Fast Composite Optimization and Statistical Recovery in Federated Learning

Yajie Bao [1] **Michael Crawshaw** [2] **Shan Luo** [1] **Mingrui Liu** [2]

Abstract

As a prevalent distributed learning paradigm, Federated Learning (FL) trains a global model on a massive amount of devices with infrequent communication. This paper investigates a class of composite optimization and statistical recovery problems in the FL setting, whose loss function consists of a data-dependent smooth loss and a non-smooth regularizer. Examples include sparse linear regression using Lasso, low-rank matrix recovery using nuclear norm regularization, etc. In the existing literature, federated composite optimization algorithms are designed only from an optimization perspective without any statistical guarantees. In addition, they do not consider commonly used (restricted) strong convexity in statistical recovery problems. We advance the frontiers of this problem from both optimization and statistical perspectives. From optimization upfront, we propose a new algorithm named *Fast Federated Dual Averaging* for strongly convex and smooth loss and establish state-of-the-art iteration and communication complexity in the composite setting. In particular, we prove that it enjoys a fast rate, linear speedup, and reduced communication rounds. From statistical upfront, for restricted strongly convex and smooth loss, we design another algorithm, namely *Multi-stage Federated Dual Averaging*, and prove a high probability complexity bound with linear speedup up to optimal statistical precision. Numerical experiments in both synthetic and real data demonstrate that our methods perform better than other baselines. To the best of our knowledge, this is the first work providing fast optimization algorithms and statistical recovery guarantees for composite problems in FL.

[1]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China [2]Department of Computer Science, George Mason University, Fairfax, VA 22030, USA. Correspondence to: Mingrui Liu <mingruil@gmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Federated Learning (FL) is a popular learning paradigm in distributed learning that enables a large number of clients to collaboratively learn a global model without sharing individual data (McMahan et al., 2017). The most well-known algorithm in FL is called Federated Averaging (FedAvg). In each round, FedAvg samples a subset of devices and runs multiple steps of Stochastic Gradient Descent (SGD) on these devices in parallel, then the central server updates the global model by aggregation at the end of the communication round and broadcasts the updated model to clients. It has been verified that FedAvg achieves similar performance with fewer communication rounds compared with parallel SGD (Li et al., 2019; Stich, 2019; Woodworth et al., 2020a).

Most of the research in FL mainly focuses on unconstrained smooth optimization problems without a regularizer and assumes each client has access to its local population distribution. However, people usually want the learned model to have some patterns, such as (group) sparsity and low rank. These desired patterns are usually achieved by solving composite optimization problems, e.g., LASSO (Tibshirani, 1996), Graphical LASSO (Friedman et al., 2008), Elastic net (Zou & Hastie, 2005), matrix completion (Candès & Recht, 2009). So it is crucial to study how to solve these composite problems under the FL environment in the current big data era. This paper considers solving a composite optimization problem in the FL paradigm where only infrequent communication is allowed. In particular, we aim to solve

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \phi(\boldsymbol{w}) := \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{w}) + h(\boldsymbol{w}), \tag{1}$$

where π_k is the weight of the k-th client, $\sum_{i=1}^{K} \pi_k = 1$, $\mathcal{L}_k(\boldsymbol{w}) = \mathbb{E}_{\xi\sim\mathcal{P}_k}[f(\boldsymbol{w};\xi)]$ is the loss function evaluated at the k-th client, $\mathcal{P}_k$ denotes the population distribution on the k-th client, and $h(\boldsymbol{w})$ is a non-smooth regularizer. The most related works which can solve this problem in FL setting are Yuan et al. (2021) and Tran-Dinh et al. (2021). Yuan et al. (2021) proposed an algorithm called Federated Dual Averaging (FedDA). In one round, each client in FedDA performs dual averaging to update its primal and dual states for several steps; then, the server aggregates the dual states and updates the global primal state by a proximal step. However, they did

Table 1. Comparison of related works under bounded heterogeneity (see Assumption 3). (R)SC and GC refer to (restricted) strongly convex and general convex respectively. (R)SM refers to (restricted) smooth. N/A means not available. K: number of clients; L: smooth parameter; μ: (restricted) strongly convex parameter; σ^2: variance of stochastic gradient; $\widehat{\boldsymbol{w}}$: global minimizer of (1); $\widehat{\boldsymbol{w}}_{\text{Fast-FedDA}}$: returned solution from Algorithm 1 after running T iterations on each client; $\widehat{\boldsymbol{w}}_{\text{MC-FedDA}}$: returned solution from Algorithm 3 after running T iterations on each client; $\boldsymbol{w}^*$: ground-truth solution (5); ϵ_{stat}: optimal statistical precision in Proposition 3.1; $\tilde{\mathcal{O}}$: hides logarithmic factors.

Algorithm	Reference	Communication rounds for linear speedup	Problem	Conditions	Convergence rate for $\phi(\widehat{\boldsymbol{w}}_{\text{Fast-FedDA}}) - \phi(\widehat{\boldsymbol{w}})$	Iteration complexity for $\|\widehat{\boldsymbol{w}}_{\text{MC-FedDA}} - \boldsymbol{w}^*\|^2 \le \epsilon_{\text{stat}}$	Guarantee
FedAvg	(Woodworth et al., 2020a)	$\mathcal{O}(T^{1/2}K^{1/2})$	unconstrained	SC, SM	$\mathcal{O}(\sigma^2/(\mu KT))$	N/A	Expectation
		$\mathcal{O}(T^{3/4}K^{3/4})$	unconstrained	GC, SM	$\mathcal{O}(\sigma/\sqrt{KT})$	N/A	Expectation
SCAFFOLD	(Karimireddy et al., 2020a)	$\tilde{\mathcal{O}}(L/\mu)$	unconstrained	SC, SM	$\mathcal{O}(\sigma^2/(\mu KT))$	N/A	Expectation
		$\mathcal{O}(T^{1/2}K^{1/2})$	unconstrained	GC, SM	$\mathcal{O}(\sigma/\sqrt{KT})$	N/A	Expectation
FedDA	(Yuan et al., 2021)	$\mathcal{O}(T^{3/4}K^{3/4})$	composite	GC, SM	$\mathcal{O}(\sigma/\sqrt{KT})$	N/A	Expectation
Fast-FedDA	Theorem 2.1	$\tilde{\mathcal{O}}(T^{1/2}K^{1/2})$	composite	SC, SM	$\mathcal{O}(\sigma^2/(\mu KT))$	N/A	Expectation
MC-FedDA	Theorem 3.2	$\tilde{\mathcal{O}}(T^{1/2}K^{1/2})$	composite	RSC, RSM	$\tilde{\mathcal{O}}(\sigma^2/(\mu KT))$	$\tilde{\mathcal{O}}(\sigma^2/(\mu K\epsilon_{\text{stat}}))$	High probability

not consider exploiting the strong convexity of the loss function and hence only ended up with a slow convergence rate (i.e., $\mathcal{O}(1/\sqrt{T})$). Tran-Dinh et al. (2021) considered non-convex loss and provided an algorithm that can converge to a point with small gradient mapping, but it does not have any global optimization guarantees. It remains unclear how to improve the convergence rate further when solving strongly convex composite problems in the FL setting. To answer this question, we propose a new algorithm, namely *Fast Federated Dual Averaging* (`Fast-FedDA`), with provable fast rate, linear speedup and almost the same communication complexity achieved by FedAvg as in the unconstrained strongly convex case without regularizer (Woodworth et al., 2020a; Karimireddy et al., 2020a).

A fundamental assumption in FL literature is that it assumes that every client can have access to its local population distribution. However, it may not be the case in practice: each client usually only has access to its local empirical distribution (Negahban et al., 2012; Agarwal et al., 2012; Wainwright, 2019). This motivates us to consider a more challenging problem in FL: statistical recovery. It is devoted to recovering the ground-truth model parameter $\boldsymbol{w}^*$ by only accessing the empirical distribution. It is much more difficult than the typical results in FL since we need to simultaneously deal with computational, statistical, and communication efficiency. Statistical recovery is usually achieved through solving a composite optimization problem as

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \phi(\boldsymbol{w}) := \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{w}) + \lambda \mathcal{R}(\boldsymbol{w}), \tag{2}$$

where $\mathcal{L}_k(\boldsymbol{w}) = \mathbb{E}_{\xi\sim\mathcal{D}_k}[f(\boldsymbol{w};\xi)]$ is the empirical loss at k-th client [1], $\mathcal{D}_k$ is the corresponding empirical distribution on the k-th client, λ is regularization parameter, and $\mathcal{R}(\cdot)$ is a non-smooth norm penalty. In addition, due to high dimensionality and small sample size in each client, the assumption of strong convexity might be demanding and unrealistic. Hence we further consider the broadly used restricted strong convexity (RSC) and restricted smooth (RSM) conditions in statistical recovery problems (Agarwal et al., 2012; Wang et al., 2014; Loh & Wainwright, 2015).

Distributed statistical recovery is an extensively studied topic in recent years (Lee et al., 2017; Wang et al., 2017; Jordan et al., 2018; Chen et al., 2020), but these works assume that all clients have the same data distribution, and they also assume that there is a closed-form solution for some non-trivial optimization subproblems. The unique features of FL are high heterogeneity in data among clients and local updates to solve these subproblems explicitly. Hence the algorithms and theoretical results of the literature mentioned above are not directly applicable in the FL regime. To address this issue, under RSC and RSM conditions, we first introduce an algorithm named *Constrained Federated Dual Averaging* (`C-FedDA`) to solve a $\mathcal{R}$-norm constrained subproblem. Then we introduce another algorithm called *Multi-stage Constrained Federated Dual Averaging* (`MC-FedDA`), which calls `C-FedDA` in multiple stages with adaptively changing hyperparameter (e.g., shrinking radius of $\mathcal{R}$-norm ball). In particular, after finishing one stage of `C-FedDA`, we use the output as a warm-start for the next stage.

We summarize our contributions in the following:

1. For federated composite optimization problems with strongly convex and smooth loss, we propose the `Fast-FedDA` algorithm for solving (1) by accessing data sampled from population distribution. Under the general bounded heterogeneity assumption, we show that `Fast-FedDA` enjoys linear speedup, and the communication complexity matches the lower bound of `FedAvg` (up to some logarithmic factors) for strongly convex problems without a regularizer (Karimireddy et al., 2020a).

2. To obtain the statistical recovery results through solving (2), we propose an algorithm, namely `MC-FedDA` in the FL setting by accessing data sampled from the

[1] We use $\mathcal{L}_k$ to denote the empirical loss in Section 3, and denote the population loss in Section 2.

empirical distribution. Under RSC and RSM conditions, we prove that MC-FedDA enjoys optimal (high probability) convergence rate $\tilde{\mathcal{O}}(\sigma^2 \log(1/\delta)/(\mu TK))$ to attain statistical error bound. We find that the typical convergence rate in expectation in FL is insufficient for achieving statistical recovery guarantees, and this is the first high probability result for composite problems in FL.

3. We conduct numerical experiments on linear regression with ℓ_1 penalty, low-rank matrix estimation with nuclear norm penalty, and multiclass logistic regression with ℓ_1 penalty. Both synthetic and real data show better performances of Fast-FedDA and MC-FedDA compared with other baselines.

A comparison of our results to related works is presented in Table 1. For more related work, please refer to Appendix A.

Notations. For a vector $\boldsymbol{w} \in \mathbb{R}^p$, we use $\|\boldsymbol{w}\|$ to denote the Euclidean norm. For a matrix $\boldsymbol{W} \in \mathbb{R}^{p_1 \times p_2}$, we use $\|\boldsymbol{W}\|_{\mathrm{F}}$ to denote the Frobenius norm and use $\|\boldsymbol{W}\|_{\mathrm{nuc}}$ to denote the nuclear norm. For two real positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists some positive constant c such that $a_n \leq cb_n$. We use $a_n = \mathcal{O}(b_n)$ to hide multiplicative absolute constant c and also use $a_n = \tilde{\mathcal{O}}(b_n)$ to hide logarithmic factors. In our paper, $\mathbb{E}_{\mathcal{P}}$ means taking expectation with the randomness from true distribution $\mathcal{P}$ and $\mathbb{E}_{\mathcal{D}}$ means taking expectation with the randomness from empirical distribution $\mathcal{D}$.

2. Fast Federated Composite Optimization

In this section, we focus on the composite optimization problem in FL environment for strongly convex and smooth loss. Given a user-specific loss function $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$, suppose there are K clients, let $\mathcal{L}_k(\boldsymbol{w}) = \mathbb{E}_{\xi \sim \mathcal{P}_k}[f(\boldsymbol{w}; \xi)]$ be the local population loss and π_k be the local weight for $k = 1, ..., K$. We consider the following composite problem

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{w}) + h(\boldsymbol{w}) \right\}, \tag{3}$$

where $\mathcal{W} \subseteq \mathbb{R}^p$ is the domain and $h : \mathbb{R}^p \to \mathbb{R}$ is a non-smooth regularizer. From now on, we denote $\mathcal{L}(\boldsymbol{w}) = \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{w})$ and write the global composite objective as $\phi(\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w}) + h(\boldsymbol{w})$.

Assumption 1. The local loss functions $\mathcal{L}_k$ for $k \in [K]$ are L-smooth and μ-strongly convex, that is for any $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathcal{W}$, there exist $0 < \mu \leq L$ such that

$$\mathcal{L}_k(\boldsymbol{w}) - \mathcal{L}_k(\boldsymbol{w}') - \langle \nabla \mathcal{L}_k(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2$$

and

$$\mathcal{L}_k(\boldsymbol{w}) - \mathcal{L}_k(\boldsymbol{w}') - \langle \nabla \mathcal{L}_k(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle \leq \frac{L}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2.$$

Algorithm 1 Fast-FedDA($\boldsymbol{w}_0$, R, E, μ, L)

1: **Input:** Initial point $\boldsymbol{w}_0$, iteration number T, constants (μ, L) and synchronized set $\mathcal{I} = \{t_r : 0 \leq r \leq R\}$.
2: **Initialize:** $\boldsymbol{w}_0^k = \boldsymbol{w}_0$ for $k \in [K]$, $\alpha_t = t + 1$, $\gamma_t = L\alpha_t$.
3: **for** Round $r = 0$ **to** R **do**
4: **for** Client $k = 1$ **to** K **do**
5: **for** $t = t_r$ **to** $t_{r+1} - 1$ **do**
6: Query $\boldsymbol{G}_t^k = \nabla f(\boldsymbol{w}_t^k; \xi_t^k)$ for $\xi_t^k \sim \mathcal{P}_k$.
7: Compute $\boldsymbol{g}_t^k = \boldsymbol{g}_{t-1}^k + \alpha_t \boldsymbol{G}_t^k$.
8: **if** $t < t_{r+1} - 1$ **then**
9: Update: $\boldsymbol{w}_{t+1}^k = \mathrm{Prox}_t(\boldsymbol{g}_t^k - \mu \widetilde{\boldsymbol{w}}_t^k/2)$ and $\widetilde{\boldsymbol{w}}_{t+1}^k = \widetilde{\boldsymbol{w}}_t^k + \alpha_{t+1} \boldsymbol{w}_{t+1}^k$.
10: **else**
11: Send $\boldsymbol{g}_{t_{r+1}-1}^k$ and $\widetilde{\boldsymbol{w}}_{t_{r+1}-1}^k$ to the server.
12: **end if**
13: **end for**
14: **end for**
15: Server aggregates: $\boldsymbol{g}_{t_{r+1}-1} = \sum_{k=1}^{K} \pi_k \boldsymbol{g}_{t_{r+1}-1}^k$ and $\widetilde{\boldsymbol{w}}_{t_{r+1}-1} = \sum_{k=1}^{k} \pi_k \widetilde{\boldsymbol{w}}_{t_{r+1}-1}^k$.
16: Server updates: $\boldsymbol{w}_{t_{r+1}} = \mathrm{Prox}_{t_{r+1}-1}(\boldsymbol{g}_{t_{r+1}-1} - \mu \widetilde{\boldsymbol{w}}_{t_{r+1}-1}/2)$ and $\widetilde{\boldsymbol{w}}_{t_{r+1}} = \widetilde{\boldsymbol{w}}_{t_{r+1}-1} + \alpha_{t_{r+1}} \boldsymbol{w}_{t_{r+1}}$.
17: Synchronization: $\boldsymbol{g}_{t_{r+1}-1}^k \leftarrow \boldsymbol{g}_{t_{r+1}-1}$ and $\widetilde{\boldsymbol{w}}_{t_{r+1}}^k \leftarrow \widetilde{\boldsymbol{w}}_{t_{r+1}}$.
18: **end for**

To solve the problem (3) under strongly convex case, we propose a new algorithm named *Fast Federated Dual Averaging* (Fast-FedDA) in Algorithm 1. The main difference between our algorithm and the FedDA algorithm in Yuan et al. (2021) is that we employed a different dual-averaging scheme in the local updates of Algorithm 1 (line 9). In particular, we not only use information on history cumulative gradient $\boldsymbol{g}_t^k$ as in Yuan et al. (2021) but also history model parameter $\widetilde{\boldsymbol{w}}_t^k$ to leverage the strong convexity.

2.1. Fast Federated Dual Averaging

We begin with defining a *proximal operator* $\mathrm{Prox}_t(\boldsymbol{z})$ for $t \geq 0$ as the solution of the following problem:

$$\min_{\boldsymbol{w} \in \mathcal{W}} \left\{ \langle \boldsymbol{w}, \boldsymbol{z} - \gamma_t \boldsymbol{w}_0 \rangle + \left(\frac{\mu A_t}{2} + \gamma_t \right) \frac{\|\boldsymbol{w}\|^2}{2} + A_t h(\boldsymbol{w}) \right\},$$

where $\boldsymbol{w}_0$ is the initial point and $A_t = \sum_{i=0}^{t} \alpha_i$ is the summation of weights. For a loss function with *strong convexity coefficient* $\mu > 0$, classical stochastic dual averaging (Tseng, 2008; Nesterov, 2009; Chen et al., 2012) updates the model parameter by $\boldsymbol{w}_{t+1} = \mathrm{Prox}_t(\boldsymbol{g}_t - \mu \widetilde{\boldsymbol{w}}_t/2)$ where $\boldsymbol{g}_t = \sum_{i=0}^{t} \alpha_i \nabla f(\boldsymbol{w}_i; \boldsymbol{\xi}_i)$ is the weighted summation of past stochastic gradients and $\widetilde{\boldsymbol{w}}_t = \sum_{i=0}^{t} \alpha_i \boldsymbol{w}_i$ is the weighted summation of past solutions. Denote the synchronized step set by $\mathcal{I} = \{t_r \mid t_r = rE \text{ for } 0 \leq r \leq R\}$, where $t_R = (R + 1)E = T$. Similar to FedAvg (McMa-

han et al., 2017) and FedDA (Yuan et al., 2021), a natural idea to develop a federated dual averaging algorithm for strongly convex loss includes following local updates and server aggregation and update:

- The k-th client updates its local solution by
$$\boldsymbol{w}_{t+1}^k = \mathrm{Prox}_t\left(\boldsymbol{g}_t^k - \mu\widetilde{\boldsymbol{w}}_t^k/2\right),$$
for $t_r \le t \le t_{r+1} - 1$.
- The server updates the global solution by
$$\boldsymbol{w}_{t_{r+1}} = \mathrm{Prox}_{t_{r+1}-1}\left(\sum_{k=1}^K \pi_k(\boldsymbol{g}_t^k - \mu\widetilde{\boldsymbol{w}}_t^k/2)\right).$$

For ease of reference, the detailed procedure of Fast-FedDA is summarized in Algorithm 1. The definitions of $\boldsymbol{g}_t^k$ and $\widetilde{\boldsymbol{w}}_t^k$ are provided in line 7 and 9 respectively.

2.2. Main Results for Fast-FedDA

To establish the convergence results for Fast-FedDA, we impose the following assumptions on the regularizer h, loss function and stochastic gradients.

Assumption 2. The regularizer $h : \mathcal{W} \to \mathbb{R}$ is a closed convex function.

Assumption 3. The global loss function is Λ-smooth, which means $\|\nabla\mathcal{L}(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w}')\| \le \Lambda\|\boldsymbol{w} - \boldsymbol{w}'\|$. In addition, there exists some positive constant H such that $\|\nabla\mathcal{L}(\boldsymbol{w}) - \nabla\mathcal{L}_k(\boldsymbol{w})\| \le H$ for any $\boldsymbol{w} \in \mathcal{W}$ and $k = 1, ..., K$.

Assumption 4. The stochastic gradient sampled from the local population distribution $\mathcal{P}_k$ satisfies that: for any $\boldsymbol{w} \in \mathcal{W}$, it holds that $\mathbb{E}_{\xi\sim\mathcal{P}_k}[\nabla f(\boldsymbol{w};\xi)] = \nabla\mathcal{L}_k(\boldsymbol{w})$ and $\mathbb{E}_{\xi\sim\mathcal{P}_k}[\|\nabla f(\boldsymbol{w};\xi) - \nabla\mathcal{L}_k(\boldsymbol{w})\|^2] \le \sigma^2$.

Assumption 2 is very common in composite optimization literature (Tseng, 2008; Nesterov, 2009; Xiao, 2010). We use Assumption 3 to bound the heterogeneity between clients, which also appears in Woodworth et al. (2020a); Yuan & Ma (2020). The next theorem provides the convergence rate of Fast-FedDA in expectation, and the proof is deferred to Appendix B.2.

Theorem 2.1. *Under Assumptions 1-4, we assume the domain is bounded by $\rho > 0$, that is $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^p : \|\boldsymbol{w}\| \le \rho\}$. We choose $\alpha_t = t + 1$ and $\gamma_t = L\alpha_t$ in Algorithm 1. Considering $\widehat{\boldsymbol{w}}_{\text{Fast-FedDA}} = \sum_{t=0}^T \alpha_t \mathrm{Prox}_{t+1}(\sum_{k=1}^K \pi_k(\boldsymbol{g}_t^k + \mu\widetilde{\boldsymbol{w}}_t^k/2))/A_T$, it satisfies that*

$$\mathbb{E}_{\mathcal{P}}[\phi(\widehat{\boldsymbol{w}}_{\text{Fast-FedDA}}) - \phi(\widehat{\boldsymbol{w}})] \lesssim \frac{LB}{T} + \frac{\bar{\sigma}^2}{\mu T} + \frac{LE\sigma^2\log T}{\mu^2T^2} + \frac{LE^2(H^2 + \Lambda^2 + \mu^2\rho^2)\log T}{\mu^2T^2}, \tag{4}$$

where $\bar{\sigma}^2 = \sum_{k=1}^K \pi_k^2\sigma^2$ and $B = \|\boldsymbol{w}_0 - \widehat{\boldsymbol{w}}\|^2$.

Remark 2.1. The first two terms in (4) are the convergence rate of the dual averaging method in the centralized setting (Lan, 2012; Chen et al., 2012), and the last two terms are incurred from infrequent communication. Now considering the equal-weighted case, that is $\pi_1 = \cdots = \pi_K = 1/K$, the weighted variance is given by $\bar{\sigma}^2 = \sigma^2/K$. In (4), we may choose $E = \widetilde{\mathcal{O}}(\sqrt{T/K})$, and then the convergence rate attains linear speedup with respective to K. Meanwhile, the communication complexity of Fast-FedDA is $\tilde{\mathcal{O}}(T^{1/2}K^{1/2})$, which matches the lower bound for FedAvg up to a logarithmic factor (see Theorem II in Karimireddy et al. (2020a)). Under the same bounded heterogeneity assumption, Yuan et al. (2021) only considered the quadratic loss. In this vein, we investigate a more general loss function in Theorem 2.1.

3. Fast Federated Statistical Recovery

In this section, we consider the statistical recovery via composite optimization in FL framework. Let $\mathcal{P}_k$ for $k = 1, 2, ..., K$ be the unknown local population distributions, then the "true parameter" is defined as

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \sum_{k=1}^K \pi_k\mathbb{E}_{\xi\sim\mathcal{P}_k}[f(\boldsymbol{w};\xi)]. \tag{5}$$

Denote the i.i.d. dataset sampled from $\mathcal{P}_k$ by $\{\xi_i : i \in \mathcal{H}_k$ and $|\mathcal{H}_k| = n_k\}$. We may obtain the sparse/low-rank estimator of $\boldsymbol{w}^*$ through solving the following composite problem

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \left\{\sum_{k=1}^K \pi_k\mathcal{L}_k(\boldsymbol{w}) + \lambda\mathcal{R}(\boldsymbol{w})\right\}, \tag{6}$$

where $\mathcal{L}_k(\boldsymbol{w}) = \sum_{i\in\mathcal{H}_k} f(\boldsymbol{w};\xi_i)/n_k$ is the local empirical loss function and $\mathcal{R}(\cdot)$ is a non-smooth norm regularizer. Here we use $\mathcal{D}_k$ to denote the *empirical distribution* on the k-th client, which means $\mathcal{L}_k(\boldsymbol{w}) = \mathbb{E}_{\xi\sim\mathcal{D}_k}[f(\boldsymbol{w};\xi)]$.

3.1. Illustrative Examples

In this subsection, we take two well known examples to illustrate the statistical recovery problems in FL.

Example 3.1 (Sparse Linear Regression). The linear model in each client is given by

$$\mathrm{y}_i^k = (\mathbf{x}_i^k)^\top\boldsymbol{w}^* + \varepsilon_i^k \quad \text{for } i \in [n_k] \text{ and } k \in [K],$$

where the covariate $\mathbf{x}_i^k$ follows some unknown distribution $\mathcal{P}_k$ and the noise $\varepsilon_i^k \sim N(0, 1)$ is independent of $\mathbf{x}_i^k$. We assume the true regression coefficient $\boldsymbol{w}^*$ is s-sparse, that is $\|\boldsymbol{w}^*\|_0 = s$, and $s \ll p$. Let $\pi_k = \frac{n_k}{N}$, then our goal is to solve the following federated Lasso problem

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \frac{1}{2N}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathrm{y}_i^k - (\mathbf{x}_i^k)^\top\boldsymbol{w})^2 + \lambda\|\boldsymbol{w}\|_1,$$

where λ is the regularization parameter and $\mathcal{W} = \{\boldsymbol{w} : \|\boldsymbol{w}\| \leq \rho\}$.

Example 3.2 (Low-Rank Matrix Estimation). Let $\boldsymbol{W}^* \in \mathbb{R}^{p_1 \times p_2}$ be an unknown matrix with low rank $r^* \ll \min\{p_1, p_2\}$. For each client, the response variable $\boldsymbol{y}_i^k$ and covariate matrix $\mathbf{X}_i^k$ are linked to the unknown matrix via

$$\mathrm{y}_i^k = \langle \mathbf{X}_i^k, \boldsymbol{W}^* \rangle + \varepsilon_i^k \quad \text{for } i \in [n_k] \text{ and } k \in [K],$$

where $\mathbf{X}_i^k$ is sampled from some unknown distribution $\mathcal{P}_k$ and the noise $\varepsilon_i^k \sim N(0,1)$ is independent of $\mathbf{x}_i^k$. Let $\pi_k = \frac{n_k}{N}$, then our goal is to solve the following federated trace regression problem

$$\widehat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W} \in \mathcal{W}} \frac{1}{2N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\boldsymbol{y}_i^k - \langle \mathbf{X}_i^k, \boldsymbol{W} \rangle)^2 + \lambda \|\boldsymbol{W}\|_{\text{nuc}},$$

where λ is the regularization parameter and $\mathcal{W} = \{\boldsymbol{W} : \|\boldsymbol{W}\|_{\text{F}} \leq \rho\}$.

3.2. Restricted Strong Convexity and Smoothness

To develop the techniques for the statistical properties of regularization in FL, we introduce the definition of decomposable regularizer (Negahban et al., 2012).

Definition 1. Given a pair of subspaces in $\mathbb{R}^p$ such that $\mathcal{M} \subseteq \bar{\mathcal{M}}$, a norm regularizer $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$ if

$$\mathcal{R}(\boldsymbol{w}+\boldsymbol{v}) = \mathcal{R}(\boldsymbol{w}) + \mathcal{R}(\boldsymbol{v}) \quad \text{for all } \boldsymbol{w} \in \mathcal{M} \text{ and } \boldsymbol{v} \in \bar{\mathcal{M}}^{\perp}.$$

The subspace Lipschitz constant with respect to the subspace $\bar{\mathcal{M}}$ is defined by

$$\Psi(\bar{\mathcal{M}}) := \sup_{\boldsymbol{u} \in \mathcal{M} \setminus \{\mathbf{0}\}} \frac{\mathcal{R}(\boldsymbol{u})}{\|\boldsymbol{u}\|}.$$

Assumption 5. The regularizer $\mathcal{R}(\cdot)$ is a norm with dual $\mathcal{R}^*(\cdot)$, which satisfies $\mathcal{R}^*(\cdot) \leq \|\cdot\| \leq \mathcal{R}(\cdot)$. There is a pair of subspace $\mathcal{M} \subseteq \bar{\mathcal{M}}$ such that the regularizer decomposes over $(\bar{\mathcal{M}}, \bar{\mathcal{M}}^{\perp})$. Moreover, we assume $\boldsymbol{w}^* \in \mathcal{M}$.

Remark 3.1. $\bar{\mathcal{M}}$ usually encodes structural information of the regularizer. For example, for sparse linear model in Example 3.1, the subspace is defined by $\bar{\mathcal{M}} \equiv \mathcal{M} := \{\boldsymbol{w} \in \mathbb{R}^p | w_j = 0 \text{ for } j \in \mathbb{S}\}$ for some subset $\mathbb{S} \subseteq [p]$. Correspondingly, the subspace Lipschitz constant is given by $\Psi(\bar{\mathcal{M}}) = \sqrt{s}$, where s is the cardinality of the support set $\mathbb{S}$. And the regularizer is $\|\cdot\|_1$, whose dual norm is $\|\cdot\|_\infty$. Clearly, Assumption 5 is satisfied for sparse linear model since $\|\cdot\|_\infty \leq \|\cdot\| \leq \|\cdot\|_1$. In Example 3.2, Assumption 5 is also satisfied. Due to space limit, we refer to Negahban et al. (2012) for more details about $\mathcal{M}$ and $\bar{\mathcal{M}}$ in low-rank matrix estimation.

In the high-dimensional setting ($p > n_k$), it is usually hard to guarantee the strong convexity for the local empirical loss $\mathcal{L}_k$. Therefore, we consider the restricted strong convexity in Assumption 6, which is widely used in statistical recovery literature (Agarwal et al., 2012; Wang et al., 2014; Loh & Wainwright, 2015; Cai et al., 2020). For $k = 1, .., K$, denote the first-order Taylor series expansion of $\mathcal{L}_k(\boldsymbol{w})$ around $\mathcal{L}_k(\boldsymbol{w}')$ by

$$\mathcal{T}_k(\boldsymbol{w}, \boldsymbol{w}') = \mathcal{L}_k(\boldsymbol{w}) - \mathcal{L}_k(\boldsymbol{w}') - \langle \nabla \mathcal{L}_k(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle.$$

Assumption 6. The local loss functions $\mathcal{L}_k$ for $k = 1, ..., K$ are convex and satisfy the restricted strongly convex (RSC) condition, that is for $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathcal{W}$, there exist $\mu > 0$ and $\tau_k \geq 0$ such that

$$\mathcal{T}_k(\boldsymbol{w}, \boldsymbol{w}') \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2 - \tau_k \mathcal{R}^2(\boldsymbol{w} - \boldsymbol{w}').$$

From Assumption 6, the global loss function $\mathcal{L}$ also satisfies the RSC condition: for any $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathcal{W}$

$$\begin{aligned} \mathcal{T}(\boldsymbol{w}, \boldsymbol{w}') &= \mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}') - \langle \nabla \mathcal{L}(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle \\ &\geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2 - \tau \mathcal{R}^2(\boldsymbol{w} - \boldsymbol{w}'), \end{aligned} \tag{7}$$

where $\tau = \sum_{k=1}^{K} \pi_k \tau_k$. We also introduce an analogous notion of restricted smoothness.

Assumption 7. The local loss functions $\mathcal{L}_k$ for $k = 1, ..., K$ satisfy the restricted smooth (RSM) condition, that is for any $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathcal{W}$ there exist $L > 0$ and $\nu_k \geq 0$ such that

$$\mathcal{T}_k(\boldsymbol{w}, \boldsymbol{w}') \leq \frac{L}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2 + \nu_k \mathcal{R}^2(\boldsymbol{w} - \boldsymbol{w}').$$

Similarly, under Assumption 7, the global loss $\mathcal{L}$ satisfies the RSM condition with coefficient L and $\nu = \sum_{k=1}^{K} \pi_k \nu_k$.

With the decomposable regularizer $\mathcal{R}$ and the RSC condition (7) for $\mathcal{L}$, the statistical recovery results via solving the composite problem (6) has been extensively investigated in the past decade (see Negahban et al. (2012); Wainwright (2019) and references therein). We present the optimal statistical error of global estimator $\widehat{\boldsymbol{w}}$ in the following proposition, which is a direct result of Corollary 1 in Negahban et al. (2012) or Theorem 9.19 in Wainwright (2019). The error bound in Proposition 3.1 is also the target precision to achieve optimal statistical recovery.

Proposition 3.1. *Under Assumptions 5 and 6. If* $\tau \Psi^2(\bar{\mathcal{M}}) \leq \frac{\mu}{64}$ *holds, with choice* $\lambda = \lambda_{opt} \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\boldsymbol{w}^*))$ *in* (6)*, the statistical error of* $\widehat{\boldsymbol{w}}$ *can be bounded by*

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\| \leq 3\epsilon_{stat} \quad \text{and} \quad \mathcal{R}(\widehat{\boldsymbol{w}} - \boldsymbol{w}^*) \leq 12\Psi(\bar{\mathcal{M}})\epsilon_{stat},$$

where $\epsilon_{stat} = \Psi(\bar{\mathcal{M}})\lambda_{opt}/\mu$.

Algorithm 2 C-FedDA($\boldsymbol{w}_0$, R, E, ϵ_0, μ, L, λ)

Input: Initial point $\boldsymbol{w}_0$, iteration number T, constants $(\epsilon_0, \gamma_r, \mu)$ and synchronized set $\mathcal{I} = \{t_r : 1 \le r \le R\}$.
Initialize: $\widetilde{\boldsymbol{w}}_0 = \bar{\boldsymbol{w}}_0 = \boldsymbol{w}_0$, $\alpha_r = r + 1$.
for Round $r = 0$ **to** R **do**
  for Client $k = 1$ **to** K **do**
    for $t = t_r$ **to** $t_{r+1} - 1$ **do**
      Query $\boldsymbol{G}_t^k = \nabla f(\boldsymbol{w}_t^k; \xi_t^k)$ for for $\xi_t^k \sim \mathcal{D}_k$.
      Update $\boldsymbol{g}_t^k = \boldsymbol{g}_{t-1}^k + \alpha_r \boldsymbol{G}_t^k$.
      if $t < t_{r+1} - 1$ **then**
        $\boldsymbol{w}_t^k = \mathrm{CProx}_r(\boldsymbol{g}_t^k - \mu E \widetilde{\boldsymbol{w}}_r/2; \boldsymbol{w}_0, \epsilon_0, \lambda)$.
      else
        Send $\boldsymbol{g}_{t_{r+1}-1}^k$ to the server.
      end if
    end for
  end for
  Server aggregates: $\boldsymbol{g}_{t_{r+1}-1} = \sum_{k=1}^K \pi_k \boldsymbol{g}_{t_{r+1}-1}^k$.
  Server updates:

$$\bar{\boldsymbol{w}}_{r+1} = \mathrm{CProx}_r(\boldsymbol{g}_{t_{r+1}-1} - \mu E \widetilde{\boldsymbol{w}}_r/2; \boldsymbol{w}_0, \epsilon_0, \lambda)$$

  and $\widetilde{\boldsymbol{w}}_{r+1} = \widetilde{\boldsymbol{w}}_r + \alpha_{r+1} \bar{\boldsymbol{w}}_{r+1}$.
  Synchronization: $\boldsymbol{g}_{t_{r+1}-1}^k \leftarrow \boldsymbol{g}_{t_{r+1}-1}$.
end for

3.3. Constrained Federated Dual Averaging

In light of Proposition 3.1, we aim to estimate the ground-truth $\boldsymbol{w}^*$ defined in (5) by solving the following composite problem:

$$\widehat{\boldsymbol{w}}_{\text{opt}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \{\mathcal{L}(\boldsymbol{w}) + \lambda_{\text{opt}} \mathcal{R}(\boldsymbol{w})\} \tag{8}$$

where $\mathcal{L}(\boldsymbol{w}) = \sum_{k=1}^K \pi_k \mathcal{L}_k(\boldsymbol{w})$ and $\lambda_{\text{opt}} \ge 2\mathcal{R}^*(\nabla \mathcal{L}(\boldsymbol{w}^*))$. Let $\widehat{\boldsymbol{w}}_{\text{Fed}}$ be the output of a federated algorithm, and we hope $\|\widehat{\boldsymbol{w}}_{\text{Fed}} - \boldsymbol{w}^*\|^2$ can attain the optimal statistical precision ϵ_{stat} with iteration complexity $\mathcal{O}(\bar{\sigma}^2/(\mu\epsilon_{\text{stat}}))$. Similar to Woodworth et al. (2020b); Yuan et al. (2021), Fast-FedDA also has a drawback. To guarantee the fast convergence rate, the final estimator of Algorithm 1 takes the weighted average of all iterations. However, we cannot obtain this estimator in the FL setting, since the server only has access to the solution $\boldsymbol{w}_t$ for $t \in \mathcal{I}$. To address this issue, we first propose a new algorithm named *Constrained Federated Dual Averaging* (C-FedDA) in Algorithm 2 in subsection 3.3. In addition, to achieve optimal statistical recovery guarantees, we propose another algorithm named *Multi-stage Constrained Federated Dual Averaging* (MC-FedDA) in Algorithm 3 in subsection 3.4, which calls Algorithm 2 as a subroutine. We provide convergence rate for Algorithm 2 and statistical recovery results for Algorithm 3, both in high probability.

For the ease of representation, we define a *constrained proximal operator* $\mathrm{CProx}_r(\boldsymbol{z}; \boldsymbol{w}_0, \epsilon_0, \lambda)$ for $r \ge 0$ as the solution of the following constrained problem:

$$\min_{\boldsymbol{w} \in \mathcal{W}(\epsilon_0; \boldsymbol{w}_0)} \Big\{ \langle \boldsymbol{w}, \boldsymbol{z} - \gamma_r E \boldsymbol{w}_0 \rangle + \left(\frac{\mu A_r}{2} + \gamma_r \right) \frac{E\|\boldsymbol{w}\|^2}{2} + A_r E \lambda \mathcal{R}(\boldsymbol{w}) \Big\},$$

where $\mathcal{W}(\epsilon_0; \boldsymbol{w}_0) := \{\boldsymbol{w} \in \mathcal{W} \mid \mathcal{R}(\boldsymbol{w} - \boldsymbol{w}_0) \le \epsilon_0\}$. Let $\widetilde{\boldsymbol{w}}_r = \sum_{j=0}^r \alpha_j \bar{\boldsymbol{w}}_j$ be the sum of past solutions obtained on the *server*. In the r-th round, each client updates the weighted cumulative gradient as $\boldsymbol{g}_t^k = \boldsymbol{g}_{t-1}^k + \alpha_r \boldsymbol{G}_t^k$ and updates the local solution by

$$\boldsymbol{w}_{t+1}^k = \mathrm{CProx}_r \left(\boldsymbol{g}_t^k - \frac{\mu E}{2} \widetilde{\boldsymbol{w}}_r; \boldsymbol{w}_0, \epsilon_0, \lambda \right),$$

for $t_r \le t \le t_{r+1} - 1$. At the end of the r-th round, the server updates the global solution by

$$\bar{\boldsymbol{w}}_{r+1} = \mathrm{CProx}_r \left(\sum_{k=1}^K \pi_k \boldsymbol{g}_{t_{r+1}-1}^k - \frac{\mu E}{2} \widetilde{\boldsymbol{w}}_r; \boldsymbol{w}_0, \epsilon_0, \lambda \right),$$

and the weighted cumulative variable $\widetilde{\boldsymbol{w}}_{r+1} = \widetilde{\boldsymbol{w}}_r + \alpha_{r+1} \bar{\boldsymbol{w}}_{r+1}$. The details of C-FedDA is stated in Algorithm 2, which can output a weighted estimator $\widehat{\boldsymbol{w}}_{\text{Fed}} = \sum_{r=0}^R \alpha_r \bar{\boldsymbol{w}}_{r+1}/A_R$ with provable convergence rate. To cope with the RSC and RSM conditions, we introduce the following light-tailed condition to perform high-probability analysis (Duchi et al., 2012; Chen et al., 2012; Lan, 2012).

Assumption 8. The stochastic gradient sampled from the local empirical distribution $\mathcal{D}_k$ satisfies: for any $\boldsymbol{w} \in \mathcal{W}$, it holds that $\mathbb{E}_{\xi \sim \mathcal{D}_k}[\nabla f(\boldsymbol{w}; \xi)] = \nabla \mathcal{L}_k(\boldsymbol{w})$ and

$$\mathbb{E}_{\xi \sim \mathcal{D}_k} \left[\exp\left(\|\nabla f(\boldsymbol{w}; \xi) - \nabla \mathcal{L}_k(\boldsymbol{w})\|^2 / \sigma^2 \right) \right] \le 1.$$

The following theorem provides a *high probability* convergence result for C-FedDA in Algorithm 2. The proof of Theorem 3.1 can be found in Appendix B.3.

Theorem 3.1. *Under Assumptions 3 and 5- 8, we assume the initial point satisfies $\mathcal{R}(\boldsymbol{w}_0 - \widehat{\boldsymbol{w}}) \le \epsilon_0$ and $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^p : \|\boldsymbol{w}\| \le \rho\}$ for $\rho > 0$. By choosing $\gamma_r = (L + \mu)\alpha_r$ and $\alpha_r = r + 1$ in Algorithm 2, with probability at least $1 - \delta$, the output $\widehat{\boldsymbol{w}}_{\text{C-FedDA}} = \sum_{r=0}^R \alpha_r \bar{\boldsymbol{w}}_{r+1}/A_R$ satisfies that*

$$\phi(\widehat{\boldsymbol{w}}_{\text{C-FedDA}}) - \phi(\widehat{\boldsymbol{w}}_{opt}) \lesssim \frac{LE\epsilon_0^2}{T} + \frac{\bar{\sigma}^2}{2\mu T} + \frac{\bar{\sigma}\epsilon_0 \sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{L \log T}{\mu^2 T^2} \left(E\sigma^2 \log(1/\delta) + E^2 (H + \Lambda\rho)^2 \right) + (\tau + \nu)\epsilon_0^2 \tag{9}$$

where $\bar{\sigma}^2 = \sum_{k=1}^K \pi_k^2 \sigma^2$.

Remark 3.2. For the R.H.S. of (9), there are 6 terms. The 1st and 2nd terms come from the parallel dual averaging, the 3rd term comes from concentration inequality, the 4th and

Algorithm 3 Multi-stage C-FedDA

Input: Initial point $\widehat{\boldsymbol{w}}_0$, number of stages M, R_m and E_m for $m \in [M]$ and initial regularization parameter λ_0.
for Stage $m = 0$ **to** $M-1$ **do**
Update: $\lambda_m = 2^{-m}\lambda_0$ and $\epsilon_m = 108\Psi^2(\bar{\mathcal{M}})\lambda_m/\mu$.
Update estimator by calling C-FedDA

$$\widehat{\boldsymbol{w}}_{m+1} = \text{C-FedDA}(\widehat{\boldsymbol{w}}_m, R_m, E_m, \epsilon_m, \mu, L, \lambda_m).$$

end for

5th terms are due to skipped communication, and the 6-th term $(\tau+\nu)\epsilon_0^2$ in (9) incurs an additional error regarding the tolerances in RSC and RSM conditions. The 2nd term $\bar{\sigma}\sqrt{\epsilon_0 \log(1/\delta)/T}$ is the best-known high probability rate of centralized dual averaging (Xiao, 2010; Lan, 2012; Chen et al., 2012). By choosing $E = \tilde{\mathcal{O}}(\bar{\sigma}T^{1/2})$ for Algorithm 2, the discrepancy (the 4th and 5th term) from local updates will be dominated by the concentration bound (the 3rd term).

3.4. Multi-stage Constrained Federated Dual Averaging

To reduce the error brought from the RSC and RSM conditions, the first attempt is solving (8) by directly using shrinking domain technique (Iouditski & Nesterov, 2014; Hazan & Kale, 2011; Lan, 2012; Liu et al., 2018) according to $\mathcal{R}(\cdot)$-norm. In each stage, we use the output of the previous stage as the initial point and shrink the radius of the $\mathcal{R}(\cdot)$-norm ball in C-FedDA. In particular, we need to guarantee that $\mathcal{R}^2(\widehat{\boldsymbol{w}}_m - \widehat{\boldsymbol{w}}_{\text{opt}})$ is also reduced with high probability through controlling $(\phi(\widehat{\boldsymbol{w}}_m) - \phi(\widehat{\boldsymbol{w}}_{\text{opt}}))/\lambda_{\text{opt}}$ at the m-th stage (see Lemma B.8), since we need to make sure that $\widehat{\boldsymbol{w}}_{\text{opt}}$ always lies into the ball with high probability. However, it can be only decreased up to $(\tau+\nu)\mathcal{R}^2(\widehat{\boldsymbol{w}}_{m-1} - \widehat{\boldsymbol{w}}_{\text{opt}})/\lambda_{\text{opt}}$ according to the last term in (9), which could be very large since λ_{opt} is usually very small. This indicates that we cannot directly employ shrinking domain technique for solving (8).

To address this issue, our solution is motivated by the homotopy continuation strategy (Xiao & Zhang, 2013; Wang et al., 2014): we select a decreasing sequence of the regularization parameter [2] $\lambda_m = \lambda_0 \cdot 2^{-m}$, where $\lambda_{\text{opt}} < \lambda_0$ and $\lambda_M = \lambda_{\text{opt}}$. At the m-th stage, we call Algorithm 2 to solve the following subproblem

$$\min_{\boldsymbol{w} \in \mathcal{W}(\widehat{\boldsymbol{w}}_{m-1}, r_m)} \{\mathcal{L}(\boldsymbol{w}) + \lambda_m \mathcal{R}(\boldsymbol{w})\},$$

where $\widehat{\boldsymbol{w}}_{m-1}$ is the output of previous stage and r_m is the current radius. By shrinking both the radius and regularization parameter in each stage, a final estimator with optimal statistical precision can be obtained. We present the detailed procedure of *Multi-stage Constrained Federated Dual Averaging* (MC-FedDA) in Algorithm 3.

[2] Here we set $1/2$ as the contraction rate for technique convenience. In practice, we may choose more flexible non-increasing sequence λ_m.

3.5. Main Results for MC-FedDA

In this subsection, we present the statistical recovery results of the algorithm MC-FedDA. The proof of Theorem 3.2 is deferred to Appendix B.4.

Assumption 9. There exists some constant $C > 0$, such that the averaged RSC and RSM coefficients satisfy $C(\tau+\nu)\Psi^2(\bar{\mathcal{M}}) \leq \mu$.

Theorem 3.2. *Under the same conditions in Theorem 3.1 and Assumption 9. We assume the initial point satisfies $\mathcal{R}(\widehat{\boldsymbol{w}}_0 - \widehat{\boldsymbol{w}}_{opt}) \leq 84\Psi^2(\bar{\mathcal{M}})\lambda_0/\mu$ and choose E_m and R_m such that $E_m^2 \lesssim \bar{\sigma}^2 T_m/\log(T_m+1)$ for $T_m = E_m R_m$. When Algorithm 3 terminates ($M = \log_2(\lambda_0/\lambda_{opt})+1$), with probability[3] at least $1-\delta$, the total number of iterations $T = \sum_{m=0}^{M} T_m$ is no more than (up to a constant factor)*

$$\frac{4\bar{\sigma}^2(\log_2(\lambda_0/\lambda_{opt})+1)}{\Psi^2(\bar{\mathcal{M}})\lambda_{opt}^2} \log\left(\frac{\log_2(\lambda_0/\lambda_{opt})+1}{\delta}\right). \tag{10}$$

Let $\widehat{\boldsymbol{w}}_{\text{MC-FedDA}} = \widehat{\boldsymbol{w}}_M$ from Algorithm 3, we can guarantee that

$$\phi(\widehat{\boldsymbol{w}}_{\text{MC-FedDA}}) - \phi(\widehat{\boldsymbol{w}}_{opt}) \leq \frac{\Psi^2(\bar{\mathcal{M}})\lambda_{opt}^2}{\mu}.$$

In addition, the estimation error can be bounded by

$$\|\widehat{\boldsymbol{w}}_{\text{MC-FedDA}} - \boldsymbol{w}^*\| \leq \frac{4\Psi(\bar{\mathcal{M}})\lambda_{opt}}{\mu}.$$

Remark 3.3. Notice that $\epsilon_{\text{stat}} = \Psi(\bar{\mathcal{M}})\lambda_{\text{opt}}/\mu$ converges to 0 as the total sample size N tends to infinity. Let $\epsilon = \mu\epsilon_{\text{stat}}^2$, if the total number of iterations satisfies $T = \sum_{m=0}^{M-1} T_m = \tilde{\mathcal{O}}(\bar{\sigma}^2/(\mu\epsilon))$, then we are guaranteed that $\phi(\widehat{\boldsymbol{w}}_{\text{MC-FedDA}}) - \phi(\widehat{\boldsymbol{w}}_{\text{opt}}) \leq \epsilon$. Up to some logarithmic factors, this is equivalent to the linear speedup convergence rate $\sigma^2/(\mu KT)$ for the equal-weighted case ($\bar{\sigma}^2 = \sigma^2/K$). Moreover, with the choice $E_m = \tilde{\mathcal{O}}(T_m^{1/2}/K^{1/2})$, the total communication complexity is bounded by $\sum_{m=0}^{M-1} T_m^{1/2}K^{1/2} = \tilde{\mathcal{O}}(T^{1/2}K^{1/2})$. In fact, the total complexity is mainly due to the complexity of the last stage.

Next, we illustrate the implications of Theorem 3.2 through Example 3.1 and 3.2 in subsection 3.1.

Sparse Linear Regression. Under some regular conditions, the RSC and RSM coefficients in each client are given by $\tau_k = c\log p/n_k$ and $\nu_k = c\log p/n_k$ for some absolute constant c (see Agarwal et al. (2012); Loh & Wainwright (2015)). With the weight choice $\pi_k = \frac{n_k}{N}$, we

[3] The randomness is from the empirical distribution $\mathcal{D} = \{\mathcal{D}_k : k = 1, ..., K\}$.

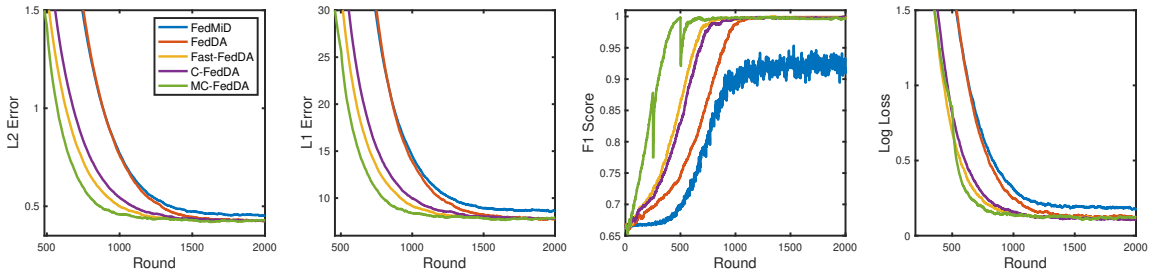

Figure 1. Recovery results for federated sparse linear regression problem with $p = 1024$ and $s = 512$. Except `FedMiD`, other methods nearly achieve perfect support recovery. Our proposed three algorithms show faster numerical convergence in four metrics.

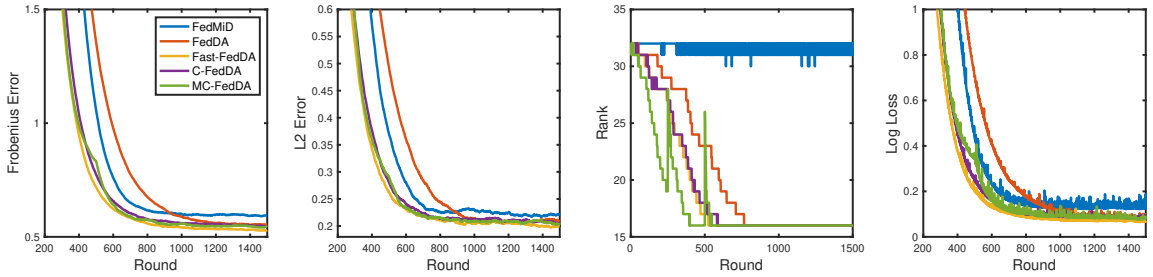

Figure 2. Recovery results for federated low-rank matrix estimation problem with $p_1 = p_2 = 32$ and $r^* = 16$. Except `FedMiD`, other methods all recover the true rank of $\boldsymbol{W}^*$. Our proposed three methods also show faster convergence than other two baselines.

have $\tau = \nu = cK \log p/N$. If the total sample size N and the number of clients K satisfies $sK \lesssim N$, then Assumption 9 will be satisfied. According to Proposition 3.1, we need to choose the regularization parameter such that $\lambda_{\text{opt}} \geq c\sqrt{\log p/N}$ to guarantee the optimal statistical convergence rate $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\| = \mathcal{O}(\sqrt{s \log p/N})$ with high probability (Raskutti et al., 2009; Ye & Zhang, 2010). Therefore, to attain the optimal statistical convergence rate, the iteration complexity in Theorem 3.2 is given by $\tilde{\mathcal{O}}(N/(sK))$.

Low-Rank Matrix Estimation. In this case, the subspace Lipschitz constant is $\Psi(\bar{\mathcal{M}}) = \sqrt{r^*}$. Under some regular conditions, the averaged RSC and RSM coefficients are both $\tau = \nu = c(p_1 \vee p_2)K/N$ (Agarwal et al., 2012; Wainwright, 2019). If the total sample size N and the number of clients K satisfies $r^*K(p_1 \vee p_2) \lesssim N$, then Assumption 9 will be satisfied. To achieve optimal statistical convergence rate $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_{\text{F}} = \mathcal{O}(\sqrt{r^*(p_1 \vee p_2)\log(p_1 \vee p_2)/N})$ with high probability (Koltchinskii et al., 2011), we choose the regularization parameter as $\lambda \geq c\sqrt{(p_1 \vee p_2)\log(p_1 \vee p_2)/N}$. Thus the iteration complexity in Theorem 3.2 will be $\tilde{\mathcal{O}}(N/(r^*(p_1 \vee p_2)K))$.

4. Numerical Experiments

In this section, we investigate the empirical performance of our proposed method with four experiments: two with synthetic data and two with real world data. For Example 3.1 and 3.2, we generate heterogeneous synthetic data for 64 clients, and each client containing 128 independent samples. For federated sparse logistic regression, we use the Federated EMNIST (Caldas et al., 2019) dataset of handwritten letters and digits. We compare our proposed three algorithms `Fast-FedDA`, `C-FedDA` and `MC-FedDA` with Federated Mirror Descent (`FedMiD`) and Federated Dual Averaging (`FedDA`) algorithms introduced in Yuan et al. (2021). The detailed parameter tuning of the experiments in this section is provided in Appendix F.

Federated Sparse Linear Regression. In this experiment, we conduct experiments to Example 3.1 on synthetic data. The true sparse regression coefficient is $\boldsymbol{w}^* = (\mathbf{1}_s^\top \mathbf{0}_{p-s}^\top)^\top$. In the k-th client, we first generate a heterogeneity vector $\boldsymbol{\delta}_k$ from $N(\mathbf{0}, \boldsymbol{I}_{p\times p})$. The covariate is generated according to $\mathbf{x}_i^k = \boldsymbol{\delta}_k + \mathbf{z}_i^k$ for $i = 1, 2, ..., n_k$, where $\mathbf{z}_i^k$ is independently sampled from $N(\mathbf{0}, \boldsymbol{\Sigma})$. The (i, j)-th element of the covariance matrix $\boldsymbol{\Sigma}$ is given by $\sigma_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. Then the response variable y_i^k is generated accordingly. At each round, we sample 10 clients to conduct local updates and the number of local updates is $K = 10$. In this experiment, the batch size is 10 and the regularization parameter is $\lambda = 0.5^5$. To evaluate the performance of different methods, we record ℓ_2 error, ℓ_1 error, F_1 score of support recovery and training loss after each round and results are reported in Figure 1.

Federated Low-Rank Matrix Estimation. In this subsection, we conduct experiments to Example 3.2 on syn-

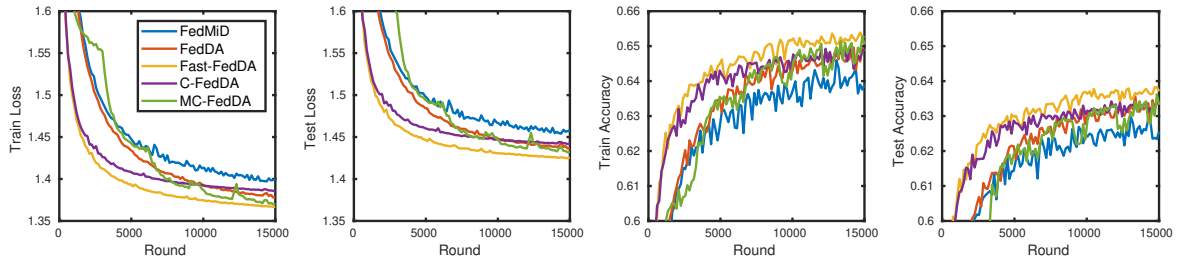

Figure 3. Results for federated sparse logistic regression on EMNIST-62 dataset. Our proposed algorithms `Fast-FedDA` and `MC-FedDA` reach a lower loss and higher accuracy than the two baselines from (Yuan et al., 2021), and `Fast-FedDA` exhibits faster convergence.

thetic data. The p by p true low-rank matrix is given by $\boldsymbol{W}^* = \text{diag}(\mathbf{1}_{r^*}, \mathbf{0}_{p-r^*})$. At the k-th client, we first construct a heterogeneity matrix $\mathbf{Z}_k \in \mathbb{R}^{p\times p}$ with each entry independently sampled from $N(0,1)$. Then we generate the covariate matrix by $\mathbf{X}_i^k = \mathbf{Z}_k + \mathbf{A}_i^k$, where each entry of $\mathbf{A}_i^k$ is also independently sampled from $N(0,1)$. As with the previous experiment, 10 clients are sampled each round to conduct local updates and $K = 10$. In this experiment, the batch size is also 10 and the regularization parameter is $\lambda = 0.1$. We choose estimation error in Frobenius norm, ℓ_2 norm (operator norm), recovery rank, and training loss to evaluate performances of different algorithms. The results are plotted in Figure 2.

From the results in Figure 1 and 2, we can see that our proposed three algorithms show faster convergence than `FedDA` and `FedMiD`, which is consisting with our linear speedup results. Except `FedMiD`, other methods nearly achieve perfect support recovery. As we expected, the evaluation metrics of `MC-FedDA` converge to the same values with other algorithms. It is worthwhile noting that F_1 score of `MC-FedDA` already converges to 1 after the first stage. The reason is that the regularization parameter is larger, which tends to output more sparse solution.

Federated Sparse Logistic Regression. We also provide experimental results on real world data, namely the Federated EMNIST dataset (Caldas et al., 2019). This dataset is a modification of the EMNIST dataset (Cohen et al., 2017) for the federated setting, in which each client's dataset consists of all characters written by a single author. In this way, the data distribution differs across clients. The complete dataset contains 800K examples across 3500 clients. We train a multi-class logisitc regression model on two versions of this dataset: EMNIST-10 (digits only, 10 classes), and EMNIST-62 (all alphanumeric characters, 62 classes). Following (Yuan et al., 2021), we use only 10% of the samples, which is sufficient to train a logistic regression model. Our subsampled EMNIST-10 dataset consists of 367 clients with an average of 99 examples each, while EMNIST-62 consists of 379 clients with an average of 194 examples each. For both experiments, we use a batch size of 25, a regularization parameter $\lambda = 10^{-4}$, and we sample 36 clients to perform local updates at each communication round. For EMNIST-10, each sampled client performs $K = 40$ updates per communication round for $R = 15000$ rounds. For EMNIST-62, $K = 10$ and $R = 75000$. Comparisons of algorithms for EMNIST-62 and EMNIST-10 are shown in Figure 3 and Figure 4 (Appendix F). We can see that our algorithms (`Fast-FedDA`, `C-FedDA`) outperforms baselines (`FedDA` and `FedMid`) in terms of convergence speed on both training and test performance.

5. Conclusion

This paper investigates the composite optimization and statistical recovery problem in FL. For the composite optimization problem, we proposed a fast dual averaging algorithm (`Fast-FedDA`), in which we prove linear speedup for strongly convex loss. For statistical recovery, we proposed a multi-stage constrained dual averaging algorithm (`MC-FedDA`). Under restricted strongly convex and smooth assumption, we provided a high probability iteration complexity to attain optimal statistical precision, equivalent to the linear speedup result for strongly convex case. Several experiments on synthetic and real data are conducted to verify the superior performance of our proposed algorithms over other baselines.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. Michael Crawshaw and Mingrui Liu are supported in part by a grant from George Mason University. Shan Luo's research is supported by National Program on Key Basic Research Project, NSFC Grant No. 12031005 (PI: Qi-Man Shao, Southern University of Science and Technology). Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu). The work of Yajie Bao was done when he was virtually visiting Mingrui Liu's research group in the Department of Computer Science at George Mason University.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, pp. 2452–2482, 2012.

Bao, Y. and Xiong, W. One-round communication efficient distributed M-estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 46–54. PMLR, 2021.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pp. 14668–14679, 2019.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352—-1382, 2018.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings. In *NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Chen, X., Lin, Q., and Pena, J. Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems*, pp. 395–403. Citeseer, 2012.

Chen, X., Liu, W., Mao, X., and Yang, Z. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21(182):1–43, 2020.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32:11082–11094, 2019.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Conference on Learning Theory*, pp. 421–436. JMLR Workshop and Conference Proceedings, 2011.

He, X. and Shao, Q.-M. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

Iouditski, A. and Nesterov, Y. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

Jordan, M. I., Lee, J. D., and Yang, Y. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020a.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143, 2020b.

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

Li, X., Liang, J., Chang, X., and Zhang, Z. Statistical estimation and inference via local SGD in federated learning. *arXiv preprint arXiv:2109.01326*, 2021.

Liu, B., Yuan, X.-T., Wang, L., Liu, Q., Huang, J., and Metaxas, D. N. Distributed inexact newton-type pursuit for non-convex sparse learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 343–352. PMLR, 2019.

Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic AUC maximization with $O(1/n)$-convergence rate. In *International Conference on Machine Learning*, pp. 3189–3197. PMLR, 2018.

Loh, P.-L. and Wainwright, M. J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16 (1):559–616, 2015.

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

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of convergence for high-dimensional regression under ℓ_q-ball sparsity. In *47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 251–257. IEEE, 2009.

Shamir, O., Srebro, N., and Zhang, T. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008. PMLR, 2014.

Spiridonoff, A., Olshevsky, A., and Paschalidis, I. C. Communication-efficient SGD: From local SGD to one-shot averaging. *arXiv preprint arXiv:2106.04759*, 2021.

Stich, S. U. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. FedDR–Randomized Douglas-Rachford splitting algorithms for nonconvex federated composite optimization. *arXiv preprint arXiv:2103.03452*, 2021.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization (Submitted)*, 2008.

Tu, J., Liu, W., and Mao, X. Byzantine-robust distributed sparse learning for M-estimation. *Machine Learning*, pp. 1–32, 2021.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pp. 3636–3645. PMLR, 2017.

Wang, Z., Liu, H., and Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of statistics*, 42(6):2164, 2014.

Woodworth, B., Patel, K. K., and Srebro, N. Minibatch vs local SGD for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*, 2020a.

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

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.

Xiao, L. and Zhang, T. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Ye, F. and Zhang, C.-H. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. *arXiv preprint arXiv:1905.03817*, 2019a.

Yu, H., Yang, S., and Zhu, S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019b.

Yuan, H. and Ma, T. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33, 2020.

Yuan, H., Zaheer, M., and Reddi, S. Federated composite optimization. In *International Conference on Machine Learning*, pp. 12253–12266. PMLR, 2021.

Zhu, X., Li, F., and Wang, H. Least squares approximation for a distributed system. *Journal of Computational and Graphical Statistics (Accepted)*, 2021.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

A. Related Work

Federated Learning. As an active research area, a tremendous amount of research has been devoted to investigating the theory and application of FL. The most popular algorithm in FL is the so-called Federated Averaging (`FedAvg`) proposed by McMahan et al. (2017). For strongly convex problems, Stich (2019) provided the first convergence analysis of `FedAvg` in a homogeneous environment and showed that the communication rounds can be reduced up to a factor of $\mathcal{O}(\sqrt{TK})$ without affecting linear speedup. Then Li et al. (2019; 2020) investigated the convergence rate in a heterogeneous environment. Karimireddy et al. (2020b) introduced a stochastic controlled averaging for FL to learn from hetergeneous data. Stich & Karimireddy (2019); Khaled et al. (2020) improved the analysis and showed $\mathcal{O}(K\text{poly}\log(T))$ rounds is sufficient to achieve linear speedup. Recently, Yuan & Ma (2020) proposed an accelerated FedAvg algorithm, which requires $\mathcal{O}(K^{1/3}\text{poly}\log(T))$ to attain linear speedup. Recently, Li et al. (2021) investigated the statistical estimation and inference problem for local SGD in FL. However, Li et al. (2021) focused on the unconstrained smooth statistical optimization, but we considered a different problem with non-smooth regularizer aiming to recover the sparse/low-rank structure of ground-truth model. For the strongly convex finite-sum problem, Mitra et al. (2021) proposed an algorithm named `FedLin` based on the variance reduction technique and obtained the linear convergence rate. However, their analysis and algorithm are not applicable in the composite setting, and they do not consider statistical recovery at all. A more recent work Spiridonoff et al. (2021) showed that the number of rounds can be independent of T under homogeneous setting. Recently, there is a line of work focusing on analyzing nonconvex problems in FL (Yu et al., 2019b;a; Basu et al., 2019; Haddadpour et al., 2019). This list is by no means complete due to the vast amount of literature in FL. For a more comprehensive survey, please refer to (Kairouz et al., 2019) and reference therein.

Distributed Statistical Recovery. With the increasing data size, statistical recovery in the distributed environment is a hot topic in recent years. These works focus on the homogeneous setting. Lee et al. (2017) proposed an one-shot debiasing method and required each client solve a composite problem using its own data. Other one-shot methods for different tasks can be found in Battey et al. (2018); Bao & Xiong (2021); Zhu et al. (2021). Motivated by the approximated Newton's method (Shamir et al., 2014), Wang et al. (2017) proposed a multi-round algorithm, where each client only needs to compute gradients and the server solves a shifted ℓ_1 penalized problem. Meanwhile, Jordan et al. (2018) developed Communication-efficient Surrogate Loss (CSL) framework for more general ℓ_1-penalized problems. A series of statistical recovery problems based on CSL scheme has also been studied (Liu et al., 2019; Chen et al., 2020; Tu et al., 2021).

B. Proof of Main Results

B.1. Concentration Inequalities for Martingale Differences

Let $\{\boldsymbol{\xi}_i \in \mathbb{R}^p\}_{i=1}^{\infty}$ be a sequence of martingale differences with respective to the filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$. It satisfies that $\mathbb{E}[\boldsymbol{\xi}_i|\mathcal{F}_i] = \mathbf{0}$ and the light-tail condition $\mathbb{E}_{\mathcal{D}}[\exp(\|\boldsymbol{\xi}_i\|^2/\sigma^2)|\mathcal{F}_i] \leq 1$ for some $\sigma > 0$. Under this condition, it follows from Jensen's inequality that

$$\exp(\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{\xi}_i\|^2|\mathcal{F}_i]/\sigma^2) \leq \mathbb{E}_{\mathcal{D}}[\exp(\|\boldsymbol{\xi}_i\|^2/\sigma^2)|\mathcal{F}_i] \leq 1.$$

Hence we have $\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{\xi}_i\|^2|\mathcal{F}_i] \leq \sigma^2$.

The following three lemmas are used throughout in our proof, we defer the proof of Lemma B.3 to Section E.

Lemma B.1 (Lemma 5 in (Duchi et al., 2012))**.** *Under the assumption of Theorem 3.1, for any positive and non-decreasing sequence $\{a_t\}_{t=0}^{\infty}$, we have*

$$\sum_{t=0}^{T}\frac{\|\boldsymbol{\xi}_t\|^2}{a_t} \geq \sum_{t=0}^{T}\frac{\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{\xi}_t\|^2]}{a_t} + \max\left\{\frac{8\sigma^2\log(1/\delta)}{a_0}, 16\sigma^2\sqrt{\sum_{t=0}^{T}\frac{\log(1/\delta)}{a_i^2}}\right\}$$

holds with probability at most $\delta \in (0,1)$.

Lemma B.2 (Lemma 6 in (Lan, 2012))**.** *Under the assumption of Theorem 3.1, for any sequence $\{\boldsymbol{w}_t\}_{t=0}^{\infty}$ such that $\boldsymbol{z}_t$ is $\mathcal{F}_{t-1}$-measurable, we have*

$$\sum_{t=0}^{T}\langle \boldsymbol{z}_t, \boldsymbol{\xi}_t\rangle \geq \sqrt{3\log(1/\delta)}\left(\sum_{t=0}^{T}\|\boldsymbol{z}_t\|^2\right)^{1/2}$$

holds with probability at most $\delta \in (0,1)$.

The following lemma is a martingale's version of Lemma 3.1 in He & Shao (2000), and the proof is deferred to Section E.

Lemma B.3. *If $\mathbb{E}[\|\boldsymbol{\xi}_i\|^2|\mathcal{F}_i] < \infty$, then for any $x > 0$ it holds that,*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{t}\boldsymbol{\xi}_i\right\| \geq x\left(B_t + \left(\sum_{i=1}^{t}\|\boldsymbol{\xi}_i\|^2\right)^{1/2}\right)\right\} \leq 8\exp(-x^2/8), \tag{11}$$

where $B_t = (\sum_{i=1}^{t}\mathbb{E}(\|\boldsymbol{\xi}_i\|^2))^{1/2}$.

From now on, we use $\mathcal{F}_t$ to denote the σ-algebra generated by prior sequence $\{\boldsymbol{w}_i^k : 0 \leq i \leq t,\ 1 \leq k \leq K\}$.

B.2. Proof of Theorem 2.1

Let $\boldsymbol{g}_t = \sum_{k=1}^{K}\pi_k\boldsymbol{g}_t^k$, $\check{\boldsymbol{w}}_t = \sum_{k=1}^{K}\pi_k\boldsymbol{w}_t^k$ and $\widetilde{\boldsymbol{w}}_t = \sum_{k=1}^{K}\pi_k\widetilde{\boldsymbol{w}}_t^k = \sum_{i=0}^{t}\alpha_i\check{\boldsymbol{w}}_i$, we define a virtual sequence:

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}\in\mathcal{W}}\left\{\langle\boldsymbol{w}, \boldsymbol{g}_t - \frac{\mu}{2}\widetilde{\boldsymbol{w}}_t - \gamma_t\boldsymbol{w}_0\rangle + \left(\frac{A_t\mu}{2} + \gamma_t\right)\frac{\|\boldsymbol{w}\|^2}{2} + A_t h(\boldsymbol{w})\right\}. \tag{12}$$

which can be also equivalently written as

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}\in\mathcal{W}}\left\{\langle\boldsymbol{w}, \boldsymbol{g}_t\rangle + \frac{\mu}{4}\sum_{i=0}^{t}\alpha_i\|\boldsymbol{w} - \check{\boldsymbol{w}}_i\|^2 + \frac{\gamma_t}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + A_t h(\boldsymbol{w})\right\}. \tag{13}$$

According to Algorithm 1, $\boldsymbol{w}_{t+1}$ is exactly the solution updated by the server for $t+1 \in \mathcal{I}$. Next we define a pseudo distance between $\boldsymbol{w}$ and $\boldsymbol{w}'$ at the t-th step as

$$\begin{aligned} D_t(\boldsymbol{w};\boldsymbol{w}') = \langle\boldsymbol{w} - \boldsymbol{w}', \boldsymbol{g}_{t-1}\rangle + \frac{\mu}{4}\sum_{i=0}^{t-1}\alpha_i\left(\|\boldsymbol{w} - \check{\boldsymbol{w}}_i\|^2 - \|\boldsymbol{w}' - \check{\boldsymbol{w}}_i\|^2\right) \\ + \frac{\gamma_t}{2}\left(\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 - \|\boldsymbol{w}' - \boldsymbol{w}_0\|^2\right) + A_{t-1}(h(\boldsymbol{w}) - h(\boldsymbol{w}')). \end{aligned} \tag{14}$$

Let $\boldsymbol{g}_{-1} = \boldsymbol{0}$, $A_{-1} = 0$ and $\sum_{i=0}^{-1} = 0$, we have $D_0(\boldsymbol{w};\boldsymbol{w}_0) = \frac{\gamma_0}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2$ for any $\boldsymbol{w} \in \mathcal{W}$. In addition, (13) also implies that $D_t(\boldsymbol{w};\boldsymbol{w}_t) \geq 0$ for any $\boldsymbol{w} \in \mathcal{W}$. The next lemma provide the one-step induction relation of Algorithm 2, which is crucial to the proof of convergence rate. The proof of Lemma B.4 is deferred to Section C.1.

Lemma B.4 (One-Step Induction Relation)**.** *Under the conditions of Theorem 2.1, it holds that*

$$\begin{aligned} \alpha_t[\phi(\boldsymbol{w}_{t+1}) - \phi(\widehat{\boldsymbol{w}})] \leq & D_t(\widehat{\boldsymbol{w}};\boldsymbol{w}_t) - D_{t+1}(\widehat{\boldsymbol{w}};\boldsymbol{w}_{t+1}) + \frac{\gamma_t - \gamma_{t-1}}{2}\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 \\ & + \alpha_t\langle\boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_t\rangle + \frac{\alpha_t^2\|\boldsymbol{\Delta}_t\|^2}{2(A_t\mu + 2\gamma_t - 2L\alpha_t)} \\ & + \alpha_t\left(\frac{\mu}{2}\sum_{k=1}^{K}\pi_k\|\boldsymbol{w}_t^k - \check{\boldsymbol{w}}_t\|^2 + \frac{3L}{2}\sum_{k=1}^{K}\pi_k\|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2\right), \end{aligned} \tag{15}$$

where $\boldsymbol{\Delta}_t = \sum_{k=1}^{K}\pi_k(\nabla f(\boldsymbol{w}_i^k;\xi_i^k) - \nabla\mathcal{L}_k(\boldsymbol{w}_i^k))$.

We impose the following lemma to bound the discrepancy caused by skipped communication, and the proof is deferred to Section D.1.

Lemma B.5. *Under the conditions in Theorem 2.1, we have*

$$\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2|\mathcal{F}_t],\ \mathbb{E}_{\mathcal{D}}[\|\boldsymbol{w}_t^k - \check{\boldsymbol{w}}_t\|^2|\mathcal{F}_t] \leq \frac{4E\sigma^2\alpha_t^2}{(\mu A_t/2 + \gamma_t)^2} + \frac{4E^2(H^2 + \Lambda^2 + \mu^2\rho^2)\alpha_t^2}{(\mu A_t/2 + \gamma_t)^2}.$$

Proof of Theorem 2.1. We first note that $\mathbb{E}[\boldsymbol{\Delta}_t|\mathcal{F}_t] = \boldsymbol{0}$ and

$$\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2] = \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2|\mathcal{F}_t] = \sum_{k=1}^{K}\pi_k^2\mathbb{E}[\|\nabla f(\boldsymbol{w}_i^k;\xi_i^k) - \nabla\mathcal{L}_k(\boldsymbol{w}_i^k)\|^2|\mathcal{F}_t] \leq \sum_{k=1}^{K}\pi_k^2\sigma^2 = \bar{\sigma}^2,$$

where the second equality follows from the independence between different clients. Taking conditional expectation on the both sides of (15) results in

$$\begin{aligned}\alpha_t\mathbb{E}[\phi(\boldsymbol{w}_{t+1})-\phi(\widehat{\boldsymbol{w}})|\mathcal{F}_t] \leq& \mathbb{E}[D_t(\widehat{\boldsymbol{w}};\boldsymbol{w}_t)-D_{t+1}(\widehat{\boldsymbol{w}};\boldsymbol{w}_{t+1})|\mathcal{F}_t]+\frac{\gamma_t-\gamma_{t-1}}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\\ &+\alpha_t\langle\mathbb{E}[\boldsymbol{\Delta}_t|\mathcal{F}_t],\widehat{\boldsymbol{w}}-\boldsymbol{w}_t\rangle+\frac{\alpha_t^2\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2|\mathcal{F}_t]}{2(A_t\mu+2\gamma_t-2L\alpha_t)}\\ &+\alpha_t\left\{\frac{L+\mu}{2}\sum_{k=1}^K\pi_k\mathbb{E}[\|\boldsymbol{w}_t^k-\check{\boldsymbol{w}}_t\|^2|\mathcal{F}_t]+L\sum_{k=1}^K\pi_k\mathbb{E}[\|\boldsymbol{w}_t^k-\boldsymbol{w}_t\|^2|\mathcal{F}_t]\right\}\\ \leq&\mathbb{E}[D_t(\widehat{\boldsymbol{w}};\boldsymbol{w}_t)-D_{t+1}(\widehat{\boldsymbol{w}};\boldsymbol{w}_{t+1})|\mathcal{F}_t]+\frac{\gamma_t-\gamma_{t-1}}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\\ &+\frac{\alpha_t^2\bar{\sigma}^2}{2(A_t\mu+2\gamma_t-2L\alpha_t)}+\frac{3L+\mu}{2}\alpha_t^3\left(\frac{4E\sigma^2}{(\mu A_t/2+\gamma_t)^2}+\frac{4E^2(H^2+\Lambda^2+\mu^2\rho^2)}{(\mu A_t/2+\gamma_t)^2}\right).\end{aligned} \tag{16}$$

In the second inequality of (16), we used Lemma B.5. By substituting $\gamma_t = L\alpha_t$ and $\alpha_t = t+1$, we have

$$\sum_{t=0}^T\frac{\alpha_t^2\bar{\sigma}^2}{2(A_t\mu+2\gamma_t-2L\alpha_t)}=\sum_{t=0}^T\frac{\alpha_t^2\bar{\sigma}^2}{2\mu A_t}=\sum_{t=0}^T\frac{(t+1)^2\bar{\sigma}^2}{\mu(t+1)(t+2)}\leq\frac{(T+1)\bar{\sigma}^2}{\mu},$$

and

$$\sum_{t=0}^T\frac{\alpha_t^3}{(\mu A_t/2+\gamma_t)^2}=\sum_{t=0}^T\frac{(t+1)^3}{(\mu(t+1)(t+2)/4+L(t+1))^2}\leq\frac{16\log(T+1)}{\mu^2}.$$

In addition, it follows from $D_{T+1}(\widehat{\boldsymbol{w}};\boldsymbol{w}_{T+1})\geq 0$ and $D_0(\widehat{\boldsymbol{w}};\boldsymbol{w}_0)=\gamma_0\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|/2$ that

$$\begin{aligned}&\sum_{t=0}^T\left\{D_t(\widehat{\boldsymbol{w}};\boldsymbol{w}_t)-D_{t+1}(\widehat{\boldsymbol{w}};\boldsymbol{w}_{t+1})+\frac{\gamma_t-\gamma_{t-1}}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\right\}\\ =&D_0(\widehat{\boldsymbol{w}};\boldsymbol{w}_0)-D_{T+1}(\widehat{\boldsymbol{w}};\boldsymbol{w}_{T+1})+\frac{\gamma_T-\gamma_{-1}}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\\ \leq&\frac{\gamma_0}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2+\frac{\gamma_T}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\leq\gamma_T\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2.\end{aligned}$$

Let $B=\|\boldsymbol{w}_0-\widehat{\boldsymbol{w}}\|^2$, telescoping (16) from time $t=0$ to $t=T$ gives rise to

$$\begin{aligned}&\frac{1}{A_T}\sum_{t=0}^T\alpha_t\mathbb{E}[\phi(\boldsymbol{w}_{t+1})-\phi(\widehat{\boldsymbol{w}})]\leq\frac{\gamma_T\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2}{A_T}+\frac{1}{A_T}\sum_{t=0}^T\frac{\alpha_t^2\bar{\sigma}^2}{2(A_t\mu+2\gamma_t-2L\alpha_t)}\\ &+\frac{3L+\mu}{2A_T}\sum_{t=0}^T\alpha_t^3\left(\frac{4E\sigma^2}{(\mu A_t/2+\gamma_t)^2}+\frac{4E^2(H^2+\Lambda^2+\mu^2\rho^2)}{(\mu A_t/2+\gamma_t)^2}\right)\\ \leq&\frac{2LB}{T+1}+\frac{2\bar{\sigma}^2}{\mu(T+1)}+32(3L+\mu)\log(T+1)\left(\frac{4E\sigma^2}{\mu^2T(T+1)}+\frac{4E^2(H^2+\Lambda^2+\mu^2\rho^2)}{\mu^2T(T+1)}\right),\end{aligned}$$

Thus the result follows from Jensen's inequality and the convexity of $\phi(\cdot)$. □

B.3. Proof of Theorem 3.1

Similar to the proof of Theorem 3.1, we define the following pseudo distance at the r-th communication step

$$\begin{aligned}D_r(\boldsymbol{w};\boldsymbol{w}')=&\langle\boldsymbol{g}_{t_r-1},\boldsymbol{w}-\boldsymbol{w}'\rangle+\frac{\mu E}{4}\sum_{j=0}^{r-1}\alpha_j(\|\boldsymbol{w}-\bar{\boldsymbol{w}}_j\|^2-\|\boldsymbol{w}'-\bar{\boldsymbol{w}}_j\|^2)\\ &+\frac{\gamma_rE}{2}(\|\boldsymbol{w}-\bar{\boldsymbol{w}}_0\|^2-\|\boldsymbol{w}'-\bar{\boldsymbol{w}}_0\|^2)+EA_{r-1}h(\boldsymbol{w}),\end{aligned} \tag{17}$$

where $\boldsymbol{g}_{t_r-1}=\sum_{k=1}^K\pi_k\boldsymbol{g}_{t_r-1}^k$. Let $\boldsymbol{g}_{-1}=\boldsymbol{0}$ and $\sum_{j=0}^{-1}=0$, then we have $D_0(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_0)=\frac{\gamma_rE}{2}\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_0\|^2\leq\frac{\gamma_r}{2}\epsilon_0$. The following lemma characterizes the one round progress of Algorithm 2, and the proof is deferred to Section C.2.

Lemma B.6 (One-Step Induction Relation)**.** *Under the conditions of Theorem 3.1, we have*

$$
\begin{aligned}
E\alpha_r[\phi(\bar{\boldsymbol{w}}_{r+1}) - \phi(\widehat{\boldsymbol{w}})] \leq\ & D_r(\widehat{\boldsymbol{w}}; \bar{\boldsymbol{w}}_r) - D_{r+1}(\widehat{\boldsymbol{w}}; \bar{\boldsymbol{w}}_{r+1}) + \frac{(\gamma_r - \gamma_{r-1})E}{2}\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 \\
& + \alpha_r \sum_{i=t_r}^{t_{r+1}-1} \langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r \rangle + 20E\alpha_r(\tau + \nu)\epsilon_0^2 \\
& + \frac{\alpha_r^2 \|\sum_{i=t_r}^{t_{r+1}-1} \boldsymbol{\Delta}_i\|^2}{2(A_r E\mu + 2\gamma_r E - 2(L+\mu)\alpha_r E)} + \frac{3L + 2\mu}{2}\alpha_r \sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^{K} \pi_k \|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2,
\end{aligned} \tag{18}
$$

where $\boldsymbol{\Delta}_i = \sum_{k=1}^K \pi_k(\boldsymbol{G}_i^k - \nabla \mathcal{L}_k(\boldsymbol{w}_i^k))$.

Next lemma provides the upper bound for the discrepancy of local updates in Algorithm 2, and the proof is in Section D.2.

Lemma B.7. *Under the conditions of Theorem 3.1, for any* $t_r \leq t \leq t_{r+1} - 1$*, we have*

$$
\|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\| \leq \frac{4\alpha_r}{E\mu A_r/2 + \gamma_t E}\left(4\sqrt{E}\sigma \log(2/\delta) + E(H + \Lambda\rho)\right)
$$

holds with with probability at least $1 - \delta$.

Proof of Theorem 3.1. Plugging the conclusion of Lemma B.7 into (18), it follows that

$$
\begin{aligned}
\sum_{r=0}^{R} \alpha_r \sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^{K} \pi_k \|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2 &\leq \sum_{r=0}^{R} \frac{16E\alpha_r^3}{(\mu A_r + 2\gamma_r)^2}\left(32E^{-1}\sigma^2 \log^2(2/\delta) + 2(\Lambda\rho + H)^2\right) \\
&\leq \sum_{r=0}^{R} \frac{16E}{\mu^2(r+1)}\left(32E^{-1}\sigma^2 \log^2(2/\delta) + 2(\Lambda\rho + H)^2\right) \\
&\leq \frac{16E\log(R+1)}{\mu^2}\left(32E^{-1}\sigma^2 \log^2(2/\delta) + 2(\Lambda\rho + H)^2\right)
\end{aligned} \tag{19}
$$

holds with probability at least $1 - \delta/3$. Meanwhile, with the choice $\gamma_r = (L+\mu)\alpha_r$, the following summation is bounded by

$$
\sum_{r=0}^{R} \frac{16\alpha_r^3 E}{(\mu A_r + 2\gamma_r)^2} \leq \sum_{r=0}^{R} \frac{16E}{\mu^2(r+1)} \leq \frac{32E\log(R+1)}{\mu^2}.
$$

Then using the concentration inequality in Lemma B.1, with probability at least $1 - \delta/3$, we have

$$
\begin{aligned}
\sum_{r=0}^{R} \alpha_r \sum_{i=t_r}^{t_{r+1}-1} \langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r \rangle &\leq \bar{\sigma}\sqrt{3\log(3/\delta)}\left(\sum_{r=0}^{R} \alpha_r^2 \sum_{i=t_r}^{t_{r+1}-1} \|\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r\|^2\right)^{1/2} \\
&\leq \bar{\sigma}\sqrt{3\log(3/\delta)}\left(E\sum_{r=0}^{R} \alpha_r^2 \mathcal{R}^2(\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r)\right)^{1/2} \\
&\leq \bar{\sigma}\sqrt{6E\log(3/\delta)}\left(\sum_{r=0}^{R} \alpha_r^2(\mathcal{R}^2(\widehat{\boldsymbol{w}} - \boldsymbol{w}_0) + \mathcal{R}^2(\bar{\boldsymbol{w}}_r - \boldsymbol{w}_0))\right)^{1/2} \\
&\leq 2\epsilon_0\bar{\sigma}\sqrt{3\log(3/\delta)E(R+1)^3}.
\end{aligned} \tag{20}
$$

In the second inequality of (20), we used the assumption $\|\cdot\| \leq \mathcal{R}(\cdot)$. And the last inequality of (20) follows from the constraint in proximal step $\mathcal{R}(\bar{\boldsymbol{w}}_r - \boldsymbol{w}_0) \leq \epsilon_0$ and the assumption $\mathcal{R}(\widehat{\boldsymbol{w}} - \boldsymbol{w}_0) \leq \epsilon_0$. Additionally, by Lemma B.3, with

probability at least $1-\delta/3$, we also have

$$\begin{aligned}\left\|\alpha_r \sum_{i=t_r}^{t_{r+1}-1} \boldsymbol{\Delta}_i\right\|^2 &\leq 8\log(6/(8\delta))\alpha_r^2\left(\sum_{i=t_r}^{t_{r+1}-1}\mathbb{E}_{\mathcal{P}}\|\boldsymbol{\Delta}_i\|^2+\sum_{i=t_r}^{t_{r+1}-1}\|\boldsymbol{\Delta}_i\|^2\right)\\ &\leq 8\log(6/(8\delta))\alpha_r^2\left(2\sum_{i=t_r}^{t_{r+1}-1}\mathbb{E}_{\mathcal{P}}\|\boldsymbol{\Delta}_i\|^2+\max\left\{8\bar{\sigma}^2\log(6/\delta),16\bar{\sigma}^2\sqrt{E\log(6/\delta)}\right\}\right)\\ &\leq 8\log(6/(8\delta))\alpha_r^2\left(2E\bar{\sigma}^2+16\bar{\sigma}^2\sqrt{E\log(6/\delta)}\right)\\ &\leq 16\bar{\sigma}^2\log(6/\delta)\alpha_r^2\left(E+8\sqrt{E\log(6/\delta)}\right),\end{aligned} \tag{21}$$

where the second inequality follows from Lemma B.1 and the fact $\mathbb{E}[\|\boldsymbol{\Delta}_i\|^2]\leq\bar{\sigma}^2$. Then (21) results in

$$\begin{aligned}\sum_{r=0}^{R}\alpha_r^2\frac{\|\sum_{i=t_r}^{t_{r+1}-1}\boldsymbol{\Delta}_i\|^2}{2(A_rE\mu+2\gamma_rE-2(L+\mu)E)} &\leq 16\bar{\sigma}^2\log(6/\delta)\left(E+8\sqrt{E\log(6/\delta)}\right)\sum_{r=0}^{R}\frac{\alpha_r^2}{2(A_rE\mu+2\gamma_rE-2(L+\mu)E)}\\ &= 16\bar{\sigma}^2\log(6/\delta)\left(E+8\sqrt{E\log(3/\delta)}\right)\sum_{r=0}^{R}\frac{(r+1)^2}{E\mu(r+1)(r+2)}\\ &\leq 16\bar{\sigma}^2\log(6/\delta)\left(1+8\sqrt{\frac{\log(6/\delta)}{E}}\right)\frac{R+1}{\mu},\end{aligned} \tag{22}$$

where the equality holds due to $\gamma_r=(L+\mu)\alpha_r$ and $\alpha_r=r+1$. In addition, it follows from $D_{R+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{R+1})\geq 0$ and $D_0(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_0)=\gamma_0\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2/2$ that

$$\begin{aligned}&\sum_{r=0}^{R}D_r(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_r)-D_{r+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{r+1})+\frac{(\gamma_r-\gamma_{r-1})E}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\\ =&D_0(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_0)-D_{R+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{R+1})+\frac{(\gamma_R-\gamma_0)E}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2\\ \leq&\gamma_RE\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_0\|^2.\end{aligned} \tag{23}$$

Telescoping the induction relation (18) from $r=0$ to $r=R$, in conjunction with bounds (19)-(23), we can guarantee with probability at least $1-\delta$

$$\begin{aligned}\frac{1}{A_R}\sum_{r=0}^{R}\alpha_r[\phi(\bar{\boldsymbol{w}}_{r+1})-\phi(\widehat{\boldsymbol{w}})] &\leq \frac{2\gamma_R\epsilon_0^2}{(R+1)(R+2)}+32\bar{\sigma}^2\log(6/\delta)\left(1+8\sqrt{\frac{\log(3/\delta)}{E}}\right)\frac{1}{\mu E(R+2)}\\ &+\frac{16(3L+2\mu)\log(R+1)}{\mu^2(R+1)(R+2)}\left(32E^{-1}\sigma^2\log^2(2/\delta)+2(\Lambda\rho+H)^2\right)\\ &+\frac{8\epsilon_0\bar{\sigma}\sqrt{3\log(3/\delta)E(R+1)^3}}{E(R+1)(R+2)}+20(\tau+\nu)\epsilon_0^2\\ &\lesssim\frac{L\epsilon_0^2E}{T}+\frac{\bar{\sigma}^2\log(1/\delta)}{\mu T}+\frac{L\log(T+1)}{\mu^2T^2}\left(E\sigma^2\log^2(1/\delta)+E^2(\Lambda\rho+H)^2\right)\\ &+\frac{\epsilon_0\bar{\sigma}\sqrt{\log(1/\delta)}}{\sqrt{T}}+(\tau+\nu)\epsilon_0^2.\end{aligned}$$

Therefore the conclusion follows from Jensen's inequality. □

B.4. Proof of Theorem 3.2

The following corollary is a direct result of Theorem 3.1

Corollary B.1. *Under the same conditions in Theorem 3.1, we choose the number of local iterations E_m such that $E_m^2 \lesssim \bar{\sigma}^2 T_m$ for $T_m = E_m R_m$. Suppose the output of previous stage satisfies $\mathcal{R}(\widehat{\boldsymbol{w}}_{m-1} - \widehat{\boldsymbol{w}}^m) \leq r_m$, then the excess risk after the m-th stage is bounded by*

$$\phi(\widehat{\boldsymbol{w}}_m) - \phi(\widehat{\boldsymbol{w}}^m) \lesssim \frac{\bar{\sigma} r_m \sqrt{\log(1/\delta)}}{\sqrt{T_m}} + (\tau + \nu) r_m^2$$

with probability at least $1 - \delta$.

The next lemma restricts the averaged optimization error to a cone-like set. The conclusion (24) is from the relation (83) in the supplementary material to Agarwal et al. (2012).

Lemma B.8 (Lemma 3 and 11, Agarwal et al. (2012), modified)**.** *Let $\widehat{\boldsymbol{w}}$ be any optimum of the following regularized M-estimator*

$$\min_{\boldsymbol{w} \in \mathcal{W}} \left\{ \sum_{k=1}^K \pi_k \mathcal{L}_k(\boldsymbol{w}) + \lambda \mathcal{R}(\boldsymbol{w}) \right\},$$

where $\lambda > \mathcal{R}^*(\sum_{k=1}^K \pi_k \nabla \mathcal{L}_k(\boldsymbol{w}^*))/2$. Denote $\upsilon := 8\Psi(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\| + 2\eta/\lambda$. If $\phi(\boldsymbol{w}) - \phi(\widehat{\boldsymbol{w}}) \leq \eta$ for some $\eta > 0$ and $\boldsymbol{w}^* \in \mathcal{W}$, then we have

$$\mathcal{R}(\boldsymbol{w} - \boldsymbol{w}^*) \leq 4\Psi(\bar{\mathcal{M}})\|\boldsymbol{w} - \boldsymbol{w}^*\| + 2\frac{\eta}{\lambda} \tag{24}$$

and

$$\left(\frac{\mu}{2} - 32\Psi^2(\bar{\mathcal{M}})\tau\right)\|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|^2 \leq 2\tau\upsilon^2 + \phi(\boldsymbol{w}) - \phi(\widehat{\boldsymbol{w}}).$$

for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \bar{\mathcal{M}}^\top)$.

Proof of Theorem 3.2. In the first stage, we consider the following optimization problem

$$\widehat{\boldsymbol{w}}^1 = \arg \min_{\boldsymbol{w} \in \mathcal{W}(\widehat{\boldsymbol{w}}_0; \epsilon_0)} \left\{ \sum_{k=1}^K \pi_k \mathcal{L}_k(\boldsymbol{w}) + \lambda_0 \mathcal{R}(\boldsymbol{w}) \right\}. \tag{25}$$

Note that $\mathcal{R}(\widehat{\boldsymbol{w}}_0 - \widehat{\boldsymbol{w}}_{\text{opt}}) \leq 84\Psi^2(\bar{\mathcal{M}})\lambda_0/\mu$, thus we have

$$\begin{aligned}
\mathcal{R}(\widehat{\boldsymbol{w}}_0 - \widehat{\boldsymbol{w}}^1) &\leq \mathcal{R}(\widehat{\boldsymbol{w}}_0 - \widehat{\boldsymbol{w}}_{\text{opt}}) + \mathcal{R}(\widehat{\boldsymbol{w}}_{\text{opt}} - \boldsymbol{w}^*) + \mathcal{R}(\widehat{\boldsymbol{w}}^1 - \boldsymbol{w}^*) \\
&\leq 84\Psi^2(\bar{\mathcal{M}})\frac{\lambda_0}{\mu} + 12\Psi^2(\bar{\mathcal{M}})\frac{\lambda_{\text{opt}}}{\mu} + 12\Psi^2(\bar{\mathcal{M}})\frac{\lambda_0}{\mu} \leq 108\Psi^2(\bar{\mathcal{M}})\frac{\lambda_0}{\mu} = \epsilon_0,
\end{aligned}$$

where we used the fact $\lambda_0 \geq \lambda_{\text{opt}}$. Choosing R_0 and E_0 in Algorithm 3 such that

$$T_0 = R_0 E_0 = \frac{8 \times 54^2 \bar{\sigma}^2 \log(8M/\delta)}{\Psi^2(\bar{\mathcal{M}})\epsilon_0^2},$$

Corollary B.1 yields that with probability at least $1 - \delta/(2M)$

$$\begin{aligned}
\phi(\widehat{\boldsymbol{w}}_1) - \phi(\widehat{\boldsymbol{w}}^1) &\leq \frac{\bar{\sigma}\epsilon_0\sqrt{\log(8M/\delta)}}{\sqrt{T_0}} + (\tau + \nu)\epsilon_0^2 \\
&\leq \frac{\bar{\sigma}\epsilon_0\sqrt{\log(8M/\delta)}}{\sqrt{T_0}} + \frac{1}{2\mu}\Psi^2(\bar{\mathcal{M}})\lambda_0 \leq \frac{1}{\mu}\Psi^2(\bar{\mathcal{M}})\lambda_0 := \eta_1,
\end{aligned}$$

where we used the assumption $\mu(\tau + \nu)\epsilon_0^2 \leq \Psi^2(\bar{\mathcal{M}})\lambda_0$. In fact, $\boldsymbol{w}^*$ is also feasible for (25) since

$$\begin{aligned}
\mathcal{R}(\boldsymbol{w}^* - \widehat{\boldsymbol{w}}_0) &\leq \mathcal{R}(\boldsymbol{w}^* - \widehat{\boldsymbol{w}}_{\text{opt}}) + \mathcal{R}(\widehat{\boldsymbol{w}}_{\text{opt}} - \widehat{\boldsymbol{w}}_0) \\
&\leq 12\Psi^2(\bar{\mathcal{M}})\frac{\lambda_{\text{opt}}}{\mu} + 84\Psi^2(\bar{\mathcal{M}})\frac{\lambda_0}{\mu} \leq \epsilon_0.
\end{aligned}$$

In addition, we assume $\lambda_{\text{opt}} > \mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{w}^*))$ in Proposition 3.1, hence $\lambda_0 > \mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{w}^*))$. Applying Lemma B.8, we can obtain that

$$\begin{aligned}
\mathcal{R}^2(\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}^*) &\leq 32\Psi^2(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}^*\|^2 + \frac{2\eta_1^2}{\lambda_0^2} \\
&\leq 64\Psi^2(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}_1 - \widehat{\boldsymbol{w}}^1\|^2 + 64\Psi^2(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}^1 - \boldsymbol{w}^*\|^2 + \frac{2\eta_1^2}{\lambda_0^2} \\
&\overset{(a)}{\leq} \frac{64\tau\Psi^2(\bar{\mathcal{M}})}{\mu}\left(8\Psi(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}^1 - \boldsymbol{w}^*\| + 2\frac{\eta_1}{\lambda_0}\right)^2 + 64\Psi^2(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}^0 - \boldsymbol{w}^*\|^2 + \frac{2\eta_1^2}{\lambda_0^2} \\
&\overset{(b)}{\leq} 128\Psi^2(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}^1 - \boldsymbol{w}^*\|^2 + \frac{4\eta_1^2}{\lambda_0^2} \\
&\overset{(c)}{\leq} 128 \times 9\Psi^4(\bar{\mathcal{M}})\frac{\lambda_0^2}{\mu^2} + \frac{4\eta_1^2}{\lambda_0^2} \leq 48^2\Psi^4(\bar{\mathcal{M}})\frac{\lambda_0^2}{\mu^2}.
\end{aligned} \tag{26}$$

where (a) follows from the second conclusion in Lemma B.8, (b) follows from $128\Psi^2(\bar{\mathcal{M}})\tau \leq \mu$ and (c) follows from Proposition 3.1. Let $\lambda_m = \lambda_0 \cdot 2^{-m}$, then we consider the following optimization problems

$$\widehat{\boldsymbol{w}}^{m+1} = \arg\min_{\boldsymbol{w}\in\mathcal{W}(\widehat{\boldsymbol{w}}_m; r_m)} \left\{\sum_{k=1}^K \pi_k\mathcal{L}_k(\boldsymbol{w}) + \lambda_m\mathcal{R}(\boldsymbol{w})\right\} \tag{27}$$

where $r_m = 108\Psi^2(\bar{\mathcal{M}})\lambda_m/\mu$ for $m \geq 0$. We define the following good events: for any $m = 0, 1, ..., M-1$

$$\mathcal{A}_m = \left\{\mathcal{R}(\widehat{\boldsymbol{w}}_{m+1} - \boldsymbol{w}^*) \leq 48\frac{\Psi^2(\bar{\mathcal{M}})\lambda_m}{\mu}\right\}.$$

Now we prove $\mathbb{P}(\mathcal{A}_m^c) \leq \frac{\delta}{2} + \frac{m\delta}{2M}$. Recall the definition of η_0, then it follows from (26) that $\mathbb{P}(\mathcal{A}_0^c) \leq \delta/2$. Now we assume $\mathbb{P}(\mathcal{A}_{m-1}^c) \leq \frac{\delta}{2} + \frac{(m-1)\delta}{2M}$ holds. Under the event $\mathcal{A}_{m-1}$, note that

$$\begin{aligned}
\mathcal{R}(\widehat{\boldsymbol{w}}_m - \widehat{\boldsymbol{w}}^{m+1}) &\leq \mathcal{R}(\widehat{\boldsymbol{w}}_m - \boldsymbol{w}^*) + \mathcal{R}(\widehat{\boldsymbol{w}}^{m+1} - \boldsymbol{w}^*) \leq 48\Psi^2(\bar{\mathcal{M}})\frac{\lambda_{m-1}}{\mu} + 12\Psi^2(\bar{\mathcal{M}})\frac{\lambda_m}{\mu} \\
&= 96\Psi^2(\bar{\mathcal{M}})\frac{\lambda_m}{\mu} + 12\Psi^2(\bar{\mathcal{M}})\frac{\lambda_m}{\mu} = r_m,
\end{aligned} \tag{28}$$

where we applied Proposition 3.1 to $\mathcal{R}(\widehat{\boldsymbol{w}}^m - \boldsymbol{w}^*)$. We may choose R_m and E_m in Algorithm 3 satisfies that

$$T_m = R_m E_m = \frac{8 \times 54^2\Psi^2(\bar{\mathcal{M}})\bar{\sigma}^2\log(2M/\delta)}{\mu^2 r_m^2},$$

then Corollary B.1 guarantees there exists some Borel set $\mathcal{C}_m$ such that $\mathbb{P}(\mathcal{C}_m^c) \leq \delta/(2M)$. Under the event $\mathcal{A}_{m-1} \cap \mathcal{C}_m$, we have

$$\begin{aligned}
\phi(\widehat{\boldsymbol{w}}_{m+1}) - \phi(\widehat{\boldsymbol{w}}^{m+1}) &\leq \frac{\bar{\sigma}r_m\sqrt{\log(2M/\delta)}}{\sqrt{T_m}} + (\tau+\nu)r_m^2 \\
&\leq \frac{\bar{\sigma}r_m\sqrt{\log(1/\delta)}}{\sqrt{T_m}} + \frac{\mu}{8 \times 54^2\Psi^2(\bar{\mathcal{M}})}r_m^2 \leq \frac{\Psi^2(\bar{\mathcal{M}})\lambda_m^2}{\mu} := \eta_{m+1}.
\end{aligned}$$

In the first inequality above, we used $\widehat{\boldsymbol{w}}_{m-1}$ is the initial point of the m-th stage and the relation (28). In the second inequality above, we used the Assumption 9. Recall that $\mathcal{R}(\widehat{\boldsymbol{w}}_{m-1} - \boldsymbol{w}^*) \leq 48\Psi^2(\bar{\mathcal{M}})\lambda_m/\mu \leq r_m$, thus $\boldsymbol{w}^*$ is also feasible for problem (27). Applying Lemma B.8 again, we have

$$\mathcal{R}^2(\widehat{\boldsymbol{w}}_m - \boldsymbol{w}^*) \leq 128 \times 9\Psi^4(\bar{\mathcal{M}})\frac{\lambda_m^2}{\mu^2} + \frac{4\eta_{m+1}^2}{\lambda_m^2} \leq 48^2\Psi^4(\bar{\mathcal{M}})\frac{\lambda_m^2}{\mu^2}. \tag{29}$$

Hence we have proved $(\mathcal{A}_{m-1} \cap \mathcal{C}_m) \subseteq \mathcal{A}_m$, then it follows that

$$\mathbb{P}(\mathcal{A}_m^c) \leq \mathbb{P}(\mathcal{A}_{M-1}^c) + \mathbb{P}(\mathcal{C}_M) \leq \frac{\delta}{2} + \frac{(m-1)\delta}{2M} + \frac{\delta}{2M} = \frac{\delta}{2} + \frac{m\delta}{2M}.$$

We choose the number of stages such that $\lambda_{M-1} = \lambda_{\text{opt}}$, which means $M = \log_2(\lambda_0/\lambda_{\text{opt}}) + 1$. In fact, at the M-th stage, $\widehat{\boldsymbol{w}}^M = \widehat{\boldsymbol{w}}$ since $\lambda_M = \lambda_{\text{opt}}$. Under the event $\mathcal{A}_M$ with $\mathbb{P}(\mathcal{A}_M^c) \leq \delta$, we are guaranteed that

$$\phi(\widehat{\boldsymbol{w}}_M) - \phi(\widehat{\boldsymbol{w}}^M) = \phi(\widehat{\boldsymbol{w}}_M) - \phi(\widehat{\boldsymbol{w}}_{\text{opt}}) \leq \frac{\Psi^2(\bar{\mathcal{M}})\lambda_{\text{opt}}^2}{\mu} := \eta_M.$$

Together with the second conclusion in Lemma B.8, we have

$$\begin{aligned}\|\widehat{\boldsymbol{w}}_M - \widehat{\boldsymbol{w}}_{\text{opt}}\|^2 &\leq \frac{6\tau}{\mu}\left(8\Psi(\bar{\mathcal{M}})\|\widehat{\boldsymbol{w}}_{\text{opt}} - \boldsymbol{w}^*\| + \frac{\eta_M^2}{\lambda_{\text{opt}}^2}\right)^2 + \frac{3}{\mu}\eta_M \\ &\leq \|\widehat{\boldsymbol{w}}_{\text{opt}} - \boldsymbol{w}^*\|^2 + \frac{6}{\mu}\eta_M = \|\widehat{\boldsymbol{w}}_{\text{opt}} - \boldsymbol{w}^*\|^2 + \frac{6\Psi^2(\bar{\mathcal{M}})\lambda_{\text{opt}}^2}{\mu^2}.\end{aligned}$$

In addition, from the definition of $\mathcal{A}_M$, we also have

$$\mathcal{R}(\widehat{\boldsymbol{w}}_M - \boldsymbol{w}^*) \leq \frac{48\Psi^2(\bar{\mathcal{M}})\lambda_M}{\mu^2} = \frac{48\Psi^2(\bar{\mathcal{M}})\lambda_{\text{opt}}}{\mu^2}.$$

Now we consider the total complexity,

$$\begin{aligned}T &= \sum_{m=0}^{M-1} T_m = \sum_{m=0}^{M-1} \frac{16 \times 54^2 \Psi^2(\bar{\mathcal{M}})\bar{\sigma}^2 \log(2M/\delta)}{\mu^2 r_m^2} \\ &\leq \sum_{m=0}^{M-1} \frac{4\bar{\sigma}^2 \log(2M/\delta)}{\Psi^2(\bar{\mathcal{M}})\lambda_m^2} \leq M \cdot 2^{2M} \frac{4\bar{\sigma}^2 \log(2M/\delta)}{\Psi^2(\bar{\mathcal{M}})\lambda_{\text{opt}}^2} \\ &\leq \frac{4\bar{\sigma}^2(\log_2(\lambda_0/\lambda_{\text{opt}}) + 1)}{\Psi^2(\bar{\mathcal{M}})\lambda_{\text{opt}}^2} \log\left(\frac{\log_2(\lambda_0/\lambda_{\text{opt}}) + 1}{\delta}\right).\end{aligned}$$

□

C. One-Step Induction Relation

Lemma C.1 (Proposition 1 in the appendix of (Chen et al., 2012))**.** *Given any proper lsc convex function $\psi(x)$ and a sequence of $\{\boldsymbol{z}_i\}_{i=0}^t$ with each $\boldsymbol{z}_i \in \mathcal{W}$, if*

$$\boldsymbol{z}_+ = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \left\{\psi(\boldsymbol{w}) + \sum_{i=0}^t \frac{\eta_i}{2}\|\boldsymbol{w} - \boldsymbol{z}_i\|^2\right\},$$

where $\{\eta_i \geq 0\}_{i=1}^t$ is a sequence of parameters, then for any $\boldsymbol{w} \in \mathcal{W}$:

$$\left(\frac{1}{2}\sum_{i=0}^t \eta_i\right)\|\boldsymbol{w} - \boldsymbol{z}_+\|^2 \leq \psi(\boldsymbol{w}) + \sum_{i=0}^t \frac{\eta_i}{2}\|\boldsymbol{w} - \boldsymbol{z}_i\|^2 - \left\{\psi(\boldsymbol{z}_+) + \sum_{i=0}^t \frac{\eta_i}{2}\|\boldsymbol{z}_+ - \boldsymbol{z}_i\|^2\right\}. \tag{30}$$

C.1. Deferred Proof of Lemma B.4

Proof of Lemma B.4. According to the definition of $D_{t+1}(\widehat{\boldsymbol{w}}; \boldsymbol{w}_{t+1})$ in (14), we note that

$$\begin{aligned}D_{t+1}(\widehat{\boldsymbol{w}}; \boldsymbol{w}_{t+1}) =& \langle \widehat{\boldsymbol{w}} - \boldsymbol{w}_{t+1}, \boldsymbol{g}_t\rangle + \frac{\mu}{4}\sum_{i=0}^t \alpha_i\left(\|\widehat{\boldsymbol{w}} - \check{\boldsymbol{w}}_i\|^2 - \|\boldsymbol{w}_{t+1} - \check{\boldsymbol{w}}_i\|^2\right) \\ &+ \frac{\gamma_t}{2}\left(\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_0\|^2\right) + A_t(h(\widehat{\boldsymbol{w}}) - h(\boldsymbol{w}_{t+1})).\end{aligned}$$

Recall the fact $A_t = A_{t-1} + \alpha_t$, then simple arrangement gives rise to the following decomposition

$$\begin{aligned}
D_{t+1}(\widehat{\boldsymbol{w}}; \boldsymbol{w}_{t+1}) =& \langle \widehat{\boldsymbol{w}} - \boldsymbol{w}_t, \boldsymbol{g}_{t-1} \rangle + \frac{\mu}{4} \sum_{i=0}^{t} \alpha_i \left(\|\widehat{\boldsymbol{w}} - \check{\boldsymbol{w}}_i\|^2 - \|\boldsymbol{w}_t - \check{\boldsymbol{w}}_i\|^2 \right) \\
&+ \frac{\gamma_{t-1}}{2} \left(\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 - \|\boldsymbol{w}_t - \boldsymbol{w}_0\|^2 \right) + A_{t-1}(h(\widehat{\boldsymbol{w}}) - h(\boldsymbol{w}_t)) \\
&- \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \boldsymbol{g}_{t-1} \rangle - \frac{\mu}{4} \sum_{i=0}^{t-1} \alpha_i \left(\|\boldsymbol{w}_{t+1} - \check{\boldsymbol{w}}_i\|^2 - \|\boldsymbol{w}_t - \check{\boldsymbol{w}}_i\|^2 \right) \\
&- \frac{\gamma_{t-1}}{2} \left(\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_0\|^2 - \|\boldsymbol{w}_t - \boldsymbol{w}_0\|^2 \right) - A_{t-1}(h(\boldsymbol{w}_{t+1}) - h(\boldsymbol{w}_t))) \\
&+ \alpha_t \left\{ \langle \boldsymbol{G}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_{t+1} \rangle + h(\widehat{\boldsymbol{w}}) + \frac{\mu}{4} \|\widehat{\boldsymbol{w}} - \check{\boldsymbol{w}}_t\|^2 - \frac{\mu}{4} \|\boldsymbol{w}_{t+1} - \check{\boldsymbol{w}}_t\|^2 - h(\boldsymbol{w}_{t+1}) \right\} \\
&+ \frac{\gamma_t - \gamma_{t-1}}{2} (\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_0\|^2).
\end{aligned}$$

From the definitions of $D_t(\widehat{\boldsymbol{w}}; \boldsymbol{w}_t)$ and $D_t(\boldsymbol{w}_{t+1}; \boldsymbol{w}_t)$ in (14), together with $\gamma_t \geq \gamma_{t-1}$ we have

$$\begin{aligned}
D_{t+1}(\widehat{\boldsymbol{w}}; \boldsymbol{w}_{t+1}) \leq & D_t(\widehat{\boldsymbol{w}}; \boldsymbol{w}_t) - D_t(\boldsymbol{w}_{t+1}; \boldsymbol{w}_t) + \alpha_t \langle \boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_{t+1} \rangle + \frac{\gamma_t - \gamma_{t-1}}{2} \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 \\
&+ \alpha_t \underbrace{\left\{ \mathcal{L}(\boldsymbol{w}_t) + \sum_{k=1}^{K} \pi_k \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \widehat{\boldsymbol{w}} - \boldsymbol{w}_t \rangle + \frac{\mu}{4} \|\check{\boldsymbol{w}}_t - \widehat{\boldsymbol{w}}\|^2 + h(\widehat{\boldsymbol{w}}) \right\}}_{A_1} \\
&- \alpha_t \underbrace{\left\{ \mathcal{L}(\boldsymbol{w}_t) + \sum_{k=1}^{K} \pi_k \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + h(\boldsymbol{w}_{t+1}) \right\}}_{A_2},
\end{aligned} \tag{31}$$

where $\boldsymbol{\Delta}_t = \sum_{k=1}^{K} \pi_k [\nabla f(\boldsymbol{w}_t^k; \xi_t^k) - \nabla \mathcal{L}_k(\boldsymbol{w}_t^k)]$. By μ-strong convexity and L-smoothness of local loss $\mathcal{L}_k$, we get

$$\begin{aligned}
\mathcal{L}_k(\boldsymbol{w}_t^k) + \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \widehat{\boldsymbol{w}} - \boldsymbol{w}_t^k \rangle + \frac{\mu}{4} \|\widehat{\boldsymbol{w}} - \check{\boldsymbol{w}}_t\|^2 &\leq \mathcal{L}_k(\widehat{\boldsymbol{w}}) + \frac{\mu}{4} \|\widehat{\boldsymbol{w}} - \check{\boldsymbol{w}}_t\|^2 - \frac{\mu}{2} \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_t^k\|^2 \\
&\leq \mathcal{L}_k(\widehat{\boldsymbol{w}}) + \frac{\mu}{2} \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_t^k\|^2 + \frac{\mu}{2} \|\check{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2 - \frac{\mu}{2} \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_t^k\|^2 \\
&= \mathcal{L}_k(\widehat{\boldsymbol{w}}) + \frac{\mu}{2} \|\check{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2,
\end{aligned}$$

and

$$\mathcal{L}_k(\boldsymbol{w}_t) + \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \boldsymbol{w}_t^k - \boldsymbol{w}_t \rangle \leq \mathcal{L}_k(\boldsymbol{w}_t^k) + \frac{L}{2} \|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2.$$

Summing the two inequalities above and taking average over k, together with the definition $\phi(\widehat{\boldsymbol{w}}) = \mathcal{L}_(\widehat{\boldsymbol{w}}) + h(\widehat{\boldsymbol{w}})$, we can bound A_1 in (31) as

$$A_1 \leq \phi(\widehat{\boldsymbol{w}}) + \frac{\mu}{2} \sum_{k=1}^{K} \pi_k \|\boldsymbol{w}_t^k - \check{\boldsymbol{w}}_t\|^2 + \frac{L}{2} \sum_{k=1}^{K} \pi_k \|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2. \tag{32}$$

Using the convexity and L-smoothness of $\mathcal{L}_k$ again, we can obtain that

$$\mathcal{L}_k(\boldsymbol{w}_t) \geq \mathcal{L}_k(\boldsymbol{w}_t^k) + \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \boldsymbol{w}_t - \boldsymbol{w}_t^k \rangle,$$

and

$$\mathcal{L}_k(\boldsymbol{w}_t^k) \geq \mathcal{L}_k(\boldsymbol{w}_{t+1}) + \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \boldsymbol{w}_t^k - \boldsymbol{w}_{t+1} \rangle - \frac{L}{2} \|\boldsymbol{w}_t^k - \boldsymbol{w}_{t+1}\|^2.$$

Summing two inequalities displayed above gives the bound of A_2, that is

$$\begin{aligned} A_2 &= \mathcal{L}(\boldsymbol{w}_t) + \sum_{k=1}^K \pi_k \langle \nabla \mathcal{L}_k(\boldsymbol{w}_t^k), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + h(\boldsymbol{w}_{t+1}) \\ &\geq \phi(\boldsymbol{w}_{t+1}) - \frac{L}{2} \sum_{k=1}^K \pi_k \|\boldsymbol{w}_t^k - \boldsymbol{w}_{t+1}\|^2 \\ &\geq \phi(\boldsymbol{w}_{t+1}) - L\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 - L \sum_{k=1}^K \pi_k \|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2. \end{aligned} \tag{33}$$

Plugging (32) and (33) into (31) results in

$$\begin{aligned} D_{t+1}(\widehat{\boldsymbol{w}}; \boldsymbol{w}_{t+1}) \leq & D_t(\widehat{\boldsymbol{w}}; \boldsymbol{w}_t) - D_t(\boldsymbol{w}_{t+1}; \boldsymbol{w}_t) + \alpha_t[\phi(\widehat{\boldsymbol{w}}) - \phi(\boldsymbol{w}_{t+1})] \\ & + \alpha_t L \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 + \langle \boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_{t+1} \rangle + \alpha_t \frac{\gamma_t - \gamma_{t-1}}{2} \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 \\ & + \alpha_t \left(\frac{\mu}{2} \sum_{k=1}^K \pi_k \|\boldsymbol{w}_t^k - \check{\boldsymbol{w}}_t\|^2 + \frac{3L}{2} \sum_{k=1}^K \pi_k \|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2 + 4(4\tau + 3\nu)\epsilon_0^2 \right). \end{aligned} \tag{34}$$

To apply Lemma C.1, we let $\psi(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{g}_t \rangle$, $\eta_i = \mu\alpha_i/2$ for $i \leq t-1$ and $\eta_t = \gamma_t/2$, $\boldsymbol{z}_i = \check{\boldsymbol{w}}_i$ for $i \leq t-1$ and $\boldsymbol{z}_t = \boldsymbol{w}_0$. Recalling the definition of $\boldsymbol{w}_t$ in (13), that is

$$\boldsymbol{w}_t = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \left\{ \psi(\boldsymbol{w}) + \sum_{i=0}^t \frac{\eta_i}{2} \|\boldsymbol{w} - \boldsymbol{z}_i\|^2 \right\},$$

which implies that

$$\begin{aligned} \left(\frac{\mu}{2} A_t + \gamma_t \right) \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 &\leq \psi(\boldsymbol{w}_{t+1}) + \sum_{i=0}^t \frac{\eta_i}{2} \|\boldsymbol{w}_{t+1} - \boldsymbol{z}_i\|^2 - \left\{ \psi(\boldsymbol{w}_t) + \sum_{i=0}^t \frac{\eta_i}{2} \|\boldsymbol{w}_t - \boldsymbol{z}_i\|^2 \right\} \\ &= D_t(\boldsymbol{w}_{t+1}; \boldsymbol{w}_t). \end{aligned}$$

In addition, using the simple inequality: $-ax^2 + bx \leq \frac{b^2}{4a}$ for $a > 0$, we have

$$\begin{aligned} & - D_t(\boldsymbol{w}_{t+1}; \boldsymbol{w}_t) + L\alpha_t \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 + \alpha_t \langle \boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_{t+1} \rangle \\ \leq & - \left(\frac{\mu}{2} A_t + \gamma_t - L\alpha_t \right) \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 + \alpha_t \|\boldsymbol{\Delta}_t\| \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| + \alpha_t \langle \boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_t \rangle \\ \leq & \frac{\alpha_t^2 \|\boldsymbol{\Delta}_t\|^2}{2(\mu A_t + 2\gamma_t - 2L\alpha_t)} + \alpha_t \langle \boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_t \rangle. \end{aligned}$$

Then plugging above inequality into (34) yields

$$\begin{aligned} \alpha_t[\phi(\boldsymbol{w}_{t+1}) - \phi(\widehat{\boldsymbol{w}})] \leq & D_t(\widehat{\boldsymbol{w}}; \boldsymbol{w}_t) - D_{t+1}(\widehat{\boldsymbol{w}}; \boldsymbol{w}_{t+1}) + \frac{\gamma_t - \gamma_{t-1}}{2} \|\widehat{\boldsymbol{w}} - \boldsymbol{w}_0\|^2 \\ & + \alpha_t \langle \boldsymbol{\Delta}_t, \widehat{\boldsymbol{w}} - \boldsymbol{w}_t \rangle + \frac{\alpha_t^2 \|\boldsymbol{\Delta}_t\|^2}{2(A_t\mu + 2\gamma_t - 2L\alpha_t)} \\ & + \alpha_t \left(\frac{\mu}{2} \sum_{k=1}^K \pi_k \|\boldsymbol{w}_t^k - \check{\boldsymbol{w}}_t\|^2 + \frac{3L}{2} \sum_{k=1}^K \pi_k \|\boldsymbol{w}_t^k - \boldsymbol{w}_t\|^2 \right). \end{aligned} \tag{35}$$

Thus we have complete the proof of Lemma B.4. □

C.2. Deferred Proof of Lemma B.6

Proof of Lemma B.6. We first recall the definition of $D_{r+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{r+1})$

$$\begin{aligned}D_{r+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{r+1}) =&\langle \boldsymbol{g}_{t_{r+1}-1}, \widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_{r+1}\rangle + \frac{\mu E}{4}\sum_{j=0}^{r}\alpha_j(\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_j\|^2 - \|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_j\|^2)\\ &+\frac{\gamma_r E}{2}(\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_0\|^2 - \|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_0\|^2) + A_r E[h(\widehat{\boldsymbol{w}})-h(\bar{\boldsymbol{w}}_{r+1})].\end{aligned}$$

Using $A_r = A_{r-1}+\alpha_r$, we may write $D_{r+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{r+1})$ as

$$\begin{aligned}D_{r+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{r+1}) =& \langle \boldsymbol{g}_{t_r-1}, \widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_r\rangle + \frac{\mu E}{4}\sum_{j=0}^{r-1}\alpha_j(\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_j\|^2 - \|\bar{\boldsymbol{w}}_r-\bar{\boldsymbol{w}}_j\|^2)\\ &+\frac{\gamma_{r-1} E}{2}(\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_0\|^2 - \|\bar{\boldsymbol{w}}_r-\bar{\boldsymbol{w}}_0\|^2) + A_{r-1}E[h(\widehat{\boldsymbol{w}})-h(\bar{\boldsymbol{w}}_{r+1})]\\ &-\langle \boldsymbol{g}_{t_r-1}, \bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_r\rangle - \frac{\mu E}{4}\sum_{j=0}^{r-1}\alpha_j(\|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_j\|^2 - \|\bar{\boldsymbol{w}}_r-\bar{\boldsymbol{w}}_j\|^2)\\ &-\frac{\gamma_{r-1} E}{2}(\|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_0\|^2 - \|\bar{\boldsymbol{w}}_r-\bar{\boldsymbol{w}}_0\|^2) - A_{r-1}E[h(\bar{\boldsymbol{w}}_{r+1})-h(\bar{\boldsymbol{w}}_{r+1})]\\ &+\langle \boldsymbol{g}_{t_{r+1}-1}-\boldsymbol{g}_{t_r-1}, \widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_{r+1}\rangle + \alpha_r E\left(\frac{\mu}{4}\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_r\|^2 - \frac{\mu}{4}\|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_r\|^2 + h(\widehat{\boldsymbol{w}})-h(\bar{\boldsymbol{w}}_{r+1})\right)\\ &+\frac{(\gamma_r-\gamma_{r-1})E}{2}(\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_0\|^2 - \|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_0\|^2).\end{aligned}$$

From the definition of $D_r(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_r)$, $D_r(\bar{\boldsymbol{w}}_r;\bar{\boldsymbol{w}}_{r+1})$ and $\boldsymbol{g}_{t_{r+1}-1}-\boldsymbol{g}_{t_r-1} = \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^{K}\pi_k\boldsymbol{G}_i^k$, we have

$$\begin{aligned}D_{r+1}(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_{r+1}) \le& D_r(\widehat{\boldsymbol{w}};\bar{\boldsymbol{w}}_r) - D_r(\bar{\boldsymbol{w}}_{r+1};\bar{\boldsymbol{w}}_r) + \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_{r+1}\rangle\\ &+\alpha_r\underbrace{\sum_{i=t_r}^{t_{r+1}-1}\left\{\mathcal{L}(\bar{\boldsymbol{w}}_r)+\sum_{k=1}^{K}\pi_k\langle\nabla\mathcal{L}_k(\boldsymbol{w}_i^k), \widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_r\rangle + \frac{\mu}{4}\|\bar{\boldsymbol{w}}_r-\widehat{\boldsymbol{w}}\|^2 + h(\widehat{\boldsymbol{w}})\right\}}_{B_1}\\ &-\alpha_r\underbrace{\sum_{i=t_r}^{t_{r+1}-1}\left\{\mathcal{L}(\bar{\boldsymbol{w}}_r)+\sum_{k=1}^{K}\pi_k\langle\nabla\mathcal{L}_k(\boldsymbol{w}_i^k), \bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_r\rangle + h(\bar{\boldsymbol{w}}_{r+1})\right\}}_{B_2},\end{aligned} \tag{36}$$

where $\boldsymbol{\Delta}_i = \sum_{k=1}^{K}\pi_k(\boldsymbol{G}_i^k - \nabla\mathcal{L}_k(\boldsymbol{w}_i^k))$. By the RSC and RSM of $\mathcal{L}_k$, it follows that for any $t_r \le i \le t_{r+1}-1$

$$\begin{aligned}&\mathcal{L}_k(\boldsymbol{w}_i^k) + \langle\nabla\mathcal{L}_k(\boldsymbol{w}_i^k), \widehat{\boldsymbol{w}}-\boldsymbol{w}_i^k\rangle + \frac{\mu}{4}\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_r\|^2\\ &\le\mathcal{L}_k(\widehat{\boldsymbol{w}}) + \frac{\mu}{4}\|\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_r\|^2 - \frac{\mu}{2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}_i^k\|^2 + \tau_k\mathcal{R}^2(\widehat{\boldsymbol{w}}-\boldsymbol{w}_i^k)\\ &\le\mathcal{L}_k(\widehat{\boldsymbol{w}}) + \frac{\mu}{2}\|\bar{\boldsymbol{w}}_r-\boldsymbol{w}_i^k\|^2 + 2\tau_k\mathcal{R}^2(\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_{r+1}) + 2\tau_k\mathcal{R}^2(\bar{\boldsymbol{w}}_{r+1}-\boldsymbol{w}_i^k)\\ &\le\mathcal{L}_k(\widehat{\boldsymbol{w}}) + \mu\|\bar{\boldsymbol{w}}_{r+1}-\boldsymbol{w}_i^k\|^2 + \mu\|\bar{\boldsymbol{w}}_r-\bar{\boldsymbol{w}}_{r+1}\|^2 + 2\tau_k\mathcal{R}^2(\widehat{\boldsymbol{w}}-\bar{\boldsymbol{w}}_{r+1}) + 2\tau_k\mathcal{R}^2(\bar{\boldsymbol{w}}_{r+1}-\boldsymbol{w}_i^k),\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}_k(\bar{\boldsymbol{w}}_r) + \langle\nabla\mathcal{L}_k(\boldsymbol{w}_i^k), \boldsymbol{w}_i^k-\bar{\boldsymbol{w}}_r\rangle \le& \mathcal{L}_k(\boldsymbol{w}_i^k) + \frac{L}{2}\|\boldsymbol{w}_i^k-\bar{\boldsymbol{w}}_r\|^2 + \nu_k\mathcal{R}^2(\boldsymbol{w}_i^k-\bar{\boldsymbol{w}}_r)\\ \le&\mathcal{L}_k(\boldsymbol{w}_i^k) + L\|\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_r\|^2 + 2\nu_k\mathcal{R}^2(\bar{\boldsymbol{w}}_{r+1}-\bar{\boldsymbol{w}}_r)\\ &+L\|\boldsymbol{w}_i^k-\bar{\boldsymbol{w}}_r\|^2 + 2\nu_k\mathcal{R}^2(\boldsymbol{w}_i^k-\bar{\boldsymbol{w}}_{r+1}).\end{aligned}$$

Summing the two inequalities above and taking average over k, together with the definition $\tau = \sum_{k=1}^K \pi_k \tau_k$, we can bound B_1 in (36) as

$$\begin{aligned} B_1 \leq & E\phi(\widehat{\boldsymbol{w}}) + E(L+\mu)\|\bar{\boldsymbol{w}}_{r+1} - \bar{\boldsymbol{w}}_r\|^2 + 2E\nu\mathcal{R}^2(\bar{\boldsymbol{w}}_{r+1} - \bar{\boldsymbol{w}}_r) + 2E\tau\mathcal{R}^2(\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_{r+1}) \\ & + (L+\mu)\sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^K \pi_k \|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2 + \sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^K \pi_k 2(\tau_k + \nu_k)\mathcal{R}^2(\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}) \\ \leq & E\phi(\widehat{\boldsymbol{w}}) + (L+\mu)E\|\bar{\boldsymbol{w}}_{r+1} - \bar{\boldsymbol{w}}_r\|^2 + (L+\mu)\sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^K \pi_k \|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2 + 16E(\tau+\nu)\epsilon_0^2. \end{aligned} \tag{37}$$

We used the constrain $\mathcal{R}(\boldsymbol{w} - \bar{\boldsymbol{w}}_0) \leq \epsilon_0$ in the proximal operator and the assumption $\mathcal{R}(\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_0) \leq \epsilon_0$ in the last inequality of (37). Applying the convexity and RSM of $\mathcal{L}_k$ again, we can obtain that

$$\mathcal{L}_k(\bar{\boldsymbol{w}}_r) \geq \mathcal{L}_k(\boldsymbol{w}_i^k) + \langle \nabla\mathcal{L}_k(\boldsymbol{w}_i^k), \bar{\boldsymbol{w}}_r - \boldsymbol{w}_i^k\rangle,$$

and

$$\mathcal{L}_k(\boldsymbol{w}_i^k) \geq \mathcal{L}_k(\bar{\boldsymbol{w}}_{r+1}) + \langle \nabla\mathcal{L}_k(\boldsymbol{w}_i^k), \boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\rangle - \frac{L}{2}\|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{r+1}\|^2 - \nu_k\mathcal{R}^2(\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{r+1}).$$

In conjunction with the definition $\nu = \sum_{k=1}^K \pi_k\nu_k$, two inequalities displayed above shows that

$$\begin{aligned} B_2 \geq & E\phi(\bar{\boldsymbol{w}}_{r+1}) - \frac{L}{2}\sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^K \pi_k\|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2 - \sum_{k=1}^K \pi_k\nu_k\mathcal{R}^2(\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}) \\ \geq & E\phi(\bar{\boldsymbol{w}}_{r+1}) - \frac{L}{2}\sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^K \pi_k\|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2 - 4E\nu\epsilon_0^2. \end{aligned} \tag{38}$$

According to Lemma C.1, we may guarantee that

$$\begin{aligned} & - D_r(\bar{\boldsymbol{w}}_{r+1}; \bar{\boldsymbol{w}}_r) + (L+\mu)E\alpha_r\|\bar{\boldsymbol{w}}_{r+1} - \bar{\boldsymbol{w}}_r\|^2 + \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_{r+1}\rangle \\ \leq & - E\left(\frac{\mu}{2}A_r + \gamma_r - (L+\mu)\alpha_r\right)\|\bar{\boldsymbol{w}}_{r+1} - \bar{\boldsymbol{w}}_r\|^2 - \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\langle \boldsymbol{\Delta}_i, \bar{\boldsymbol{w}}_{r+1} - \bar{\boldsymbol{w}}_r\rangle + \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r\rangle \\ \leq & \frac{\alpha_r^2\|\sum_{i=t_r}^{t_{r+1}-1}\boldsymbol{\Delta}_i\|^2}{2(A_rE\mu + 2\gamma_rE - 2(L+\mu)E)} + \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r\rangle, \end{aligned} \tag{39}$$

where we used the inequality $-ax^2 + bx \leq \frac{b^2}{4a}$ for $a > 0$ in the last inequality. Plugging three upper bounds (37), (38) and (39) into (36), we have

$$\begin{aligned} D_{r+1}(\widehat{\boldsymbol{w}}; \bar{\boldsymbol{w}}_{r+1}) - D_r(\widehat{\boldsymbol{w}}; \bar{\boldsymbol{w}}_r) \leq & E\alpha_r[\phi(\widehat{\boldsymbol{w}}) - \phi(\bar{\boldsymbol{w}}_{r+1})] + \frac{(\gamma_r - \gamma_{r-1})E}{2}\|\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_0\|^2 \\ & + \alpha_r\sum_{i=t_r}^{t_{r+1}-1}\langle \boldsymbol{\Delta}_i, \widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}_r\rangle + \frac{\|\alpha_r^2\sum_{i=t_r}^{t_{r+1}-1}\boldsymbol{\Delta}_i\|^2}{2(A_rE\mu + 2\gamma_rE - 2(L+\mu)E)} \\ & + \alpha_r\frac{3L+2\mu}{2}\sum_{i=t_r}^{t_{r+1}-1}\sum_{k=1}^K \pi_k\|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\|^2 + +20\alpha_rE(\tau+\nu)\epsilon_0^2. \end{aligned} \tag{40}$$

□

D. Upper Bound for Discrepancy

Lemma D.1 (Proposition B.5, (Yuan et al., 2021)). *Let $\omega : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed μ_ω-strongly convex function, for $\boldsymbol{z} \in \mathbb{R}^d$ we define*

$$\nabla(\omega + h)^*(\boldsymbol{z}) = \arg\min_{\boldsymbol{w}} \{\langle -\boldsymbol{z}, \boldsymbol{w}\rangle + \omega(\boldsymbol{w}) + h(\boldsymbol{w})\},$$

then it holds that

$$\|\nabla(\omega + h)^*(\boldsymbol{z}) - \nabla(\omega + h)^*(\boldsymbol{y})\| \leq 1/\mu_\omega \|\boldsymbol{z} - \boldsymbol{y}\|_*.$$

D.1. Deferred Proof of Lemma B.5

Proof of Lemma B.5. Reacll the definitions of $\boldsymbol{w}_t^k$ and $\boldsymbol{w}_t$

$$\boldsymbol{w}_t^k = \arg\min_{\boldsymbol{w}\in\mathcal{W}(\epsilon_0;\boldsymbol{w}_0)} \left\{\langle \boldsymbol{w}, \boldsymbol{g}_{t-1}^k - \frac{\mu}{2}\widetilde{\boldsymbol{w}}_{t-1}^k - \gamma_{t-1}\boldsymbol{w}_0\rangle + \left(\frac{A_{t-1}\mu}{2} + \gamma_{t-1}\right)\frac{\|\boldsymbol{w}\|^2}{2} + A_{t-1}h(\boldsymbol{w})\right\}$$

and

$$\boldsymbol{w}_t = \arg\min_{\boldsymbol{w}\in\mathcal{W}(\epsilon_0;\boldsymbol{w}_0)} \left\{\langle \boldsymbol{w}, \boldsymbol{g}_{t-1} - \frac{\mu}{2}\widetilde{\boldsymbol{w}}_{t-1} - \gamma_{t-1}\boldsymbol{w}_0\rangle + \left(\frac{A_{t-1}\mu}{2} + \gamma_{t-1}\right)\frac{\|\boldsymbol{w}\|^2}{2} + A_{t-1}h(\boldsymbol{w})\right\}.$$

Since the synchronization at step t_r, we have $\boldsymbol{g}_{t-1}^k - \boldsymbol{g}_{t-1} = \sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{G}_i^k - \sum_{l=1}^K \pi_l \boldsymbol{G}_i^l)$ and $\widetilde{\boldsymbol{w}}_{t-1}^k - \widetilde{\boldsymbol{w}}_{t-1} = \sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{w}_i^k - \sum_{l=1}^K \pi_l \boldsymbol{w}_i^l)$ for $t_r \leq t-1 \leq t_{r+1}-1$. Then applying Lemma D.1, it holds that

$$\begin{aligned}
\|\boldsymbol{w}_t^k - \boldsymbol{w}_t\| &\leq \frac{1}{\mu A_{t-1}/2 + \gamma_{t-1}}\left(\|\boldsymbol{g}_{t-1}^k - \boldsymbol{g}_{t-1}\| + \frac{\mu}{2}\|\widetilde{\boldsymbol{w}}_{t-1}^k - \widetilde{\boldsymbol{w}}_{t-1}\|\right) \\
&\leq \frac{1}{\mu A_{t-1}/2 + \gamma_{t-1}}\left(\left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{G}_i - \boldsymbol{G}_i^k)\right\| + \frac{\mu}{2}\left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{w}_i^k - \check{\boldsymbol{w}}_i)\right\|\right) \\
&\leq \frac{1}{\mu A_{t-1}/2 + \gamma_{t-1}}\left(\sum_{l=1}^K \pi_l \left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{G}_i^l - \boldsymbol{G}_i^k)\right\| + \mu\rho(A_{t-1} - A_{t_r-1})\right),
\end{aligned} \tag{41}$$

where we used ρ-bounded domain in the last inequality. Let $\boldsymbol{\Delta}_i^k = \boldsymbol{G}_i^k - \nabla\mathcal{L}_k(\boldsymbol{w}_i^k)$, then we may decompose the difference of local stochastic gradients as

$$\begin{aligned}
\left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{G}_i^l - \boldsymbol{G}_i^k)\right\| &\leq \left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^k\right\| + \left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^l\right\| + \sum_{i=t_r}^{t-1}\alpha_i\left\|\nabla\mathcal{L}_l(\boldsymbol{w}_i^l) - \nabla\mathcal{L}_k(\boldsymbol{w}_i^k)\right\| \\
&\leq \left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^k\right\| + \left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^l\right\| + \sum_{i=t_r}^{t-1}\alpha_i\left\|\nabla\mathcal{L}_l(\boldsymbol{w}_i^l) - \nabla\mathcal{L}(\boldsymbol{w}_i^l)\right\| \\
&\quad + \sum_{i=t_r}^{t-1}\alpha_i\left\|\nabla\mathcal{L}_k(\boldsymbol{w}_i^k) - \nabla\mathcal{L}(\boldsymbol{w}_i^k)\right\| + \sum_{i=t_r}^{t-1}\alpha_i\left\|\nabla\mathcal{L}(\boldsymbol{w}_i^l) - \nabla\mathcal{L}(\boldsymbol{w}_i^k)\right\| \\
&\leq \left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^k\right\| + \left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^l\right\| + 2(A_{t-1} - A_{t_r-1})(H + \Lambda\rho),
\end{aligned} \tag{42}$$

where the third inequality follows from the bounded heterogeneity assumption and Λ-smoothness of global loss $\mathcal{L}$. By the conditional dependence, we have

$$\mathbb{E}_{\mathcal{D}}\left[\left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^k\right\|^2 |\mathcal{F}_{t_r}\right] = \sum_{i=t_r}^{t-1}\alpha_i^2\mathbb{E}_{\mathcal{D}}[\mathbb{E}(\|\boldsymbol{\Delta}_i^k\|^2|\mathcal{F}_i)|\mathcal{F}_{t_r}] \leq E\alpha_t^2\sigma^2$$

Taking expectation on both sides of (42) and using the relation above, we have

$$\mathbb{E}_{\mathcal{D}}\left[\left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{G}_i^l-\boldsymbol{G}_i^k)\right\|^2\Big|\mathcal{F}_{t_r}\right]\le 2\mathbb{E}_{\mathcal{D}}\left[\left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^k\right\|^2+\left\|\sum_{i=t_r}^{t-1}\alpha_i\boldsymbol{\Delta}_i^l\right\|^2\Big|\mathcal{F}_{t_r}\right]+4(A_{t-1}-A_{t_r-1})^2(H+\Lambda\rho)^2$$
$$\le 4E\alpha_t^2\sigma^2+4\alpha_t^2E^2(H+\Lambda\rho)^2, \tag{43}$$

where the last inequality follows from $(A_{t-1}-A_{t_r-1})/\alpha_t\le E$. Combining (41) and (43), together with $\|\boldsymbol{w}_i^k-\boldsymbol{w}_i^l\|\le 2\rho$, we are guaranteed that

$$\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{w}_t^k-\boldsymbol{w}_t\|^2]\le\frac{4E\sigma^2\alpha_t^2}{(\mu A_t/2+\gamma_t)^2}+\frac{4\alpha_t^2E^2(H+\Lambda\rho)^2}{(\mu A_t/2+\gamma_t)^2}.$$

Similarly, $\mathbb{E}_{\mathcal{D}}[\|\check{\boldsymbol{w}}_t-\boldsymbol{w}_t^k\|^2]$ shares the same upper bound with $\mathbb{E}[\|\boldsymbol{w}_t^k-\boldsymbol{w}_t\|^2]$ due to the following relation

$$\|\boldsymbol{w}_t^k-\widetilde{\boldsymbol{w}}_t\|\le\sum_{l=1}^K\pi_l\|\boldsymbol{w}_t^k-\boldsymbol{w}_t^l\|$$
$$\le\frac{1}{\mu A_{t-1}/2+\gamma_{t-1}}\sum_{l=1}^K\pi_l\left(\left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{G}_i^k)-\boldsymbol{G}_i^l\right\|+\frac{\mu}{2}\left\|\sum_{i=t_r}^{t-1}\alpha_i(\boldsymbol{w}_i^k-\boldsymbol{w}_i^l)\right\|\right).$$

□

D.2. Deferred Proof of Lemma B.7

Proof of Lemma B.7. Recalling the definitions of $\bar{\boldsymbol{w}}_r$ and $\boldsymbol{w}_i^k$ for $t_r\le i\le t_{r+1}-1$:

$$\bar{\boldsymbol{w}}_{r+1}=\arg\min_{\boldsymbol{w}\in\mathcal{W}(\epsilon_0;\boldsymbol{w}_0)}\left\{\langle\boldsymbol{w},\boldsymbol{g}_{t_{r+1}-1}-\frac{\mu E}{2}\sum_{j=0}^r\alpha_j\bar{\boldsymbol{w}}_j-\gamma_rE\bar{\boldsymbol{w}}_0\rangle+\left(\frac{A_r\mu}{2}+\gamma_r\right)E\frac{\|\boldsymbol{w}\|^2}{2}+A_rEh(\boldsymbol{w})\right\}$$
$$\boldsymbol{w}_i^k=\arg\min_{\boldsymbol{w}\in\mathcal{W}(\epsilon_0;\boldsymbol{w}_0)}\left\{\langle\boldsymbol{w},\boldsymbol{g}_{i-1}^k-\frac{\mu E}{2}\sum_{j=0}^r\alpha_j\bar{\boldsymbol{w}}_j-\gamma_rE\bar{\boldsymbol{w}}_0\rangle+\left(\frac{A_r\mu}{2}+\gamma_r\right)E\frac{\|\boldsymbol{w}\|^2}{2}+A_rEh(\boldsymbol{w})\right\},$$

where $\boldsymbol{g}_{t_{r+1}-1}=\boldsymbol{g}_{t_r-1}+\sum_{j=t_r}^{t_{r+1}-1}\alpha_r\boldsymbol{G}_j$ and $\boldsymbol{g}_i^k=\boldsymbol{g}_{t_r-1}+\sum_{j=t_r}^i\alpha_r\boldsymbol{G}_j^k$. Using Lemma D.1 and similar decomposition in (42), we have

$$\|\boldsymbol{w}_i^k-\bar{\boldsymbol{w}}_{r+1}\|\le\frac{1}{A_rE\mu/2+\gamma_rE}\|\boldsymbol{g}_{i-1}^k-\boldsymbol{g}_{t_{r+1}-1}\|\le\frac{1}{A_rE\mu/2+\gamma_rE}\left\|\sum_{j=i}^{t_{r+1}-1}\alpha_r(\boldsymbol{G}_j-\boldsymbol{G}_j^k)\right\|$$
$$\le\frac{1}{A_rE\mu/2+\gamma_rE}\sum_{l=1}^K\pi_l\left\|\sum_{j=i}^{t_{r+1}-1}\alpha_r(\boldsymbol{G}_j^l-\boldsymbol{G}_j^k)\right\| \tag{44}$$
$$\le\frac{\alpha_r}{A_rE\mu/2+\gamma_rE}\sum_{l=1}^K\pi_l\left(\left\|\sum_{j=i}^{t_{r+1}-1}\boldsymbol{\Delta}_j^k\right\|+\left\|\sum_{j=i}^{t_{r+1}-1}\boldsymbol{\Delta}_j^l\right\|+2(t_{r+1}-i)(H+\Lambda\rho)\right).$$

Applying Lemma B.3, we can obtain that

$$\left\|\sum_{j=i}^{t_{r+1}-1}\boldsymbol{\Delta}_j^k\right\|\le 2\sqrt{2\log(1/(4\delta))}\left\{\left(\sum_{j=i}^{t_{r+1}-1}\mathbb{E}\|\boldsymbol{\Delta}_j^k\|^2\right)^{1/2}+\left(\sum_{j=i}^{t_{r+1}-1}\|\boldsymbol{\Delta}_j^k\|^2\right)^{1/2}\right\} \tag{45}$$

holds with probability at least $1-\delta/2$. Using the light-tailed assumption and Lemma B.1, we are guaranteed that $\mathbb{E}\|\boldsymbol{\Delta}_j^k\|^2\le\sigma^2$ and

$$\sum_{j=i}^{t_{r+1}-1}\|\boldsymbol{\Delta}_j^k\|^2\le E\sigma^2+\max\left\{8\sigma^2\log(2/\delta),16\sigma^2\sqrt{E\log(2/\delta)}\right\} \tag{46}$$

holds with probability at least $1-\delta/2$. Substituting (45) and (46) into (44), it holds that

$$\begin{aligned}\|\boldsymbol{w}_i^k - \bar{\boldsymbol{w}}_{r+1}\| &\leq \frac{2\alpha_r}{E\mu A_r/2+\gamma_t E}\left(\sqrt{2\log(1/(4\delta))}\left(2\sqrt{E}\sigma + 4\sqrt{E}\sigma\sqrt{\log(2/\delta)}\right) + 2E(H+\Lambda\rho)\right)\\ &\leq \frac{2\alpha_r}{E\mu A_r/2+\gamma_t E}\left(8\sqrt{E}\sigma\log(2/\delta) + 2E(H+\Lambda\rho)\right)\end{aligned}$$

with probability at least $1-\delta$. □

E. Proof of Lemma B.3

This lemma is a martingale's version of Lemma 3.1 in (He & Shao, 2000), we provide the proof for completeness.

Proof of Lemma B.3. Without loss of generality, we consider that $x > 16$. If $x \leq 16$, we may modify the tail probability in Lemma B.3 as $100\exp(-x^2/100)$. Let $\{\zeta_i\}_{i=1}^{\infty}$ be an independent copy of $\{\xi_i\}_{i=1}^{\infty}$, which is also adapted to $\{\mathcal{F}_i\}_{i=1}^{\infty}$. By Chebyshev's inequality, we have

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{i=1}^t \zeta_i\right\| \leq 2B_t,\ \sum_{i=0}^t \|\zeta_i\|^2 \leq 2B_t^2\right) \geq& 1 - \mathbb{P}\left(\left\|\sum_{i=1}^t \zeta_i\right\| > 2B_t\right) - \mathbb{P}\left(\sum_{i=0}^t \|\zeta_i\|^2 > 2B_t^2\right)\\ \geq& 1 - \mathbb{P}\left(\left\|\sum_{i=1}^t \zeta_i\right\| > 2B_t\right) - \frac{1}{2}\\ \geq& 1 - \frac{\mathbb{E}\left\|\sum_{i=1}^t \zeta_i\right\|^2}{4B_t^2} - 1/2\\ \geq& 1 - 1/4 - 1/2 = 1/4,\end{aligned} \tag{47}$$

where the last inequality follows from $B_t^2 = \mathbb{E}\|\sum_{i=1}^t \zeta_i\|^2 = \sum_{i=1}^t \mathbb{E}\|\zeta_i\|^2$ due to the martingale property. Let $\{\varepsilon_i\}_{i=1}^t$ be a Rademacher sequence independent of $\{\xi_i\}_{i=1}^t$ and $\{\zeta_i\}_{i=1}^t$. With slightly abusing notation, we denote $S_t = (\sum_{i=1}^t(\|\xi_i - \zeta_i\|^2))^{1/2}$. We assume the following event holds

$$\left\{\left\|\sum_{i=1}^t \xi_i\right\| \geq x\left(B_t + \left(\sum_{i=1}^t \|\xi_i\|^2\right)^{1/2}\right),\ \left\|\sum_{i=1}^t \zeta_i\right\| \leq 2B_t,\ \sum_{i=0}^t \|\zeta_i\|^2 \leq 2B_t^2\right\}.$$

Then we notice that

$$\begin{aligned}\left\|\sum_{i=1}^t \xi_i - \zeta_i\right\| \geq \left\|\sum_{i=1}^t \xi_i\right\| - \left\|\sum_{i=1}^t \zeta_i\right\| &\geq x\left(B_t + \left(\sum_{i=1}^t \|\xi_i\|^2\right)^{1/2}\right) - 2B_t\\ &\geq (x-2)B_t + \frac{x}{2}\left(\sum_{i=1}^t \|\xi_i\|^2\right)^{1/2} + \frac{x}{2}\left(\sum_{i=1}^t \|\zeta_i\|^2\right)^{1/2} - \frac{\sqrt{2}x}{2}B_t\\ &\overset{(a)}{\geq} \left(\frac{x}{4} - 2\right)B_t + \frac{x}{2}\left(\sum_{i=1}^t \|\xi_i - \zeta_i\|^2\right)^{1/2}\\ &\overset{(b)}{\geq} \frac{x}{2}\left(\sum_{i=1}^t \|\xi_i - \zeta_i\|^2\right)^{1/2},\end{aligned}$$

where the inequality (a) follows from the triangle inequality and the inequality (b) holds since $x > 16$. The relation above

implies that

$$\left\{\left\|\sum_{i=1}^{t}\xi_i\right\| \geq x\left(B_t+\left(\sum_{i=1}^{t}\|\xi_i\|^2\right)^{1/2}\right),\ \left\|\sum_{i=1}^{t}\zeta_i\right\| \leq 2B_t,\ \sum_{i=0}^{t}\|\zeta_i\|^2 \leq 2B_t^2\right\} \subseteq \left\{\left\|\sum_{i=1}^{t}\xi_i-\zeta_i\right\| \geq \frac{x}{2}S_t\right\}. \tag{48}$$

Using the dependence of ξ_i and ζ_i, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=1}^{t}\xi_i\right\| \geq x\left(B_t+\left(\sum_{i=1}^{t}\|\xi_i\|^2\right)^{1/2}\right)\right)\\
=&\mathbb{P}\left(\left\|\sum_{i=1}^{t}\xi_i\right\| \geq x\left(B_t+\left(\sum_{i=1}^{t}\|\xi_i\|^2\right)^{1/2}\right),\ \left\|\sum_{i=1}^{t}\zeta_i\right\| \leq 12B_t,\ \sum_{i=0}^{t}\|\zeta_i\|^2 \leq 2B_t^2\right)\\
\times&\mathbb{P}\left(\left\|\sum_{i=1}^{t}\zeta_i\right\| \leq 12B_t,\ \sum_{i=0}^{t}\|\zeta_i\|^2 \leq 2B_t^2\right)\\
\leq&4\mathbb{P}\left(\left\|\sum_{i=1}^{t}\xi_i-\zeta_i\right\| \geq \frac{x}{2}S_t\right),
\end{aligned} \tag{49}$$

where the last inequality follows from (47) and (48). Note that $\{\xi_i-\zeta_i\}_{i=1}^t$ is a symmetric martingale difference sequence. Then using double expectation (given $\mathcal{F}_t$, ξ_i and ζ_i for $1\leq i\leq t$ are fixed), we have

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{i=1}^{t}\xi_i-\zeta_i\right\| \geq \frac{x}{2}S_t\right) &= \mathbb{P}\left(\left\|\sum_{i=1}^{t}(\xi_i-\zeta_i)\varepsilon_i\right\| \geq \frac{x}{2}S_t\right)\\
&= \mathbb{E}\left\{\mathbb{P}\left(\left\|\sum_{i=1}^{t}(\xi_i-\zeta_i)\varepsilon_i\right\| \geq \frac{x}{2}S_t\Big|\mathcal{F}_t\right)\right\}\\
&\leq \mathbb{E}\left\{2\exp\left(-\frac{x^2S_t^2}{8\sum_{i=1}^{t}\|\xi_i-\zeta_i\|^2}\right)\right\}\\
&\leq 2\exp\left(-\frac{x^2}{8}\right),
\end{aligned}$$

where the first inequality follows from the exponential inequality for Rademacher sequence (see, e.g., Ledoux & Talagrand (1991), p.101). Plugging this upper bound into (49), we are guaranteed that

$$\mathbb{P}\left(\left\|\sum_{i=1}^{t}\xi_i\right\| \geq x\left(B_t+\left(\sum_{i=1}^{t}\|\xi_i\|^2\right)^{1/2}\right)\right) \leq 8\exp\left(-\frac{x^2}{8}\right).$$

□

F. Additional Results in Section 4

Federated sparse linear regression. For `MC-FedDA`, we set the number of stages $M=3$ and use the regularization sequence $\{0.5^3, 0.5^4, 0.5^5\}$ for regularization parameters in 3 stages. For other methods, the regularization parameter is $\lambda=0.5^5$. The hyperparameters for `Fast-FedDA` are $\mu=0.1$ and $L=550$. For `C-FedDA` and `MC-FedDA`, we choose $\mu=0.1$ and $L=600$. For `FedDA` and `FedMiD`, we set the server learning rate $\eta_s=1.0$ and tuned the client learning rate η_c by selecting the best performing value over the set $\{0.0001, 0.001, 0.01, 0.1\}$, which was 0.001 for both baselines.

Federated low-rank matrix estimation. For `MC-FedDA`, we set the number of stages $M=3$ and use the sequence $\{0.3, 0.15, 0.1\}$ for regularization parameters in 3 stages. For other methods, the regularization parameter is $\lambda=0.1$. The choices for hyperparameters follow the same setting in sparse linear regression.

Federated sparse logistic regression. The experimental results on EMNIST-10 and EMNIST-62 are reported in Figure 4 and Figure 3 respectively. For the two baselines (`FedMid` and `FedDA`), we set the server learning rate $\eta_s = 1.0$ and tuned the client learning rate η_c by selecting the best performing value over the set $\{0.001, 0.003, 0.01, 0.03, 0.1\}$, which was $\eta_c = 0.01$ for both baselines. For our proposed algorithms, we tuned μ and γ by selecting the best performing values over the sets $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$ and $\{10, 25, 50, 100\}$, respectively. The best values were $\mu = 0.001$ and $\gamma = 25$ for all proposed algorithms. For `MC-FedDA`, we use the regularization sequence $\{0.000225, 0.00015, 0.0001, 0.0001, 0.0001\}$.

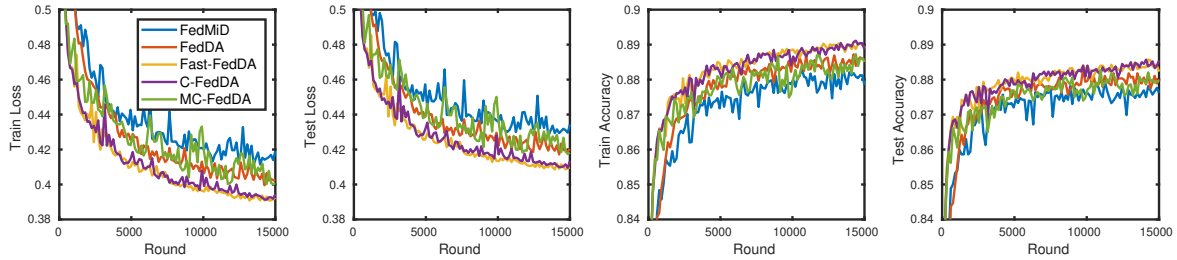

Figure 4. Results for federated sparse logistic regression on EMNIST-10 dataset. Our proposed algorithms `Fast-FedDA` and `C-FedDA` reach a lower loss and higher accuracy than the two baselines from (Yuan et al., 2021), and exhibit faster convergence.