ON THE MINIMAX EXCESS RISK AND $O(1)$-COMMUNICATION ALGORITHM FOR FEDERATED LEARNING UNDER STATISTICAL HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the minimax excess risk in federated learning (FL) with strongly convex and smooth loss functions under statistical heterogeneity. Previous work established a minimax excess risk lower bound that depends on the degree of statistical heterogeneity, but leaves large gaps with existing upper bounds in terms of problem-dependent parameters. In this work, we establish sharper lower bounds, which improve previous lower bound by $O(\mu^{-2})$ in the stochastic gradient regime where μ is strong convexity constant, is applicable to the exact gradient regime, and match existing upper bounds, resolving an open problem raised by (Chen et al., 2023). We also design a two-stage FL algorithm that is nearly optimal and only requires a $O(\kappa \ln M)$ communication rounds where κ is the condition number and M is the number of clients, while previous algorithms require a $O(T)$ communication rounds where T is the local sample size. The lower bound analysis requires a non-trivial extension of the local Fano's method proposed for the minimax risk analysis of centralized learning to deal with FL. The algorithmic novelties include a multistage stochastic gradient and a two-stage federated training strategy coupled with a model initialization scheme. Our results fully characterize the condition under which FL is better than non-cooperative learning, showing its advantages and limitations. Our algorithm provides a simple and communication-efficient FL framework that holds independent interest.

1 INTRODUCTION

Federated learning (FL) is an emerging machine learning paradigm, which enables multiple physically isolated devices/clients to collaboratively train models with the coordination of a central server, without direct access to local training datasets, which differentiates from typical distributed learning due to the heterogeneous data distributions across clients, massive clients and limited communication (McMahan et al., 2017; Li et al., 2020; Kairouz et al., 2021). Suppose there are M clients, each with a local training set $\mathcal{T}_k = \{\mathbf{z}_1, \ldots, \mathbf{z}_{T_k}\}$ independently sampled from a sample space $\mathcal{Z}$ following a distribution $\mathcal{D}_k$, $k = 1, \ldots, M$. In many real-world applications, due to feature distribution skew or label distribution skew, the data distributions across clients tend to be diverse, which we call it *statistical heterogeneity*. Let $\mathcal{W}$ be an unconstrained parameter space, such as $\mathbb{R}^d$ or a Hilbert space, and $f(\cdot, \cdot) : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. On each client k, FL aims to collaboratively solve the stochastic optimization problem (1) (Zhao et al., 2023; Yu et al., 2025).

$$\min_{\mathbf{w} \in \mathcal{W}} F_k(\mathbf{w}), \quad F_k(\mathbf{w}) := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_k} \left[f\left(\mathbf{w}, \mathbf{z}\right) \right]. \tag{1}$$

In this work, the loss function is restricted to be μ-strongly convex and β-smooth. Let $\mathbf{w}^k$ be a data-dependent estimator. From the system-wise perspective, the statistical performance of $\mathbf{w}^1, \ldots, \mathbf{w}^M$ is measured by the *weighted excess risk* (Chen et al., 2023) defined by

$$\mathcal{E}_{\mathbf{p}}\left(\mathbf{w}^1, \ldots, \mathbf{w}^M\right) = \sum\nolimits_{k=1}^{M} p_k \cdot F_k\left(\mathbf{w}^k\right) - \sum\nolimits_{k=1}^{M} p_k \cdot \min_{\mathbf{w} \in \mathcal{W}} F_k\left(\mathbf{w}\right), \tag{2}$$

where $p_k \in [0, 1]$ is a weight satisfying $\sum_{k=1}^{K} p_k = 1$. Besides, in a FL system, the wireless links between clients and server have limited bandwidths (Rothchild et al., 2020; Kairouz et al., 2021),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

making it necessary to minimize the communication cost typically measured by the communication rounds, denoted by R (Woodworth et al., 2021; Hou et al., 2022). The primary goal is to develop an algorithm that simultaneously minimizes the weighted excess risk (2) and the communication cost.

To characterize the weighted excess risk, it is critical to quantify the statistical heterogeneity. As an extreme case, suppose all of the local distributions are the same, i.e., $\mathcal{D}_k = \mathcal{D}$ for all $k = 1, \ldots, M$, then the weighted excess risk is equal to the *global excess risk* defined as follows,

$$\mathcal{E}_{\mathbf{p}}^{g}\left(\mathbf{w}^1, \ldots, \mathbf{w}^M\right) = \sum\nolimits_{k=1}^{M} p_k \cdot F_k(\mathbf{w}^k) - \min_{\mathbf{w}\in\mathcal{W}} \sum\nolimits_{k=1}^{M} p_k \cdot F_k\left(\mathbf{w}\right). \tag{3}$$

Most of existing FL algorithms measure their statistical performance by the global excess risk, such as FedAvg (McMahan et al., 2017), SCAFFOLD (Karimireddy et al., 2020), FedAdam (Reddi et al., 2021), FedChain (Hou et al., 2022). However, under heterogeneous distributions, the global excess risk is weaker than the weighted excess risk (2). There is a gap between them, denoted by Δ,

$$\Delta := \min_{\mathbf{w}\in\mathcal{W}} \sum\nolimits_{k=1}^{M} p_k \cdot F_k(\mathbf{w}) - \sum\nolimits_{k=1}^{M} p_k \cdot \min_{\mathbf{w}\in\mathcal{W}} F_k\left(\mathbf{w}\right). \tag{4}$$

This gap measures how the statistical performance of the globally optimal model differs from that of the M locally optimal models, capturing the degree of statistical heterogeneity (or termed heterogeneity level) (Li et al., 2020). If the data distributions are identical, then $\Delta = 0$. Otherwise, $\Delta > 0$. By (4), SCAFFOLD and FedChain separately enjoy a $O(\exp(-O(1)\frac{R}{\kappa^{\alpha}})\Upsilon(\mathbf{w}_{-1}) + \frac{\sigma^2}{\mu MT} + \Delta)$ weighted excess risk bound with $\alpha = 1$ and $\alpha = 0.5$, in which $\Upsilon(\mathbf{w}_{-1})$ measures the statistical performance of an initial model $\mathbf{w}_{-1}$, σ^2 is a weighted gradient variance and $\kappa = \frac{\beta}{\mu}$. If the heterogeneity level is large enough, such as $\Delta = \Omega(\frac{\sigma^2}{\mu T})$, a trivial approach that independently runs SGD or its variant on each client, which we call pure local training (PLT), is better than FL. Specifically, PLT with SGD achieves a $O(\beta\exp(-\frac{T}{4\kappa})\sum_{k=1}^{M}\frac{1}{M}\|\mathbf{w}_{-1}-\mathbf{w}_*^k\|_2^2+\frac{\sigma^2}{\mu T})$ weighted excess risk (Stich, 2019) without communication cost ($R = 0$). However, there is not a lower bound demonstrating the dichotomous strategy that makes a choice between FL algorithms and PLT is optimal.

There is a heterogeneity level equivalent to Δ (Appendix B), denoted by Δ_1 (Chen et al., 2023)

$$\Delta_1 := \sum\nolimits_{k=1}^{M} p_k \cdot \left\| \mathbf{w}_*^k - \sum\nolimits_{i=1}^{M} p_i \cdot \mathbf{w}_*^i \right\|_2^2, \quad \mathbf{w}_*^k = \arg\min_{\mathbf{w}\in\mathcal{W}} F_k\left(\mathbf{w}\right), \tag{5}$$

under which FedAvg enjoys a $O(\frac{\kappa}{MT} + \beta\Delta_1)$ weighted excess risk bound in a $O(T)$ communication rounds. If $\Delta_1 = \Omega(\frac{\sigma^2}{\beta\mu T})$, PLT with SGD is better than FedAvg. Thus a better approach is the dichotomous strategy that chooses either PLT with SGD or a FL algorithm such as FedAvg, SCAFFOLD, FedChain, by the heterogeneity level. Another approach is personalized federated learning, which collaboratively trains $\mathbf{w}^k$ for each client, such as pFedMe (Dinh et al., 2020), AFA (Even et al., 2022), FedCLUP (Yu et al., 2025) and so on. FedCLUP achieves a $O(\beta\exp(-O(1)\frac{R}{\kappa})\Upsilon(\mathbf{w}_{-1}) + \frac{\kappa\sigma^2}{\mu MT} + \Xi\min\{\Delta_1, \frac{1}{T}\})$ weighted excess risk bound [1] where Ξ depends on μ, β and σ^2. The dichotomous strategy attains the same excess risk bound with FedCLUP. There is also a $\Omega(\frac{\mu\sigma^2}{MT} + \mu\min\{\Delta_1, \frac{\sigma^2}{T}\})$ lower bound (Chen et al., 2023), showing the dichotomous strategy and FedCLUP are optimal in terms of T when $\sigma^2 \gg 0$. We identify a loose condition under which FL outperforms PLT, i.e., $\Delta_1 = O(\frac{\sigma^2}{M^{c_1}T^{1+c_2}})$ where $c_1 + c_2 > 0$ and $c_1 c_2 \geq 0$.

However, a complete characterization on the minimax weighted excess risk for FL under statistical heterogeneity remains absent. (i) In the case $\sigma^2 \gg 0$, there are large gaps between the lower bound and existing upper bounds in terms of μ and β. (ii) The lower bound in the case $\sigma^2 = 0$ is unknown, while the excess risk bounds of SCAFFOLD, FedChain and FedCLUP decay exponentially with R. As the optimal minimax weighted excess risk remains unknown, so do the conditions under which FL outperforms PLT, limiting our understanding of the advantages and limitations of FL.

Besides, previous FL algorithms need a $O(T)$ communication rounds to achieve weighted excess risk bounds that decay exponentially when $\sigma^2 = 0$. Although they can reduce the communication cost to $R = O(\ln T)$ or $O(1)$, their excess risk bounds also become much worse. No FL algorithm can achieve both the exponentially decaying convergence rate and $O(1)$ communication cost.

[1] The original paper of FedCLUP only gives an upper bound on the individualized estimation error $\|\mathbf{w}^k - \mathbf{w}_*^k\|_2^2$. For smooth and strongly convex loss functions and unconstrained $\mathcal{W}$, the estimation error bound is converted to weighted excess risk bound by multiplying by β.

1.1 MAIN RESULTS

We establish sharper lower bounds on the minimax weighted excess risk, containing the cases $\sigma^2 > 0$ (stochastic gradient) and $\sigma^2 = 0$ (exact gradient), and design a two-stage FL algorithm, named TWS-Fed that is nearly optimal and only requires a $O(\kappa \ln M)$ communication rounds. We first propose a multistage stochastic gradient (M-SG), and independently execute M-SG on all clients at the first stage, yielding M local models. At the second stage, we propose a federated stochastic gradient (SG) for learning a global model. The key is to initialize the federated SG by the M local models returned from the first stage. Our main results are summarized in Table 1.

(i) Sharper Lower bounds. In the stochastic gradient regime, we prove a $\Omega(\frac{\sigma^2}{\mu MT} + \min\{\Delta, \frac{\sigma^2}{\mu T}\})$ lower bound. Our proof constructs a hard problem instance satisfying $\Delta = \frac{\mu}{2}\Delta_1$. By this equivalence, we improve the $\Omega(\frac{\mu\sigma^2}{MT} + \mu\min\{\Delta_1, \frac{\sigma^2}{T}\})$ lower bound (Chen et al., 2023) by a factor of $O(\mu^{-2})$. In the exact gradient regime, we prove a $\Omega(\mu\exp(-\frac{4T}{\sqrt{\kappa}-1})\sum_{k=1}^{M}\frac{1}{M}\|\mathbf{w}_{-1}-\mathbf{w}_*^k\|_2^2)$ lower bound for any $\kappa > 1$, which is independent of Δ. By independently running the Nesterov's accelerated gradient descent (Nesterov, 2004) on each client, we attain a matched upper bound. In terms of the dependence on problem-dependent constants, such as μ and β, our lower bounds match with existing upper bounds, resolving the first open problem raised by Chen et al. (2023).

(ii) Minimax excess risk of M-SG. On each client k, M-SG enjoys a $O(\mu\exp(\frac{-O(T)}{\kappa})\|\mathbf{w}_{-1} - \mathbf{w}_*^k\|_2^2 + \frac{\sigma^2}{\mu T})$ excess risk bound. Compared with our lower bounds with $M = 1$, M-SG is optimal up to $O(\sqrt{\kappa})$ factor appearing in the exponential decay term. M-SG also improves the excess risk bound of SGD (Stich, 2019) by $O(\kappa)$ in the exponential decay term that is $O(\beta\exp(\frac{-O(T)}{\kappa})\|\mathbf{w}_{-1}-\mathbf{w}_*^k\|_2^2$.

(iii) A nearly optimal and $O(1)$-communication FL algorithm [2] Our algorithm, TWS-Fed, enjoys a $O(\mu\exp(\frac{-O(T)}{\kappa})\sum_{k=1}^{M}\frac{1}{M}\|\mathbf{w}_{-1}-\mathbf{w}_*^k\|_2^2 + \sigma^2\frac{\ln M}{\mu MT} + \Delta)$ upper bound. Compared with our lower bounds, in the case of $\Delta = O(\frac{\sigma^2}{\mu T})$, our algorithm is optimal up to $O(\sqrt{\kappa})$ and $O(\ln M)$ factor. By Table 1, our algorithm only requires a $O(\kappa \ln M)$ communication rounds, while SCAFFOLD, FedChain and FedCLUP require a $O(T)$ communication rounds for achieving the same excess risk bound, or require the same communication cost achieving excess risk bounds that are far from optimal. Besides, our algorithm require a weaker condition than SCAFFOLD and FedChain.

(iv) Optimal condition under which FL is better than non-cooperative learning. The dichotomous strategy that makes a choice between TWS-Fed and M-SG is nearly optimal, and improves the excess risk bound of FedCLUP in terms of μ and β. Our results fully characterize the minimax weighted excess risk of FL under statistical heterogeneity, giving the optimal condition under which FL is better than non-cooperative learning, i.e., $\Delta = O(\frac{M^{-c_1}\sigma^2}{\mu T^{1+c_2}})$ where $c_1 + c_2 > 0$ and $c_1 c_2 \geq 0$.

1.2 TECHNICAL CHALLENGES ON PROVING THE LOWER BOUNDS

In the stochastic gradient regime, our lower bound is $\Omega(\frac{\sigma^2}{\mu MT} + \min\{\Delta, \frac{\sigma^2}{\mu T}\})$ that depends on the strongly convex parameter μ and Δ. The first challenge is to precisely control the value of μ, which is the smallest eigenvalue of the Hessian matrix $\mathbb{E}[\nabla^2 f(\mathbf{w}, \mathbf{z})]$. Previous work (Chen et al., 2023) constructed a distributed logistic regression where $f(\mathbf{w}, \mathbf{z}) = \ln(1 + \exp(-y\mathbf{w}^\top\mathbf{x}))$ and $y \in \{-1, 1\}$ satisfying $\nabla^2 f(\mathbf{w}, \mathbf{z}) = \mathbf{x}\mathbf{x}^\top \exp(y\mathbf{w}^\top\mathbf{x}) \cdot (1 + \exp(y\mathbf{w}^\top\mathbf{x}))^{-2}$. As $\exp(y\mathbf{w}^\top\mathbf{x})$ is hard to control, it is also hard to get a precise μ, making their lower bound sub-optimal in terms of μ. To address this issue, we construct a distributed regression where $f(\mathbf{w}, \mathbf{z}) = (\mathbf{w}^\top\mathbf{x} - y)^2$ satisfying $\nabla^2 f(\mathbf{w}, \mathbf{z}) = 2\mathbf{x}\mathbf{x}^\top$. Let $\mathbf{x}$ be sampled from $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$ uniformly where $\mathbf{e}_i$ is the standard basis in $\mathbb{R}^d$. We have $\mu = \frac{2}{d}$. Then we extend the local Fano's method (Khas'minskii, 1979; Duchi et al., 2013) proposed for minimax risk analysis of centralized learning to deal with federated learning, and prove a lower bound that depends on d, naturally transferring to μ.

The second challenge lies in constructing hard instances such that Δ varies from 0 to a positive constant. We consider three cases. (i) $\Delta = O(\frac{\sigma^2}{\mu MT})$. The lower bound is $\Omega(\frac{\sigma^2}{\mu MT})$. (ii) $\omega(\frac{\sigma^2}{\mu MT}) = \Delta = O(\frac{\sigma^2}{\mu T})$. The lower bound is $\Omega(\Delta)$. (iii) $\Delta = \omega(\frac{\sigma^2}{\mu T})$. The lower bound is $\Omega(\frac{\sigma^2}{\mu T})$. Previous

[2] $O(1)$-communication means the communication cost is independent of T, but can depend on κ and M.

Algorithm / Lower Bound	Weighted Excess Risk $\mathcal{E}_{\mathbf{p}}(\mathbf{w}^1, \ldots, \mathbf{w}^M)$	Commu	Assumptions
Chen et al. (2023)	$\Omega\left(\frac{\mu\sigma^2}{MT} + \mu\min\{\Delta_1, \frac{\sigma^2}{T}\}\right)$	-	(a), (b)
Ours	$\Omega\left(\frac{\sigma^2}{\mu MT} + \min\{\Delta, \frac{\sigma^2}{\mu T}\}\right), \quad \Delta = \frac{\mu}{2}\Delta_1$	-	(a), (b)
	$\Omega\left(\mu\exp(-\frac{4T}{\sqrt{\kappa}-1})\Gamma_{\mathbf{p}}\right)$	-	(a), (d)
SCAFFOLD (Karimireddy et al., 2020)	$O\left(\mu\exp(-\frac{R}{162\kappa})\Gamma + \frac{\sigma^2}{\mu MT} + \Delta\right)$	R	(a), (c)
FedChain (Hou et al., 2022)	$O\left(\exp(\frac{-R}{\sqrt{\kappa}})\mathcal{E}^g_{\mathbf{p}}(\mathbf{w}_{-1}) + \frac{\sigma^2}{\mu MT} + \Delta\right)$	R	(a), (c)
FedAvg (Chen et al., 2023)	$O\left(\frac{\kappa}{MT} + \beta\Delta_1\right)$	$O(T)$	(a), (b)
FedCLUP (Yu et al., 2025)	$O\left(\beta\exp(-O(1)\frac{R}{\kappa})\Gamma_{\mathbf{p}} + \frac{\kappa\sigma^2}{\mu MT} + \Xi\min\{\Delta_1, \frac{1}{T}\}\right)$	R	(a), (b)
TWS-Fed (Ours)	$O\left(\mu\exp(-\frac{aT}{16\kappa})\Gamma_{\mathbf{p}} + \frac{\sigma^2\ln M}{(1-a)\mu MT} + \Delta\right), a \in (0,1)$	$O(\kappa\ln M)$	(a), (b)

Table 1: Comparison of weighted excess risk and communication rounds. "Commu" denotes "Communication". $\Gamma = \|\mathbf{w}_{-1} - \mathbf{w}_*\|_2^2$ where $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}}\frac{1}{M}\sum_{k=1}^M F_k(\mathbf{w})$, $\Gamma_{\mathbf{p}} = \frac{1}{M}\sum_{k=1}^M \|\mathbf{w}_{-1} - \mathbf{w}_*^k\|_2^2$, $\kappa = \beta/\mu$, $\Xi = \beta(\kappa^2(1+\frac{\sigma^2}{\mu^2}) + \frac{\sigma^4}{\mu^4})$. Assumptions: (a) β-smooth and μ-strongly convex loss function, (b) $\frac{1}{M}\sum_{k=1}^M \mathbb{E}[\|\nabla f(\mathbf{w}_*^k, \mathbf{z})\|_2^2] \le \sigma^2$, (c) $\forall k, \mathbf{w}, \mathbb{E}[\|\nabla f(\mathbf{w}, \mathbf{z}) - \nabla F_k(\mathbf{w})\|_2^2] \le \sigma^2$, (d) $\sigma = 0$ in Assumption (c). **Assumption (c) is stronger than (b)**.

work (Chen et al., 2023) constructed an optimal parameter $\mathbf{w}_*^k = \delta_k\mathbf{v}_k$ for each client $k \in [M]$ where $\mathbf{v}_k$ is uniformly sampled from a packing set of $\{-1, 1\}^d$. However, their analyses will yield a suboptimal lower bound in the first case. Specifically, if $0 < \Delta = O(\frac{\sigma^2}{\mu MT})$, their analysis will yield a $\Omega(\Delta)$ lower bound, rather than $\Omega(\frac{\sigma^2}{\mu MT})$. The reason is that they use a same $\delta_k = \delta$ for each client, so the lower bound is of the same order as Δ. To address this issue, we uniformly sample a $\mathbf{v}$ from a packing set of $\{-1, 1\}^d$, and construct $\mathbf{w}_*^k = \delta_k\mathbf{v}$ for each client. We set $\delta_k = \delta + \varepsilon$ for $k = 1, \ldots, (M+1)/2$ and $\delta_k = \delta - \varepsilon$ for others where $\delta = \Theta(\sqrt{\sigma^2/(\mu MT)})$ and $\varepsilon < \delta$. In this way, we can make Δ approach 0 and maintain a $\Omega(\frac{\sigma^2}{\mu MT})$ lower bound simultaneously.

1.3 Technical Challenges on Designing the Algorithm

By our lower bounds, if $\Delta = O(\frac{\sigma^2}{\mu T})$, it is enough to execute standard FL paradigm. The challenge is that existing FL algorithms, such as FedAvg, SCAFFOLD, etc, can not achieve nearly optimal excess risk at small communication cost. By Table 1, SCAFFOLD requires $R = \Theta(T)$ communication rounds for achieving a $O(\mu\exp(-\frac{O(T)}{\kappa})\Gamma + \frac{\sigma^2}{\mu MT} + \Delta)$ weighted excess risk bound. To address this issue, we propose a two-stage FL algorithm, coupled with an initialization scheme. In the first stage, we learn a local model $\hat{\mathbf{w}}_*^k$ on each client, and construct $\mathbf{w}_1 = \frac{1}{M}\sum_{k=1}^M \hat{\mathbf{w}}_*^k$. In the second stage, we perform federated SG starting from $\mathbf{w}_1$. If these local models satisfy the condition (6),

$$\forall k \in [M], \quad \mathbb{E}\left[F_k\left(\hat{\mathbf{w}}_*^k\right) - F_k\left(\mathbf{w}_*^k\right)\right] = O\left(\mu\exp\left(-\frac{O(T)}{\kappa}\right)\cdot\left\|\mathbf{w}_{-1} - \mathbf{w}_*^k\right\|_2^2 + \frac{\sigma_k^2}{\mu T}\right), \quad (6)$$

in which σ_k^2 is a local gradient variance, then the proposed TWS-Fed can achieve a nearly optimal excess risk bound in $O(\kappa\ln M)$ communication rounds.

The second technical challenge is to construct $\hat{\mathbf{w}}_*^k$ that satisfies (6). We can not run SGD or its existing variants (Needell et al., 2016; Stich, 2019; Vaswani et al., 2022; Wang et al., 2023) on each client, due to sub-optimal convergence rates (refer to Section 4.4.1). We propose a multistage stochastic gradient, similar to Epoch-GD (Hazan & Kale, 2011), Epoch-GD-F (Zhang & Zhou, 2019) and M-ASG (Aybat et al., 2019), and a non-uniform weight scheme aggregating the models generated on each stage. Compared with the uniform weight scheme adopted by Epoch-GD and Epoch-GD-F, and the last weight scheme adopted by M-ASG, which only yield the sub-optimal $O(\exp(-\frac{O(T)}{\kappa})\cdot[F_k(\mathbf{w}_{-1}) - F_k(\mathbf{w}_*^k)] + \frac{\sigma_k^2}{\mu T})$ convergence rate, our non-uniform weights utilize a negative term in the per-round convergence rate analysis, yielding the desired convergence rate.

2 PROBLEM SETTING

Let $\mathcal{Z}$ be a sample space, $\mathcal{W}$ be an unconstrained parameter space endowed with norm $\|\cdot\|_2$, such as $\mathbb{R}^d$ or Hilbert space, and $f(\cdot,\cdot):\mathcal{W}\times\mathcal{Z}\to\mathbb{R}$ be a loss function. Let $\nabla f(\cdot,\cdot)$ be the gradient in its first parameter. For an integer $n>0$, let $[n]=\{1,\dots,n\}$. Let $\mathbf{1}_M\in\mathbb{R}^M$ be a vector of ones.

We consider federated learning over M clients. For each client $k\in[M]$, suppose there is an i.i.d. training set $\mathcal{T}_k=\{\mathbf{z}_t : t\in[T]\}$ where $\mathbf{z}_t\in\mathcal{Z}$ follows a distribution $\mathcal{D}_k$. Let $F_k(\cdot)=\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_k}\left[f\left(\cdot,\mathbf{z}\right)\right]$ be expected loss function, and $\mathbf{w}_*^k=\arg\min_{\mathbf{w}\in\mathcal{W}}F_k(\mathbf{w})$. Suppose there are $i,k\in[M]$ such that $\mathcal{D}_k\neq\mathcal{D}_i$ which we call it *statistical heterogeneity*. In this work, we define federated learning as cooperatively solving the stochastic optimization problem (1) across all clients. Let $\mathcal{A}:\mathcal{Z}\to\mathcal{W}$ denote any algorithm (FL algorithm or non-cooperative learning algorithm) that takes $\{\mathcal{T}_k\}_{k=1}^M$ as input and outputs $\mathbf{w}^k$ on each client k. We measure the statistical performance of $\mathcal{A}$ by the weighted excess risk $\mathcal{E}_{\mathbf{p}}\left(\mathbf{w}^1,\dots,\mathbf{w}^M\right)$ (2). Note that it is possible $\mathbf{w}^k=\bar{\mathbf{w}}$ for all $k\in[M]$.

2.1 REGULARITY ASSUMPTIONS

Assumption 1 (β-smooth)**.** *$f(\cdot,\cdot)$ is β-smooth in its first parameter, that is,*

$$\forall\mathbf{w},\mathbf{v}\in\mathcal{W},\quad \|\nabla f(\mathbf{w},\cdot)-\nabla f(\mathbf{v},\cdot)\|_2\leq\beta\cdot\|\mathbf{w}-\mathbf{v}\|_2.$$

Assumption 2 (μ-strongly convex)**.** *$f(\cdot,\cdot)$ is convex in its first parameter. Besides, for any $k=1,2,\dots,M$, $F_k(\cdot)$ is μ-strongly convex, that is,*

$$\forall\mathbf{w},\mathbf{v}\in\mathcal{W},\quad F_k(\mathbf{w})-F_k(\mathbf{v})\geq\langle\nabla F_k(\mathbf{v}),\mathbf{w}-\mathbf{v}\rangle+\frac{\mu}{2}\|\mathbf{w}-\mathbf{v}\|_2^2. \tag{7}$$

Let $\kappa=\beta/\mu$ be the condition number. Many regularized risk minimization problems satisfy Assumption 1 and 2, such as $f(\mathbf{w},\mathbf{z})=l(\mathbf{w},\mathbf{z})+\lambda\cdot\|\mathbf{w}\|_2^2$ where $l(\mathbf{w},\mathbf{z})$ is the square loss function.

Assumption 3 (Bounded Gradient Variance)**.** *$\forall k\in[M]$, there is a constant $\sigma_k\geq0$ such that: (i) either $\forall\mathbf{w}\in\mathcal{W}$, $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_k}[\|\nabla f(\mathbf{w},\mathbf{z})-\nabla F_k(\mathbf{w})\|_2^2]\leq\sigma_k^2$, (ii) or $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_k}[\|\nabla f(\mathbf{w}_*^k,\mathbf{z})\|_2^2]\leq\sigma_k^2$.*

Assumption 3 (i) requires bounded gradient variance at any $\mathbf{w}\in\mathcal{W}$, which is general in stochastic optimization (Nemirovski et al., 2009; Ghadimi & Lan, 2012; 2013) and federated/distributed learning (Woodworth et al., 2021; Zhao et al., 2023). Assumption 3 (ii) is weaker than the first one, only requires bounded variance at $\mathbf{w}_*^k$ where $\nabla F_k\left(\mathbf{w}_*^k\right)=0$, and is also general in stochastic optimization (Bach & Moulines, 2011; Needell et al., 2016) and federated/distributed learning (Woodworth et al., 2020; Chen et al., 2023). Our results hold under Assumption 3 (ii).

2.2 QUANTIFYING THE DEGREE OF STATISTICAL HETEROGENEITY

Definition 1 (Degree of Statistical Heterogeneity (Li et al., 2020))**.** *There is a constant $\Delta\geq0$ such that $\sum_{k=1}^M p_k\left[F_k(\mathbf{w}_*)-F_k(\mathbf{w}_*^k)\right]=\Delta$ where $\mathbf{w}_*=\arg\min_{\mathbf{w}\in\mathcal{W}}\sum_{k=1}^M p_kF_k(\mathbf{w})$.*

In this work, we consider the case $p_k=\frac{1}{M}$. If there is no statistical heterogeneity, i.e., $\mathcal{D}_k=\mathcal{D}$ for all $k\in[M]$, then $\mathbf{w}_*^k=\mathbf{w}_*$ and $\Delta=0$. Otherwise, Δ may be very large. An equivalent definition with Δ is Δ_1 (5) (Chen et al., 2023). Lemma 9 (refer to Appendix B) shows the equivalence.

3 LOWER BOUNDS ON THE MINIMAX WEIGHTED EXCESS RISK

We separately establish a lower bound on the minimax weighted excess risk in the case of exact gradients (i.e., Assumption 3 (i) with $\sigma_k^2=0$) and stochastic gradients (i.e., Assumption 3 (ii) with $\sigma_k^2\neq0$).

Theorem 1. *Let $d\geq32$, $3\leq M\leq\exp(d/8)$ be odd, $\mathcal{W}=\mathbb{R}^d$, and $\sigma^2=M^{-1}\sum_{k=1}^M\sigma_k^2$. There are problem instances satisfying Assumption 2, Assumption 3 (ii) with $\sigma_k^2\neq0$, and Definition 1 with $\Delta\geq0$. On these problem instances, any algorithm with outputs $\mathbf{w}^1,\dots,\mathbf{w}^M$ satisfies*

$$\mathbb{E}\left[\mathcal{E}_{\frac{1}{M}\mathbf{1}_M}\left(\mathbf{w}^1,\dots,\mathbf{w}^M\right)\right]=\Omega\left(\min\left\{\Delta,\frac{\sigma^2}{\mu T}\right\}+\frac{\sigma^2}{\mu MT}\right), \tag{8}$$

where the expectation is taken over the randomness of the training sets.

Let $M \geq 2$ and $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{+\infty} : \|\mathbf{w}\|_2^2 < +\infty\}$. For any initial parameter $\mathbf{w}_{-1}$, there is a problem instance satisfying Assumption 1-2 with $\beta > \mu$, and Assumption 3 (i) with $\sigma_k^2 = 0$, and on this problem instance, any first-order algorithm with outputs $\mathbf{w}^1, \ldots, \mathbf{w}^M$ satisfies

$$\mathcal{E}_{\frac{1}{M}\mathbf{1}_M}\left(\mathbf{w}^1, \ldots, \mathbf{w}^M\right) \geq \frac{\mu}{2}\exp\left(-\frac{4T}{\sqrt{\kappa}-1}\right)\frac{1}{M}\sum\nolimits_{k=1}^{M}\left\|\mathbf{w}_{-1}-\mathbf{w}_*^k\right\|_2^2. \tag{9}$$

We compare with prior lower bounds. If Assumption 3 (ii) with $\sigma_k^2 \neq 0$ holds, there is a lower bound of $\Omega(\mu\min\{\Delta_1, \frac{\sigma^2}{T}\} + \mu\frac{\sigma^2}{MT})$ (Chen et al., 2023). To prove (8), we construct hard problem instances where $\Delta = \frac{\mu}{2}\Delta_1$. By this equivalence, we improve their lower bound by $O(\mu^{-2})$. If Assumption 3 (i) with $\sigma_k^2 = 0$ holds, then there is a lower bound on the global excess risk $\mathcal{E}_{\mathbf{p}}^g(\bar{\mathbf{w}}) \geq \exp(-4T)\mathcal{E}_{\mathbf{p}}^g(\mathbf{w}_{-1})$ (Mitra et al., 2021), which is not applicable to personalized FL and weaker than ours by $O(1/\sqrt{\kappa})$. There are lower bounds under diverse definitions on the heterogeneity level (Woodworth et al., 2020; Hou et al., 2022) that are also not applicable to personalized FL.

The lower bound (8) shows how the statistical performance of FL algorithm (including Personalized FL), depends on the heterogeneity level. If $\Delta = O(\frac{\sigma^2}{\mu MT})$, the lower bound is $\Omega(\frac{\sigma^2}{\mu MT})$ and FL improves the excess risk bound of PLT that is $O(\frac{\sigma^2}{\mu T})$ by a factor of $O(\frac{1}{M})$. As Δ increases, the lower bound is $\Omega(\Delta)$ and the improvement becomes restricted. When $\Delta = \Omega(\frac{\sigma^2}{\mu T})$, then PLT is optimal, and thus there is no need to execute FL algorithm. It is interesting that the lower bound (9) is independent of Δ. By independently running the Nesterov's accelerated gradient descent (Nesterov, 2004) on each client, we attain a matched upper bound. This result shows that federated learning is also unnecessary if exact gradients are observed.

4 A Nearly Optimal and $O(1)$ Communication Algorithm

4.1 Overview of Algorithm

The global excess risk of existing FL algorithms that output a global parameter $\bar{\mathbf{w}}$ (Woodworth et al., 2020; Karimireddy et al., 2020; Hou et al., 2022; Yu et al., 2025) consists of two parts, defined by

$$\sum\nolimits_{k=1}^{M} p_k \cdot [F_k(\bar{\mathbf{w}}) - F_k(\mathbf{w}_*)] = O\left(\exp\left(-O(R)/\kappa^{\alpha}\right)\right)\cdot\Gamma(\mathbf{w}_1) + \sigma^2 \cdot O\left(T^{-1}\right), \quad \alpha \in \{0.5, 1\},$$

where $\Gamma(\mathbf{w}_1) = \sum_k p_k \cdot [F_k(\mathbf{w}_1) - F_k(\mathbf{w}_*)]$ or $\Gamma(\mathbf{w}_1) = \|\mathbf{w}_1 - \mathbf{w}_*\|_2^2$, $\mathbf{w}_1$ is an initial parameter, and R is the communication rounds. It seems that $R = \Theta(T)$ is necessary for ensuring the first term is $O\left(\exp\left(-O(T)/\kappa^{\alpha}\right)\right)$. However, it was omitted that a tight upper bound can be obtained by reducing $\Gamma(\mathbf{w}_1)$. For instance, if $\Gamma(\mathbf{w}_1) = O(\exp(-O(T)/\kappa^{\alpha}))$, then we can set $R = \Theta(1)$. Inspired by this observation, we first learn a good initialization $\mathbf{w}_1$ with a fast convergence rate.

At the first stage, our algorithm executes PLT to learn $\mathbf{w}_1$. We first introduce M-SG, a variant of M-ASG (multistage Nesterov's accelerated stochastic gradient) (Aybat et al., 2019). Specifically, we set the momentum parameter to zero in M-ASG and redefine the initial model of the new stage as a weighted average of the models of the previous stage. On each client i, we divide its training set $\mathcal{T}_i$ into two disjoint subsets $\mathcal{T}_{i,1}$ and $\mathcal{T}_{i,2}$, and run M-SG on $\mathcal{T}_{i,1}$, learning a local $\hat{\mathbf{w}}_*^i$. Then we enter **the second stage**, in which we propose a federated SG with the initialization $\mathbf{w}_1 = \frac{1}{M}\sum_{i=1}^{M}\hat{\mathbf{w}}_*^i$. Benefit from the initialization, there is no need to run the more complex federated M-SG.

4.2 Stage One: Pure Local Training

We give M-SG in Algorithm 1. On each client i, let the size of $\mathcal{T}_{i,1}$ and $\mathcal{T}_{i,2}$ be T_1 and T_2, respectively, in which $T_1 + T_2 = T$. We further divide $\mathcal{T}_{i,1}$ into $K+1$ disjoint subsets, denoted by $\mathcal{T}_{i,1} = \cup_{k=0}^{K}\mathcal{T}_{i,1,k}$. For each $k \in \{0, 1, \ldots, K\}$, let s_k and e_k be the start and end point of $\mathcal{T}_{i,1,k}$, respectively, and let the size of $\mathcal{T}_{i,1,k}$ be $T_{1,k}$. The main idea of M-SG is to run stochastic gradient method on each stage (i.e., $\mathcal{T}_{i,1,k}$) with different hyper-parameters. Specifically, at each round $t \in \mathcal{T}_{i,1,k}$, let $\mathbf{w}_{k,t}^i$ be the current parameter on the i-th client. It was updated as follows

$$\forall t \in \mathcal{T}_{i,1,k}, \quad \mathbf{w}_{k,t+1}^i = \mathbf{w}_{k,t}^i - \eta_k \cdot \nabla f\left(\mathbf{w}_{k,t}^i, \mathbf{z}_{k,t}^i\right). \tag{10}$$

At the $(k+1)$-th stage, $\mathbf{w}^i_{k+1,s_{k+1}}$ is initialized as follows

$$\mathbf{w}^i_{k+1,s_{k+1}} := \bar{\mathbf{w}}^i_k = \sum_{t=s_k}^{e_k} \frac{\varepsilon_k^{e_k-t}}{\sum_{\tau=s_k}^{e_k}\varepsilon_k^{e_k-\tau}}\mathbf{w}^i_{k,t}, \quad \text{where} \quad \varepsilon_k = 1-\frac{1}{2}\eta_k\mu. \tag{11}$$

Let $\mathbf{w}_{-1} \in \mathcal{W}$ be an initialization. We set $\mathbf{w}^i_{0,s_0} = \mathbf{w}_{-1}$.

Our initialization (11) uses non-uniform weights to aggregate the models in previous stage, which is different from M-ASG that uses the last model in previous stage and Epoch-GD (Hazan & Kale, 2011) and Epoch-GD-F (Zhang & Zhou, 2019) that use uniform weights. Our non-uniform weights are derived from the negative term, $-\frac{1}{2}\eta_k\mu\|\mathbf{w}^i_{k,t}-\mathbf{w}^i_*\|_2^2$, in per-round convergence rate analysis,

$$\left\|\mathbf{w}^i_{k,t+1}-\mathbf{w}^i_*\right\|_2^2 \le \left\|\mathbf{w}^i_{k,t}-\mathbf{w}^i_*\right\|_2^2 - \frac{1}{2}\eta_k\mu\left\|\mathbf{w}^i_{k,t}-\mathbf{w}^i_*\right\|_2^2 - \eta_k\left(F_i\left(\mathbf{w}^i_{k,t}\right)-F_i\left(\mathbf{w}^i_*\right)\right)+\Gamma_{k,t},$$

where $\Gamma_{k,t}$ contains a variance term and some error terms. The analyses of M-ASG, Epoch-GD and Epoch-GD-F did not utilize this negative term. Benefit from the non-uniform weights, M-SG achieves the desired exponential decay term $O(\mu\exp(-O(T)\cdot\kappa^{-1})\cdot\left\|\mathbf{w}_{-1}-\mathbf{w}^i_*\right\|_2^2)$ (refer to Theorem 1 with $M=1$), rather than a weaker one $O(\exp(-O(T)\cdot\kappa^{-1})\cdot[F_i(\mathbf{w}_{-1})-F_i(\mathbf{w}^i_*)])$.

Algorithm 1 M-SG on the i-th client

Require: c_0, K, β, μ, T_1, $\mathbf{w}_{-1}$
Ensure: $\mathbf{w}^i_{0,s_0}=\mathbf{w}_{-1}$, $\eta_0=\eta_1=\frac{1}{2\beta}$, $T_{1,0}=\frac{1}{c_0}T_1$, $T_{1,1}=\lceil\frac{1}{\eta_1\mu}2^{1.5}\rceil$
1: **for** $k=0,1,\ldots,K$ **do**
2: **for** $t=s_k,s_k+1,\ldots,e_k$ **do**
3: Set $\mathbf{w}^i_{k,t+1}=\mathbf{w}^i_{k,t}-\eta_k\cdot\nabla f\left(\mathbf{w}^i_{k,t},\mathbf{z}^i_{k,t}\right)$
4: **end for**
5: Set $\mathbf{w}^i_{k+1,s_{k+1}}=\bar{\mathbf{w}}^i_k$ following (11)
6: **end for**
7: Return $\hat{\mathbf{w}}^i_* := \bar{\mathbf{w}}^i_K$

4.3 Stage Two: Federated Training

We execute federated training on the training sets $\{\mathcal{T}_{i,2}\}_{i=1}^M$. We limit the rounds of communication to $R \le T$, and randomly select $S>1$ clients at each communication round. Suppose M is divisible by S. We divide $[M]$ into M/S disjoint subsets with equal size S, denoted by $\mathcal{S}_1,\ldots,\mathcal{S}_{M/S}$. At each communication round $r \le R$, server randomly samples a subset $\mathcal{S}_{I_r}$ where $I_r \in \{1,\ldots,M/S\}$. To maintain nearly optimal excess risk bounds and achieve a $O(1)$ communication rounds simultaneously, the key is to construct a good initialization, denoted by $\mathbf{w}_1$. As $\hat{\mathbf{w}}^i_*$ approximates $\mathbf{w}^i_*$ well for all $i=1,\ldots,M$, it is natural to construct $\mathbf{w}_1$ by these local parameters.

Let N and R be two positive integers satisfying $T_2 = NR$. We divide $\{1,\ldots,T_2\}$ into R disjoint epochs, denoted by $\{E_r\}_{r=1}^R$ where $E_r=\{s_r,s_r+1\ldots,e_r\}$, $s_r=(r-1)N+1$ and $e_r=r\cdot N$ is the start and the end point, respectively. Server first initializes $\mathbf{w}_1$ as follows

$$\mathbf{w}_1 = \frac{1}{M}\cdot\sum\nolimits_{i=1}^M \hat{\mathbf{w}}^i_*. \tag{12}$$

At the beginning of E_r, server samples a subset of clients $\mathcal{S}_{I_r}$ and broadcasts $\mathbf{w}_r$ to the selected clients. For each client $i\in\mathcal{S}_{I_r}$, initializing $\mathbf{w}^i_{r,s_r}=\mathbf{w}_r$. Then clients update local parameters following (13), and server updates global parameter following (14).

$$\forall t\in E_r,\quad i\in\mathcal{S}_{I_r},\quad \mathbf{w}^i_{r,t+1}=\mathbf{w}^i_{r,t}-\eta_l\cdot\nabla f\left(\mathbf{w}^i_{r,t},\mathbf{z}^i_{r,t}\right), \tag{13}$$

$$\mathbf{w}_{r+1}=\mathbf{w}_r-\eta_g\cdot\frac{1}{S}\sum\nolimits_{i\in\mathcal{S}_{I_r}}\eta_l\cdot\sum\nolimits_{t=s_r}^{e_r}\nabla f\left(\mathbf{w}^i_{r,t},\mathbf{z}^i_{r,t}\right), \tag{14}$$

in which η_l is the local learning rate and η_g is the global learning rate.

At the end of the R-th round, server outputs $\bar{\mathbf{w}}$ defined as follows,

$$\bar{\mathbf{w}}=\sum_{r=1}^R\frac{\xi^{R-r}}{\sum_{r=1}^R\xi^{R-r}}\mathbf{w}_r, \quad \text{where} \quad \xi=1-\frac{1}{2}N\mu\eta_l\eta_g. \tag{15}$$

We name this algorithm TWS-Fed (TWo-Stage Federated Learning) shown in Algorithm 2.

Algorithm 2 TWS-Fed

Require: c_0, K, β, μ, T_1, T_2, $\mathbf{w}_{-1}$, R, S
Ensure: $\mathbf{w}^i_{0,s_0} = \mathbf{w}_{-1}$, $\eta_0 = \eta_1 = \frac{1}{2\beta}$, $T_{1,0} = \frac{1}{c_0}T_1$, $T_{1,1} = \lceil \frac{1}{\eta_1 \mu} 2^{1.5} \rceil$, $\kappa = \beta/\mu$
1: **for** $i = 1, \ldots, M$ **do**
2: The client runs M-SG (Algorithm 1) independently on $\mathcal{T}_{i,1}$ and obtains $\hat{\mathbf{w}}^i_*$
3: The client sends $\hat{\mathbf{w}}^i_*$ to server
4: **end for**
5: Enter **Stage Two** (Algorithm 3)

Algorithm 3 TWS-Fed (**Stage Two**)

Require: $\{\hat{\mathbf{w}}^i_*, i = 1, \ldots, M\}$
1: Server initializes $\mathbf{w}_1$ following (12)
2: **for** $r = 1, \ldots, R$ **do**
3: Server selects $\mathcal{S}_{I_r}$ where I_r is sampled from $\{1, \ldots, \frac{M}{S}\}$ uniformly
4: Server broadcasts $\mathbf{w}_r$ to $\mathcal{S}_{I_r}$
5: **for** $i \in \mathcal{S}_{I_r}$ in parallel **do**
6: The client initializes $\mathbf{w}^i_{r,s_r} = \mathbf{w}_r$
7: **for** $t = s_r, s_r + 1, \ldots, e_r$ **do**
8: The client computes $\mathbf{w}^i_{r,t+1}$ following (13)
9: **end for**
10: The client sends $\mathbf{w}_r - \mathbf{w}^i_{r,e_r+1} = \eta_l \cdot \sum_{t=s_r}^{e_r} \nabla f\left(\mathbf{w}^i_{r,t}, \mathbf{z}^i_{r,t}\right)$ to server
11: **end for**
12: Server computes $\mathbf{w}_{r+1}$ following (14)
13: **end for**
14: Server computes $\bar{\mathbf{w}}$ following (15), and broadcasts it to all clients

4.4 Excess Risk Bounds

4.4.1 Excess Risk of M-SG

Theorem 2 (Excess Risk of M-SG)**.** *Suppose Assumption 1-2, and Assumption 3 (ii) hold, and* $T_1 \geq \lceil 38\kappa \rceil$. *There is a constant* $c_0 \in [\frac{7}{5}, 4]$ *and an integer* $K \geq 2$ *such that*

$$\begin{cases} T_{1,0} = T_1/c_0, \quad T_{1,1} = \left\lceil 2^{1.5}/(\eta_1 \mu) \right\rceil, \quad T_{1,k} = 2^{k-1} T_{1,1}, \quad k = 2, \ldots, K, \\ \eta_0 = \eta_1 = 1/(2\beta), \quad \eta_k = 2^{-(k-1)} \cdot \eta_1, \quad k = 2, \ldots, K. \end{cases}$$

The excess risk of M-SG satisfies

$$\forall i = 1, \ldots, M, \quad \mathbb{E}^i \left[F_i \left(\hat{\mathbf{w}}^i_* \right) \right] - F_i \left(\mathbf{w}^i_* \right) \leq \frac{3\mu}{10} \exp\left(-\frac{T_1}{16\kappa} \right) \cdot \left\| \mathbf{w}_{-1} - \mathbf{w}^i_* \right\|_2^2 + \frac{200\sigma_i^2}{\mu T_1}, \quad (16)$$

where the expectation $\mathbb{E}^i[\cdot]$ *is w.r.t. the randomness of* $\mathcal{T}_{i,1}$.

If Assumption 3 (i) with $\sigma_i^2 = 0$ holds, there is a $\Omega(\frac{\mu}{2} \exp(-O(1)\frac{T_1}{\sqrt{\kappa}}) \left\| \mathbf{w}_{-1} - \mathbf{w}^i_* \right\|_2^2)$ lower bound for first-order algorithms (Nemirovski & Yudin, 1983) (or Theorem 2.1.13 (Nesterov, 2004) and Theorem 3.15 (Bubeck, 2015)). If Assumption 3 (ii) with $\sigma_i^2 \neq 0$ holds, there is a $\Omega(\frac{\sigma_i^2}{\mu T_1})$ lower bound (Raginsky & Rakhlin, 2011; Nguyen et al., 2019). Thus M-SG is optimal up to $O(\sqrt{\kappa})$. It is interesting to study whether M-ASG is optimal under Assumption 3 (ii).

Under assumptions similar to those in Theorem 2, SGD with carefully designed learning rates (Stich, 2019) achieves a $O(\beta \exp(-\frac{T}{4\kappa}) \cdot \|\mathbf{w}_{-1} - \mathbf{w}^i_*\|_2^2 + \frac{\sigma_i^2}{\mu T})$ excess risk bound. In the case $\sigma_i^2 = 0$, M-SG improves its excess risk bound by a factor of $O(\kappa)$. SGD with exponential decay learning rates (Vaswani et al., 2022) achieves $\mathbb{E}[\|\hat{\mathbf{w}}^i_* - \mathbf{w}^i_*\|_2^2] = O(\exp(-\frac{T}{\kappa \ln T}) \cdot \|\mathbf{w}_{-1} - \mathbf{w}^i_*\|_2^2 + \frac{\kappa \sigma_i^2 \ln^2 T}{\mu^2 T})$, and SGD or its variants with Polyak learning rates (Loizou et al., 2021; Wang et al., 2023) achieves

$\mathbb{E}[\|\hat{\mathbf{w}}_*^i - \mathbf{w}_*^i\|_2^2] = O(\exp(-\frac{T}{\kappa}) \cdot \|\mathbf{w}_{-1} - \mathbf{w}_*^i\|_2^2 + \frac{\sigma_i^2}{\mu^2})$. By the smoothness inequality $F_i\left(\hat{\mathbf{w}}_*^i\right) - F_i\left(\mathbf{w}_*^i\right) \leq \frac{\beta}{2} \cdot \|\hat{\mathbf{w}}_*^i - \mathbf{w}_*^i\|_2^2$, the two types of excess risk bounds are worse than ours in terms of both T and problem-dependent constants.

4.4.2 Weighted Excess Risk and Communication Cost of TWS-Fed

Theorem 3. *Suppose Assumption 1-2, and Assumption 3 (ii) hold, and there are two positive integers N and R such that $NR = T_2$. Let $\sigma^2 = M^{-1}\sum_{k=1}^M \sigma_k^2$, $\mathbf{w}_1$ follow (12), $12SR\beta^2N^2\eta_l^2 = 1$ and $\eta = \eta_l\eta_g$. Suppose $12N\beta\eta_l \leq 1$ and $12\eta N\beta \leq 1$. The global excess risk of TWS-Fed satisfies*

$$N\eta \cdot \mathbb{E}\left[\sum_{r=1}^R \left(1 - \frac{N\eta\mu}{2}\right)^{R-r} \cdot \frac{1}{M}\sum\nolimits_{k=1}^M \left[F_k(\mathbf{w}_r) - F_k\left(\mathbf{w}_*\right)\right]\right]$$
$$\leq \left(1 - \frac{N\eta\mu}{2}\right)^R \cdot \|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \frac{\sigma^2}{ST_2\beta\mu} + \frac{8\eta\sigma^2}{S\mu} + \frac{2}{\mu}\Delta,$$

where the expectation is w.r.t. the randomness $\{\mathcal{T}_{k,2}\}_{k=1}^M$ and $\{\mathcal{S}_{I_r}\}_{r=1}^R$.

We take a close look at the global excess risk bound. The first term depends on the estimation error $\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2$, and decays exponentially with the communication rounds R. The second term and the third term depend on the variance σ^2 and the local sample size T_2. The last term depends on the heterogeneity level Δ. It seems that $R = T_2$ is required for making the first term decay exponentially at a rate $\exp(-\frac{N\eta\mu}{2}T_2)$. However, it omits $\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2] = O(\exp(-O(1)T_1/\kappa))$ which can be obtained from Theorem 2. If we set both $T_1 = \Theta(T)$ and $T_2 = \Theta(T)$, then it is enough to set $R = \Theta(1)$. Next we give the weighted excess risk bound by tuning η and R.

Corollary 1 (Weighted Excess Risk of TWS-Fed)**.** *Let $T_1 = a \cdot T$ and $T_2 = T - T_1$ where $a \in (0,1)$. Suppose $T_1 \geq \lceil 38\kappa \rceil$, $M \leq \exp(T_2/(24\kappa))$ and the conditions in Theorem 3 hold. Let*

$$\eta = \frac{2}{NR\mu}\ln S, \quad S > 1, \quad \textit{and} \quad R = \lceil 24\kappa \ln S \rceil. \tag{17}$$

The weighted excess risk of TWS-Fed satisfies

$$\mathbb{E}\left[\mathcal{E}_{\frac{\mathbf{1}_M}{M}}\left(\bar{\mathbf{w}}\right)\right] \leq \frac{3\mu}{4S}\sum_{k=1}^M \exp\left(\frac{-aT}{16\kappa}\right)\frac{\left\|\mathbf{w}_{-1} - \mathbf{w}_*^k\right\|_2^2}{M} + \left[\frac{450}{a} + \frac{1 + 16\ln S}{1-a}\right]\frac{\sigma^2}{ST\mu} + \frac{18 + 3S}{S}\Delta,$$

where the expectation is w.r.t. the randomness of $\{\mathcal{T}_k\}_{k=1}^M$ and $\{\mathcal{S}_{I_r}\}_{r=1}^R$.

TWS-Fed only requires $R = \lceil 24\kappa \ln S \rceil = O(1)$. By the lower bounds in Theorem 1, in the case $\Delta = O(\frac{\sigma^2}{\mu T})$, TWS-Fed is optimal up to $O(\sqrt{\kappa})$ on the exponential decay term and $O(\ln M)$ on the variance-dependent term. We naturally obtain an optimal learning strategy: if $\Delta = \Omega(\frac{\sigma^2}{\mu T})$, independently running M-SG on each client; otherwise, running TWS-Fed. This dichotomous strategy attains a $O\left(\mu\exp(-\frac{aT}{16\kappa})\Gamma_{\mathbf{p}} + \frac{\sigma^2\ln M}{\mu MT} + \min\{\Delta, \frac{\sigma^2}{\mu T}\}\right)$ weighted excess risk bound, nearly matching our lower bounds. Our results characterize the minimax weighted excess risk for heterogeneous FL, and give the optimal condition under which FL is better than non-cooperative learning.

5 Conclusion

In this paper, we have characterized the minimax weighted excess risk for FL under statistical heterogeneity, revealing how the heterogeneity level influences the statistical performance of FL algorithms, and giving the optimal condition under which FL is better than non-cooperative learning. We also propose the first nearly optimal and $O(1)$-communication FL algorithm. Our work contributes to a deeper understanding of both the advantages and limitations of FL, and a simple and communication-efficient two-stage FL framework.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on this paper. We have not used generative AI tools for other tasks with required disclosure. We also used generative AI tools to assist sentence-level text refinement for improving the writing. We take responsibility for the text produced with the aid of generative AI in this work.

ETHICS STATEMENT

This paper is purely theoretical. We identify no ethical concerns associated with this paper.

REPRODUCIBILITY STATEMENT

All of the assumptions required for our theoretical results are given in the main text. The full proofs for all theorems and corollaries are given in the Appendix.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Necdet Serhat Aybat, Alireza Fallah, Mert Gürbüzbalaban, and Asuman E. Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *Advances in Neural Information Processing Systems*, 32:8523–8534, 2019.

Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24:451–459, 2011.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, 2004.

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

Shuxiao Chen, Qinqing Zheng, Qi Long, and Weijie J. Su. Minimax estimation for personalized federated learning: An alternative between fedavg and local training? *Journal of Machine Learning Research*, 24:262:1–262:59, 2023.

Canh T. Dinh, Nguyen Hoang Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

John Duchi. Lecture notes on statistics and information theory. December 2023. URL `http://web.stanford.edu/class/stats311/lecture-notes.pdf`.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pp. 429–438, 2013.

Mathieu Even, Laurent Massoulié, and Kevin Scaman. On sample optimality in personalized collaborative and federated learning. *Advances in Neural Information Processing Systems*, 35:212–225, 2022.

Saeed Ghadimi and Guanghui 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.

Saeed Ghadimi and Guanghui 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.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 421–436, 2011.

Charlie Hou, Kiran Koshy Thekumparampil, Giulia Fanti, and Sewoong Oh. FedChain: Chained algorithms for near-optimal communication cost in federated learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2): 1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5132–5143, 2020.

R. Z. Khas'minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Junkang Liu, Yuanyuan Liu, Fanhua Shang, Hongying Liu, Jin Liu, and Wei Feng. Improving generalization in federated learning with highly heterogeneous data via momentum-based stochastic controlled weight averaging. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An Adaptive Learning Rate for Fast Convergence. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 1306–1314, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Ion Necoara, Yurii E. Nesterov, and François Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.

Deanna Needell, Nathan Srebro, and Rachel A. Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 155(1-2):549–573, 2016.

Arkadi Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Arkadii Semenovich Nemirovski and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

Phuong Ha Nguyen, Lam M. Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in SGD. *Advances in Neural Information Processing Systems*, 32:3660–3669, 2019.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. FetchSGD: Communication-efficient federated learning with sketching. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8253–8265, 2020.

Yue Sheng and Edgar Dobriban. One-shot distributed ridge regression in high dimensions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8763–8772, 2020.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *CoRR*, abs/1907.04232v2, 2019. URL https://arxiv.org/pdf/1907.04232v2.

Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 22015–22059, 2022.

Xiaoyu Wang, Mikael Johansson, and Tong Zhang. Generalized polyak step size for first order optimization with momentum. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 35836–35863, 2023.

Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs Local SGD for Heterogeneous Distributed Learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

Blake E. Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Proceedings of the 34th Annual Conference on Learning Theory*, pp. 4386–4437, 2021.

Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. FedCM: Federated learning with client-level momentum. *CoRR*, abs/2106.10874, 2021. URL https://arxiv.org/abs/2106.10874.

Dingzhi Yu, Yunuo Cai, Wei Jiang, and Lijun Zhang. Efficient algorithms for empirical group distributionally robust optimization and beyond. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 57384–57414, 2024.

Xin Yu, Zelin He, Ying Sun, Lingzhou Xue, and Runze Li. Understanding the statistical accuracy-communication trade-off in personalized federated learning with minimax guarantees. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Lijun Zhang and Zhi-Hua Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the $o(1/T)$ convergence rate. In *Proceedings of the 32nd Annual Conference on Learning Theory*, pp. 3160–3179, 2019.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

Xuyang Zhao, Huiyuan Wang, and Wei Lin. The aggregation-heterogeneity trade-off in federated learning. In *Proceedings of the 36th Annual Conference on Learning Theory*, pp. 5478–5502, 2023.

Martin Zinkevich, Markus Weimer, Alexander J. Smola, and Lihong Li. Parallelized stochastic gradient descent. *Advances in Neural Information Processing Systems*, 23:2595–2603, 2010.

A Related Work

Although the idea of two-stage learning paradigm is not new for federated learning (Hou et al., 2022; Chen et al., 2023) and stochastic optimization (Yu et al., 2024), the implementation of our proposed two-stage strategy is totally different from existing ones. FedChain (Hou et al., 2022) and FedProx (Chen et al., 2023) execute federated training at the first stage, obtaining a good initial model, but requiring a $O(T)$ communication rounds. Our algorithm independently executes M-SG on each client at the first stage, only requiring a single communication round. This procedure is same with parallelized training algorithms (Zinkevich et al., 2010; Zhang et al., 2013; Sheng & Dobriban, 2020). Then our algorithm executes federated training in a $O(\kappa \ln M)$ communication rounds. ALEM (Yu et al., 2024) is a two-stage algorithm for solving distributionally robust optimization. While ALEM and our algorithm share a similar strategy in the first stage, they employ distinct strategies in the second stage. Specifically, ALEM only uses the estimation of the local minimal empirical risks to estimate the local excess risk, but does not average the local models (refer to (12)) returned from the first stage. Thus the implementation of our proposed two-stage strategy is also different from that of ALEM.

Much work attempted to address the heterogeneous challenge by designing more elaborate algorithms to learn a global model. FedDyn (Acar et al., 2021) incorporates a dynamic regularizer for each client. FedCM (Xu et al., 2021) corrects the local updates by incorporating a momentum-like term. FedMoSWA (Liu et al., 2025) estimates the difference between local and global updating direction and uses it to correct the local updates. Our work shows, for strongly convex and smooth loss functions, these elaborate techniques can not improve the weighted excess risk bound compared with the simple federated stochastic gradient in FedAvg and our proposed algorithm. Besides, all of these algorithms can also not adapt to the heterogeneity level.

Finally, it is worth mentioning that there is an impossible result on establishing algorithms adapting to unknown heterogeneity level (Zhao et al., 2023). Our result does not contradict with this pessimistic result. The first reason is that they define the heterogeneity level from the client-wide perspective by the decay rate of the distances between local parameters, i.e., $\|\mathbf{w}_*^k - \mathbf{w}_*^i\|_2$ for $k \neq i$ (refer to the discussion on Definition 1), while our definition follows the system-wide perspective. The second one is that they measure the individualized estimation error of an algorithm, while we measure its weighted excess risk, which are not equivalent under statistical heterogeneity.

B Technical Lemmas

In this section, we give some important technical lemmas.

Lemma 1 (first-order optimality condition (Boyd & Vandenberghe, 2004)). *Let $\psi(\cdot) : \mathcal{W} \to \mathbb{R}$ be a convex and differential function, $\mathcal{W}$ be a convex domain, and $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \psi(\mathbf{w})$. Then one has*

$$\forall \mathbf{v} \in \mathcal{W}, \quad \langle \nabla \psi(\mathbf{w}_*), \mathbf{v} - \mathbf{w}_* \rangle \geq 0. \tag{18}$$

Lemma 2. *If $f(\cdot, \cdot)$ is β-smooth in its first parameter, then one has*

$$\forall \mathbf{w}, \mathbf{v}, \quad f(\mathbf{w}, \cdot) - f(\mathbf{v}, \cdot) \leq \langle \nabla f(\mathbf{v}, \cdot), \mathbf{w} - \mathbf{v} \rangle + \frac{\beta}{2} \|\mathbf{w} - \mathbf{v}\|_2^2.$$

Lemma 3 (Lemma 3.5 in Bubeck (2015)). *If $f(\cdot, \cdot)$ is convex and β-smooth in its first parameter, then one has*

$$\forall \mathbf{w}, \mathbf{v}, \quad f(\mathbf{w}, \cdot) - f(\mathbf{v}, \cdot) \leq \langle \nabla f(\mathbf{w}, \cdot), \mathbf{w} - \mathbf{v} \rangle - \frac{1}{2\beta} \|\nabla f(\mathbf{w}, \cdot) - \nabla f(\mathbf{v}, \cdot)\|_2^2. \tag{19}$$

Lemma 4 (Bubeck (2015)). *If $f(\cdot, \cdot)$ is convex and β-smooth in its first parameter, then one has*

$$\forall \mathbf{w}, \mathbf{v}, \quad \langle \nabla f(\mathbf{w}, \cdot) - \nabla f(\mathbf{v}, \cdot), \mathbf{w} - \mathbf{v} \rangle \geq \frac{1}{\beta} \|\nabla f(\mathbf{w}, \cdot) - \nabla f(\mathbf{v}, \cdot)\|_2^2. \tag{20}$$

Lemma 5 (relaxed triangle inequality (Karimireddy et al., 2020)). *For any $\mathbf{w}, \mathbf{v} \in \mathcal{W}$,*

$$\forall a > 0, \quad \|\mathbf{w} + \mathbf{v}\|_2^2 \leq (1 + a)\|\mathbf{w}\|_2^2 + \left(1 + \frac{1}{a}\right) \|\mathbf{v}\|_2^2. \tag{21}$$

Lemma 6. *For any $\mathbf{w} \in \mathcal{W}$, let $F(\mathbf{w}) = \mathbb{E}_{\mathbf{z}\in\mathcal{Z}}[f(\mathbf{w},\mathbf{z})]$. Let $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. If $f(\cdot,\cdot)$ is convex and β-smooth in its first parameter, then one has*

$$\begin{aligned}\forall \mathbf{w}\in\mathcal{W}, \mathbf{z}\in\mathcal{Z},\ &\|\nabla f(\mathbf{w},\mathbf{z})\|_2^2 \\ \leq& 2\|\nabla f(\mathbf{w}_*,\mathbf{z})\|_2^2 + 2\beta\langle \nabla F(\mathbf{w}), \mathbf{w}-\mathbf{w}_*\rangle + \\ & 2\beta\langle \nabla f(\mathbf{w},\mathbf{z}) - \nabla F(\mathbf{w}) + \nabla F(\mathbf{w}_*) - \nabla f(\mathbf{w}_*,\mathbf{z}), \mathbf{w}-\mathbf{w}_*\rangle .\end{aligned} \tag{22}$$

Lemma 6 shows that the empirical gradient variance at any $\mathbf{w}\in\mathcal{W}$ can be bounded by the empirical gradient variance at $\mathbf{w}_*$. This inequality is critical for analyzing the excess risk under Assumption 3 (ii), rather than the stronger Assumption 3 (i).

Proof. By Lemma 1 and Lemma 4, one has,

$$\begin{aligned}\|\nabla f(\mathbf{w},\mathbf{z})\|_2^2 &\overset{(21)}{\leq} 2\|\nabla f(\mathbf{w},\mathbf{z}) - \nabla f(\mathbf{w}_*,\mathbf{z})\|_2^2 + 2\|\nabla f(\mathbf{w}_*,\mathbf{z})\|_2^2 \\ &\overset{(20)}{\leq} 2\beta\langle \nabla f(\mathbf{w},\mathbf{z}) - \nabla f(\mathbf{w}_*,\mathbf{z}), \mathbf{w}-\mathbf{w}_*\rangle + 2\|\nabla f(\mathbf{w}_*,\mathbf{z})\|_2^2 \\ &= 2\beta\langle \nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*), \mathbf{w}-\mathbf{w}_*\rangle + 2\|\nabla f(\mathbf{w}_*,\mathbf{z})\|_2^2 + \\ &\quad 2\beta\langle \nabla f(\mathbf{w},\mathbf{z}) - \nabla F(\mathbf{w}) + \nabla F(\mathbf{w}_*) - \nabla f(\mathbf{w}_*,\mathbf{z}), \mathbf{w}-\mathbf{w}_*\rangle \\ &\overset{(18)}{\leq} 2\beta\langle \nabla F(\mathbf{w}), \mathbf{w}-\mathbf{w}_*\rangle + 2\|\nabla f(\mathbf{w}_*,\mathbf{z})\|_2^2 + \\ &\quad 2\beta\langle \nabla f(\mathbf{w},\mathbf{z}) - \nabla F(\mathbf{w}) + \nabla F(\mathbf{w}_*) - \nabla f(\mathbf{w}_*,\mathbf{z}), \mathbf{w}-\mathbf{w}_*\rangle ,\end{aligned}$$

which concludes the proof. □

Lemma 7 (Theorem 2.1.8 (Nesterov, 2004))**.** *Let $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. If $F(\mathbf{w})$ is μ-strongly convex in $\mathbf{w}$, then*

$$\forall \mathbf{w}\in\mathcal{W}, \quad F(\mathbf{w}) - F(\mathbf{w}_*) \geq \frac{\mu}{2}\|\mathbf{w}-\mathbf{w}_*\|_2^2 . \tag{23}$$

Lemma 7 is also referred to as the quadratic growth condition (Necoara et al., 2019). It is easy to prove this lemma by combining Lemma 1 and the μ-strongly convex property of $F(\cdot)$.

Lemma 8 (perturbed strong convexity, Lemma 5 in Karimireddy et al. (2020))**.** *If $f(\cdot,\cdot)$ is β-smooth and μ-strongly convex in its first parameter, then one has*

$$\langle \nabla f(\mathbf{u},\cdot), \mathbf{w}-\mathbf{v}\rangle \geq f(\mathbf{w},\cdot) - f(\mathbf{v},\cdot) + \frac{\mu}{4}\|\mathbf{w}-\mathbf{v}\|_2^2 - \beta\|\mathbf{u}-\mathbf{w}\|_2^2. \tag{24}$$

Next we establish two equivalent definitions on the heterogeneity level defined by Definition 1 under the conditions of Assumption 1-2.

Lemma 9 (Equivalent Definitions on the Heterogeneity Level)**.** *Let $\mathcal{W}$ be an unconstraint parameter space, $\bar{\mathbf{w}}_* = \frac{1}{M}\sum_{k=1}^M \mathbf{w}_*^k$ and Δ follow Definition 1. Under Assumptions 1-2, one has*

$$\frac{2}{\beta}\Delta \leq \sum_{k=1}^M p_k \left\|\mathbf{w}_*^k - \mathbf{w}_*\right\|_2^2 \leq \frac{2}{\mu}\Delta, \quad \text{and} \quad \frac{2}{\beta}\Delta \leq \sum_{k=1}^M p_k \left\|\mathbf{w}_*^k - \bar{\mathbf{w}}_*\right\|_2^2 \leq \frac{8}{\mu}\Delta. \tag{25}$$

By Lemma 9, we can equivalently define the heterogeneity level by the distances $\|\mathbf{w}_*^k - \mathbf{w}_*\|_2^2$ or $\|\mathbf{w}_*^k - \bar{\mathbf{w}}_*\|_2^2$, $k = 1,\dots,M$. Previous work defines the heterogeneity level by $\Delta_1 = \sum_{k=1}^M p_k \left\|\mathbf{w}_*^k - \bar{\mathbf{w}}_*\right\|_2^2$ (Chen et al., 2023), which is equivalent to ours.

Proof of Lemma 9. By the μ-strong convexity of $\{F_k(\cdot)\}_{k=1}^M$ (Assumption 2), one has

$$\Delta = \sum_{k=1}^M p_k \left[F_k(\mathbf{w}_*) - F_k\left(\mathbf{w}_*^k\right)\right] \overset{(23)}{\geq} \sum_{k=1}^M p_k \cdot \frac{\mu}{2}\left\|\mathbf{w}_* - \mathbf{w}_*^k\right\|_2^2 .$$

Similarly, by the β-smooth of $f(\cdot,\cdot)$ (Assumption 1), one has

$$F_k(\mathbf{w}_*) - F_k\left(\mathbf{w}_*^k\right) \leq \left\langle \nabla F_k\left(\mathbf{w}_*^k\right), \mathbf{w}_* - \mathbf{w}_*^k\right\rangle + \frac{\beta}{2}\left\|\mathbf{w}_* - \mathbf{w}_*^k\right\|_2^2 \leq \frac{\beta}{2}\left\|\mathbf{w}_* - \mathbf{w}_*^k\right\|_2^2 ,$$

in which $\nabla F_k\left(\mathbf{w}_*^k\right) = \mathbf{0}$ for unconstrained $\mathcal{W}$. Summing over $k = 1, \ldots, M$ yields,

$$\sum_{k=1}^{M} p_k \left\|\mathbf{w}_* - \mathbf{w}_*^k\right\|_2^2 \geq \frac{2}{\beta}\Delta.$$

Next we prove the second equivalence. By the β-smooth property of $f(\cdot, \cdot)$,

$$\Delta = \sum_{k=1}^{M} p_k \left[F_k\left(\mathbf{w}_*\right) - F_k\left(\mathbf{w}_*^k\right)\right] \leq \sum_{k=1}^{M} p_k \left[F_k\left(\bar{\mathbf{w}}_*\right) - F_k\left(\mathbf{w}_*^k\right)\right] \leq \frac{\beta}{2} \sum_{k=1}^{M} p_k \left\|\bar{\mathbf{w}}_* - \mathbf{w}_*^k\right\|_2^2.$$

Besides, by the definition of $\bar{\mathbf{w}}_*$ and the Jensen's inequality, there is an upper bound as follows,

$$\sum_{k=1}^{M} p_k \left\|\mathbf{w}_*^k - \bar{\mathbf{w}}_*\right\|_2^2 \leq \sum_{i=1}^{M} p_i \cdot \sum_{k=1}^{M} p_k \left\|\mathbf{w}_*^k - \mathbf{w}_*^i\right\|_2^2 \leq 4 \sum_{k=1}^{M} p_k \left\|\mathbf{w}_*^k - \mathbf{w}_*\right\|_2^2 \leq \frac{8\Delta}{\mu}.$$

We conclude the proof. □

Lemma 10. *Let $\mathcal{V} \subseteq \{-1, 1\}^d$ be a packing of the d-dimensional hypercube $\{-1, 1\}^d$ such that $\forall \mathbf{v}_1 \neq \mathbf{v}_2 \in \mathcal{V}$, $\|\mathbf{v}_1 - \mathbf{v}_2\|_1 \geq \frac{d}{2}$. Then for any integer $S \geq 2$ and $S \leq \exp(d/8)$, any $\delta_1 \geq 0, \ldots, \delta_S \geq 0$, and any S different elements in $\mathcal{V}$, one has,*

$$\forall \mathbf{v}_1 \neq \mathbf{v}_2 \neq \ldots \neq \mathbf{v}_S \in \mathcal{V}, \quad \left\|\frac{1}{S}\sum_{i=1}^{S} \delta_i \mathbf{v}_i\right\|_2^2 \leq \frac{d}{S^2}\sum_{i=1}^{S} \delta_i^2 + \frac{d}{2S^2}\sum_{i \neq j} \delta_i \delta_j. \tag{26}$$

Proof of Lemma 10. The Gilbert-Varshamov bound (refer to Lemma 8.2.3 (Duchi, 2023)) guarantees that $\mathcal{V}$ exists and $|\mathcal{V}| \geq \exp\left(\frac{d}{8}\right)$. Since $S \leq \exp(d/8)$, there exists $\mathbf{v}_1, \ldots, \mathbf{v}_S \in \mathcal{S}$.

$$\forall i \neq j, \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \frac{1}{2}\left(\|\mathbf{v}_i\|_2^2 + \|\mathbf{v}_j\|_2^2 - \|\mathbf{v}_i - \mathbf{v}_j\|_2^2\right) = \frac{1}{2}(2d - 2 \cdot \|\mathbf{v}_i - \mathbf{v}_j\|_1) \leq \frac{d}{2}.$$

Then

$$\begin{aligned}\left\|\frac{1}{S}\sum_{i=1}^{S} \delta_i \cdot \mathbf{v}_i\right\| =& \frac{1}{S^2}\sum_{i=1}^{S} \delta_i^2 \|\mathbf{v}_i\|_2^2 + \frac{1}{S^2}\sum_{i \neq j} \delta_i \delta_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle \\ \leq& \frac{d}{S^2}\sum_{i=1}^{S} \delta_i^2 + \frac{d}{2S^2}\sum_{i \neq j} \delta_i \delta_j,\end{aligned}$$

which concludes the proof. □

Lemma 11. *Let $\mathcal{V} \subseteq \{-1, 1\}^d$ be a packing of the d-dimensional hypercube $\{-1, 1\}^d$ such that $\forall \mathbf{v}_1 \neq \mathbf{v}_2 \in \mathcal{V}$, $\|\mathbf{v}_1 - \mathbf{v}_2\|_1 \geq \frac{d}{2}$. Then for any integer $S \geq 2$ and $S \leq \exp(d/8)$, one has*

$$\forall \mathbf{v}_1 \neq \mathbf{v}_2 \neq \ldots \neq \mathbf{v}_S \in \mathcal{V}, \quad \left\|\frac{1}{S}\sum_{i=1}^{S} \mathbf{v}_i\right\|_2^2 \leq \frac{3}{4}d.$$

Proof of Lemma 11. By Lemma 10, let $\delta_1 = \ldots = \delta_S = 1$, one has

$$\left\|\frac{1}{S}\sum_{i=1}^{S} \mathbf{v}_i\right\| \leq \frac{d}{S} + \frac{1}{S^2} S(S-1)\frac{d}{2} \leq \frac{3}{4}d,$$

which concludes the proof. □

C Proof of Theorem 1

Proof of Theorem 1. We borrow some techniques from the proof of Example 8.4.5 in Duchi (2023), specifically extending the local Fano's method (Duchi et al., 2013) to dealing with heterogeneous datasets, and the proof of Theorem 2.1.13 in Nesterov (2004) (or Theorem 3.15 in Bubeck (2015)).

C.1 Stochastic Gradients

We consider a supervised learning problem. Let $\mathbf{z} = (\mathbf{x}, y)$ where $\mathbf{x} \in \mathbb{R}^d$ is a feature vector and $y \in \mathbb{R}$ is the true output. Let $f(\mathbf{w}, \mathbf{z})$ be the square loss function defined by

$$\forall \mathbf{w} \in \mathcal{W}, \quad \forall \mathbf{z} = (\mathbf{x}, y), \quad f(\mathbf{w}, \mathbf{z}) = \left(\langle \mathbf{w}, \mathbf{x} \rangle - y\right)^2.$$

For each $i \in [d]$, denote by $\mathbf{e}_i = (0, 0, \ldots, 1, \ldots, 0)$ the canonical basis of $\mathbb{R}^d$. For each $k \in [M]$ and $t \in [T]$, let $\mathbf{z}_t^k = (\mathbf{x}_t^k, y_t^k)$ where we sample $\mathbf{x}_t^k$ uniformly and independently from $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$. Let $\mathcal{N}(0, \gamma^2)$ be the Gaussian distribution. For each client $k = 1, \ldots, M$, suppose there is a true regressor $\mathbf{w}_*^k \in \mathcal{W}$ that determines y_t^k in a way defined by

$$\forall k = 1, \ldots, M, \quad \forall t = 1, \ldots, T, \quad y_t^k = \left\langle \mathbf{w}_*^k, \mathbf{x}_t^k \right\rangle + \varepsilon_t^k, \quad \varepsilon_t^k \sim \mathcal{N}(0, \gamma^2).$$

y_t^k follows the Gaussian distribution $\mathcal{N}(\left\langle \mathbf{w}_*^k, \mathbf{x}_t^k \right\rangle, \gamma^2)$. The local expected loss function is

$$\forall k = 1, 2, \ldots, M, \quad F_k(\mathbf{w}) = \mathbb{E}_{\mathbf{x}, y}\left[f(\mathbf{w}, \mathbf{z})\right] = \frac{1}{d}\left\|\mathbf{w} - \mathbf{w}_*^k\right\|_2^2 + \gamma^2.$$

Next we solve the global optimal parameter $\mathbf{w}_*$. Let $F(\mathbf{w})$ be the global expected loss function,

$$F(\mathbf{w}) = \sum_{k=1}^M \frac{1}{M} F_k(\mathbf{w}) = \frac{1}{d}\sum_{k=1}^M \frac{1}{M}\left\|\mathbf{w} - \mathbf{w}_*^k\right\|_2^2 + \gamma^2.$$

Let $\nabla F(\mathbf{w}) = 0$. We obtain $\mathbf{w}_* = \sum_{k=1}^M \frac{1}{M}\mathbf{w}_*^k$. By Definition 1, the heterogeneity level is

$$\Delta = \sum_{k=1}^M \frac{F_k(\mathbf{w}_*) - F_k(\mathbf{w}_*^k)}{M} = \frac{1}{Md}\sum_{k=1}^M \left\|\mathbf{w}_*^k - \mathbf{w}_*\right\|_2^2 = \frac{1}{d} \cdot \underbrace{\frac{1}{M}\sum_{k=1}^M \left\|\mathbf{w}_*^k - \frac{1}{M}\sum_{i=1}^M \mathbf{w}_*^i\right\|_2^2}_{:=\Delta_1}, \quad (27)$$

in which Δ_1 is the heterogeneity level defined by (Chen et al., 2023).

We further analyze the problem-dependent constants μ in Assumption 2 and σ_k^2 in Definition 1.

$$\nabla f(\mathbf{w}, \mathbf{z}) = 2\left(\langle \mathbf{w}, \mathbf{x} \rangle - y\right) \cdot \mathbf{x}, \quad \nabla f\left(\mathbf{w}_*^k, \mathbf{z}\right) = 2\varepsilon \cdot \mathbf{x}, \quad \nabla^2 f(\mathbf{w}, \mathbf{z}) = 2\mathbf{x}\mathbf{x}^\top,$$

where $\varepsilon \sim \mathcal{N}(0, \gamma^2)$. Taking expectation w.r.t. $\mathbf{x}$ and ε yields

$$\mathbb{E}_{\mathbf{x}, \varepsilon}\left[\left\|\nabla f(\mathbf{w}_*^k, \mathbf{z})\right\|_2^2\right] = 4\gamma^2, \quad \text{and} \quad \nabla^2 F_k(\mathbf{w}) = 2 \cdot \mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\right] = \frac{2}{d} \cdot \mathbf{I}_{d \times d}.$$

Thereby, we have

$$\mu = \frac{2}{d}, \quad \sigma_k^2 = 4\gamma^2, \quad \text{and} \quad \sigma^2 = \frac{1}{M}\sum\nolimits_{k=1}^M \sigma_k^2 = 4\gamma^2. \quad (28)$$

Let $\hat{\mathbf{w}}_*^1, \ldots, \hat{\mathbf{w}}_*^M$ be the output of any algorithm. Let $1 \leq S \leq M$ be an integer. Suppose M is divisible by S. We divide the clients into S disjoint groups with equal size, denoted by $G_1, \ldots, G_S \subseteq [M]$. Let $\mathcal{V} \subseteq \{-1, 1\}^d$ be a packing of the d-dimensional hypercube $\{-1, 1\}^d$ such that $\forall \mathbf{v}_1 \neq \mathbf{v}_2 \in \mathcal{V}$, $\|\mathbf{v}_1 - \mathbf{v}_2\|_1 \geq \frac{d}{2}$. The Gilbert-Varshamov bound (refer to Lemma 8.2.3 (Duchi, 2023)) guarantees that $\mathcal{V}$ exists and $|\mathcal{V}| \geq \exp\left(\frac{d}{8}\right)$. We uniformly and independently sample $\mathbf{v}_1, \ldots, \mathbf{v}_S \in \mathcal{V}$. For each $s \in [S]$ and each client $k \in G_s$, let $\mathbf{w}_*^k = \delta_{k,s} \cdot \mathbf{v}_s$, in which $\delta_{k,s}$ is a constant. Then

$$\forall \mathbf{v}, \mathbf{v}' \in \mathcal{V}, \quad \left\|\delta_{k,s} \cdot \mathbf{v} - \delta_{k,s} \cdot \mathbf{v}'\right\|_2^2 = 2\delta_{k,s}^2 \cdot \left\|\mathbf{v} - \mathbf{v}'\right\|_1 \geq \delta_{k,s}^2 \cdot d. \quad (29)$$

Let $\mathbf{Y}_k = (y_1^k, \ldots, y_T^k)^\top \in \mathbb{R}^{T \times 1}$ be the observed output vector on the k-th client, and $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$. Let $\mathbf{X}_k = [\mathbf{x}_1^k, \ldots, \mathbf{x}_T^k]^\top \in \mathbb{R}^{T \times d}$ be the feature matrix, and $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$. Let $\mathcal{P}$ denote a class of distributions on the sample space $\mathcal{Z}$, $P \in \mathcal{P}$ be the distribution of the observations $\mathbf{X}$ and $\mathbf{Y}$. For any s and $k \in G_s$, we define a testing function $\Psi_{k,s}(\cdot) : \mathbf{Y} \to \mathcal{V}$ by

$$\forall s = 1, \ldots, S, \quad \Psi_{k,s}(\mathbf{w}(\mathbf{Y})) := \operatorname{argmin}_{\mathbf{v} \in \mathcal{V}} \left\|\mathbf{w}(\mathbf{Y}) - \delta_{k,s}\mathbf{v}\right\|_2^2.$$

Following Proposition 8.2.1 (Duchi, 2023), the minimax weighted excess risk is lower bounded by

$$
\begin{aligned}
&\inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \sup_{P \in \mathcal{P}} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\sum_{k=1}^M \frac{1}{M} \cdot \left(F_k(\hat{\mathbf{w}}_*^k) - F_k(\mathbf{w}_*^k) \right) \right] \\
\geq & \frac{1}{d} \cdot \inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \sup_{\{\mathbf{v}_s \in \mathcal{V}\}_{s=1}^S} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\sum_{k=1}^M \frac{1}{M} \cdot \left\| \hat{\mathbf{w}}_*^k - \mathbf{w}_*^k \right\|_2^2 \right] \\
\geq & \mathbb{E}_{\mathbf{X}} \left[\frac{1}{d} \cdot \inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \mathbb{E}_{\{\mathbf{v}_s \in \mathcal{V}\}_{s=1}^S, \mathbf{Y}|\mathbf{X}} \left[\frac{1}{M} \cdot \sum_{s=1}^S \sum_{k \in G_s} \left\| \hat{\mathbf{w}}_*^k - \mathbf{w}_*^k \right\|_2^2 \right] \right] \\
\geq & \mathbb{E}_{\mathbf{X}} \left[\frac{1}{d} \cdot \inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \frac{1}{M} \cdot \sum_{s=1}^S \sum_{k \in G_s} \frac{\delta_{k,s}^2 \cdot d}{4} \cdot \mathbb{P} \left[\left\| \hat{\mathbf{w}}_*^k - \mathbf{w}_*^k \right\|_2^2 \geq \frac{\delta_{k,s}^2 \cdot d}{4} \right] \right] \\
\geq & \mathbb{E}_{\mathbf{X}} \left[\frac{1}{d} \cdot \inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \frac{1}{M} \cdot \sum_{s=1}^S \sum_{k \in G_s} \frac{\delta_{k,s}^2 \cdot d}{4} \cdot \mathbb{P} \left[\Psi_{k,s}(\hat{\mathbf{w}}_*^k) \neq \mathbf{v}_s \right] \right] \\
= & \mathbb{E}_{\mathbf{X}} \left[\frac{1}{M} \sum_{s=1}^S \sum_{k \in G_s} \frac{1}{4} \delta_{k,s}^2 \cdot \inf_{\hat{\mathbf{w}}_*^k \in \mathcal{W}} \mathbb{P} \left[\Psi_{k,s}(\hat{\mathbf{w}}_*^k) \neq \mathbf{v}_s \right] \right].
\end{aligned}
$$

Next we prove the second last inequality. Recalling that $\mathbf{w}_*^k = \delta_{k,s} \cdot \mathbf{v}_s$. Assume that $\|\hat{\mathbf{w}}_*^k - \mathbf{w}_*^k\|_2 < \frac{1}{2}\sqrt{\delta_{k,s}^2 d}$. Then for any $\mathbf{v}' \neq \mathbf{v}_s$, the triangle inequality gives

$$
\left\| \delta_{k,s}\mathbf{v}' - \hat{\mathbf{w}}_*^k \right\|_2 \geq \left\| \delta_{k,s}\mathbf{v}' - \delta_{k,s}\mathbf{v}_s \right\|_2 - \left\| \delta_{k,s}\mathbf{v}_s - \hat{\mathbf{w}}_*^k \right\|_2 \overset{(29)}{\geq} \frac{1}{2}\sqrt{\delta_{k,s}^2 d}.
$$

In this case, it must be $\Psi_{k,s}(\hat{\mathbf{w}}_*^k) = \mathbf{v}_s$. Equivalently, $\Psi_{k,s}(\hat{\mathbf{w}}_*^k) \neq \mathbf{v}_s$ implies $\|\hat{\mathbf{w}}_*^k - \mathbf{w}_*^k\|_2 \geq \frac{1}{2}\sqrt{\delta_{k,s}^2 d}$, by which we obtain the second last inequality.

For two random variables A, B, denote by $I(A, B)$ their mutual information. For each $s = 1, \ldots, S$, let $\mathbf{Y}^s = \{\mathbf{Y}_r, r \in G_s\}$. Given $\mathbf{X}$, for any $k \in G_s$, $\Psi_{k,s}(\hat{\mathbf{w}}_*^k)$ depends on $\mathbf{Y}$ and $\mathbf{Y}$ depends on $\mathbf{v}_s$. Thus $\mathbf{v}_s \to \mathbf{Y} \to \Psi_{k,s}(\hat{\mathbf{w}}_*^k)$ is a Markov chain. Since $\mathbf{v}_s$ is uniformly sampled from $\mathcal{V}$, and $\mathbf{Y}_k$ is independent of $\mathbf{v}_s$ for all $k \notin G_s$, then by Corollary 8.4.2 in Duchi (2023), we can obtain

$$
\begin{aligned}
\inf_{\hat{\mathbf{w}}_*^k \in \mathcal{W}} \mathbb{P} \left[\Psi_{k,s}(\hat{\mathbf{w}}_*^k) \neq \mathbf{v}_s \right] &\geq 1 - \frac{I(\mathbf{v}_s, \mathbf{Y}) + \ln 2}{\ln |\mathcal{V}|} \\
&= 1 - \frac{\sum_{k=1}^M I(\mathbf{v}_s, \mathbf{Y}_k | \mathbf{Y}_1, \ldots, \mathbf{Y}_{k-1}) + \ln 2}{\ln |\mathcal{V}|} \\
&= 1 - \frac{\sum_{k \in G_s} I(\mathbf{v}_s, \mathbf{Y}_k | \mathbf{Y}_1, \ldots, \mathbf{Y}_{k-1}) + \sum_{k \notin G_s} I(\mathbf{v}_s, \mathbf{Y}_k | \mathbf{Y}_1, \ldots, \mathbf{Y}_{k-1}) + \ln 2}{\ln |\mathcal{V}|} \\
&= 1 - \frac{\sum_{k \in G_s} I(\mathbf{v}_s, \mathbf{Y}_k | \mathbf{Y}_r, r < k, r \in G_s) + \ln 2}{\ln |\mathcal{V}|} \\
&\geq 1 - \frac{I(\mathbf{v}_s, \mathbf{Y}^s) + \ln 2}{d/8}.
\end{aligned}
$$

The key is to analyze $I(\mathbf{v}_s, \mathbf{Y}^s)$. Let $P_{\mathbf{Y}^s, \mathbf{v}_s}$ be the distribution of $\mathbf{Y}^s$ given $\mathbf{v}_s$, which is the product distribution of $\frac{M}{S}$ independent Gaussian random vectors, i.e., $\{\mathcal{N}\left(\mathbf{X}_k \mathbf{w}_*^k, \gamma^2 \mathbf{I}_{T \times T}\right)\}_{k \in G_s}$. The probability density function of $\mathbf{Y}^s$, denoted by $h_{\mathbf{Y}^s, \mathbf{v}_s}$, is

$$
h_{\mathbf{Y}^s, \mathbf{v}_s} = \prod_{k \in G_s} \frac{1}{(2\pi)^{\frac{T}{2}} \gamma^T} \exp\left(-\frac{1}{2\gamma^2} \left\| \mathbf{Y}_k - \mathbf{X}_k \mathbf{w}_*^k \right\|_2^2 \right) := \prod_{k \in G_s} h_{\mathbf{Y}_k, \mathbf{v}_s}.
$$

Let $\mathrm{D}_{\mathrm{KL}}(P, Q)$ be the Kullback-Leibler (KL) divergence between distribution P and distribution Q. By Eq (8.4.5) in Duchi (2023), and taking expectation w.r.t. $\mathbf{X}_k, k \in G_s$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{X}_k, k\in G_s}[I(\mathbf{v}_s, \mathbf{Y}_s)] = &\mathbb{E}_{\mathbf{X}_k, k\in G_s}\left[\frac{1}{|\mathcal{V}|}\sum_{\mathbf{v}} \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Y}^s,\mathbf{v}}, \frac{1}{|\mathcal{V}|}\sum_{\mathbf{v}'} P_{\mathbf{Y}^s,\mathbf{v}'}\right)\right] \\
\leq &\mathbb{E}_{\mathbf{X}_k, k\in G_s}\left[\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{v},\mathbf{v}'} \mathrm{D}_{\mathrm{KL}}\left(P_{\mathbf{Y}^s,\mathbf{v}}, P_{\mathbf{Y}^s,\mathbf{v}'}\right)\right] \\
= &\mathbb{E}_{\mathbf{X}_k, k\in G_s}\left[\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{v},\mathbf{v}'} \int_{\mathbf{Y}_k, k\in G_s} \prod_{k\in G_s} h_{\mathbf{Y}_k,\mathbf{v}} \ln \frac{\prod_{k\in G_s} h_{\mathbf{Y}_k,\mathbf{v}}}{\prod_{k\in G_s} h_{\mathbf{Y}_k,\mathbf{v}'}} \mathrm{d} \prod_{k\in G_s} \mathbf{Y}_k\right] \\
= &\mathbb{E}_{\mathbf{X}_k, k\in G_s}\left[\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{v},\mathbf{v}'}\sum_{i\in G_s} \left(\int_{\mathbf{Y}_k, k\in G_s} \prod_{k\in G_s} h_{\mathbf{Y}_k,\mathbf{v}} \ln \frac{h_{\mathbf{Y}_i,\mathbf{v}}}{h_{\mathbf{Y}_i,\mathbf{v}}} \mathrm{d} \prod_{k\in G_s} \mathbf{Y}_k\right)\right] \\
= &\mathbb{E}_{\mathbf{X}_k, k\in G_s}\left[\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{v},\mathbf{v}'}\sum_{i\in G_s} \left(\int_{\mathbf{Y}_i} h_{\mathbf{Y}_i,\mathbf{v}} \ln \frac{h_{\mathbf{Y}_i,\mathbf{v}}}{h_{\mathbf{Y}_i,\mathbf{v}'}} \mathrm{d}\, \mathbf{Y}_i\right)\right] \\
= &\mathbb{E}_{\mathbf{X}_k, k\in G_s}\left[\sum_{i\in G_s}\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{v},\mathbf{v}'} \left[\frac{1}{2\gamma^2}\left(\delta_{i,s}\mathbf{X}_i\mathbf{v} - \delta_{i,s}\mathbf{X}_i\mathbf{v}'\right)^\top \left(\delta_{i,s}\mathbf{X}_i\mathbf{v} - \delta_{i,s}\mathbf{X}_i\mathbf{v}'\right)\right]\right] \\
= &\sum_{i\in G_s}\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{v},\mathbf{v}'}\frac{1}{2\gamma^2}\delta_{i,s}^2 \cdot \frac{T}{d} \cdot \|\mathbf{v} - \mathbf{v}'\|_2^2 \\
\leq &\frac{2T}{\gamma^2}\sum_{i\in G_s}\delta_{i,s}^2.
\end{aligned}
$$

Now a lower bound on the minimax weighted excess risk is given as follows,

$$
\begin{aligned}
&\inf_{\{\hat{\mathbf{w}}_*^k\in\mathcal{W}\}_{k=1}^M} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}}\left[\sum_{k=1}^M \frac{1}{M}\cdot\left(F_k(\hat{\mathbf{w}}_*^k) - F_k(\mathbf{w}_*^k)\right)\right] \\
\geq &\frac{1}{M}\sum_{s=1}^S\sum_{k\in G_s}\frac{\delta_{k,s}^2}{4}\cdot\max\left\{1 - \frac{\frac{2T}{\gamma^2}\cdot\sum_{k\in G_s}\delta_{k,s}^2 + \ln 2}{d/8}, 0\right\},
\end{aligned}
\tag{30}
$$

where the $\max\{\cdot,\cdot\}$ function comes from the natural constraint on the probability $\inf_{\hat{\mathbf{w}}_*^k\in\mathcal{W}}\mathbb{P}[\cdot] \geq 0$. This constraint is critical to the proof of lower bound when $\Delta = \Omega(\frac{\sigma^2}{\mu T})$. By (27), we rewrite Δ as

$$
\Delta = \frac{1}{Md}\sum_{k=1}^M \left\|\mathbf{w}_*^k\right\|_2^2 - \frac{1}{d}\left\|\mathbf{w}_*\right\|_2^2 = \frac{1}{M}\sum_{s=1}^S\sum_{k\in G_s}\delta_{k,s}^2 - \frac{1}{d}\left\|\frac{1}{M}\sum_{s=1}^S\sum_{k\in G_s}\delta_{k,s}\mathbf{v}_s\right\|_2^2.
$$

Next we achieve three levels of heterogeneity by adjusting the value of S. Suppose $d \geq 32$ and $3 \leq M \leq \exp(d/8)$ is odd.

- **Case 1**: $\Delta = \omega(\frac{d\gamma^2}{MT})$ and $\Delta = O(\frac{d\gamma^2}{T})$.
  We can attain these Δ by considering specific problem instances with $1 < S < M$, $\mathbf{v}_1 \neq \mathbf{v}_2 \neq \ldots \neq \mathbf{v}_S$ and

  $$
  \forall s = 1, \ldots, S, \quad \forall k \in G_s, \quad \delta_{k,s} = \delta = \sqrt{\frac{Sd\gamma^2}{32MT}}.
  $$

  Then the heterogeneity level is

  $$
  \Delta = \delta^2 - \frac{\delta^2}{d}\left\|\frac{1}{S}\sum_{s=1}^S \mathbf{v}_s\right\|_2^2.
  $$

By Lemma 11, it must be

$$\frac{1}{4}\delta^2 \leq \Delta \leq \delta^2 \Rightarrow \frac{1}{4}\frac{Sd\gamma^2}{32MT} \leq \Delta \leq \frac{Sd\gamma^2}{32MT}.$$

The minimax weighted excess risk is lower bounded as follows

$$\begin{aligned}
&\inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \left[\sum_{k=1}^M \frac{1}{M}\cdot\left(F_k(\hat{\mathbf{w}}_*^k) - F_k(\mathbf{w}_*^k)\right)\right] \\
&\overset{(30)}{\geq} \frac{1}{M}\sum_{s=1}^S \sum_{k\in G_s} \frac{\delta_{k,s}^2}{4}\cdot\left(1 - \frac{\frac{2T}{\gamma^2}\cdot\sum_{k\in G_s}\delta_{k,s}^2 + \ln 2}{d/8}\right) \geq \frac{Sd\gamma^2}{400MT} \geq \frac{2}{25}\Delta.
\end{aligned}$$

This result shows that if $\Delta = \omega(\frac{d\gamma^2}{MT})$ and $\Delta = O(\frac{d\gamma^2}{T})$, there is a $\Omega(\Delta)$ lower bound.

- **Case 2**: $\Delta \geq 0$ and $\Delta = O(\frac{d\gamma^2}{MT})$.
We can attain these Δ by considering specific problem instances with $S = 1$ and $\mathbf{v}_1 = \mathbf{v}$. We rewrite $\delta_{k,s}$ as δ_k. Let $\delta = \frac{\sqrt{d}\gamma}{\sqrt{32MT}}$ and $\varepsilon = c_1\delta$ where $0 \leq c_1 \leq \frac{1}{6}$ is a tunable parameter. Let δ_k satisfy

$$\forall k = 1,\ldots,\frac{M+1}{2}, \quad \delta_k = \delta + \varepsilon; \quad \text{and} \quad \delta_k = \delta - \varepsilon, \quad \forall k \geq \frac{M+1}{2}+1,\ldots,M.$$

Then we have

$$\left|\frac{1}{M}\cdot\sum_{k=1}^M \delta_k\right|^2 = \left(\delta + \frac{\varepsilon}{M}\right)^2 \geq \delta^2, \quad \text{and} \quad \sum_{k=1}^M \delta_k^2 = \left(1 + c_1^2 + \frac{2c_1}{M}\right)M\delta^2.$$

The heterogeneity level can be upper bounded as follows,

$$0 \leq \Delta = \sum_{k=1}^M \frac{1}{M}\delta_k^2 - \left|\frac{1}{M}\cdot\sum_{i=1}^M \delta_i\right|_2^2 = \varepsilon^2 - \frac{\varepsilon^2}{M^2} < \frac{c_1^2 d\gamma^2}{32MT} \leq \frac{d\gamma^2}{1152MT}.$$

Note that we can tune c_1 such that Δ approaches 0. The minimax weighted excess risk is lower bounded as follows

$$\begin{aligned}
&\inf_{\{\hat{\mathbf{w}}_*^k \in \mathcal{W}\}_{k=1}^M} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \left[\sum_{k=1}^M \frac{1}{M}\cdot\left(F_k(\hat{\mathbf{w}}_*^k) - F_k(\mathbf{w}_*^k)\right)\right] \\
&\overset{(30)}{\geq} \left(1 + c_1^2 + \frac{2c_1}{M}\right)\frac{\delta^2}{4}\left(1 - \frac{\frac{2T}{\gamma^2}\cdot(1 + c_1^2 + \frac{2c_1}{M})M\delta^2 + \ln 2}{d/8}\right) \geq \frac{d\gamma^2}{440MT}.
\end{aligned}$$

This result shows that if $\Delta \geq 0$ and $\Delta = O(\frac{d\gamma^2}{MT})$, there is a $\Omega(\frac{d\gamma^2}{MT})$ lower bound.

- **Case 3**: $\Delta = \omega(\frac{d\gamma^2}{T})$.
We can attain these Δ by considering specific problem instances with $S = M$ and $\mathbf{v}_1 \neq \mathbf{v}_2 \neq \ldots \neq \mathbf{v}_M$. We define $\frac{M-1}{2}$ clients that are far from the other clients, such that the heterogeneity level is large enough. We rewrite $\delta_{k,s}$ as δ_k for all $s = 1,\ldots,S$. Let

$$\begin{aligned}
\forall k = 1,2,\ldots,\frac{M-1}{2}, \quad \delta_k &= \sqrt{\frac{d\gamma^2}{32T^\nu}}, \quad \nu \in [0,1), \\
\forall k = \frac{M+1}{2},\ldots,M, \quad \delta_k &= \sqrt{\frac{d\gamma^2}{32T}}.
\end{aligned}$$

By Lemma 10, we have

$$
\begin{aligned}
&\Delta \\
=&\frac{1}{M}\left(\frac{M-1}{2}\cdot\delta_1^2+\frac{M+1}{2}\cdot\delta_M^2\right)-\frac{1}{dM^2}\left\|\sum_{s=1}^{\frac{M-1}{2}}\delta_1\mathbf{v}_s+\sum_{s=\frac{M+1}{2}}^{M}\delta_M\mathbf{v}_s\right\|_2^2 \\
\overset{(26)}{\geq}&\frac{M-1}{2M}\cdot\delta_1^2+\frac{M+1}{2M}\cdot\delta_M^2-\frac{1}{2M^2}\left(\frac{M-1}{2}\delta_1^2+\frac{M+1}{2}\delta_M^2\right)- \\
&\frac{1}{2M^2}\left[\frac{M-1}{2}\cdot\left(\frac{M-3}{2}\delta_1^2+\frac{M+1}{2}\delta_1\delta_M\right)+\frac{M+1}{2}\left(\frac{M-1}{2}\delta_1\delta_M+\frac{M-1}{2}\delta_M^2\right)\right] \\
\geq&\frac{3M^2-2M-1}{8M^2}\delta_1^2+\frac{3M^2+2M-1}{8M^2}\delta_M^2-\frac{M^2-1}{4M^2}\delta_1\delta_M \\
\geq&\frac{3M^2-2M-1}{8M^2}\delta_1^2+\frac{3M^2+2M-1}{8M^2}\delta_M^2-\frac{M^2-1}{4M^2}\frac{\delta_1^2+\delta_M^2}{2} \\
>&\frac{2M^2-2M}{8M^2}\delta_1^2 \\
\geq&\frac{d\gamma^2}{256T^\nu}.
\end{aligned}
$$

We can also tune ν such that

$$
\forall k=1,2,\ldots,\frac{M-1}{2},\quad \frac{\delta_k^2}{4}\cdot\left(1-\frac{\frac{2T}{\gamma^2}\cdot\delta_k^2+\ln 2}{d/8}\right)<0.
$$

The minimax weighted excess risk is lower bounded as follows

$$
\begin{aligned}
&\inf_{\{\hat{\mathbf{w}}_*^k\in\mathcal{W}\}_{k=1}^M}\sup_{P\in\mathcal{P}}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}}\left[\sum_{k=1}^M\frac{1}{M}\cdot\left(F_k(\hat{\mathbf{w}}_*^k)-F_k(\mathbf{w}_*^k)\right)\right] \\
\overset{(30)}{\geq}&\frac{1}{M}\sum_{k=\frac{M+1}{2}}^M\frac{\delta_k^2}{4}\cdot\left(1-\frac{\frac{2T}{\gamma^2}\cdot\delta_k^2+\ln 2}{d/8}\right) \\
>&\frac{1}{2}\cdot\frac{d\gamma^2}{128T}\left(1-\frac{\frac{2T}{\gamma^2}\cdot\frac{d\gamma^2}{32T}+\ln 2}{d/8}\right)>\frac{d\gamma^2}{800T}.
\end{aligned}
$$

This result shows that if $\Delta=\omega(\frac{d\gamma^2}{T})$, there is a $\Omega(\frac{d\gamma^2}{T})$ lower bound.

Combining the above three lower bounds, we have

$$
\begin{aligned}
\inf_{\{\hat{\mathbf{w}}_*^k\}_{k=1}^M}\sup_{\{\mathbf{w}_*^k\in\mathcal{W}\}_{k=1}^M}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}}\left[\sum_{k=1}^M\frac{1}{M}\cdot\left(F_k(\hat{\mathbf{w}}_*^k)-F_k(\mathbf{w}_*^k)\right)\right]=&\Omega\left(\min\left\{\frac{d\gamma^2}{T},\Delta\right\}+\frac{d\gamma^2}{MT}\right) \\
\overset{(28)}{=}&\Omega\left(\min\left\{\frac{\sigma^2}{\mu T},\Delta\right\}+\frac{\sigma^2}{\mu MT}\right).
\end{aligned}
$$

We recovery the first lower bound in Theorem 1.

Our proof is fundamentally different from the proof of Theorem 8 in (Chen et al., 2023): (i) We uniformly sample $1\leq S\leq M$ elements $\mathbf{v}_1,\ldots,\mathbf{v}_S\in\mathcal{V}$, while their analysis uniformly samples $\mathbf{v}_1,\ldots,\mathbf{v}_M\in\{-1,1\}^d$. As the discussions in Section 1.2, it is necessary to set $S=1$ for proving a $\Omega(\frac{\sigma^2}{\mu MT})$ lower bound in the case $0<\Delta=O(\frac{\sigma^2}{\mu MT})$. (ii) Our analysis is applicable for unconstrained $\mathcal{W}$, while their analysis is applicable for constrained $\mathcal{W}$. (iii) We use the square loss function, while their analysis uses the logistic loss function, which must restrict the size of $\mathcal{W}$ for ensuring μ-strongly convex property. Our lower bounds show sharper dependence on the problem-dependent constant μ.

C.2 EXACT GRADIENTS

Let $M \geq 2$. Suppose Assumption 3 (i) with $\sigma_k^2 = 0$ holds, that is, the exact gradients $\{\nabla F_k(\mathbf{w})\}_{k=1}^M$ are observed. For any algorithm with outputs $\hat{\mathbf{w}}_*^1, \ldots, \hat{\mathbf{w}}_*^M$, its weighted excess risk satisfies

$$\frac{1}{M}\sum_{k=1}^{M}\left[F_k\left(\hat{\mathbf{w}}_*^k\right) - F_k\left(\mathbf{w}_*^k\right)\right] \overset{(23)}{\geq} \frac{\mu}{2}\frac{1}{M}\sum_{k=1}^{M}\underbrace{\left\|\hat{\mathbf{w}}_*^k - \mathbf{w}_*^k\right\|_2^2}_{\text{estimation error}}.$$

Thus we just need to analyze the estimation error.

Let $F(\mathbf{w}) = \frac{1}{M}\sum_{k=1}^M F_k(\mathbf{w})$ be the global loss function. We must carefully design $F(\mathbf{w})$ and $F_k(\mathbf{w})$. Following the proof of Theorem 2.1.13 (Nesterov, 2004), $F(\mathbf{w})$ is defined by

$$\begin{aligned}\forall \mathbf{w} \in \mathcal{W}, \quad F(\mathbf{w}) =& \frac{\mu(\kappa-1)}{8}\left(\|\mathbf{w}\|_{\mathbf{A}}^2 - 2\langle \mathbf{e}_1, \mathbf{w}\rangle\right) + \frac{\mu}{2}\|\mathbf{w}\|_2^2, \\ \mathcal{W} =& \left\{\mathbf{w} = (w_1, \ldots, w_{+\infty}) \in \mathbb{R}^{+\infty} : \sum_{i=1}^{+\infty} w_i^2 < +\infty\right\},\end{aligned} \tag{31}$$

where $\kappa > 1$, $\mathbf{e}_1 = (1, 0, \ldots, 0)$ is the canonical basis of $\mathcal{W}$ and $\mathbf{A} : \mathcal{W} \to \mathcal{W}$ is a linear operator,

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & \ddots \\ 0 & 0 & 0 & \ddots & \ddots \end{pmatrix}.$$

We define individualized loss function $F_k(\cdot)$ as follows,

$$\begin{aligned}\forall k \in [M], \quad F_k(\mathbf{w}) =& \frac{\mu(\kappa-1)}{8}\left(\|\mathbf{w}\|_{\mathbf{A}}^2 - a_k\langle \mathbf{e}_1, \mathbf{w}\rangle\right) + \frac{\mu}{2}\|\mathbf{w}\|_2^2, \\ a_k =& 2 - \alpha\left(k - \frac{M+1}{2}\right),\end{aligned}$$

where $\alpha \geq 0$ controls the heterogeneity level Δ. If $\alpha = 0$, then $\Delta = 0$. Otherwise, $\Delta > 0$. It is easy to verify that $F(\mathbf{w}) = \frac{1}{M}\sum_{k=1}^M F_k(\mathbf{w})$.

Let $\mathbf{I} : \mathcal{W} \to \mathcal{W}$ be the unit operator. For each k, the Hessian matrixces of $F_k(\mathbf{w})$ and $F(\mathbf{w})$ are

$$\forall k = 1, \ldots, M, \quad \nabla^2 F_k(\mathbf{w}) = \nabla^2 F(\mathbf{w}) = \frac{\mu(\kappa-1)}{4}\mathbf{A} + \mu\mathbf{I}.$$

Since $\mathbf{0} \preccurlyeq \mathbf{A} \preccurlyeq 4\mathbf{I}$ (Nesterov, 2004), it must be

$$\forall \mathbf{w} \in \mathcal{W}, \quad \mu \cdot \mathbf{I} \preccurlyeq \frac{\mu(\kappa-1)}{4}\mathbf{A} + \mu\mathbf{I} \preccurlyeq \beta \cdot \mathbf{I}. \tag{32}$$

Thus $\{F_k(\cdot)\}_{k=1}^M$ and $F(\mathbf{w})$ satisfy Assumption 1-2. By (32), both $\mathbf{w}_*$ and $\{\mathbf{w}_*^k\}_{k=1}^M$ exist. Next we analyze these optimal parameters.

Let $\nabla F(\mathbf{w}) = 0$, i.e.,

$$\left(\frac{\mu(\kappa-1)}{4}\mathbf{A} + \mu\mathbf{I}\right)\mathbf{w} = \frac{\mu(\kappa-1)}{4}\mathbf{e}_1. \tag{33}$$

By Eq. (2.1.25) (Nesterov, 2004), solving this equation, each coordinate of $\mathbf{w}_*$ satisfies

$$i = 1, \ldots, +\infty, \quad w_{*,i} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i := q^i.$$

Similarly, the gradient of local loss function is given as follows,

$$\forall k = 1, \ldots, M, \quad \nabla F_k(\mathbf{w}) = \frac{\mu(\kappa-1)}{4}\left(\mathbf{A}\mathbf{w} - \frac{a_k}{2}\mathbf{e}_1\right) + \mu \cdot \mathbf{w}. \tag{34}$$

Let $\nabla F_k(\mathbf{w}) = 0$. Combined with (33), it is easy to prove that

$$\forall k = 1, \ldots, M, \quad \mathbf{w}_*^k = \frac{a_k}{2}\mathbf{w}_*.$$

Let $\mathcal{A}$ be a first-order algorithm. Given an initial global parameter $\mathbf{w}_1 \in \mathcal{W}$, after the T-iteration, $\mathcal{A}$ generates a sequence of points $\mathbf{w}_1^k, \ldots, \mathbf{w}_T^k$ on each client k. Note that we allow $\mathbf{w}_\tau^1 = \ldots = \mathbf{w}_\tau^M$ for all $\tau = 1, 2, \ldots, T$. For any first-order algorithm, it must be

$$\forall \tau \in [T], \quad \forall k \in [M], \quad \mathbf{w}_\tau^k = \mathbf{w}_1 + \operatorname{span}\left(\{\nabla F_i(\mathbf{w}_1)\}_{i=1}^M, \ldots, \{\nabla F_i(\mathbf{w}_{\tau-1}^i)\}_{i=1}^M\right), \tag{35}$$

in which $\operatorname{span}(\cdot)$ is the linear span operator.

Now we analyze the structure of $\mathbf{w}_1^k, \ldots, \mathbf{w}_T^k$. Without loss of generality, let $\mathbf{w}_1 = \mathbf{0}$. Note that if $\mathbf{w}_1 \neq \mathbf{0}$, then we can redefine a loss function $\bar{F}_k(\mathbf{w}) = F_k(\mathbf{w} - \mathbf{w}_1)$. By (34), for $\tau = 1$, $\mathcal{A}$ returns

$$\forall k = 1, \ldots, M, \quad \nabla F_k(\mathbf{w}_1) = -\frac{\mu(\kappa - 1)}{8} a_k \mathbf{e}_1.$$

Thus $\mathbf{w}_1^k \in \operatorname{span}(\mathbf{e}_1)$ for all $k = 1, \ldots, M$. By (35) and (34), we conclude by induction that

$$\forall \tau \leq T, \quad \forall k = 1, \ldots, M, \quad \mathbf{w}_\tau^k \in \operatorname{span}(\mathbf{e}_1, \ldots, \mathbf{e}_\tau).$$

Suppose $\mathcal{A}$ outputs $\hat{\mathbf{w}}_*^k$ for each client $k = 1, \ldots, M$. Note that $\hat{\mathbf{w}}_*^k$ can be a function of $\{\mathbf{w}_t^k\}_{t=1}^T$. Then a lower bound on the estimation error is given as follows

$$\forall k \in [M], \quad \left\|\hat{\mathbf{w}}_*^k - \mathbf{w}_*^k\right\|_2^2 \geq \sum_{i=T+1}^{+\infty} \left(w_{*,i}^k\right)^2 = \frac{a_k^2}{4} \cdot \sum_{i=T+1}^{+\infty} \left(w_{*,i}\right)^2 = \frac{a_k^2}{4} \cdot \frac{q^{2(T+1)}}{1 - q^2},$$

in which $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

Due to $\mathbf{w}_1 = \mathbf{0}$, it must be

$$\forall k \in [M], \quad \left\|\mathbf{w}_1 - \mathbf{w}_*^k\right\|_2^2 = \sum_{i=1}^{+\infty} \left(w_{*,i}^k\right)^2 = \frac{a_k^2}{4} \sum_{i=1}^{+\infty} (w_{*,i})^2 = \frac{a_k^2}{4} \cdot \frac{q^2(1 - q^{+\infty})}{1 - q^2} = \frac{a_k^2}{4} \cdot \frac{q^2}{1 - q^2}.$$

It is easy to prove that $1 - a \geq \exp(-\frac{a}{1-a})$ for all $a \in (0, 1)$. Recalling that $\kappa > 1$. The weighted excess risk is lower bounded by

$$\begin{aligned}
\frac{1}{M}\sum_{k=1}^M \left[F_k\left(\hat{\mathbf{w}}_*^k\right) - F_k\left(\mathbf{w}_*^k\right)\right] \geq & \frac{\mu}{2} \cdot \frac{1}{M} \sum_{k=1}^M \frac{a_k^2}{4} \cdot \frac{q^{2(T+1)}}{1 - q^2} \\
= & \frac{\mu}{2} \cdot q^{2T} \cdot \frac{1}{M} \sum_{k=1}^M \left\|\mathbf{w}_1 - \mathbf{w}_*^k\right\|_2^2 \\
\geq & \frac{\mu}{2} \exp\left(-\frac{4T}{\sqrt{\kappa} - 1}\right) \cdot \frac{1}{M} \sum_{k=1}^M \left\|\mathbf{w}_1 - \mathbf{w}_*^k\right\|_2^2,
\end{aligned}$$

which recovers the second lower bound in Theorem 1.

Combining the lower bounds in the case of stochastic gradients and exact gradients, we conclude the proof. □

D Proof of Theorem 2

Lemma 12. *Suppose $T_1 \geq \lceil 38\kappa \rceil$. There is a $c_0 \in [\frac{7}{5}, 4]$ and an integer $K \geq 2$ such that*

$$T_{1,0} + \sum_{k=1}^K T_{1,k} = T_1,$$

where $T_{1,0} = \frac{1}{c_0} T_1$, $T_{1,1} = \left\lceil 2^{2.5}\kappa \right\rceil$, $T_{1,k} = 2^{k-1} T_{1,1}$, $k = 2, \ldots, K$.

Proof of Lemma 12. Without loss of generality, suppose $T_{1,0}$ is an integer. To be specific, we can always make $T_{1,0}$ an integer by choosing a suitable c_0 or by subtracting a small constant from T_1. Substituting into the values of $T_{1,0}$ and $T_{1,k}$, we obtain

$$T_{1,0}+\sum_{k=1}^{K}T_{1,k}=T_1 \Leftrightarrow 2^K=\frac{c_0-1}{c_0}\frac{T_1}{T_{1,1}}+1.$$

We rewrite $T_1=\lceil\nu\kappa\rceil$ where $\nu\geq 38$. Let $h(c)=\frac{c-1}{c}\cdot\frac{T_1}{T_{1,1}}+1$. The minimal value and maximal value of $h(c)$ is given by

$$\min_c h(c)=h\left(\frac{7}{5}\right)=\frac{2}{7}\cdot\frac{\lceil\nu\kappa\rceil}{\lceil 2^{2.5}\kappa\rceil}+1>2,\quad \max_c h(c)=h(4)=\frac{3}{4}\cdot\frac{\lceil\nu\kappa\rceil}{\lceil 2^{2.5}\kappa\rceil}+1.$$

We need to prove that there is an integer $K\geq 2$ such that $\min_c h(c)\leq 2^K\leq\max_c h(c)$, which is equivalent to

$$\log_2 h\left(\frac{7}{5}\right)\leq K\leq\log_2 h\left(4\right).$$

If $\log_2 h\left(\frac{7}{5}\right)$ is an integer, then let $K=\log_2 h\left(\frac{7}{5}\right)\geq 2$ and $c_0=\frac{7}{5}$. This lemma holds. Otherwise, we need to prove $\log_2 h\left(4\right)>\log_2 h\left(\frac{7}{5}\right)+1$. This is obvious true, since

$$\frac{h\left(4\right)}{h\left(\frac{7}{5}\right)}=\frac{\frac{3}{4}\cdot\frac{\lceil\nu\kappa\rceil}{\lceil 2^{2.5}\kappa\rceil}+1}{\frac{2}{7}\cdot\frac{\lceil\nu\kappa\rceil}{\lceil 2^{2.5}\kappa\rceil}+1}=\frac{21}{8}-\frac{13}{8}\cdot\frac{1}{\frac{2}{7}\cdot\frac{\lceil\nu\kappa\rceil}{\lceil 2^{2.5}\kappa\rceil}+1}>2,$$

in which $\frac{\lceil\nu\kappa\rceil}{\lceil 2^{2.5}\kappa\rceil}\geq\frac{\nu\kappa}{2^{2.5}\kappa+1}>\frac{38}{2^{2.5}+1}$. Combining the two cases, we recovery the desired result. □

Next we establish a progress lemma which gives the expected excess risk of M-SG in one stage.

Lemma 13 (Progress Lemma)**.** *Let $c_0\in[\frac{7}{5},4]$ satisfy Lemma 12.*

$$\begin{cases}T_{1,0}=\dfrac{1}{c_0}T_1,\quad T_{1,1}=\left\lceil\dfrac{1}{\eta_1\mu}2^{1.5}\right\rceil,\quad T_{1,k}=2^{k-1}T_{1,1},\quad k\geq 2,\\ \eta_0=\eta_1=\dfrac{1}{2\beta},\quad \eta_k=\dfrac{\eta_1}{2^{k-1}},\quad k\geq 2.\end{cases}\tag{36}$$

At any stage $k=0,1,\ldots,K$, the expected excess risk of M-SG satisfies: for each $i=1,\ldots,M$,

$$\mathbb{E}_k^i\left[F_i\left(\bar{\mathbf{w}}_k^i\right)\right]-F_i\left(\mathbf{w}_*^i\right)\leq\frac{\mu}{2}\exp\left(-\frac{\mu}{2}T_{1,k}\eta_k\right)\frac{\left\|\mathbf{w}_{k,s_k}^i-\mathbf{w}_*^i\right\|_2^2}{1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right)}+\frac{2\eta_k\sigma_i^2}{1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right)},$$

where $\mathbb{E}_k^i[\cdot]$ is the expectation taken over $\mathbf{z}_{k,s_k}^i,\ldots,\mathbf{z}_{k,e_k}^i$.

By Lemma 13, the expected excess risk bound of $\bar{\mathbf{w}}_k^i$ consists of two components. The first one decays exponentially at a rate $\frac{\mu}{2}\exp\left(-\frac{\mu}{2}T_{1,k}\eta_k\right)$ and the second one depends on the gradient variance σ_i^2. By this progress lemma, we can easily derive the desired excess risk bound by carefully designing the learning rate η_k and the length of each stage $T_{1,k}$.

Proof of Lemma 13. We just consider a fix client $i\in\{1,\ldots,M\}$ and a fix stage $k\in\{0,1,\ldots,K\}$. For simplicity, the superscript i and subscript k will be omitted in $\mathbf{w}_{k,t}^i$ and $\mathbf{z}_{k,t}^i$. Specifically, $\mathbf{w}_{k,t}^i$

and $\mathbf{z}_{k,t}^i$ are rewritten as $\mathbf{w}_t$ and $\mathbf{z}_t$, respectively.

$$
\begin{aligned}
&\left\|\mathbf{w}_{t+1}-\mathbf{w}_*^i\right\|_2^2-\left\|\mathbf{w}_t-\mathbf{w}_*^i\right\|_2^2 \\
=&\left\|\mathbf{w}_t-\eta_k\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)-\mathbf{w}_*^i\right\|_2^2-\left\|\mathbf{w}_t-\mathbf{w}_*^i\right\|_2^2 \\
=&-2\eta_k\left\langle\mathbf{w}_t-\mathbf{w}_*^i,\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)\right\rangle+\eta_k^2\left\|\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)\right\|_2^2 \\
=&-2\eta_k\left[\left\langle\mathbf{w}_t-\mathbf{w}_*^i,\nabla F_i\left(\mathbf{w}_t\right)\right\rangle+\left\langle\mathbf{w}_t-\mathbf{w}_*^i,\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)-\nabla F_i\left(\mathbf{w}_t\right)\right\rangle\right]+\eta_k^2\left\|\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)\right\|_2^2 \\
\overset{(22)}{\leq}&-2\eta_k\left[\left\langle\mathbf{w}_t-\mathbf{w}_*^i,\nabla F_i\left(\mathbf{w}_t\right)\right\rangle+\left\langle\mathbf{w}_t-\mathbf{w}_*^i,\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)-\nabla F_i\left(\mathbf{w}_t\right)\right\rangle\right]+2\eta_k^2\left\|\nabla f\left(\mathbf{w}_*^i,\mathbf{z}_t\right)\right\|_2^2+ \\
&2\beta\eta_k^2\left[\left\langle\nabla F_i(\mathbf{w}_t),\mathbf{w}_t-\mathbf{w}_*^i\right\rangle+\left\langle\nabla f(\mathbf{w}_t,\mathbf{z}_t)-\nabla F_i(\mathbf{w}_t)+\nabla F_i(\mathbf{w}_*^i)-\nabla f\left(\mathbf{w}_*^i,\mathbf{z}_t\right),\mathbf{w}_t-\mathbf{w}_*^i\right\rangle\right] \\
=&-2\eta_k\left(1-\beta\eta_k\right)\cdot\left\langle\mathbf{w}_t-\mathbf{w}_*^i,\nabla F_i\left(\mathbf{w}_t\right)\right\rangle+2\eta_k^2\left\|\nabla f(\mathbf{w}_*^i,\mathbf{z}_t)\right\|_2^2+ \\
&\underbrace{2\eta_k\left\langle(\beta\eta_k-1)\cdot\left(\nabla f\left(\mathbf{w}_t,\mathbf{z}_t\right)-\nabla F_i(\mathbf{w}_t)\right)+\beta\eta_k(\nabla F_i\left(\mathbf{w}_*^i\right)-\nabla f\left(\mathbf{w}_*^i\right),\mathbf{z}_t),\mathbf{w}_t-\mathbf{w}_*^i\right\rangle}_{:=a_{k,t}} \\
\leq&-2\eta_k\left(1-\beta\eta_k\right)\cdot\underbrace{\left(F_i(\mathbf{w}_t)-F_i(\mathbf{w}_*^i)+\frac{\mu}{2}\left\|\mathbf{w}_t-\mathbf{w}_*^i\right\|_2^2\right)}_{\geq 0}+2\eta_k^2\left\|\nabla f\left(\mathbf{w}_*^i,\mathbf{z}_t\right)\right\|_2^2+a_{k,t} \\
\overset{(36)}{\leq}&-\frac{1}{2}\eta_k\mu\left\|\mathbf{w}_t-\mathbf{w}_*^i\right\|_2^2-\eta_k\left(F_i\left(\mathbf{w}_t\right)-F_i\left(\mathbf{w}_*^i\right)\right)+2\eta_k^2\left\|\nabla f\left(\mathbf{w}_*^i,\mathbf{z}_t\right)\right\|_2^2+a_{k,t},
\end{aligned}
$$

in which $\beta\eta_k\leq\frac{1}{2}$. For simplicity, we introduce some simplified notations.

$$
\begin{cases}
\varepsilon_k=1-\frac{1}{2}\eta_k\mu, \\
b_{k,t}=-\eta_k\left(F_i\left(\mathbf{w}_t\right)-F_i\left(\mathbf{w}_*^i\right)\right), \\
c_{k,t}=2\eta_k^2\left\|\nabla f\left(\mathbf{w}_*^i,\mathbf{z}_t\right)\right\|_2^2.
\end{cases}
$$

Rearranging terms and summing over $t=s_k,s_k+1,\ldots,e_k$ yields

$$
\begin{aligned}
&\left\|\mathbf{w}_{e_k+1}-\mathbf{w}_*^i\right\|_2^2 \\
\leq&\varepsilon_k\cdot\left\|\mathbf{w}_{e_k}-\mathbf{w}_*^i\right\|_2^2+a_{k,e_k}+b_{k,e_k}+c_{k,e_k} \\
\leq&\varepsilon_k^2\cdot\left\|\mathbf{w}_{e_k-1}-\mathbf{w}_*^i\right\|_2^2+\varepsilon_k\cdot\left(a_{k,e_k-1}+b_{k,e_k-1}+c_{k,e_k-1}\right)+a_{k,e_k}+b_{k,e_k}+c_{k,e_k} \\
\leq&\cdots \\
\leq&\varepsilon_k^{e_k-s_k+1}\cdot\left\|\mathbf{w}_{s_k}-\mathbf{w}_*^i\right\|_2^2+\sum\nolimits_{t=s_k}^{e_k}\varepsilon_k^{e_k-t}\cdot\left(a_{k,t}+b_{k,t}+c_{k,t}\right).
\end{aligned}
$$

Recalling that (or refer to (11)),

$$
\bar{\mathbf{w}}_k^i=\sum_{t=s_k}^{e_k}\frac{\varepsilon_k^{e_k-t}}{\sum_{\tau=s_k}^{e_k}\varepsilon_k^{e_k-\tau}}\cdot\mathbf{w}_t.
$$

Rearranging terms and by the convexity of $F_i(\cdot)$ yields

$$
\begin{aligned}
&F_i\left(\bar{\mathbf{w}}_k^i\right)-F_i\left(\mathbf{w}_*^i\right) \\
\leq&\frac{1}{\sum_{\tau=s_k}^{e_k}\varepsilon_k^{e_k-\tau}}\cdot\left(\sum_{t=s_k}^{e_k}\varepsilon_k^{e_k-t}F_i\left(\mathbf{w}_t\right)-\sum_{t=s_k}^{e_k}\varepsilon_k^{e_k-t}F_i\left(\mathbf{w}_*^i\right)\right) \\
=&\frac{1}{\sum_{\tau=s_k}^{e_k}\varepsilon_k^{e_k-\tau}}\sum\nolimits_{t=s_k}^{e_k}\varepsilon_k^{e_k-t}\frac{-b_{k,t}}{\eta_k} \\
\leq&\frac{1}{\sum_{\tau=s_k}^{e_k}\varepsilon_k^{e_k-\tau}}\cdot\frac{\varepsilon_k^{T_{1,k}}\cdot\left\|\mathbf{w}_{s_k}-\mathbf{w}_*^i\right\|_2^2-\left\|\mathbf{w}_{e_k+1}-\mathbf{w}_*^i\right\|_2^2+\sum_{t=s_k}^{e_k}\varepsilon_k^{e_k-t}\cdot\left(a_{k,t}+c_{k,t}\right)}{\eta_k}.
\end{aligned}
$$

By (36), it can be proved that

$$\begin{aligned}\sum_{\tau=s_k}^{e_k} \varepsilon_k^{e_k-\tau} =& \frac{1-\left(1-\frac{1}{2}\eta_k\mu\right)^{T_{1,k}}}{1-\left(1-\frac{1}{2}\eta_k\mu\right)} \leq \frac{2}{\eta_k\mu},\\ \sum_{\tau=s_k}^{e_k} \varepsilon_k^{e_k-\tau} =& \frac{1-\left(1-\frac{1}{2}\eta_k\mu\right)^{T_{1,k}}}{1-\left(1-\frac{1}{2}\eta_k\mu\right)} \geq 2\frac{1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right)}{\eta_k\mu}.\end{aligned} \tag{37}$$

Let $\mathbb{E}_k^i[\cdot]$ be the expectation taken over $\mathbf{z}_{k,s_k}^i, \ldots, \mathbf{z}_{k,e_k}^i$. Then the expectation of $F_i\left(\bar{\mathbf{w}}_k^i\right)$ satisfies

$$\begin{aligned}&\mathbb{E}_k^i\left[F_i\left(\bar{\mathbf{w}}_k^i\right)\right] - F_i\left(\mathbf{w}_*^i\right)\\ \leq& \frac{1}{\sum_{\tau=s_k}^{e_k}\varepsilon_k^{e_k-\tau}} \cdot \frac{1}{\eta_k} \cdot \mathbb{E}_k^i\left[\varepsilon_k^{T_{1,k}} \cdot \left\|\mathbf{w}_{s_k} - \mathbf{w}_*^i\right\|_2^2 + \sum\nolimits_{t=s_k}^{e_k} \varepsilon_k^{e_k-t} \cdot (a_{k,t} + c_{k,t})\right]\\ \overset{(37)}{\leq}& \frac{1}{2}\frac{\eta_k\mu}{1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right)} \cdot \frac{1}{\eta_k}\left(\varepsilon_k^{T_{1,k}} \cdot \left\|\mathbf{w}_{s_k} - \mathbf{w}_*^i\right\|_2^2 + \eta_k\frac{4\sigma_i^2}{\mu}\right) \quad \text{by Assumption 3 } (ii)\\ \leq& \frac{\mu}{2}\exp\left(-\frac{T_{1,k}\eta_k\mu}{2}\right)\frac{1}{1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right)}\left\|\mathbf{w}_{s_k} - \mathbf{w}_*^i\right\|_2^2 + \frac{2\eta_k\sigma_i^2}{1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right)},\end{aligned}$$

which concludes the proof. □

Next we prove Theorem 2.

Proof of Theorem 2. We first consider $k = 0$. Recalling that $T_1 \geq \lceil 38\kappa \rceil$. By (36), we have

$$1-\exp\left(-\frac{1}{2}T_{1,0}\eta_0\mu\right) = 1-\exp\left(-\frac{1}{2}\frac{T_1}{c_0}\frac{1}{2\beta}\mu\right) \geq 1-\exp\left(-9.5/c_0\right) > 1-\exp\left(-2\right),$$

where $c_0 \in [\frac{7}{5}, 4]$. By Lemma 13, we have

$$\mathbb{E}_0^i\left[F_i\left(\bar{\mathbf{w}}_0^i\right) - F_i\left(\mathbf{w}_*^i\right)\right] \leq \frac{3}{5}\mu\exp\left(-\frac{T_1}{4c_0\kappa}\right) \cdot \left\|\mathbf{w}_{-1} - \mathbf{w}_*^i\right\|_2^2 + \frac{1}{\beta}\frac{\sigma_i^2}{1-\exp\left(-2\right)}, \tag{38}$$

in which $\mathbf{w}_{0,s_0}^i = \mathbf{w}_{-1}$.

Next we consider $k = K, K-1, \ldots, 1$. We have

$$1-\exp\left(-\frac{1}{2}T_{1,k}\eta_k\mu\right) = 1-\exp\left(-\frac{1}{2}T_{1,1}\eta_1\mu\right) \geq 1-\exp\left(-2^{0.5}\right) > \frac{3}{4}.$$

For any $k \geq 0$, let $\mathbb{E}^i_{0:k}[\cdot]$ be the expectation taken over the training sets $\mathcal{T}_{i,1,0}, \ldots, \mathcal{T}_{i,1,k}$. By Lemma 13, and the fact $\mathbf{w}^i_{k,s_k} = \bar{\mathbf{w}}^i_{k-1}$ for all $k \geq 1$, we have

$$\begin{aligned}
\mathbb{E}^i_{0:K}\left[F_i\left(\bar{\mathbf{w}}^i_K\right) - F_i\left(\mathbf{w}^i_*\right)\right] \leq & \frac{\mu}{2}\cdot\frac{4}{3}\exp\left(-\frac{T_{1,K}\eta_K\mu}{2}\right)\cdot\mathbb{E}^i_{0:K-1}\left[\left\|\mathbf{w}^i_{K,s_K} - \mathbf{w}^i_*\right\|_2^2\right] + \eta_K\frac{8\sigma_i^2}{3} \\
= & \frac{\mu}{2}\cdot\frac{4}{3}\exp\left(-\frac{T_{1,K}\eta_K\mu}{2}\right)\cdot\mathbb{E}^i_{0:K-1}\left[\left\|\bar{\mathbf{w}}^i_{K-1} - \mathbf{w}^i_*\right\|_2^2\right] + \eta_K\frac{8\sigma_i^2}{3} \\
\overset{(23)}{\leq} & \frac{4}{3}\exp\left(-\frac{T_{1,1}\eta_1\mu}{2}\right)\cdot\mathbb{E}^i_{0:K-1}\left[F_i\left(\bar{\mathbf{w}}^i_{K-1}\right) - F_i\left(\mathbf{w}^i_*\right)\right] + \frac{\eta_1\cdot 8\sigma_i^2}{3\cdot 2^{K-1}} \\
\leq & \frac{4}{3}\exp\left(-2^{0.5}\right)\cdot\mathbb{E}^i_{0:K-1}\left[F_i\left(\bar{\mathbf{w}}^i_{K-1}\right) - F_i\left(\mathbf{w}^i_*\right)\right] + \eta_1\frac{8\sigma_i^2}{3\cdot 2^{K-1}} \\
\leq & 2^{-1.5}\cdot\mathbb{E}^i_{0:K-1}\left[F_i\left(\bar{\mathbf{w}}^i_{K-1}\right) - F_i\left(\mathbf{w}^i_*\right)\right] + \eta_1\frac{8\sigma_i^2}{3\cdot 2^{K-1}} \\
\leq & 2^{-3}\cdot\mathbb{E}^i_{0:K-2}\left[F_i\left(\bar{\mathbf{w}}^i_{K-2}\right) - F_i\left(\mathbf{w}^i_*\right)\right] + \eta_1\frac{8\sigma_i^2}{3\cdot 2^{1.5+K-2}} + \frac{\eta_1\cdot 8\sigma_i^2}{3\cdot 2^{K-1}} \\
\leq & \cdots \\
\leq & 2^{-1.5K}\cdot\mathbb{E}^i_0\left[F_i\left(\bar{\mathbf{w}}^i_0\right) - F_i\left(\mathbf{w}^i_*\right)\right] + \sum_{k=0}^{K-1}\frac{1}{2^{1.5k+K-k-1}}\cdot\eta_1\frac{8\sigma_i^2}{3} \\
= & 2^{-1.5K}\cdot\mathbb{E}^i_0\left[F_i\left(\bar{\mathbf{w}}^i_0\right) - F_i\left(\mathbf{w}^i_*\right)\right] + 2^{1-K}\frac{1-2^{-0.5K}}{1-2^{-0.5}}\cdot\eta_1\frac{8\sigma_i^2}{3}.
\end{aligned}$$

By (36) and Lemma 12, we have

$$\left(1-\frac{1}{c_0}\right)T_1 = \sum_{k=1}^{K}T_{1,k} = \sum_{k=1}^{K}T_{1,1}\cdot 2^{k-1} = T_{1,1}\cdot\frac{1-2^K}{1-2} < 2^K T_{1,1}.$$

Rearranging terms yields an upper bound on 2^{-K}. We further obtain

$$\begin{aligned}
& \mathbb{E}^i_{0:K}\left[F_i\left(\bar{\mathbf{w}}^i_K\right)\right] - F_i\left(\mathbf{w}^i_*\right) \\
\leq & \left(\frac{c_0 T_{1,1}}{(c_0-1)T_1}\right)^{1.5}\cdot\mathbb{E}^i_0\left[F_i\left(\bar{\mathbf{w}}^i_0\right) - F_i\left(\mathbf{w}^i_*\right)\right] + \frac{2}{1-2^{-0.5}}\eta_1\frac{8\sigma_i^2}{3}\cdot\frac{c_0 T_{1,1}}{(c_0-1)T_1} \\
\overset{(38)}{\leq} & \left(\frac{7}{2T_1}\left\lceil\frac{2^{1.5}}{\eta_1\mu}\right\rceil\right)^{1.5}\left[\frac{3}{5}\mu\exp\left(\frac{-T_1}{4c_0\kappa}\right)\left\|\mathbf{w}_{-1} - \mathbf{w}^i_*\right\|_2^2 + \frac{\sigma_i^2/\beta}{1-\exp(-2)}\right] + \frac{56\sqrt{2}\sigma_i^2\eta_1}{3(\sqrt{2}-1)T_1}\left\lceil\frac{2^{1.5}}{\eta_1\mu}\right\rceil \\
\leq & \frac{3\mu}{10}\exp\left(-\frac{T_1}{16\kappa}\right)\cdot\left\|\mathbf{w}_{-1} - \mathbf{w}^i_*\right\|_2^2 + 200\frac{\sigma_i^2}{\mu T_1},
\end{aligned}$$

in which we require $T_1 \geq \lceil 38\kappa \rceil$. We conclude the proof. □

E Bounding the Drift Terms

Our algorithm executes local updating for N times after each communication round, posing a technical challenge on analyzing the weighted excess risk. Specifically, after the r-th communication, the selected clients execute N local updates with the initial model $\mathbf{w}_r$, and obtain S trajectories,

$$\forall k \in \mathcal{S}_{I_r}, \quad \mathbf{w}_r = \mathbf{w}^k_{r,s_r} \to \mathbf{w}^k_{r,s_r+1} \to \ldots \to \mathbf{w}^k_{r,e_r+1}.$$

Therefore, there are $(N+1)\times S$ local models between $\mathbf{w}_{r+1}$ and $\mathbf{w}_r$. Similar to the analysis of Theorem 2, we must upper bound $\|\mathbf{w}_{r+1}-\mathbf{w}\|_2^2$ by $\|\mathbf{w}_r-\mathbf{w}\|_2^2$. To this end, it is necessary to analyze $\left\|\mathbf{w}^k_{r,\tau} - \mathbf{w}_r\right\|_2^2$ for all $\tau = s_r, \ldots, e_r+1$, which we call *drift terms*. Previous work (Karimireddy et al., 2020) uses the relaxed triangle inequality (5) to bound the drift term, leading to an iteration of the form $\|\mathbf{w}^k_{r,\tau+1} - \mathbf{w}_r\|_2^2 \leq (1+\frac{1}{N-1})\|\mathbf{w}^k_{r,\tau} - \mathbf{w}_r\|_2^2 + N\cdot h_1(\nabla f(\mathbf{w}^k_{r,\tau}, \mathbf{z}^k_{r,\tau}))$ where $h_1(\cdot)$ is a function of stochastic gradient. There is a factor $1+\frac{1}{N-1} > 1$. To remove this undesired factor, we do not use (5) and make use of the β-smooth and μ-strongly convex of loss function. In this way, we obtain an iteration of the form $\|\mathbf{w}^k_{r,\tau+1} - \mathbf{w}_r\|_2^2 \leq \|\mathbf{w}^k_{r,\tau} - \mathbf{w}_r\|_2^2 + h_2(\nabla f(\mathbf{w}^k_{r,\tau}, \mathbf{z}^k_{r,\tau}))$.

Lemma 14. *At any round $r = 1, \ldots, R$, let*

$$\begin{aligned}
\forall \mathbf{w} \in \mathcal{W}, \quad \delta_{r,\tau}^k(\mathbf{w}) =& f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right) - F_k\left(\mathbf{w}\right), \\
A_{r,\tau}^k =& \delta_{r,\tau}^k(\mathbf{w}_r) - \delta_{r,\tau}^k(\mathbf{w}) - \left\langle \nabla f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right), \mathbf{w}_r - \mathbf{w} \right\rangle, \\
B_{r,\tau}^k =& \delta_{r,\tau}^k(\mathbf{w}) - \delta_{r,\tau}^k\left(\mathbf{w}_*^k\right) - \left\langle \nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right), \mathbf{w} - \mathbf{w}_*^k \right\rangle, \\
C_{r,\tau}^k =& 8\beta\eta_l^2 \left[A_{r,\tau}^k + B_{r,\tau}^k + \delta_{r,\tau}^k(\mathbf{w}_r) - \delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right)\right], \\
D_{r,\tau}^k =& 2\eta_l \left(1 - 4\beta\eta_l\right) \left\langle \mathbf{w}_{r,\tau}^k - \mathbf{w}_r, \nabla F_k\left(\mathbf{w}_{r,\tau}^k\right) - \nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right) \right\rangle.
\end{aligned} \tag{39}$$

The trajectory $\mathbf{w}_{r,\tau}^k, \tau = s_r, \ldots, e_r + 1$, generated by TWS-Fed, satisfies

$$\begin{aligned}
\forall \tau =& s_r, \ldots, e_r, \quad \left\|\mathbf{w}_{r,\tau+1}^k - \mathbf{w}_r\right\|_2^2 \\
\leq& \sum_{t=s_r}^{\tau} \left[4\eta_l^2 \left\|\nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,t}^k\right)\right\|_2^2 - \left(2\eta_l + 8\beta\eta_l^2\right)\left[F_k\left(\mathbf{w}_*^k\right) - F_k\left(\mathbf{w}_r\right)\right] + C_{r,t}^k + D_{r,t}^k\right].
\end{aligned}$$

We take a close look at the upper bound of the drift term. Let $\mathbf{w} = \mathbf{w}_*$ be the global optimal parameter. If we take expectation w.r.t. $\mathbf{z}_{r,\tau}^k$, then $\mathbb{E}[D_{r,\tau}^k] = 0$ and

$$\mathbb{E}\left[\left\|\nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,t}^k\right)\right\|_2^2\right] \leq \sigma_k^2.$$

We further take expectation w.r.t. the random subset $\mathcal{S}_{I_r}$, then $\mathbb{E}[C_{r,\tau}^k] \leq 0$. Thus the drift term can be controlled by the gradient variance σ_k^2 and the excess risk $F_k\left(\mathbf{w}_r\right) - F_k\left(\mathbf{w}_*^k\right)$.

Proof of Lemma 14. We just consider a fix client k. By the local updating rule, we have

$$\begin{aligned}
&\left\|\mathbf{w}_{r,\tau+1}^k - \mathbf{w}_r\right\|_2^2 - \left\|\mathbf{w}_{r,\tau}^k - \mathbf{w}_r\right\|_2^2 \\
=& \eta_l^2 \left\|\nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 - 2\eta_l \left\langle \mathbf{w}_{r,\tau}^k - \mathbf{w}_r, \nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right) \right\rangle.
\end{aligned} \tag{40}$$

We need carefully analyze the squared norm of the stochastic gradient.

$$\begin{aligned}
&\left\|\nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 \\
\leq& 4\left\|\nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right) - \nabla f\left(\mathbf{w}_r, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 + 4\left\|\nabla f\left(\mathbf{w}_r, \mathbf{z}_{r,\tau}^k\right) - \nabla f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 + \\
& 4\left\|\nabla f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right) - \nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 + 4\left\|\nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 \\
\overset{(19)}{\leq}& 8\beta\left[f\left(\mathbf{w}_r, \mathbf{z}_{r,\tau}^k\right) - f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right) - \left\langle \nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right), \mathbf{w}_r - \mathbf{w}_{r,\tau}^k \right\rangle\right] + \\
& 8\beta\left[f\left(\mathbf{w}_r, \mathbf{z}_{r,\tau}^k\right) - f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right) - \left\langle \nabla f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right), \mathbf{w}_r - \mathbf{w} \right\rangle\right] + \\
& 8\beta\left[f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right) - f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right) - \left\langle \nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right), \mathbf{w} - \mathbf{w}_*^k \right\rangle\right] + 4\left\|\nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2 \\
=& 8\beta\left[F_k\left(\mathbf{w}_r\right) - F_k\left(\mathbf{w}_{r,\tau}^k\right) + \delta_{r,\tau}^k(\mathbf{w}_r) - \delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right) + \left\langle \mathbf{w}_{r,\tau}^k - \mathbf{w}_r, \nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right) \right\rangle\right] + \\
& 8\beta\left[F_k\left(\mathbf{w}_r\right) - F_k\left(\mathbf{w}\right) + \delta_{r,\tau}^k(\mathbf{w}_r) - \delta_{r,\tau}^k(\mathbf{w}) - \left\langle \nabla f\left(\mathbf{w}, \mathbf{z}_{r,\tau}^k\right), \mathbf{w}_r - \mathbf{w} \right\rangle\right] + \\
& 8\beta\left[F_k\left(\mathbf{w}\right) - F_k\left(\mathbf{w}_*^k\right) + \delta_{r,\tau}^k(\mathbf{w}) - \delta_{r,\tau}^k\left(\mathbf{w}_*^k\right) - \left\langle \nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right), \mathbf{w} - \mathbf{w}_*^k \right\rangle\right] + \\
& 4\left\|\nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right)\right\|_2^2,
\end{aligned}$$

in which $\delta_{r,\tau}^k(\mathbf{w})$ follows (39). Let η_l satisfy

$$1 - 4\beta\eta_l > 0. \tag{41}$$

Substituting the upper bound into (40) and simplifying some notations by (39), we obtain

$$
\begin{aligned}
&\left\|\mathbf{w}_{r,\tau+1}^k-\mathbf{w}_r\right\|_2^2-\left\|\mathbf{w}_{r,\tau}^k-\mathbf{w}_r\right\|_2^2\\
\leq&8\beta\eta_l^2\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_{r,\tau}^k\right)+\delta_{r,\tau}^k(\mathbf{w}_r)-\delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right)\right]+\\
&8\beta\eta_l^2\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*^k\right)+A_{r,\tau}^k+B_{r,\tau}^k\right]+4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+\\
&2\eta_l\left(1-4\beta\eta_l\right)\left\langle\mathbf{w}_{r,\tau}^k-\mathbf{w}_r,\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)\right\rangle-\\
&2\eta_l\left(1-4\beta\eta_l\right)\left\langle\mathbf{w}_{r,\tau}^k-\mathbf{w}_r,\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)\right\rangle\\
\overset{(7)}{\leq}&8\beta\eta_l^2\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_{r,\tau}^k\right)+\delta_{r,\tau}^k(\mathbf{w}_r)-\delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right)\right]+\\
&8\beta\eta_l^2\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*^k\right)+A_{r,\tau}^k+B_{r,\tau}^k\right]+4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+\\
&2\eta_l\left(1-4\beta\eta_l\right)\left\langle\mathbf{w}_{r,\tau}^k-\mathbf{w}_r,\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)\right\rangle-\\
&\left(2\eta_l-8\beta\eta_l^2\right)\cdot\left[F_k\left(\mathbf{w}_{r,\tau}^k\right)-F_k\left(\mathbf{w}_r\right)+\frac{\mu}{2}\left\|\mathbf{w}_{r,\tau}^k-\mathbf{w}_r\right\|_2^2\right]\\
\overset{(41)}{\leq}&8\beta\eta_l^2\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*^k\right)+A_{r,\tau}^k+B_{r,\tau}^k+\delta_{r,\tau}^k(\mathbf{w}_r)-\delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right)\right]+\\
&2\eta_l\left(1-4\beta\eta_l\right)\left\langle\mathbf{w}_{r,\tau}^k-\mathbf{w}_r,\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)\right\rangle+\\
&4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2-2\eta_l\left[F_k\left(\mathbf{w}_{r,\tau}^k\right)-F_k\left(\mathbf{w}_r\right)\right]\\
=&8\beta\eta_l^2\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*^k\right)+A_{r,\tau}^k+B_{r,\tau}^k+\delta_{r,\tau}^k(\mathbf{w}_r)-\delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right)\right]+\\
&2\eta_l\left(1-4\beta\eta_l\right)\left\langle\mathbf{w}_{r,\tau}^k-\mathbf{w}_r,\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)\right\rangle+\\
&4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2-2\eta_l\left[F_k\left(\mathbf{w}_{r,\tau}^k\right)-F_k\left(\mathbf{w}_*^k\right)\right]-2\eta_l\left[F_k\left(\mathbf{w}_*^k\right)-F_k\left(\mathbf{w}_r\right)\right]\\
\leq&\underbrace{8\beta\eta_l^2\left[A_{r,\tau}^k+B_{r,\tau}^k+\delta_{r,\tau}^k(\mathbf{w}_r)-\delta_{r,\tau}^k\left(\mathbf{w}_{r,\tau}^k\right)\right]}_{:=C_{r,\tau}^k}+4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+\\
&\underbrace{2\eta_l\left(1-4\beta\eta_l\right)\left\langle\mathbf{w}_{r,\tau}^k-\mathbf{w}_r,\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)\right\rangle}_{:=D_{r,\tau}^k}-\\
&\left(2\eta_l+8\beta\eta_l^2\right)\left[F_k\left(\mathbf{w}_*^k\right)-F_k\left(\mathbf{w}_r\right)\right].
\end{aligned}
$$

Summing over $t=s_r,\ldots,\tau$ terms yields

$$
\begin{aligned}
&\left\|\mathbf{w}_{r,\tau+1}^k-\mathbf{w}_r\right\|_2^2-\left\|\mathbf{w}_{r,s_r}^k-\mathbf{w}_r\right\|_2^2\\
\leq&\sum_{t=s_r,\ldots,\tau}\left[4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,t}^k\right)\right\|_2^2+C_{r,t}^k+D_{r,t}^k-\left(2\eta_l+8\beta\eta_l^2\right)\left(F_k\left(\mathbf{w}_*^k\right)-F_k\left(\mathbf{w}_r\right)\right)\right].
\end{aligned}
$$

We conclude the proof by $\mathbf{w}_{r,s_r}^k=\mathbf{w}_r$. □

F Proof of Theorem 3

We again establish a progress lemma which measures the progress of $\mathbf{w}_r$ towards a desired $\mathbf{w}\in\mathcal{W}$ after an iteration of federated aggregation.

Lemma 15 (Progress Lemma). *For $r=1,\ldots,R$, let $\bar{\nabla}_r=\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)$ and $\nabla_r=\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla F_k\left(\mathbf{w}_{r,\tau}^k\right)$. Let $\eta=\eta_l\cdot\eta_g$ and $2\eta\beta+4\beta^2N\eta^2\leq 3\eta\beta$. Let $A_{r,\tau}^k$, $B_{r,\tau}^k$,*

$C_{r,\tau}^k$ and $D_{r,\tau}^k$ follow (39). For any $r = 1, \ldots, R$, TWS-Fed achieves

$$
\begin{aligned}
\forall \mathbf{w} \in \mathcal{W}, \quad & \|\mathbf{w}_{r+1} - \mathbf{w}\|_2^2 - \left(1 - \frac{1}{2} N\mu\eta\right) \|\mathbf{w}_r - \mathbf{w}\|_2^2 \\
& \leq -2\eta \left\langle \mathbf{w}_r - \mathbf{w}, \bar{\nabla}_r - \nabla_r \right\rangle + 3\eta\beta \frac{1}{S} \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r-1} \sum_{t=s_r}^{\tau} 4\eta_l^2 \left\| \nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,t}^k\right) \right\|_2^2 + \\
& \quad 3\eta\beta \frac{1}{S} \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r-1} \sum_{t=s_r}^{\tau} \left[C_{r,t}^k + D_{r,t}^k - \left(2\eta_l + 8\beta\eta_l^2\right) \left(F_k\left(\mathbf{w}_*^k\right) - F_k\left(\mathbf{w}_r\right)\right) \right] + \\
& \quad 8\eta^2\beta N \frac{1}{S} \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r} \left[A_{r,\tau}^k + B_{r,\tau}^k \right] + \frac{4\eta^2}{S^2} \left\| \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r} \nabla f\left(\mathbf{w}_*^k, \mathbf{z}_{r,\tau}^k\right) \right\|_2^2 + \\
& \quad 8\eta^2\beta N^2 \frac{1}{S} \sum_{k \in \mathcal{S}_{I_r}} \left[F_k\left(\mathbf{w}\right) - F_k\left(\mathbf{w}_*^k\right) \right] - 2N\eta(1 - 4\beta\eta N) \frac{1}{S} \sum_{k \in \mathcal{S}_{I_r}} \left[F_k(\mathbf{w}_r) - F_k\left(\mathbf{w}\right) \right].
\end{aligned}
$$

Let $\mathbf{w} = \mathbf{w}_*$. Taking expectation w.r.t. the random subset $\mathcal{S}_{I_r}$ and the random examples $\{\mathbf{z}_{r,\tau}^k\}$, $\tau = s_r, \ldots, e_r$, then the right-hand side of the inequality only depends on the gradient variance σ^2, the heterogeneity level Δ and the global excess risk $\frac{1}{M}\sum_{k\in[M]}\left[F_k(\mathbf{w}_r) - F_k\left(\mathbf{w}_*\right)\right]$. By carefully designing the values of η_l, η and R, it is easy to derive the final weighted excess risk by Lemma 15.

Proof of Lemma 15. Let $\eta = \eta_l \cdot \eta_g$. For any $r = 1, 2, \ldots, R$,

$$
\begin{aligned}
& \|\mathbf{w}_{r+1} - \mathbf{w}\|_2^2 - \|\mathbf{w}_r - \mathbf{w}\|_2^2 \\
= & -2\eta \left\langle \mathbf{w}_r - \mathbf{w}, \nabla_r \right\rangle - 2\eta \left\langle \mathbf{w}_r - \mathbf{w}, \bar{\nabla}_r - \nabla_r \right\rangle + \eta^2 \left\| \bar{\nabla}_r \right\|_2^2 \\
= & -\frac{2\eta}{S} \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r} \left\langle \mathbf{w}_r - \mathbf{w}, \nabla F_k\left(\mathbf{w}_{r,\tau}^k\right) \right\rangle - 2\eta \left\langle \mathbf{w}_r - \mathbf{w}, \bar{\nabla}_r - \nabla_r \right\rangle + \eta^2 \left\| \bar{\nabla}_r \right\|_2^2 \\
\overset{(24)}{\leq} & -\frac{2\eta}{S} \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r} \left[F_k(\mathbf{w}_r) - F_k(\mathbf{w}) + \frac{\mu}{4} \|\mathbf{w}_r - \mathbf{w}\|_2^2 - \beta \left\| \mathbf{w}_{r,\tau}^k - \mathbf{w}_r \right\|_2^2 \right] - \\
& 2\eta \left\langle \mathbf{w}_r - \mathbf{w}, \bar{\nabla}_r - \nabla_r \right\rangle + \eta^2 \left\| \bar{\nabla}_r \right\|_2^2 \\
= & -2\eta \left[\frac{N}{S} \sum_{k \in \mathcal{S}_{I_r}} \left(F_k(\mathbf{w}_r) - F_k\left(\mathbf{w}\right)\right) + \frac{N\mu}{4} \|\mathbf{w}_r - \mathbf{w}\|_2^2 - \frac{\beta}{S} \sum_{k \in \mathcal{S}_{I_r}} \sum_{\tau=s_r}^{e_r} \left\| \mathbf{w}_{r,\tau}^k - \mathbf{w}_r \right\|_2^2 \right] - \\
& 2\eta \left\langle \mathbf{w}_r - \mathbf{w}, \bar{\nabla}_r - \nabla_r \right\rangle + \eta^2 \left\| \bar{\nabla}_r \right\|_2^2. \qquad (42)
\end{aligned}
$$

Next we give an upper bound on $\left\|\bar{\nabla}_r\right\|_2^2$.

$$
\begin{aligned}
&\left\|\bar{\nabla}_r\right\|_2^2 \\
\overset{(21)}{\leq}&\frac{4}{3}\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right]\right\|_2^2+4\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2 \\
=&\frac{4}{3}\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[\nabla f\left(\mathbf{w}_{r,\tau}^k,\mathbf{z}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_r,\mathbf{z}_{r,\tau}^k\right)+\nabla f\left(\mathbf{w}_r,\mathbf{z}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)\right]+\right. \\
&\left.\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right]\right\|_2^2+4\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2 \\
\leq&\frac{4N\beta^2}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left\|\mathbf{w}_{r,\tau}^k-\mathbf{w}_r\right\|_2^2+4\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+ \\
&\frac{4N}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[\left\|\nabla f\left(\mathbf{w}_r,\mathbf{z}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+\left\|\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)-\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2\right] \\
\overset{(19)}{\leq}&\frac{4N\beta^2}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left\|\mathbf{w}_{r,\tau}^k-\mathbf{w}_r\right\|_2^2+4\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+ \\
&\frac{4N}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}2\beta\left[f\left(\mathbf{w}_r,\mathbf{z}_{r,\tau}^k\right)-f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)-\left\langle\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right),\mathbf{w}_r-\mathbf{w}\right\rangle\right]+ \\
&\frac{4N}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}2\beta\left[f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)-f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)-\left\langle\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right),\mathbf{w}-\mathbf{w}_*^k\right\rangle\right] \\
=&\frac{4N\beta^2}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left\|\mathbf{w}_{r,\tau}^k-\mathbf{w}_r\right\|_2^2+4\left\|\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+ \\
&\frac{8N\beta}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}\right)+\underbrace{\delta_{r,\tau}^k(\mathbf{w}_r)-\delta_{r,\tau}^k(\mathbf{w})-\left\langle\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right),\mathbf{w}_r-\mathbf{w}\right\rangle}_{=A_{r,\tau}^k}\right]+ \\
&\frac{8N\beta}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_*^k\right)+\underbrace{\delta_{r,\tau}^k(\mathbf{w})-\delta_{r,\tau}^k\left(\mathbf{w}_*^k\right)-\left\langle\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right),\mathbf{w}-\mathbf{w}_*^k\right\rangle}_{=B_{r,\tau}^k}\right].
\end{aligned}
$$

Let η satisfy

$$
2\eta\beta+4\beta^2N\eta^2\leq 3\eta\beta. \tag{43}
$$

Substituting the upper bound into (42), and by Lemma 14, we obtain

$$
\begin{aligned}
&\|\mathbf{w}_{r+1}-\mathbf{w}\|_2^2-\left(1-\frac{1}{2}N\mu\eta\right)\|\mathbf{w}_r-\mathbf{w}\|_2^2\\
\leq&-2\eta\left\langle\mathbf{w}_r-\mathbf{w},\bar{\nabla}_r-\nabla_r\right\rangle+\left(2\eta\beta+4\beta^2N\eta^2\right)\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\underbrace{\left\|\mathbf{w}_{r,\tau}^k-\mathbf{w}_r\right\|_2^2}_{\text{drift term}}+\\
&8\eta^2\beta N\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[A_{r,\tau}^k+B_{r,\tau}^k\right]+\frac{4\eta^2}{S^2}\left\|\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+\\
&8\eta^2\beta N^2\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_*^k\right)\right]-2N\eta(1-4\beta\eta N)\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}\right)\right]\\
\overset{(43)}{\leq}&-2\eta\left\langle\mathbf{w}_r-\mathbf{w},\bar{\nabla}_r-\nabla_r\right\rangle+3\eta\beta\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r-1}\sum_{t=s_r}^{\tau}4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,t}^k\right)\right\|_2^2+\\
&3\eta\beta\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r-1}\sum_{t=s_r}^{\tau}\left[C_{r,t}^k+D_{r,t}^k-\left(2\eta_l+8\beta\eta_l^2\right)\left(F_k\left(\mathbf{w}_*^k\right)-F_k\left(\mathbf{w}_r\right)\right)\right]+\\
&8\eta^2\beta N\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[A_{r,\tau}^k+B_{r,\tau}^k\right]+\frac{4\eta^2}{S^2}\left\|\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2+\\
&8\eta^2\beta N^2\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_*^k\right)\right]-2N\eta(1-4\beta\eta N)\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}\right)\right],
\end{aligned}
$$

which concludes the proof. □

Next we prove Theorem 3.

Proof of Theorem 3. To start with, we introduce some simplified notations.

$$
\begin{aligned}
\xi=&1-\frac{1}{2}N\mu\eta,\\
a_r=&3\eta\beta\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r-1}\sum_{t=s_r}^{\tau}\left[C_{r,t}^k+D_{r,t}^k\right]+\eta^2\frac{8\beta N}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left[A_{r,\tau}^k+B_{r,\tau}^k\right],\\
b_r=&3\eta\beta\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r-1}\sum_{t=s_r}^{\tau}4\eta_l^2\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,t}^k\right)\right\|_2^2+\frac{4\eta^2}{S^2}\left\|\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2,\\
c_r=&-2N\eta\left[1-4\beta\eta N-3\beta\eta_l(1+4\beta\eta_l)\frac{N-1}{2}\right]\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}\right)\right],\\
d_r=&\eta\beta N\frac{8\eta N+3\eta_l(1+4\beta\eta_l)(N-1)}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_*^k\right)\right]-2\eta\left\langle\mathbf{w}_r-\mathbf{w},\bar{\nabla}_r-\nabla_r\right\rangle.
\end{aligned}
$$

By Lemma 15, one has

$$
\begin{aligned}
&\|\mathbf{w}_{r+1}-\mathbf{w}\|_2^2\\
\leq&\xi\cdot\|\mathbf{w}_r-\mathbf{w}\|_2^2-2\eta\left\langle\mathbf{w}_r-\mathbf{w},\bar{\nabla}_r-\nabla_r\right\rangle+a_r+b_r-\\
&3\eta\beta\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r-1}(\tau-s_r+1)\left(2\eta_l+8\beta\eta_l^2\right)\left[F_k\left(\mathbf{w}_*^k\right)-F_k\left(\mathbf{w}_r\right)\right]+\\
&8\eta^2\beta N^2\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_*^k\right)\right]-2N\eta(1-4\beta\eta N)\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}\right)\right]\\
=&\xi\cdot\|\mathbf{w}_r-\mathbf{w}\|_2^2-2\eta\left\langle\mathbf{w}_r-\mathbf{w},\bar{\nabla}_r-\nabla_r\right\rangle+a_r+b_r-\\
&3\eta\beta\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\frac{N(N-1)}{2}\left(2\eta_l+8\beta\eta_l^2\right)\left[F_k\left(\mathbf{w}_*^k\right)-F_k\left(\mathbf{w}\right)+F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_r\right)\right]+\\
&8\eta^2\beta N^2\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}\right)-F_k\left(\mathbf{w}_*^k\right)\right]-2N\eta(1-4\beta\eta N)\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}\right)\right]\\
=&\xi\cdot\|\mathbf{w}_r-\mathbf{w}\|_2^2+a_r+b_r+c_r+d_r.
\end{aligned}
$$

Summing over $r=1,2,\ldots,R$, we obtain

$$
\begin{aligned}
&\|\mathbf{w}_{R+1}-\mathbf{w}\|_2^2\\
\leq&\xi\cdot\|\mathbf{w}_R-\mathbf{w}\|_2^2+a_R+b_R+c_R+d_R\\
\leq&\xi\cdot\left(\xi\|\mathbf{w}_{R-1}-\mathbf{w}\|_2^2+a_{R-1}+b_{R-1}+c_{R-1}+d_{R-1}\right)+a_R+b_R+c_R+d_R\\
=&\xi^2\cdot\|\mathbf{w}_{R-1}-\mathbf{w}\|_2^2+\xi\left(a_{R-1}+b_{R-1}+c_{R-1}+d_{R-1}\right)+a_R+b_R+c_R+d_R\\
\leq&\cdots\\
\leq&\xi^R\cdot\|\mathbf{w}_1-\mathbf{w}\|_2^2+\sum_{r=1}^{R}\xi^{R-r}\left(a_r+b_r+c_r+d_r\right).
\end{aligned}
$$

Recalling that the second stage of TWS-Fed is a single-pass federated stochastic gradient and the training sets $\mathcal{T}_{1,2},\ldots,\mathcal{T}_{M,2}$ are independent of $\mathbf{w}_1$. Thus for any $r\in\{1,\ldots,R\}$ and for any $\tau\in\{s_r,\ldots,e_r\}$, $\mathbf{z}_{r,\tau}^k$ is independent of $\mathbf{w}_1,\ldots,\mathbf{w}_r$, and $\mathbf{w}_{r,t}^k$, $t=s_r,\ldots,\tau-1$. Next we consider the expectation of a_r,b_r,c_r and d_r.

- $\mathbb{E}[a_r]$
  Suppose $\mathbf{z}_{r,\tau}^k$ is independent (or conditionally independent) with $\mathbf{w}\in\mathcal{W}$ and $\mathbf{v}\in\mathcal{W}$. Then taking expectation w.r.t. $\mathbf{z}_{r,\tau}^k$ yields

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[\delta_{r,\tau}^k(\mathbf{w})\right]=\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right)\right]-F_k\left(\mathbf{w}\right)=0,\\
&\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[\left\langle\nabla f\left(\mathbf{w},\mathbf{z}_{r,\tau}^k\right),\mathbf{v}-\mathbf{w}\right\rangle\right]=\left\langle\nabla F_k\left(\mathbf{w}\right),\mathbf{v}-\mathbf{w}\right\rangle.
\end{aligned}
$$

  Then by (39), one has

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[A_{r,\tau}^k\right]=-\left\langle\nabla F_k\left(\mathbf{w}\right),\mathbf{w}_r-\mathbf{w}\right\rangle,\\
&\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[B_{r,\tau}^k\right]=-\left\langle\nabla F_k\left(\mathbf{w}_*^k\right),\mathbf{w}-\mathbf{w}_*^k\right\rangle\overset{(18)}{\leq}0,\\
&\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[C_{r,\tau}^k\right]\leq-8\beta\eta_l^2\cdot\left\langle\nabla F_k\left(\mathbf{w}\right),\mathbf{w}_r-\mathbf{w}\right\rangle,\\
&\mathbb{E}_{\mathbf{z}_{r,\tau}^k}\left[D_{r,\tau}^k\right]=0.
\end{aligned}
$$

Let $\mathbf{w} = \mathbf{w}_*$. Taking expectation w.r.t. $\mathcal{S}_{I_r}$ yields

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_{I_r}}\left[\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left\langle\nabla F_k\left(\mathbf{w}_*\right),\mathbf{w}_r-\mathbf{w}_*\right\rangle\right]=&\frac{S}{M}\sum_{I_r=1}^{\frac{M}{S}}\cdot\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left\langle\nabla F_k\left(\mathbf{w}_*\right),\mathbf{w}_r-\mathbf{w}_*\right\rangle\\=&\frac{1}{M}\sum\nolimits_{k=1}^{M}\left\langle\nabla F_k\left(\mathbf{w}_*\right),\mathbf{w}_r-\mathbf{w}_*\right\rangle\overset{(18)}{\geq}0.\end{aligned}$$

Given $\mathbf{w}_r$, taking expectation w.r.t. $\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}, k\in\mathcal{S}_{I_r}$ yields

$$\mathop{\mathbb{E}}_{\mathcal{S}_{I_r},\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}[a_r]=\mathop{\mathbb{E}}_{\mathcal{S}_{I_r},\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}\left[\frac{3\eta\beta}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r-1}\sum_{t=s_r}^{\tau}C_{r,t}^k+\eta^2\frac{8\beta N}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}A_{r,\tau}^k\right]\leq 0.$$

The expected value of a_r is non-positive.

- $\mathbb{E}[b_r]$

Recalling that $\sigma^2=\frac{1}{M}\sum_{k=1}^{M}\sigma_k^2$. One has

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{S}_{I_r},\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}[b_r]\\=&12\eta\beta\eta_l^2\cdot\frac{N(N-1)}{2}\sigma^2+\mathop{\mathbb{E}}_{\mathcal{S}_{I_r},\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}\left[\frac{4\eta^2}{S^2}\left\|\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2\right]\\\leq&12\eta\beta\eta_l^2\cdot\frac{N(N-1)}{2}\sigma^2+\frac{4\eta^2}{S^2}\mathop{\mathbb{E}}_{\mathcal{S}_{I_r},\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}\left[\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\left\|\nabla f\left(\mathbf{w}_*^k,\mathbf{z}_{r,\tau}^k\right)\right\|_2^2\right]+\\&\frac{4\eta^2}{S^2}\sum_{i\neq j\in\mathcal{S}_{I_r}}\sum_{t\neq\tau=s_r}^{e_r}2\left\langle\nabla F_j\left(\mathbf{w}_*^j\right),\nabla F_i\left(\mathbf{w}_*^i\right)\right\rangle\\\leq&6\eta\beta\eta_l^2\cdot N(N-1)\sigma^2+\frac{4}{S}N\sigma^2\eta^2.\end{aligned}$$

The expected value of b_r only depends on the weighted gradient variance σ^2.

- $\mathbb{E}[c_r]$

Let $\mathbf{w} = \mathbf{w}_*$. Taking expectation w.r.t. $\mathcal{S}_{I_r}$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_{I_r}}\left[\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*\right)\right]\right]=&\frac{S}{M}\sum\nolimits_{r=1}^{\frac{M}{S}}\cdot\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*\right)\right]\\=&\frac{1}{M}\sum\nolimits_{k=1}^{M}\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*\right)\right]\geq 0.\end{aligned}$$

Recalling Theorem 3, we require that η_l and η satisfy

$$\begin{cases}0<\beta\eta_l\leq\frac{1}{12N}, & N\geq 1,\\ \eta N\beta\leq\frac{1}{12}.\end{cases}\tag{44}$$

By this inequality, one has

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_{I_r}}\left[c_r\right]=&-2N\eta\left[1-4\beta\eta N-3\beta\eta_l(1+4\beta\eta_l)\frac{N-1}{2}\right]\frac{1}{M}\sum\nolimits_{k=1}^{M}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}_*\right)\right]\\\leq&-2N\eta\left(\frac{5}{6}-4\beta\eta N\right)\frac{1}{M}\sum\nolimits_{k=1}^{M}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}_*\right)\right]\\\leq&-N\eta\cdot\frac{1}{M}\sum\nolimits_{k=1}^{M}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}_*\right)\right].\end{aligned}$$

The expected value of c_r only depends on the excess risk and is non-positive.

- $\mathbb{E}[d_r]$
  For any $\tau = s_r, \ldots, e_r$, $\mathbf{w}_{r,\tau}^k$ is independent of $\mathbf{z}_{r,\tau}^k$. Taking expectation w.r.t. $\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}$ yields

$$\mathop{\mathbb{E}}_{\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}\left[\bar{\nabla}_\tau\right] = \mathop{\mathbb{E}}_{\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}\left[\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\sum_{\tau=s_r}^{e_r}\nabla f\left(\mathbf{w}_{r,\tau}^k, \mathbf{z}_{r,\tau}^k\right)\right] = \nabla_\tau.$$

  Thus we have

$$\mathop{\mathbb{E}}_{\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}\left[\left\langle \mathbf{w}_r - \mathbf{w}, \bar{\nabla}_r - \nabla_r\right\rangle\right] = 0.$$

  Let $\mathbf{w} = \mathbf{w}_*$. Conditioned on $\mathbf{w}_1, \ldots, \mathbf{w}_r$, taking expectation w.r.t. $\mathcal{S}_{I_r}$, we obtain

$$\mathbb{E}_{\mathcal{S}_{I_r}}\left[\frac{1}{S}\sum_{k\in\mathcal{S}_{I_r}}\left[F_k\left(\mathbf{w}_*\right) - F_k\left(\mathbf{w}_*^k\right)\right]\right] = \frac{1}{M}\sum_{k=1}^{M}\left[F_k\left(\mathbf{w}_*\right) - F_k\left(\mathbf{w}_*^k\right)\right] = \Delta,$$

  in which we use Definition 1. Thus we have

$$\mathop{\mathbb{E}}_{\mathcal{S}_{I_r},\{\mathbf{z}_{r,\tau}^k\}_{\tau=s_r}^{e_r}}[d_r] \le \eta\beta N\left(8\eta N + 3\eta_l(1+4\beta\eta_l)(N-1)\right)\cdot\Delta \overset{(44)}{\le} 2\eta N\left(4\eta\beta N + \frac{1}{6}\right)\cdot\Delta.$$

  The expected value of d_r only depends on the heterogeneity level.

Recalling Theorem 3, η_l, η and R are defined by

$$\eta_l^2 = \frac{1}{12SR\beta^2N^2}. \tag{45}$$

Let $\mathbb{E}_{\{\mathcal{T}_{k,2}\}_{k=1}^M,\{\mathcal{S}_{I_r}\}_{r=1}^R}[\cdot]$ be the expectation taken over $\{\mathcal{T}_{k,2}\}_{k=1}^M$ and $\{\mathcal{S}_{I_r}\}_{r=1}^R$. Substituting into the expected values of a_r, b_r, c_r and d_r, we obtain

$$\begin{aligned}
&\mathbb{E}_{\{\mathcal{T}_{k,2}\}_{k=1}^M,\{\mathcal{S}_{I_r}\}_{r=1}^R}\left[\left\|\mathbf{w}_{R+1} - \mathbf{w}_*\right\|_2^2\right] \\
\le&\xi^R\cdot\left\|\mathbf{w}_1 - \mathbf{w}_*\right\|_2^2 + \sum_{r=1}^{R}\xi^{R-r}\left[6\eta\beta\eta_l^2N(N-1)\sigma^2 + \frac{4}{S}N\sigma^2\eta^2 + 2\eta N\left(4\eta\beta N + \frac{1}{6}\right)\Delta\right] - \\
&N\eta\cdot\mathbb{E}_{\{\mathcal{T}_{k,2}\}_{k=1}^M,\{\mathcal{S}_{I_r}\}_{r=1}^R}\left[\sum_{r=1}^{R}\xi^{R-r}\cdot\frac{1}{M}\sum_{k=1}^{M}\left[F_k(\mathbf{w}_r) - F_k\left(\mathbf{w}_*\right)\right]\right] \\
\le&\left(1 - \frac{N\eta\mu}{2}\right)^R\cdot\left\|\mathbf{w}_1 - \mathbf{w}_*\right\|_2^2 + \frac{1-\xi^R}{1-\xi}\left[\frac{\eta\sigma^2}{2SR\beta} + \frac{4}{S}N\sigma^2\eta^2 + 2\eta N\left(4\eta\beta N + \frac{1}{6}\right)\Delta\right] - \\
&N\eta\cdot\mathbb{E}_{\{\mathcal{T}_{k,2}\}_{k=1}^M,\{\mathcal{S}_{I_r}\}_{r=1}^R}\left[\sum_{r=1}^{R}\xi^{R-r}\cdot\frac{1}{M}\sum_{k=1}^{M}\left[F_k(\mathbf{w}_r) - F_k\left(\mathbf{w}_*\right)\right]\right].
\end{aligned}$$

Simplifying the right-hand side of the inequality yields the desired result. □

G Proof of Corollary 1

Proof of Corollary 1. Recalling (17) in which

$$\begin{cases} \eta = \eta_l\cdot\eta_g = \frac{2\ln S}{NR\mu}, & S > 1, \\ R = \lceil 24\kappa\ln S\rceil. \end{cases}$$

It is easy to verify that the two conditions in (44) hold.

Note that $\xi = 1 - \frac{1}{2}N\eta\mu$. By (17), we have

$$\xi^R = \left(1 - \frac{\ln S}{R}\right)^R \le \frac{1}{S} \le \frac{1}{2}, \quad \text{and} \quad \frac{1}{N\mu\eta} \le \sum_{r=1}^{R}\xi^{R-r} = \frac{1-\xi^R}{1-\xi} \le \frac{2}{N\mu\eta}. \tag{46}$$

By Theorem 3, we have

$$\begin{aligned} & N\eta \cdot \mathbb{E}_{\{\mathcal{T}_{k,2}\}_{k=1}^{M},\{\mathcal{S}_{I_r}\}_{r=1}^{R}}\left[\sum_{r=1}^{R}\xi^{R-r}\cdot\frac{1}{M}\sum_{k=1}^{M}\left[F_k(\mathbf{w}_r)-F_k\left(\mathbf{w}_*\right)\right]\right] \\ \leq & \frac{\left\|\mathbf{w}_1-\mathbf{w}_*\right\|_2^2}{S}+\left(\frac{1}{\beta}+\frac{16\ln S}{\mu}\right)\frac{\sigma^2}{ST_2\mu}+\frac{2}{\mu}\Delta. \end{aligned} \tag{47}$$

By (15), TWS-Fed outputs a global model $\bar{\mathbf{w}}$ that is a weighted combination of $\mathbf{w}_1,\ldots,\mathbf{w}_R$, i.e,

$$\bar{\mathbf{w}}=\frac{1}{\sum_{r=1}^{R}\xi^{R-r}}\sum_{r=1}^{R}\xi^{R-r}\cdot\mathbf{w}_r.$$

Next we analyze the weighted excess risk of $\bar{\mathbf{w}}$.

Let $\mathbb{E}[\cdot]$ be the expectation taken over the randomness of $\{\mathcal{T}_k\}_{k=1}^{M}$ and $\{\mathcal{S}_{I_r}\}_{r=1}^{R}$ and $\mathbb{E}_{\{\mathcal{T}_{k,1}\}_{k=1}^{M}}[\cdot]$ be the expectation taken over $\{\mathcal{T}_{k,1}\}_{k=1}^{M}$. Then we have

$$\begin{aligned} & \mathbb{E}\left[\mathcal{E}_{\frac{1}{M}\mathbf{1}_M}\left(\bar{\mathbf{w}}\right)\right] \\ = & \mathbb{E}\left[\frac{1}{M}\sum_{k=1}^{M}F_k\left(\bar{\mathbf{w}}\right)-\frac{1}{M}\sum_{k=1}^{M}F_k\left(\mathbf{w}_*\right)\right]+\underbrace{\frac{1}{M}\sum_{k=1}^{M}F_k\left(\mathbf{w}_*\right)-\frac{1}{M}\sum_{k=1}^{M}F_k\left(\mathbf{w}_*^k\right)}_{\text{By Definition 1}} \\ = & \mathbb{E}\left[\frac{1}{M}\sum_{k=1}^{M}F_k\left(\sum_{r=1}^{R}\frac{\xi^{R-r}}{\sum_{r=1}^{R}\xi^{R-r}}\mathbf{w}_r\right)-\frac{1}{M}\sum_{k=1}^{M}F_k\left(\mathbf{w}_*\right)\right]+\Delta \\ \leq & \frac{1}{\sum_{r=1}^{R}\xi^{R-r}}\mathbb{E}\left[\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{R}\xi^{R-r}F_k\left(\mathbf{w}_r\right)-\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{R}\xi^{R-r}F_k\left(\mathbf{w}_*\right)\right]+\Delta \\ = & \frac{1}{N\eta}\cdot\frac{1}{\sum_{r=1}^{R}\xi^{R-r}}\cdot\mathbb{E}\left[N\eta\cdot\sum_{r=1}^{R}\xi^{R-r}\cdot\frac{1}{M}\sum_{k=1}^{M}\left[F_k\left(\mathbf{w}_r\right)-F_k\left(\mathbf{w}_*\right)\right]\right]+\Delta \\ \overset{(46),(47)}{\leq} & \mu\cdot\left[\frac{1}{S}\cdot\mathbb{E}_{\{\mathcal{T}_{k,1}\}_{k=1}^{M}}\left[\left\|\mathbf{w}_1-\mathbf{w}_*\right\|_2^2\right]+\left(\frac{1}{\beta}+\frac{16\ln S}{\mu}\right)\frac{\sigma^2}{ST_2\mu}+\frac{2}{\mu}\Delta\right]+\Delta \\ \leq & \mu\cdot\frac{1}{S}\cdot\mathbb{E}_{\{\mathcal{T}_{k,1}\}_{k=1}^{M}}\left[\frac{1}{M}\sum_{k=1}^{M}\left\|\hat{\mathbf{w}}_*^k-\mathbf{w}_*\right\|_2^2\right]+\left(1+16\ln S\right)\frac{\sigma^2}{ST_2\mu}+3\Delta \\ \leq & \frac{3\mu}{4MS}\sum_{k=1}^{M}\exp\left(\frac{-T_1}{16\kappa}\right)\cdot\left\|\mathbf{w}_{-1}-\mathbf{w}_*^k\right\|_2^2+\frac{450\sigma^2}{ST_1\mu}+\frac{1+16\ln S}{ST_2\mu}\sigma^2+\left(\frac{18}{S}+3\right)\Delta, \end{aligned}$$

in which $\mathbf{w}_1=\hat{\mathbf{w}}_*=\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{w}}_*^k$ and

$$\begin{aligned} & \mathbb{E}_{\{\mathcal{T}_{k,1}\}_{k=1}^{M}}\left[\frac{1}{M}\sum_{k=1}^{M}\left\|\hat{\mathbf{w}}_*^k-\mathbf{w}_*\right\|_2^2\right] \\ \overset{(21)}{\leq} & \mathbb{E}_{\{\mathcal{T}_{k,1}\}_{k=1}^{M}}\left[\frac{1}{M}\sum_{k=1}^{M}\frac{9}{8}\left\|\hat{\mathbf{w}}_*^k-\mathbf{w}_*^k\right\|_2^2\right]+9\frac{1}{M}\sum_{k=1}^{M}\left\|\mathbf{w}_*^k-\mathbf{w}_*\right\|_2^2 \\ \overset{(25)}{\leq} & \mathbb{E}_{\{\mathcal{T}_{k,1}\}_{k=1}^{M}}\left[\frac{9}{4\mu}\frac{1}{M}\sum_{k=1}^{M}\left[F_k\left(\hat{\mathbf{w}}_*^k\right)-F_k\left(\mathbf{w}_*^k\right)\right]\right]+\frac{18}{\mu}\frac{1}{M}\sum_{k=1}^{M}\left[F_k\left(\mathbf{w}_*\right)-F_k\left(\mathbf{w}_*^k\right)\right] \\ \overset{(16)}{\leq} & \frac{9}{4\mu}\frac{1}{M}\sum_{k=1}^{M}\left[\frac{3\mu}{10}\exp\left(-\frac{T_1}{16\kappa}\right)\cdot\left\|\mathbf{w}_{-1}-\mathbf{w}_*^k\right\|_2^2+200\frac{\sigma_k^2}{\mu T_1}\right]+\frac{18}{\mu}\Delta. \end{aligned}$$

Finally, we explain that there is an upper bound on M. As the maximal rounds of communication is T_2 by setting $N=1$. Thus we require $R\leq T_2$, implies the condition $24\kappa\ln M\leq T_2$. Solving this inequality yields $M\leq\exp(\frac{T_2}{24\kappa})$. In this way, we conclude the proof. □