Psi}} \ln\left(\frac{\Delta\bar{\mu}M'}{d\Psi}\right), \frac{\Phi^2/K + \Upsilon^2}{d\Psi}\right\} & \textit{if } \frac{\Psi}{M'} \leqslant \frac{\psi_{\max}}{M} \\ \max\left\{\sqrt{\frac{ML^2}{d\psi_{\max}}} \ln\left(\frac{\Delta\bar{\mu}M}{d\psi_{\max}}\right), \frac{\phi_{\max}^2/K + \Upsilon^2}{d\psi_{\max}}\right\} & \textit{otherwise} \end{cases}$ implies

$$\mathbb{E}\widehat{F}(w_R^{ag}) - \widehat{F}(w^*) \lesssim \frac{d}{\bar{\mu}}\min\left\{\frac{\Psi}{M'}, \frac{\psi_{\max}}{M}\right\}, \tag{43}$$

provided $R \geqslant 1$.

Proof. The proof is very similar to the proof of Theorem 2.1. We apply Theorem D.2 to the β-smooth objective $\widetilde{F}_\beta$ (for convex case) or $\widehat{F}_\beta$ (for strongly convex) and use Lemma D.9. This ensures that excess risk with respect to $\widehat{F}$ increases by at most L^2/β, which is bounded by the smooth convex (or strongly convex, respectively) excess risk bound by our choice of β. □

E Lower Bounds for LDP FL

E.1 Example of LDP Algorithm that is Not Compositional

This example is a simple modification of [39, Example 2.2] (adapted to our definition of compositionality for $\delta_0 > 0$). Given any $C > 0$, set $d := 2C^2$ and let $\mathcal{X} = \{e_1, \cdots e_d\} \subset \{0,1\}^d$ be the standard basis for $\mathbb{R}^d$. Let $n = 1$ and $\mathbf{X} = (x_1, \cdots, x_N) \in \mathcal{X}^N$. For all $i \in [N]$ let $\mathcal{Q}^{(i)} : \mathcal{X} \to \mathcal{X}$ be the randomized response mechanism that outputs $\mathcal{Q}^{(i)}(x_i) = x_i$ with probability $\frac{e^{\epsilon_0}}{e^{\epsilon_0}+d-1}$ and otherwise outputs a uniformly random element of $\mathcal{X}\backslash\{x_i\}$. Note that $\mathcal{Q}^{(i)}$ is ϵ_0-DP, hence (ϵ_0, δ_0)-DP for any $\delta_0 > 0$. Consider the following d-round algorithm $\mathcal{A} : \mathcal{X}^N \to \mathcal{Z}^{d \times N}$ where $\mathcal{Z} = \mathbb{R}^d$.

Algorithm 3 LDP Algorithm that is not C-compositional

```
for $r \in [d]$ do
    for $i \in N$ do
        if $x_i = e_r$ then
            $\mathcal{R}_r^{(i)}(x_i) := \mathcal{Q}^{(i)}(x_i)$.
        else
            $\mathcal{R}_r^{(i)}(x_i) := 0 \in \mathbb{R}^d$.
        end if
    end for
end for
return $\{\mathcal{R}_r^{(i)}(x_i)\}_{i \in [N], r \in [d]}$.
```

Since each client's data is only referenced once and $\mathcal{Q}^{(i)}$ is ϵ_0-DP, we have $\epsilon_0^r = \epsilon_0$ and $\mathcal{A}$ is (ϵ_0, δ_0)-DP. However, $\sqrt{\sum_{r=1}^d (\epsilon_0^r)^2} = \sqrt{d\epsilon_0^2} = \sqrt{2}C\epsilon_0 > C\epsilon_0$, so that $\mathcal{A}$ is not C-compositional.

E.2 Proof of Theorem 3.1

To prove Theorem 3.1, we first analyze the central privacy guarantees of $\mathcal{A} \in \mathbb{A}_{(\epsilon_0, \delta_0)}$ when client data sets $X_1, \cdots, X_N$ are shuffled each round before the randomizers are applied, showing that privacy amplifies to $\epsilon = \widetilde{O}(\frac{\epsilon_0}{\sqrt{N}})$ (Theorem E.1). This is an extension of [25, Theorem 3.8] to $n > 1$ and fully interactive compositional algorithm. The second step is to apply the CDP lower bounds of [8, Appendix C] to $\mathcal{A}_s$, the "shuffled" version of $\mathcal{A}$. [10] This implies that the shuffled algorithm $\mathcal{A}_s$ has excess population loss that is lower bounded as in Theorem 3.1. The final step in the proof is to observe that the i.i.d. assumption implies that $\mathcal{A}_s$ and $\mathcal{A}$ have the same expected population loss.

Step 1: Privacy amplification by shuffling. We begin by stating and proving the amplification by shuffling result that we will leverage to obtain Theorem 3.1:

Theorem E.1. *Let $\mathcal{A} \in \mathbb{A}_{(\epsilon_0, \delta_0)}$ such that $\epsilon_0 \in (0, \sqrt{N}]$ and $\delta_0 \in (0,1)$. Assume that in each round, the local randomizers $\mathcal{R}_r^{(i)}(\mathbf{Z}_{(1:r-1)}, \cdot) : \mathcal{X}^n \to \mathcal{Z}$ are $(\epsilon_0^r, \delta_0^r)$-DP for all $i \in [N]$, $r \in [R]$, $\mathbf{Z}_{(1:r-1)} \in \mathcal{Z}^{r-1 \times N}$ with $\epsilon_0^r \leqslant \frac{1}{n}$. If $\mathcal{A}$ is compositional, then assume $\delta_0^r \in [\frac{2e^{-N/16}}{Nn}, \frac{1}{14nNR}]$ and denote $\delta := 14Nn\sum_{r=1}^R \delta_0^r$; if instead $\mathcal{A}$ is sequentially interactive, then assume $\delta_0 = \delta_0^r \in [\frac{2e^{-N/16}}{Nn}, \frac{1}{7Nn}]$ and denote $\delta := 7Nn\delta_0$. Let $\mathcal{A}_s : \mathbb{X} \to \mathcal{W}$ be the same algorithm as $\mathcal{A}$ except that in each round r, $\mathcal{A}_s$ draws a random permutation π_r of $[N]$ and applies $\mathcal{R}_r^{(i)}$ to $X_{\pi_r(i)}$ instead of X_i. Then, $\mathcal{A}_s$ is (ϵ, δ)-CDP, where $\epsilon = O\left(\frac{\epsilon_0 \ln\left(1/nN\delta_0^{\min}\right)}{\sqrt{N}}\right)$, and $\delta_0^{\min} := \min_{r \in [R]} \delta_0^r$. Note that for sequentially interactive $\mathcal{A}$, $\delta_0^{\min} = \delta_0$.*

To the best of our knowledge, the restriction on ϵ_0^r is needed to obtain $\epsilon = \widetilde{O}(\epsilon_0/\sqrt{N})$ in all works that have analyzed privacy amplification by shuffling [22, 25, 6, 14, 7], but these works focus on

[10] While [8] does not explicitly prove lower bounds for strongly convex CDP SCO, their proof technique easily extends to strongly convex loss, implying that $\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{\Omega}\left(\left(\frac{L^2}{\mu n N} + LD\frac{d}{\epsilon^2 n^2 N^2}\right)\right)$ for (ϵ, δ) CDP algorithms with $\delta = o(1/nN)$, by [9, Theorem 5.5].

the sequentially interactive case with $n = 1$, so the restriction amounts to $\epsilon_0 \lesssim 1$ (or $\epsilon_0 = \widetilde{O}(1)$). Theorem E.1 will follow as a pair of corollaries (Corollary E.1 and Corollary E.2) from the following result which analyzes the privacy amplification in each round:

Theorem E.2 (Single round privacy amplification by shuffling)**.** *Let $\epsilon_0^r \leqslant \ln\left(\frac{N}{16\ln(2/\delta^r)}\right)/n,\ r \in \mathbb{N}$ and let $\mathcal{R}_r^{(i)}(\mathbf{Z}, \cdot) : \mathcal{X}^n \to \mathcal{Z}$ be an $(\epsilon_0^r, \delta_0^r)$-DP local randomizer for all $\mathbf{Z} = Z_{(1:r-1)}^{(1:N)} \in \mathcal{Z}^{(r-1)\times N}$ and $i \in [N]$, where $\mathcal{X}$ is an arbitrary set. Given a distributed data set $\mathbf{X} = (X_1, \cdots, X_N) \in \mathcal{X}^{N\times n}$ and $\mathbf{Z} = Z_{(1:r-1)}^{(1:N)}$, consider the shuffled algorithm $\mathcal{A}_s^r : \mathcal{X}^{n\times N} \times \mathcal{Z}^{(r-1)\times N} \to \mathcal{Z}^N$ that first samples a random permutation π of $[N]$ and then computes $Z_r = (Z_r^{(1)}, \cdots, Z_r^{(N)})$, where $Z_r^{(i)} := \mathcal{R}_r^{(i)}(\mathbf{Z}, X_{\pi(i)})$. Then, $\mathcal{A}_s^r$ is $(\epsilon^r, \widetilde{\delta}^r)$-CDP, where*

$$\epsilon^r := \ln\left[1 + \left(\frac{e^{\epsilon_0^r} - 1}{e^{\epsilon_0^r} + 1}\right)\left(\frac{8\sqrt{e^{n\epsilon_0^r}\ln(4/\delta^r)}}{\sqrt{N}} + \frac{8e^{n\epsilon_0^r}}{N}\right)\right], \tag{44}$$

and $\widetilde{\delta}^r := \delta^r + 2Nne^{(n-1)\epsilon_0^r}\delta_0^r$. In particular, if $\epsilon_0^r = O\left(\frac{1}{n}\right)$, then

$$\epsilon^r = O\left(\frac{\epsilon_0^r\sqrt{\ln(1/\delta^r)}}{\sqrt{N}}\right). \tag{45}$$

Further, if $\epsilon_0^r \leqslant 1/n$, then setting $\delta^r := Nn\delta_0^r$ implies that

$$\epsilon^r = O\left(\frac{\epsilon_0^r\sqrt{\ln(1/nN\delta_0^r)}}{\sqrt{N}}\right) \tag{46}$$

and $\widetilde{\delta}^r \leqslant 7Nn\delta_0^r$, which is in $(0, 1)$ if we assume $\delta_0^r \in (0, \frac{1}{7Nn})$.

We sometimes refer to the algorithm $\mathcal{A}_s^r$ as the shuffled algorithm derived from the randomizers $\{\mathcal{R}_r^{(i)}\}$. From Theorem E.2, we obtain:

Corollary E.1 (R-round privacy amplification for compositional algorithms)**.** *Let $\mathcal{A} : \mathcal{X}^{n\times N} \to \mathcal{Z}^{R\times N}$ be an R-round (ϵ_0, δ_0)-LDP and C-compositional algorithm such that $\epsilon_0 \in (0, \sqrt{N}]$ and $\delta_0 \in (0, 1)$, where $\mathcal{X}$ is an arbitrary set. Assume that in each round, the local randomizers $\mathcal{R}_r^{(i)}(\mathbf{Z}_{(1:r-1)}, \cdot) : \mathcal{X}^n \to \mathcal{Z}$ are $(\epsilon_0^r, \delta_0^r)$-DP for $i \in [N], r \in [R]$, where $\epsilon_0^r \leqslant \frac{1}{n}$, and $\delta_0^r \in [\frac{2e^{-N/16}}{Nn}, \frac{1}{14nNR}]$. Then, the shuffled algorithm $\mathcal{A}_s : \mathcal{X}^{n\times N} \to \mathcal{Z}^{R\times N}$ derived from $\{\mathcal{R}_r^{(i)}(\mathbf{Z}_{(1:r-1)}, \cdot)\}_{i\in[N], r\in[R]}$ (i.e. $\mathcal{A}_s$ is the composition of the R shuffled algorithms $\mathcal{A}_s^r$ defined in Theorem E.2) is (ϵ, δ)-CDP, where $\delta \leqslant 14Nn\sum_{r=1}^R \delta_0^r$ and $\epsilon = O\left(\frac{\epsilon_0\ln\left(1/nN\delta_0^{\min}\right)}{\sqrt{N}}\right)$, where $\delta_0^{\min} := \min_{r\in[R]}\delta_0^r$.*

Proof. Let $\delta' := \sum_r Nn\delta_0^r$ and $\delta^r := Nn\delta_0^r$. Then the (central) privacy loss of the full R-round shuffled algorithm is bounded as

$$\begin{aligned}
\epsilon &\leqslant 2\sum_r (\epsilon^r)^2 + \sqrt{2\sum_r (\epsilon^r)^2\ln(1/\delta')} \\
&= O\left(\sum_r\left(\frac{(\epsilon_0^r)^2\ln(1/\delta^r)}{N}\right) + \sqrt{\sum_r \frac{(\epsilon_0^r)^2\ln(1/\delta^r)\ln(1/\delta')}{N}}\right) \\
&= O\left(\frac{\epsilon_0\ln(1/nN\delta_0^{\min})}{\sqrt{N}}\right),
\end{aligned}$$

where the three (in)equalities follow in order from the Advanced Composition Theorem [20], (46) in Theorem E.2, and C-compositionality of $\mathcal{A}$ combined with the assumption $\epsilon_0 \lesssim \sqrt{N}$. Also, $\delta = \delta' + \sum_r \widetilde{\delta}^r$ by the Advanced Composition Theorem, where $\widetilde{\delta}_r \leqslant 7Nn\delta_0^r$ by Theorem E.2. Hence $\delta \leqslant 14Nn\sum_r \delta_0^r$. □

Remark E.1. *The upper bounds assumed on δ_0^r and δ^r in Theorem 3.1 ensure that $\delta \in (0,1)$ and that the lower bounds of [9] apply (see Theorem E.3). These assumptions are not very restrictive in practice, since $\delta_0^r, \delta_0 \ll 1$ is needed for meaningful privacy guarantees (see e.g. [20, Chapter 2]) and R must be polynomial for the algorithm to run. Also, since $N \gg 1$ is the regime of interest (otherwise if $N = \widetilde{O}(1)$, the CDP lower bounds of [8] already match our upper bounds up to logarithms), the requirement that N be larger than $16\ln(2/\delta_0^{\min} n)$ is unimportant.*[11]

Corollary E.2. *(R round privacy amplification for sequentially interactive algorithms) Let $\mathcal{A} : \mathcal{X}^{n\times N} \to \mathcal{Z}^{R\times N}$ be an R-round (ϵ_0, δ_0)-LDP sequentially interactive with $\epsilon_0 \leqslant 1/n$ and $\delta_0 \in [\frac{2e^{-N/16}}{Nn}, \frac{1}{7Nn}]$. Then the shuffled algorithm derived from $\mathcal{A}$ is (ϵ, δ)-CDP, where*

$$\epsilon = O\left(\frac{\epsilon_0\sqrt{\ln(1/nN\delta_0)}}{\sqrt{N}}\right),$$

and $\delta \leqslant 7Nn\delta_0$.

Proof. For sequentially interactive algorithms, we have $\epsilon_0^r = \epsilon_0$ and $\delta_0^r = \delta_0$ since each client's local data is only referenced once throughout the algorithm. Likewise, $\epsilon^r = \epsilon$ and $\delta^r = \delta$ represent the total central privacy loss of each client during the algorithm. Thus, the result is immediate from (46) in Theorem E.2. □

We now turn to the proof of Theorem E.2, which uses the techniques from [25]. First, we'll need some more notation. The privacy relation in (3) between random variables P and Q can be characterized by the *hockey-stick divergence*: $D_{e^\epsilon}(P\|Q) := \int \max\{0, p(x) - e^\epsilon q(x)\}dx$, where p and q denote the probability density or mass functions of P and Q respectively. Then $P \underset{(\epsilon,\delta)}{\simeq} Q$ iff $\max\{D_{e^\epsilon}(P\|Q), D_{e^\epsilon}(Q\|P)\} \leqslant \delta$. Second, recall the *total variation distance* between P and Q is given by $TV(P,Q) = \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)|dx$. Third, we recall the notion of group privacy:

Definition 6 (Group DP). *A randomized algorithm $\mathcal{A} : \mathcal{X}^{\mathcal{N}} \to \mathcal{Z}$ is (ϵ, δ) group DP for groups of size $\mathcal{N}$ if $\mathcal{A}(\mathbf{X}) \underset{(\epsilon,\delta)}{\simeq} \mathcal{A}(\mathbf{X}')$ for all $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^{\mathcal{N}}$.*

We'll also need the following stronger version of a decomposition from [41] and [57, Lemma 3.2].

Lemma E.1 ([41]). *Let $\mathcal{R}_0, \mathcal{R}_1 : \mathcal{X}^n \to \mathcal{Z}$ be local randomizers such that $\mathcal{R}_0(X_0)$ and $\mathcal{R}_1(X_1)$ are $(\epsilon, 0)$ indistinguishable. Then, there exists a randomized algorithm $U : \{X_0, X_1\} \to \mathcal{Z}$ such that $R_0(X_0) = \frac{e^\epsilon}{e^\epsilon+1}U(X_0) + \frac{1}{e^\epsilon+1}U(X_1)$ and $R_1(X_1) = \frac{1}{e^\epsilon+1}U(X_0) + \frac{e^\epsilon}{e^\epsilon+1}U(X_1)$.*

Lemma E.1 follows from the proof of [57, Lemma 3.2], noting that the weaker hypothesis assumed in Lemma E.1 sufficient for all steps to go through.

Definition 7 (Deletion Group DP). *Algorithm $\mathcal{R} : \mathcal{X}^n \to \mathcal{Z}$ is (ϵ, δ) deletion group DP for groups of size n if there exists a reference distribution ρ such that $\mathcal{R}(X) \underset{(\epsilon,\delta)}{\simeq} \rho$ for all $X \in \mathcal{X}^n$.*

It's easy to show that if $\mathcal{R}$ is *deletion* group DP for groups of size n, then $\mathcal{R}$ is $(2\epsilon, (1+e^\epsilon)\delta)$ group DP for groups of size n. In addition, we have the following result:

Lemma E.2. *Let $X_0 \in \mathcal{X}^n$. If $\mathcal{R} : \mathcal{X}^n \to \mathcal{Z}$ is an (ϵ, δ)-DP local randomizer, then $\mathcal{R}$ is $(n\epsilon, ne^{(n-1)\epsilon}\delta)$ deletion group DP for groups of size n with reference distribution $\mathcal{R}(X_0)$ (i.e. $\mathcal{R}(X) \underset{(\widetilde{\epsilon},\widetilde{\delta})}{\simeq} \mathcal{R}(X_0)$ for all $X \in \mathcal{X}^n$, where $\widetilde{\epsilon} = n\epsilon$ and $\widetilde{\delta} = ne^{(n-1)\epsilon}\delta$).*

Proof. By group privacy (see e.g. [42, Theorem 10]), and the assumption that $\mathcal{R}$ is (ϵ, δ)-DP, it follows that $\mathcal{R}(X)$ and $\mathcal{R}(X')$ are $(n\epsilon, ne^{(n-1)\epsilon}\delta)$ indistinguishable for all $X, X' \in \mathcal{X}^n$. In particular, taking $X' := X_0$ completes the proof. □

Lemma E.3. *Let $\mathcal{R}^{(i)} : \mathcal{X}^n \to \mathcal{Z}$ be randomized algorithms ($i \in [N]$) and let $\mathcal{A}_s : \mathcal{X}^{n\times N} \to \mathcal{Z}^N$ be the shuffled algorithm $\mathcal{A}_s(\mathbf{X}) := (\mathcal{R}^{(1)}(X_{\pi(1)}), \cdots \mathcal{R}^{(N)}(X_{\pi(N)}))$ derived from $\{\mathcal{R}^{(i)}\}_{i\in[N]}$*

[11] This assumption is needed to ensure that the condition on ϵ_0^r in Theorem E.2 is satisfied; it is inherited from [25, Theorem 3.8].

for $\mathbf{X} = (X_1, \cdots, X_N)$*, where* π *is a uniformly random permutation of* $[N]$*. Let* $\mathbf{X}_0 = (X_1^0, X_2, \cdots, X_N)$ *and* $\mathbf{X}_1 = (X_1^1, X_2, \cdots, X_N)$*,* $\delta \in (0,1)$ *and* $p \in [\frac{16\ln(2/\delta)}{N}, 1]$*. Suppose that for all* $i \in [N], X \in \mathcal{X}^n \backslash \{X_1^1, X_1^0\}$*, there exists a distribution* $LO^{(i)}(X)$ *such that*

$$\mathcal{R}^{(i)}(X) = \frac{p}{2}\mathcal{R}^{(i)}(X_1^0) + \frac{p}{2}\mathcal{R}^{(i)}(X_1^1) + (1-p)LO^{(i)}(X).$$

Then $\mathcal{A}_s(\mathbf{X}_0) \underset{(\epsilon,\delta)}{\simeq} \mathcal{A}_s(\mathbf{X}_1)$*, where*

$$\epsilon \leqslant \ln\left(1 + \frac{8\sqrt{\ln(4/\delta)}}{\sqrt{pN}} + \frac{8}{pN}\right).$$

Proof. The proof mirrors the proof of [25, Lemma 3.3] closely, replacing their notation with ours. Observe that the DP assumption in [25, Lemma 3.3] is not actually needed in the proof. □

Lemma E.4. *Let* $\mathcal{R} : \mathcal{X}^n \to \mathcal{Z}$ *be* (ϵ, δ) *deletion group DP for groups of size* n *with reference distribution* ρ*. Then there exists a randomizer* $\mathcal{R}' : \mathcal{X}^n \to \mathcal{Z}$ *such that:*
(i) $\mathcal{R}'$ *is* $(\epsilon, 0)$ *deletion group DP for groups of size* n *with reference distribution* ρ*; and*
(ii) $TV(\mathcal{R}(X), \mathcal{R}'(X)) \leqslant \delta$*.*
In particular, $\mathcal{R}'$ *is* $(2\epsilon, (1+e^\epsilon)\delta)$ *group DP for groups of size* n *(by i).*

Proof. The proof is nearly identical to the proof of [25, Lemma 3.7]. □

We also need the following stronger version of [25, Lemma 3.7]:

Lemma E.5. *If* $\mathcal{R}(X_1^0) \underset{(\epsilon_0,\delta_0)}{\simeq} \mathcal{R}(X_1^1)$*, then there exists a randomizer* $\mathcal{R}' : \mathcal{X}^n \to \mathcal{Z}$ *such that* $\mathcal{R}'(X_1^1) \underset{(\epsilon_0,0)}{\simeq} \mathcal{R}(X_1^0)$ *and* $TV(\mathcal{R}'(X_1^1), \mathcal{R}(X_1^1)) \leqslant \delta_0$*.*

Proof. The proof follows the same techniques as [25, Lemma 3.7], noting that the weaker hypothesis in Lemma E.5 is sufficient for all the steps to go through and that the assumption of $n = 1$ in [25] is not needed in the proof. □

Lemma E.6 ([20], Lemma 3.17)**.** *Given random variables* P, Q, P' *and* Q'*, if* $D_{e^\epsilon}(P', Q') \leqslant \delta$*,* $TV(P, P') \leqslant \delta'$*, and* $TV(Q, Q') \leqslant \delta'$*, then* $D_{e^\epsilon}(P, Q) \leqslant \delta + (e^\epsilon + 1)\delta'$*.*

Lemma E.7 ([25], Lemma 2.3)**.** *Let* P *and* Q *be distributions satisfying* $P = (1-q)P_0 + qP_1$ *and* $Q = (1-q)P_0 + qQ_1$ *for some* $q \in [0,1]$*. Then for any* $\epsilon > 0$*, if* $\epsilon' = \log(1 + q(e^\epsilon - 1))$*, then*

$$D_{e^{\epsilon'}}(P||Q) \leqslant q\max\{D_{e^\epsilon}(P_1||P_0), D_{e^\epsilon}(P_0||Q_1)\} \leqslant qD_{e^\epsilon}(P_1||Q_1).$$

We are now ready to prove Theorem E.2:

Proof of Theorem E.2. Let $\mathbf{X}_0, \mathbf{X}_1 \in \mathcal{X}^{n \times N}$ be adjacent (in the CDP sense) distributed data sets (i.e. $|\mathbf{X}_0 \Delta \mathbf{X}_1| \leqslant 1$). Assume WLOG that $\mathbf{X}_0 = (X_1^0, X_2, \cdots, X_N)$ and $\mathbf{X}_1 = (X_1^1, X_2, \cdots, X_N)$, where $X_1^0 = (x_{1,0}, x_{1,2}, \cdots, x_{1,n}) \neq (x_{1,1}, x_{1,2}, \cdots, x_{1,n})$. We can also assume WLOG that $X_j \notin \{X_1^0, X_1^1\}$ for all $j \in \{2, \cdots, N\}$ by re-defining $\mathcal{X}$ and $\mathcal{R}_r^{(i)}$ if necessary; details omitted here.

Fix $i \in [N], r \in [R], \mathbf{Z} = \mathbf{Z}_{1:r-1} = Z_{(1:r-1)}^{(1:N)} \in \mathcal{Z}^{(r-1)\times N}$, denote $\mathcal{R}(X) := \mathcal{R}_r^{(i)}(\mathbf{Z}, X)$ for $X \in \mathcal{X}^n$, and $\mathcal{A}_s(\mathbf{X}) := \mathcal{A}_s^r(\mathbf{Z}_{1:r-1}, \mathbf{X})$. Draw π uniformly from the set of permutations of $[N]$. Now, since $\mathcal{R}$ is (ϵ_0, δ_0)-DP, $\mathcal{R}(X_1^1) \underset{(\epsilon_0^r,\delta_0^r)}{\simeq} \mathcal{R}(X_1^0)$, so by Lemma E.5, there exists a local randomizer $\mathcal{R}'$ such that $\mathcal{R}'(X_1^1) \underset{(\epsilon_0^r,0)}{\simeq} \mathcal{R}(X_1^0)$ and $TV(\mathcal{R}'(X_1^1), \mathcal{R}(X_1^1)) \leqslant \delta_0^r$.

Hence, by Lemma E.1, there exist distributions $U(X_1^0)$ and $U(X_1^1)$ such that

$$\mathcal{R}(X_1^0) = \frac{e^{\epsilon_0^r}}{e^{\epsilon_0^r} + 1}U(X_1^0) + \frac{1}{e^{\epsilon_0^r} + 1}U(X_1^1) \tag{47}$$

and

$$\mathcal{R}'(X_1^1) = \frac{1}{e^{\epsilon_0^r} + 1}U(X_1^0) + \frac{e^{\epsilon_0^r}}{e^{\epsilon_0^r} + 1}U(X_1^1). \tag{48}$$

Denote $\widetilde{\epsilon_0} := n\epsilon_0^r$ and $\widetilde{\delta_0} := ne^{(n-1)\epsilon_0^r}\delta_0^r$. By convexity of hockey-stick divergence and the hypothesis that $\mathcal{R}$ is $(\epsilon_0^r, \delta_0^r)$-DP (hence $\mathcal{R}(X) \underset{(\widetilde{\epsilon_0}, \widetilde{\delta_0})}{\simeq} \mathcal{R}(X_1^0), \mathcal{R}(X_1^1)$ for all X by Lemma E.2), we have $\mathcal{R}(X) \underset{(\widetilde{\epsilon_0}, \widetilde{\delta_0})}{\simeq} \frac{1}{2}(\mathcal{R}(X_1^0) + \mathcal{R}(X_1^1)) := \rho$ for all $X \in \mathcal{X}^n$. That is, $\mathcal{R}$ is $(\widetilde{\epsilon_0}, \widetilde{\delta_0})$ deletion group DP for groups of size n with reference distribution ρ. Thus, Lemma E.4 implies that there exists a local randomizer $\mathcal{R}''$ such that $\mathcal{R}''(X)$ and ρ are $(\widetilde{\epsilon_0}, 0)$ indistinguishable and $TV(\mathcal{R}''(X), \mathcal{R}(X)) \leqslant \widetilde{\delta_0}$ for all X. Then by the definition of $(\widetilde{\epsilon_0}, 0)$ indistinguishability, for all X there exists a "left-over" distribution $LO(X)$ such that $\mathcal{R}''(X) = \frac{1}{e^{\widetilde{\epsilon_0}}}\rho + (1 - 1/e^{\widetilde{\epsilon_0}})LO(X) = \frac{1}{2e^{\widetilde{\epsilon_0}}}(\mathcal{R}(X_1^0) + \mathcal{R}(X_1^1)) + (1 - 1/e^{\widetilde{\epsilon_0}})LO(X)$.

Now, define a randomizer $\mathcal{L}$ by $\mathcal{L}(X_1^0) := \mathcal{R}(X_1^0)$, $\mathcal{L}(X_1^1) := \mathcal{R}'(X_1^1)$, and

$$\begin{aligned} \mathcal{L}(X) &:= \frac{1}{2e^{\widetilde{\epsilon_0}}}\mathcal{R}(X_1^0) + \frac{1}{2e^{\widetilde{\epsilon_0}}}\mathcal{R}'(X_1^1) + (1 - 1/e^{\widetilde{\epsilon_0}})LO(X) \\ &= \frac{1}{2e^{\widetilde{\epsilon_0}}}U(X_1^0) + \frac{1}{2e^{\widetilde{\epsilon_0}}}U(X_1^1) + (1 - 1/e^{\widetilde{\epsilon_0}})LO(X) \end{aligned} \tag{49}$$

for all $X \in \mathcal{X}^n \backslash \{X_1^0, X_1^1\}$. (The equality follows from (47) and (48).) Note that $TV(\mathcal{R}(X_1^0), \mathcal{L}(X_1^0)) = 0$, $TV(\mathcal{R}(X_1^1), \mathcal{L}(X_1^1)) \leqslant \delta_0^r$, and for all $X \in \mathcal{X}^n \backslash \{X_1^0, X_1^1\}$, $TV(\mathcal{R}(X), \mathcal{L}(X)) \leqslant TV(\mathcal{R}(X), \mathcal{R}''(X)) + TV(\mathcal{R}''(X), \mathcal{L}(X)) \leqslant \widetilde{\delta_0} + \frac{1}{2e^{\widetilde{\epsilon_0}}}TV(\mathcal{R}'(X_1^1), \mathcal{R}(X_1^1)) = (ne^{(n-1)\epsilon_0^r} + \frac{1}{2e^{n\epsilon_0^r}})\delta_0^r \leqslant (2ne^{(n-1)\epsilon_0^r})\delta_0^r = 2\widetilde{\delta_0}$.

Keeping r fixed (omitting r scripts everywhere), for any $i \in [N]$ and $\mathbf{Z} := \mathbf{Z}_{1:r-1} \in \mathcal{Z}^{(r-1)\times N}$, let $\mathcal{L}^{(i)}(\mathbf{Z}, \cdot)$, $U^{(i)}(\mathbf{Z}, \cdot)$, and $LO^{(i)}(\mathbf{Z}, \cdot)$ denote the randomizers resulting from the process described above. Let $\mathcal{A}_{\mathcal{L}} : \mathcal{X}^{n\times N} \to \mathcal{Z}^N$ be defined exactly the same way as $\mathcal{A}_s^r := \mathcal{A}_s$ (same π) but with the randomizers $\mathcal{R}^{(i)}$ replaced by $\mathcal{L}^{(i)}$. Since $\mathcal{A}_s$ applies each randomizer $\mathcal{R}^{(i)}$ exactly once and $\mathcal{R}^{(1)}(\mathbf{Z}, X_{\pi(1)}, \cdots \mathcal{R}^{(N)}(\mathbf{Z}, X_{\pi(N)})$ are independent (conditional on $\mathbf{Z} = \mathbf{Z}_{1:r-1}$) [12], we have $TV(\mathcal{A}_s(\mathbf{X}_0), \mathcal{A}_{\mathcal{L}}(\mathbf{X}_0) \leqslant N(2ne^{(n-1)\epsilon_0^r})\delta_0^r$ and $TV(\mathcal{A}_s(\mathbf{X}_1), \mathcal{A}_{\mathcal{L}}(\mathbf{X}_1) \leqslant N(2ne^{(n-1)\epsilon_0^r})\delta_0^r$ (see [23]). Now we claim that $\mathcal{A}_{\mathcal{L}}(\mathbf{X}_0)$ and $\mathcal{A}_{\mathcal{L}}(\mathbf{X}_1)$ are (ϵ^r, δ^r) indistinguishable for any $\delta^r \geqslant 2e^{-Ne^{-n\epsilon_0^r}/16}$. Observe that this claim implies that $\mathcal{A}_s(\mathbf{X}_0)$ and $\mathcal{A}_s(\mathbf{X}_1)$ are $(\epsilon^r, \widetilde{\delta}^r)$ indistinguishable by Lemma E.6 (with $P' := \mathcal{A}_{\mathcal{L}}(\mathbf{X}_0), Q' := \mathcal{A}_{\mathcal{L}}(\mathbf{X}_1), P := \mathcal{A}_s(\mathbf{X}_0), Q := \mathcal{A}_s(\mathbf{X}_1)$.) Therefore, it remains to prove the claim, i.e. to show that $D_{e^{\epsilon^r}}(\mathcal{A}_{\mathcal{L}}(\mathbf{X}_0), \mathcal{A}_{\mathcal{L}}(\mathbf{X}_1) \leqslant \delta^r$ for any $\delta^r \geqslant 2e^{-Ne^{-n\epsilon_0^r}/16}$.

Now, as in the proof of [25, Theorem 3.1], define $\mathcal{L}_U^{(i)}(\mathbf{Z}, X) := \begin{cases} U^{(i)}(\mathbf{Z}, X_1^0) & \text{if } X = X_1^0 \\ U^{(i)}(\mathbf{Z}, X_1^1) & \text{if } X = X_1^1 \\ \mathcal{L}^{(i)}(\mathbf{Z}, X) & \text{otherwise.} \end{cases}$. For any inputs $\mathbf{Z}, \mathbf{X}$, let $\mathcal{A}_U(\mathbf{Z}, \mathbf{X})$ be defined exactly the same as $\mathcal{A}_s(\mathbf{Z}, \mathbf{X})$ (same π) but with the randomizers $\mathcal{R}^{(i)}$ replaced by $\mathcal{L}_U^{(i)}$. Then by (47) and (48),

$$\mathcal{A}_{\mathcal{L}}(\mathbf{X}_0) = \frac{e^{\epsilon_0^r}}{e^{\epsilon_0^r}+1}\mathcal{A}_U(\mathbf{X}_0) + \frac{1}{e^{\epsilon_0^r}+1}\mathcal{A}_U(\mathbf{X}_1) \text{ and } \mathcal{A}_{\mathcal{L}}(\mathbf{X}_1) = \frac{1}{e^{\epsilon_0^r}+1}\mathcal{A}_U(\mathbf{X}_0) + \frac{e^{\epsilon_0^r}}{e^{\epsilon_0^r}+1}\mathcal{A}_U(\mathbf{X}_1). \tag{50}$$

Then by (49), for any $X \in \mathcal{X}^n \backslash \{X_1^0, X_1^1\}$ and any $\mathbf{Z} = \mathbf{Z}_{1:r-1} \in \mathcal{Z}^{(r-1)\times N}$, we have $\mathcal{L}_U^{(i)}(\mathbf{Z}, X) = \frac{1}{2e^{\widetilde{\epsilon_0}}}\mathcal{L}_U^{(i)}(\mathbf{Z}, X_1^0) + \frac{1}{2e^{\widetilde{\epsilon_0}}}\mathcal{L}_U^{(i)}(\mathbf{Z}, X_1^1) + (1 - e^{-\widetilde{\epsilon_0}})LO^{(i)}(\mathbf{Z}, X)$. Hence, Lemma E.3 (with $p := e^{-\widetilde{\epsilon_0}} = e^{-n\epsilon_0^r}$ implies that $\mathcal{A}_U(\mathbf{X}_0)$ and $\mathcal{A}_U(\mathbf{X}_1))$ are

$$\left(\log\left(1 + \frac{8\sqrt{e^{\widetilde{\epsilon_0}}\ln(4/\delta^r)}}{\sqrt{N}} + \frac{8e^{\widetilde{\epsilon_0}}}{N}\right), \delta^r\right)$$

indistinguishable for any $\delta^r \geqslant 2e^{-Ne^{-n\epsilon_0^r}/16}$. Applying Lemma E.7 with $P := \mathcal{A}_{\mathcal{L}}(\mathbf{X}_0)$, $Q = \mathcal{A}_{\mathcal{L}}(\mathbf{X}_1)$, $q = \frac{e^{\epsilon_0^r}-1}{e^{\epsilon_0^r}+1}$, $P_1 = \mathcal{A}_U(\mathbf{X}_0)$, $Q_1 = \mathcal{A}_U(\mathbf{X}_1)$, and $P_0 = \frac{1}{2}(P_1 + Q_1)$ yields that $\mathcal{A}_{\mathcal{L}}(\mathbf{X}_0)$ and $\mathcal{A}_{\mathcal{L}}(\mathbf{X}_1)$ are (ϵ^r, δ^r) indistinguishable, as desired. This proves the claim and hence (by Lemma E.6, as described earlier) the theorem. □

[12] This follows from the assumption given in the lead up to Definition 2 that $\mathcal{R}^{(i)}(\mathbf{Z}_{1:r-1}, X)$ is conditionally independent of X' given $\mathbf{Z}_{1:r-1}$ for all $\mathbf{Z}_{1:r-1}$ and $X \neq X'$.

Step 2: Combine Theorem E.1 with the following CDP SCO lower bounds which follow from [9, Theorems 5.3/5.5], [8, Appendix C], and the non-private SCO lower bounds (see [59, 3]) [13]:

Theorem E.3. *[8, 9] Let $\mu, D, \epsilon > 0$, $L \geqslant \mu D$, and $\delta = o(1/nN)$. Consider $\mathcal{X} := \{\frac{-D}{\sqrt{d}}, \frac{D}{\sqrt{d}}\}^d \subset \mathbb{R}^d$ and $\mathcal{W} := B_2(0, D) \subset \mathbb{R}^d$. Let $\mathcal{A} : \mathcal{X}^{nN} \to \mathcal{W}$ be any (ϵ, δ)-CDP algorithm. Then:*
1. There exists a ($\mu = 0$) convex, linear (β-smooth for any β), L-Lipschitz loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and a distribution $\mathcal{D}$ on $\mathcal{X}$ such that the expected loss of $\mathcal{A}$ is lower bounded as

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{\Omega}\left(LD\left(\frac{1}{\sqrt{Nn}} + \min\left\{1, \frac{\sqrt{d}}{\epsilon n N}\right\}\right)\right).$$

2. There exists a μ-strongly convex, μ-smooth, L-Lipschitz loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and a distribution $\mathcal{D}$ on $\mathcal{X}$ such that the expected loss of $\mathcal{A}$ is lower bounded as

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{\Omega}\left(\frac{L^2}{\mu n N} + LD\min\left\{1, \frac{d}{\epsilon^2 n^2 N^2}\right\}\right).$$

Namely, if $\mathcal{A}$ is (ϵ_0, δ_0)-LDP, then (under the hypotheses of Theorem 3.1) $\mathcal{A}_s$ is (ϵ, δ)-CDP for $\epsilon = \widetilde{O}(\epsilon_0/\sqrt{N})$, so Theorem E.3 implies that the excess loss of $\mathcal{A}_s$ is lower bounded as in Theorem E.3 with ϵ replaced by $\epsilon_0/\sqrt{N}$.

Step 3: We simply observe that when the expectation is taken over the randomness in sampling $\mathbf{X} \sim \mathcal{D}^{n \times N}$, the expected excess loss of $\mathcal{A}_s$ is identical to that of $\mathcal{A}$ (using the i.i.d. assumption). This completes the proof of Theorem 3.1.

E.3 Lower bounds for LDP Federated ERM

Formally, define the algorithm class $\mathbb{B}_{\epsilon_0, \delta_0} := \mathbb{B}$ to consist of those algorithms $\mathcal{A} \in \mathbb{A}_{\epsilon_0, \delta_0} = \mathbb{A}$ such that for any $\mathbf{X} \in \mathbb{X}$, $f \in \mathcal{F}_{L,D}$, the expected empirical loss of the shuffled algorithm $\mathcal{A}_s$ derived from $\mathcal{A}$ is upper bounded by the expected loss of $\mathcal{A}$: $\mathbb{E}_{\mathcal{A}, \{\pi_r\}_r}\widehat{F}(\mathcal{A}_s(\mathbf{X})) \lesssim \mathbb{E}_{\mathcal{A}}\widehat{F}(\mathcal{A}(\mathbf{X}))$. Here $\mathcal{A}_s$ denotes the algorithm that applies the randomizer $\mathcal{R}_r^{(i)}$ to $X_{\pi_r(i)}$ for all i, r, but otherwise behaves exactly like $\mathcal{A}$. This is not a very constructive definition but we will describe examples of algorithms in $\mathbb{B}$. $\mathbb{B}$ includes all compositional or sequentially interactive LDP algorithms that are symmetric with respect to each of the N clients, meaning that the aggregation functions g_r are symmetric (i.e. $g_r(Z_1, \cdots, Z_N) = g_r(Z_{\pi(1)}, \cdots Z_{\pi(N)})$ for all permutations π) and in each round r the randomizers $\mathcal{R}_r^{(i)} = \mathcal{R}_r$ are the same for all clients $i \in [N]$. ($\mathcal{R}_r^{(i)}$ can still change with r though.) For example, the three algorithms presented in Section 2 are all in $\mathbb{B}$. This is because the aggregation functions used in each round are simple averages of the $M = N$ noisy gradients received from all clients (and they are compositional) and the randomizers in round r are identical when $\epsilon_i = \epsilon_0, \delta_i = \delta_0, n_i = n, \mathcal{X}_i = \mathcal{X}$: each adds the same gaussian noise to the stochastic gradients. $\mathbb{B}$ also includes sequentially interactive algorithms that choose the order in which clients are processed uniformly at random. This is because the distributions of the updates of $\mathcal{A}$ and $\mathcal{A}_s$ are both averages over all permutations of $[N]$ of the conditional (on π) distributions of the randomizers applied to the π-permuted database.

Theorem E.4. *Let $n, d, N, R \in \mathbb{N}$, $\epsilon_0 \in (0, \sqrt{N}]$, $\delta_0 \in (0, 1)$ and $\mathcal{A} \in \mathbb{B}_{(\epsilon_0, \delta_0)}$ such that in every round $r \in [R]$, the local randomizers $\mathcal{R}_r^{(i)}(\mathbf{Z}_{(1:r-1)}, \cdot) : \mathcal{X}^n \to \mathcal{Z}$ are $(\epsilon_0^r, \delta_0^r)$-DP for all $i \in [N]$, $\mathbf{Z}_{(1:r-1)} \in \mathcal{Z}^{r-1 \times N}$, with $\epsilon_0^r \leqslant \frac{1}{n}$, and $\delta_0^r \geqslant \frac{2e^{-N/16}}{Nn}$. Assume moreover that $\sum_r \delta_0^r = o(1/n^2N^2)$ if $\mathcal{A}$ is compositional; if $\mathcal{A}$ is sequentially interactive, assume instead that $\delta_0 = o(1/n^2N^2)$. Then there exists a (linear, hence β-smooth $\forall \beta \geqslant 0$) loss function $f \in \mathcal{F}_{L,D}$ and a database $\mathbf{X} \in \mathcal{X}^{nN}$ for some $\mathcal{X}$ such that the excess empirical loss of $\mathcal{A}$ is lower bounded as:*

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{\Omega}\left(LD\min\left\{1, \frac{\sqrt{d}}{\epsilon_0 n \sqrt{N}}\right\}\right).$$

[13] Part 2 of Theorem E.3 follows from the alternate rescaling of [9]'s hard instance in which $g(w, x) = \frac{1}{2}\|w - x\|^2$ on $B_2(0, 1) \times \mathcal{X}$ is scaled to $f(w, x) = \mu g(w, x)$ on $\mathcal{W} \times \mathcal{X}$ given in Theorem E.3. Then f is μD-Lipschitz, μ-smooth, μ-strongly convex, and the excess empirical risk in [9, Theorem 5.5] is scaled by $\mu D^2 = LD$. See also [52, Proposition 2.7] and its proof.

Furthermore, there exists another (μ-smooth) $f \in \mathcal{G}_{\mu,L,D}$ such that

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{\Omega}\left(LD\min\left\{1, \frac{d}{\epsilon_0^2 n^2 N}\right\}\right).$$

Here, the $\widetilde{\Omega}$ notation hides logarithmic factors depending on δ_0^r, n, and N.

Proof. **Step 1** is identical to Step 1 of the proof of Theorem E.4. **Step 2** is very similar, but uses Theorem E.5 (below) instead of Theorem E.3 to lower bound the excess empirical loss of $\mathcal{A}_s$. Finally, the definition of $\mathbb{B}$ implies that the excess risk of $\mathcal{A}$ is the same as that of $\mathcal{A}_s$, hence the lower bound also applies to $\mathcal{A}$. □

Theorem E.5. *[9] Let $\mu, D, \epsilon > 0$, $L \geqslant \mu D$, and $\delta = o(1/nN)$. Consider $\mathcal{X} := \{\frac{-D}{\sqrt{d}}, \frac{D}{\sqrt{d}}\}^d \subset \mathbb{R}^d$ and $\mathcal{W} := B_2(0, D) \subset \mathbb{R}^d$. Let $\mathcal{A} : \mathcal{X}^{nN} \to \mathcal{W}$ be any (ϵ, δ)-CDP algorithm. Then:*
1. There exists a ($\mu = 0$) convex, linear (β-smooth for any β), L-Lipschitz loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and a database $\mathbf{X} \in \mathcal{X}^{nN}$ such that the expected empirical loss of $\mathcal{A}$ is lower bounded as

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{\Omega}\left(LD\min\left\{1, \frac{\sqrt{d}}{\epsilon n N}\right\}\right).$$

2. There exists a μ-strongly convex, μ-smooth, L-Lipschitz loss $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and a database $\mathbf{X} \in \mathcal{X}^{nN}$ such that the expected empirical loss of $\mathcal{A}$ is lower bounded as

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{\Omega}\left(LD\min\left\{1, \frac{d}{\epsilon^2 n^2 N^2}\right\}\right).$$

F Proof of Theorem 4.1

For this result, we will just prove the stated version with balanced data and same privacy needs across clients, and non-random $M_r = M \leqslant N$ (same setup as [32]).

Theorem 4.1 [Complete Version] *Let $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^d$ be β-smooth, L-Lipschitz, and μ-strongly convex (with $\mu = 0$ for convex case). Assume $\epsilon \leqslant \ln(2/\delta)$, $\delta \in (0, 1)$, and $M \geqslant 16\ln(18RM^2/N\delta)$ for R specified below. Then, there is a constant $C > 0$ such that setting $\sigma_i^2 := \frac{CL^2 RM\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}{n^2N^2\epsilon^2}$ ensures that the shuffled version of Algorithm 1 is (ϵ, δ)-CDP. Moreover, there exist $\eta_r = \eta$ and $\{\gamma_r\}_{r=0}^{R-1}$ such that the shuffled version of Algorithm 1 achieves the following upper bounds on excess loss:*
1. (Convex) Setting $R := \max\left(\frac{n^2N^2\epsilon^2}{M}, \frac{N}{M}, \min\left\{n, \frac{\epsilon^2 n^2 N^2}{dM}\right\}, \frac{\beta D}{L}\min\left\{\sqrt{nM}, \frac{\epsilon n N}{\sqrt{d}}\right\}\right)$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = O\left(LD\left(\frac{1}{\sqrt{nM}} + \frac{\sqrt{d\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}}{\epsilon n N}\right)\right). \tag{51}$$

2. (Strongly convex) $R := \max\left(\frac{n^2N^2\epsilon^2}{M}, \frac{N}{M}, \frac{8\beta}{\mu}\ln\left(\frac{\beta D^2\mu\epsilon^2 n^2 N^2}{dL^2}\right), \min\left\{n, \frac{\epsilon^2 n^2 N^2}{dM}\right\}\right)$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\frac{L^2}{\mu}\left(\frac{1}{nM} + \frac{d\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}{\epsilon^2 n^2 N^2}\right)\right). \tag{52}$$

Proof of Theorem 4.1. We fix $K = 1$ for simplicity, but note that $K > 1$ can also be used (see [32, Lemma 3] for details), which would improve the communication complexity of our algorithm by a factor of K in some parameter regimes.

Privacy: The privacy proof is similar to the proof of [32, Theorem 1], except we replace their use of [6] (for pure DP randomizers) with [25, Theorem 3.8] for our approximate DP randomizer (gaussian mechanism). Observe that in each round r, the model updates of the shuffled algorithm $\mathcal{A}_s^r$ can be viewed as post-processing of the composition $\mathcal{M}_r(\mathbf{X}) = \mathcal{S}_M \circ samp_{M,N}(Z_r^{(1)}, \cdots, Z_r^{(N)})$, where $\mathcal{S}_M$ uniformly randomly shuffles the M received reports, $samp_{M,N}$ is the mechanism that chooses M reports uniformly at random from N, and $Z_r^{(i)} = samp_{1,n}(\widehat{\mathcal{R}}_r(x_{i,1}), \cdots \widehat{\mathcal{R}}_r(x_{i,n}))$, where $\widehat{\mathcal{R}}(x) := \nabla f(w_r, x) + u$ and $u \sim N(0, \sigma^2 \mathbf{I}_d)$. Recall ([20, Theorem A.1]) that $\sigma^2 = \frac{8L^2 \ln(2/\widehat{\delta_0})}{\widehat{\epsilon_0}^2}$ suffices to ensure that $\widehat{\mathcal{R}}_r$ is $(\widehat{\epsilon_0}, \widehat{\delta_0})$-DP if $\widehat{\epsilon_0} \leqslant 1$. Now note that $\mathcal{M}_r(\mathbf{X}) = \widetilde{\mathcal{R}}^M(\mathcal{S}_M \circ samp_{M,N}(X_1, \cdots X_N)$, where $\widetilde{\mathcal{R}} : \mathcal{X}^n \to \mathcal{Z}$ is given by $X \mapsto samp_{1,n}(\widehat{\mathcal{R}}(x_1), \cdots, \widehat{\mathcal{R}}(x_n))$ and $\widetilde{\mathcal{R}}^M : \mathcal{X}^{nM} \to \mathcal{Z}^M$ is given by $\mathbf{X} \mapsto (\widetilde{\mathcal{R}}(X_1), \cdots \widetilde{\mathcal{R}}(X_M))$ for any $\mathbf{X} = (X_1, \cdots, X_M) \in \mathcal{X}^{nM}$. This is because we are applying the same randomizer (same additive Gaussian noise) across clients and the operators $\mathcal{S}_M$ and $\widetilde{\mathcal{R}}^M$ commute. (Also, applying a randomizer to all N clients and then randomly choosing M reports is equivalent to randomly choosing M clients and then applying the same randomizer to all M of these clients.) Therefore, conditional on the random subsampling of M out of N clients (denoted $(X_1, \cdots, X_M)$ for convenience), [25, Theorem 3.8] implies that $(\widehat{\mathcal{R}}(x_{\pi(1),1}), \cdots, \widehat{\mathcal{R}}(x_{\pi(1),n}), \cdots, \cdots, \widehat{\mathcal{R}}(x_{\pi(M),1}), \cdots, \widehat{\mathcal{R}}(x_{\pi(M),n}))$ is $(\widehat{\epsilon}, \widehat{\delta})$-CDP, where $\widehat{\epsilon} = O\left(\frac{\widehat{\epsilon_0}\sqrt{\ln(1/M\widehat{\delta_0})}}{\sqrt{M}}\right)$ and $\widehat{\delta} = 9M\widehat{\delta_0}$, provided $\widehat{\epsilon_0} \leqslant 1$ and $M \geqslant 16\ln(2/\widehat{\delta_0})$ (which we will see is satisfied by our assumption on M). Next, privacy amplification by subsampling (see [71] and [32, Lemma 3]) clients and local samples implies that $\mathcal{M}_r$ is (ϵ^r, δ^r)-CDP, where $\epsilon^r = \frac{2\widehat{\epsilon}M}{nN} = O\left(\widehat{\epsilon_0}\frac{\sqrt{M\ln(1/M\widehat{\delta_0})}}{nN}\right)$ and $\delta^r = \frac{M}{nN}\widehat{\delta} = \frac{9M^2}{nN}\widehat{\delta_0}$. Finally, by the advanced composition theorem [20, Theorem 3.20], to ensure $\mathcal{A}_s$ is (ϵ, δ)-CDP, it suffices to make each round $(\epsilon^r := \frac{\epsilon}{2\sqrt{2R\ln(1/\delta)}}, \delta^r := \delta/2R)$-CDP. Using the two equations to solve for $\widehat{\epsilon_0} = \frac{CnN\epsilon}{\sqrt{R\ln(1/\delta)\ln(RM/nN\delta)M}}$ for some $C > 0$ and $\widehat{\delta_0} = \frac{nN\delta}{18RM^2}$, we see that $\sigma^2 = O\left(\frac{L^2\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta))RM}{n^2N^2\epsilon^2}\right)$ ensures that $\mathcal{A}_s$ is (ϵ, δ)-CDP, i.e. that $\mathcal{A}$ is (ϵ, δ)-SDP. Note that our choices of R in the theorem (specifically $R \geqslant N/M$ and $R \geqslant \frac{n^2N^2\epsilon^2}{M}$) ensure that $\widehat{\delta_0}, \delta \leqslant 1$ and $\widehat{\epsilon_0} \lesssim 1$, so that [25, Theorem 3.8] indeed gives us the amplification by shuffling result used above.

Excess risk: The proof is very similar to the proof of Theorem D.1, except σ^2 is now smaller. *Convex case:* Set $\gamma_r = \gamma = 1/R$ for all r. Now (26), Lemma D.2, and Lemma D.1 together imply for any $\eta \leqslant 1/4\beta$ that

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) \lesssim \frac{L^2R\eta}{nM} + \frac{D^2}{\eta R} + \eta\left(L^2/MK + \frac{d\sigma^2}{M}\right).$$

Now plugging in $\eta := \min\left\{1/4\beta, \frac{D\sqrt{M}}{LR}\min\left\{\sqrt{n}, \frac{\epsilon nN}{\sqrt{dM\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}}\right\}\right\}$ yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) \lesssim LD\left(\max\left\{1/\sqrt{nM}, \frac{\sqrt{d\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}}{\epsilon nN}\right\}\right)$$
$$+ \frac{LD}{R\sqrt{M}}\min\left\{\sqrt{n}, \frac{\epsilon nN}{\sqrt{dM\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}}\right\} + \frac{\beta D^2}{R}.$$

Then one can verify that plugging in the prescribed R yields the stated excess population loss bound. *μ-strongly convex case:* Let $\mathbf{X} \in \mathcal{X}^{nN}$ and denote the empirical risk minimizer by $\widetilde{w}$. Then for any $\eta_t \leqslant 1/4\beta$, by (23) and the i.i.d assumption (so $\Upsilon^2 = 0$), we have (for all t)

$$\mathbb{E}\|w_{t+1} - \widetilde{w}\|^2 \leqslant (1 - \mu\eta_t)\mathbb{E}\|w_t - \widetilde{w}\|^2 - \eta_t(\mathbb{E}\widehat{F}(w_t) - \widehat{F}^*) + 2\eta_t^2\left(\frac{4L^2}{M} + \frac{d\sigma^2}{2M}\right).$$

Then by Lemma D.8 with $a = \mu$, $b = 1$, $c = 2\left(\frac{4L^2}{M} + \frac{dCL^2R\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}{\epsilon^2n^2N^2}\right)$, $g = 4\beta$, and $T = R$, there exists a constant stepsize $\widetilde{\eta}$ and averaging weights γ_r such that

$$\mathbb{E}\widehat{F}(\widehat{w}_R) - \widehat{F}(w^*) = \widetilde{O}\left(\beta D^2\exp\left(\frac{-\mu R}{4\beta}\right) + \frac{L^2}{\mu}\left(\frac{1}{MR} + \frac{d\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}{\epsilon^2n^2N^2}\right)\right)$$

by Jensen's inequality. (See Lemma D.8 for the explicit $\widetilde{\eta}$ and γ_r.) Hence taking $\eta = \min\{1/4\beta, \widetilde{\eta}\}$ and applying Lemma D.1 and Lemma D.2 yields

$$\mathbb{E}F(\widehat{w}_R) - F(w^*) = \widetilde{O}\left(\beta D^2\exp\left(\frac{-\mu R}{4\beta}\right) + \frac{L^2}{\mu}\left(\frac{1}{MR} + \frac{d\ln(RM^2/N\delta)\ln(R/\delta)\ln(1/\delta)}{\epsilon^2n^2N^2}\right) + \frac{L^2}{\mu Mn}\right).$$

Then one verifies that the prescribed R is large enough to achieve the stated excess population loss bound. □

G Experimental Details and Additional Results

Code for all of the experiments in this paper can be found at:
`https://github.com/lowya/Locally-Differentially-Private-Federated-Learning`

G.1 Linear Regression with Health Insurance Data

Data set: The data (`https://www.kaggle.com/mirichoi0218/insurance`), which is available under an Open Database license, consists of $\widetilde{N} = 1338$ observations. The target variable y is medical charges. There are $d - 1 = 6$ features: age, sex, BMI, number of children, smoker, and geographic region.

Experimental setup: For a given N, we grouped data into N (almost balanced) clients by sorting y in ascending order and then dividing into N groups, the first $N - 1$ of size $\lceil 1338/N \rceil$ and the remaining points in the last client. For each N, we ran experiments with $R = 35$. We ran 20 trials, each with a fresh random train/test (80/20) split. We recorded test error for $\epsilon \in \{0.5, 1, 2.5, 5, 7.5, 10\}$. We fixed $\delta_i = 1/n_i^2$ for all experiments.

To estimate υ_*^2, we followed the procedure used in [75], using Newton's method to compute w^* and then averaging $\|\nabla \widehat{F}_i(w^*)\|^2$ over all $i \in [N]$.

Preprocessing: We first numerically encoded the categorical variables and then standardized the numerical features *age* and *BMI* to have zero mean and unit variance.

Gradient clipping: In the absence of a reasonable a priori bound on the Lipschitz parameter of the squared loss (as is typical for unconstrained linear regression problems with potentially unbounded data), we incorporated gradient clipping [2] into the algorithms. We then calibrated the noise to the clip threshold L to ensure LDP. For fairness of comparison, we also allowed for clipping for the non-private algorithms (if it helped their performance).

Hyperparameter tuning: For each trial and each algorithm, we swept through a log-scale grid of 10 stepsizes and 5 clip thresholds 3 times, selected the parameter w that minimized average (over 3 repetitions) training error (among all 10 x 5 = 50), and computed the corresponding average test error. The stepsize grids we used ranged from e^{-8} and e^{1} for (LDP) MB-SGD and from e^{-10} to 1 for (LDP) Local SGD. The excess risk (train and test) we computed was for the normalized objective function $F(w, X, Y) = \|Y - wX\|^2/2N_0$ where $N_0 \in \{1070, 268\}$ (1070 for train, 268 for test) and X is $N_0 \times d$ with $d = 7$ (including a column of all 1s) and $Y \in \mathbb{R}^{N_0}$. The clip threshold grids were $\{100, 10000, 1000000, 100000000, 99999999999999999999999999999999\}$, with the last element corresponding to effectively no clipping.

It is important to note that pre-processing and hyperparameter tuning (and estimation of L) were not done in an LDP manner, since we did not want to detract focus from evaluation of LDP FL algorithms. [14]As a consequence, the overall privacy loss for the entire experimental process is higher than the ϵ indicated in the plots, which solely reflects the privacy loss from running the FL algorithms with fixed hyperparameters and (pre-processed) data. This remark also applies to the Logistic Reggression with MNIST experiment in the next section.

Choice of σ^2 and K: We used $\sigma_i^2 = \sigma^2 = \frac{256 L^2 \ln\left(\frac{2.5RK}{\delta n}\right) \ln\left(\frac{2}{\delta}\right) R}{n^2 \epsilon^2}$, where L is the clip threshold. This choice provides LDP by Theorem D.1. [15] We chose the smallest batch size that guarantees LDP, namely $K_i = \frac{\epsilon n_i}{4\sqrt{2R\ln(2/\delta)}}$.

[14] See [2, 49, 62] and the references therein for discussion of DP PCA and DP hyperparameter tuning.

[15] The logarithmic term here is slightly tighter than that used in the theoretical portion of this paper. By [71], it still ensures LDP.

G.2 Logistic Regression with MNIST

The data set can be downloaded from http://yann.lecun.com/exdb/mnist/. Our code does this for you automatically. The results of our experiment are shown in Fig. 3. We see that Algorithm 1 continues to outperform LDP Local SGD (FedAvg) across all privacy levels in this experimental setup. Furthermore, Algorithm 1 even outperforms *non-private* Local SGD for large $\epsilon \gtrsim\approx 12$. Additionally, in the MNIST experiments, we explore the role of communication unreliability ($M < N$) in performance. Below we present results when $M = 18$ and $M = 12$ clients are available in each round.

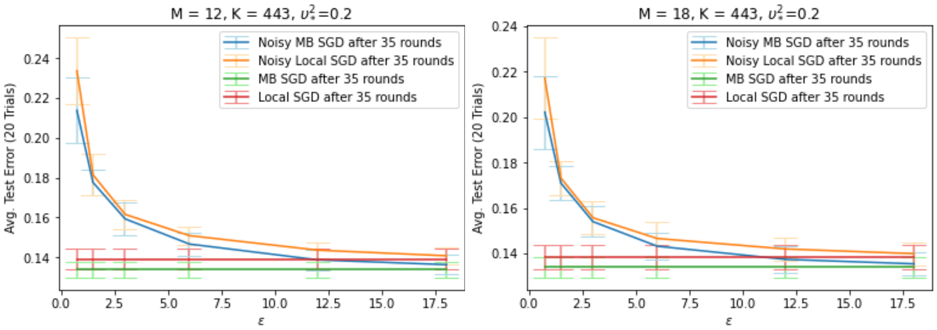

Figure 3: Test error vs. ϵ for linear regression on heterogeneous health insurance data. We display 90% error bars over the 20 trials (train/test splits). $\delta = 1/n^2$.

Experimental setup: To divide the data into $N = 25$ clients and for preprocessing, we borrow code from [75], which can be downloaded from:
`https://papers.nips.cc/paper/2020/hash/45713f6ff2041d3fdfae927b82488db8-Abstract.html`. It is available under a Creative Commons Attribution-Share Alike 3.0 license. There are $n_i = n = 8673$ training and 2168 test examples per client; to expedite training, we use only $1/7$ of the MNIST samples ($n = 1238$ training examples per client). We fix $\delta_i = \delta = 1/n^2$ and test $\epsilon \in \{0.75, 1.5, 3, 6, 12, 18\}$. The maximum υ_*^2 is about 0.17 for this problem (corresponding to each client having disjoint local data sets/pairs of digits).

Preprocessing: We used PCA to reduce the dimensionality to $d = 50$. We used an $80/20$ train/test split for all clients. To improve numerical stability, we clipped the input $\langle w, x \rangle$ (i.e. projected it onto $[-15, 15]$) before feeding into logistic loss.

Hyperparameter tuning: For each algorithm and each setting of $\epsilon, R, K, \upsilon_*^2$, we swept through a range of constant stepsizes and ran 3 trials to find the (approximately) optimal stepsize for that particular algorithm and experiment. We then used the corresponding w_R (averaged over the 3 runs) to compute test error. For (LDP) MB-SGD, the stepsize grid consisted of 10 evenly spaced points between e^{-6} and 1. For (LDP) Local SGD, the stepsizes were between e^{-8} and e^{-1}. We repeated this entire process twice for two fresh train/test splits of the data and reported the average test error in our plots.

Choice of σ^2 and K: In this experiment, we used smaller noise to get better utility (at the cost of larger K, hence larger computational cost, which is needed for privacy): $\sigma_i^2 = \sigma^2 = \frac{8L^2 \ln(1/\delta) R}{n^2 \epsilon^2}$, which provides LDP by [2, Theorem 1] if $K = \frac{n\sqrt{\epsilon}}{2\sqrt{R}}$ (c.f. [8, Theorem 3.1]). Here $L = 2 \max_{x \in X} \|x\|$ is an upper bound on the Lipschitz parameter of the logistic loss and was computed directly from the training data.

To estimate υ_*^2, we followed the procedure used in [75], using Newton's method to compute w^* and then averaging $\|\nabla \widehat{F}_i(w^*)\|^2$ over all $i \in [N]$.