Minimax Estimation for Personalized Federated Learning: An Alternative between FedAvg and Local Training?

Shuxiao Chen[⋆] shuxiaoc@wharton.upenn.edu
Qinqing Zheng[†] zhengqinqing@gmail.com
Qi Long[⋆] qlong@upenn.edu
Weijie J. Su[⋆] suw@wharton.upenn.edu
University of Pennsylvania[⋆]
Meta AI Research[†]

Editor: Ambuj Tewari

Abstract

A widely recognized difficulty in federated learning arises from the statistical heterogeneity among clients: local datasets often originate from distinct yet not entirely unrelated probability distributions, and personalization is, therefore, necessary to achieve optimal results from each individual's perspective. In this paper, we show how the excess risks of personalized federated learning using a smooth, strongly convex loss depend on data heterogeneity from a minimax point of view, with a focus on the FedAvg algorithm (McMahan et al., 2017) and pure local training (i.e., clients solve empirical risk minimization problems on their local datasets without any communication). Our main result reveals an *approximate* alternative between these two baseline algorithms for federated learning: the former algorithm is minimax rate optimal over a collection of instances when data heterogeneity is small, whereas the latter is minimax rate optimal when data heterogeneity is large, and the threshold is sharp up to a constant.

As an implication, our results show that from a worst-case point of view, a dichotomous strategy that makes a choice between the two baseline algorithms is rate-optimal. Another implication is that the popular FedAvg following by local fine tuning strategy is also minimax optimal under additional regularity conditions. Our analysis relies on a new notion of algorithmic stability that takes into account the nature of federated learning.

Keywords: Empirical Risk Minimization, Federated Learning, Personalization, Data Heterogeneity, Minimax Rates, Algorithmic Stability

1. Introduction

As one of the most important ingredients driving the success of machine learning, data are being generated and subsequently stored in an increasingly decentralized fashion in many real-world applications. For example, mobile devices will in a single day collect an unprecedented amount of data from users. These data commonly contain sensitive information such as web search histories, online shopping records, and health information, and thus are often not available to service providers (Poushter, 2016). This decentralized nature of (sensitive) data poses substantial challenges to many machine learning tasks.

To address this issue, McMahan et al. (2017) proposed a new learning paradigm, which they termed *federated learning*, for collaboratively training machine learning models on data that are locally possessed by multiple clients with the coordination of the central server (e.g., service provider), without having direct access to the local datasets. In its simplest form, federated learning considers a pool of m clients, where the i-th client has a local dataset S_i of size n_i, consisting of i.i.d. samples $\{z_j^{(i)} : j \in [n_i]\}$ (denote $[n] := \{1, 2, \ldots, n\}$) from some unknown distribution $\mathcal{D}_i$. Letting $\ell(\boldsymbol{w}, z)$ be a loss function, where $\boldsymbol{w}$ denotes the model parameter, the optimal local model for the i-th client is given by

$$\boldsymbol{w}_\star^{(i)} \in \operatorname*{argmin}_{\boldsymbol{w}} \mathbb{E}_{Z_i \sim \mathcal{D}_i} \ell(\boldsymbol{w}, Z_i). \tag{1}$$

From the *client-wise* perspective, any data-dependent estimator $\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})$, with $\boldsymbol{S} = \{S_i\}_{i=1}^m$ denoting the collection of all samples, can be evaluated based on its individualized excess risk:

$$\mathrm{IER}_i := \mathbb{E}_{Z_i \sim \mathcal{D}_i}[\ell(\widehat{\boldsymbol{w}}^{(i)}, Z_i) - \ell(\boldsymbol{w}_\star^{(i)}, Z_i)],$$

where the expectation is taken over a fresh sample $Z_i \sim \mathcal{D}_i$. At a high level, this learning paradigm of federated learning aims to obtain possibly different trained models for each client such that the individualized excess risks are low (see, e.g., Kairouz et al. 2019).

From a statistical viewpoint, perhaps the most crucial factor in determining the effectiveness of federated learning is *data heterogeneity*. When the data distribution $\mathcal{D}_i$ is (approximately) homogeneous across different clients, presumably a *single* global model would lead to small IER_i for all i. In this regime, indeed, McMahan et al. (2017) proposed the *federated averaging* algorithm (FedAvg, see Algorithm 1), which can be regarded as an instance of local stochastic gradient descent (SGD) for solving (Mangasarian and Solodov, 1993; Stich, 2019)

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i \in [m]} n_i L_i(\boldsymbol{w}, S_i), \tag{2}$$

where $L_i(\boldsymbol{w}, S_i) := \sum_{j \in [n_i]} \ell(\boldsymbol{w}, z_j^{(i)})/n_i$ is the empirical risk minimization (ERM) objective of the i-th client and $N = n_1 + \cdots + n_m$ denotes the total number of training samples. Translating Algorithm 1 into words, FedAvg in effect learns a shared global model using gradients from each client and outputs a single model as an estimate of $\boldsymbol{w}_\star^{(i)}$ for all clients. When the distributions $\{\mathcal{D}_i\}$ coincide with each other, FedAvg with a strongly convex loss achieves a weighted average excess risk of $\mathcal{O}(1/N)$, which is minimax optimal up to a constant factor (Shalev-Shwartz et al., 2009; Agarwal et al., 2012), see the formal statement in Theorem 6.

However, it is an entirely different story in the presence of data heterogeneity. FedAvg has been recognized to give inferior performance when there is a significant departure from complete homogeneity (see, e.g., Bonawitz et al. 2019). To better understand this point, consider the extreme case where the data distributions $\{\mathcal{D}_i\}$ are entirely unrelated. This roughly amounts to saying that the model parameters $\{\boldsymbol{w}_\star^{(i)}\}$ can be arbitrarily different from each other. In such a "completely heterogeneous" scenario, the objective function (2) simply has no clear interpretation, and any single global model—for example, the output of FedAvg—would lead to unbounded risks for most, if not all, clients. As a matter of fact, it is not difficult to see that the optimal training strategy for federated learning in this regime

Algorithm 1: FedAvg (McMahan et al., 2017)

```
Input: initialize $\boldsymbol{w}_0^{(\text{global})}$, number of communication rounds $T$, step sizes $\{\eta_t\}_{t=0}^{T-1}$
for $t = 0, 1, \ldots, T-1$ do
    Randomly sample a batch of clients $\mathcal{C}_t \subseteq [m]$
    for client $i \in \mathcal{C}_t$ do
        Obtain $\boldsymbol{w}_{t+1}^{(i)}$ by running several steps of SGD on $S_i$ using $\boldsymbol{w}_t^{(\text{global})}$ as the initialization
    $\boldsymbol{w}_{t+1}^{(\text{global})} \leftarrow \boldsymbol{w}_t^{(\text{global})} - \frac{m\eta_t}{N|\mathcal{C}_t|}\sum_{i\in\mathcal{C}_t} n_i(\boldsymbol{w}_t^{(\text{global})} - \boldsymbol{w}_{t+1}^{(i)})$
Output: $\widehat{\boldsymbol{w}}^{(i)} = \boldsymbol{w}_T^{(\text{global})}$, $i \in [m]$
```

is arguably PureLocalTraining, which lets each client separately run SGD to minimize its own local ERM objective

$$\min_{\boldsymbol{w}^{(i)}} L_i(\boldsymbol{w}^{(i)}, S_i) \tag{3}$$

without any communication. Indeed, PureLocalTraining is minimax rate optimal in the completely heterogeneous regime, just as FedAvg in the completely homogeneous regime (see Theorem 4).

The level of data heterogeneity in practical federated learning problems is apparently neither complete homogeneity nor complete heterogeneity. Thus, the foregoing discussion raises a pressing question of what would happen if we are in the *wide* middle ground of the two extremes. This underlines the essence of *personalized federated learning*, which seeks to develop algorithms that perform well over a wide spectrum of data heterogeneity. Despite a venerable line of work on personalized federated learning (see, e.g., Kulkarni et al. 2020), the literature remains relatively silent on how the *fundamental* limits of personalized federated learning depend on data heterogeneity, as opposed to two extreme cases where both the minimax optimal rates and algorithms are known.

1.1 Main Contributions

The present paper takes a step toward understanding the statistical limits of personalized federated learning by establishing the minimax rates of convergence for both individualized excess risks and their weighted average with smooth strongly convex losses. We briefly summarize our main contributions below.

1. We prove that if the client-wise sample sizes are relatively balanced, then there exists a problem instance on which the IER_i's of any algorithm are lower bounded by

$$\begin{cases} \Omega(1/N + R^2) & \text{if } R^2 = \mathcal{O}(m/N) \\ \Omega(m/N) & \text{if } R^2 = \Omega(m/N), \end{cases} \tag{4}$$

   where R is the minimum quantity satisfying $\min_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[m]} n_i \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}\|^2/N \leq R^2$, i.e., it measures the maximum level of heterogeneity among clients (here $\|\cdot\|$ throughout the paper denotes the Euclidean distance). Meanwhile, we show that the IER_i's of FedAvg are upper bounded by $\mathcal{O}(1/N + R^2)$, whereas the guarantee for

PureLocalTraining is $\mathcal{O}(m/N)$, regardless of the specific value of R. Moreover, we also establish similar upper and lower bounds for a weighted average of the IER_i's under a weaker condition.

2. A closer look at the above-mentioned bounds reveals a perhaps surprising phenomenon: for a given collection of problem instances with a specified maximum level of heterogeneity, exactly one of FedAvg or PureLocalTraining is minimax optimal.

3. The established minimax results suggest that the naïve dichotomous strategy of (1) running FedAvg when $R^2 = \mathcal{O}(m/N)$, and (2) running PureLocalTraining when $R^2 = \Omega(m/N)$, attains the lower bound (4). Moreover, for supervised problems, this dichotomous strategy can be implemented without knowing R by (1) running both FedAvg and PureLocalTraining, (2) evaluating the test errors of the two algorithms in a distributed fashion, and (3) deploying the algorithm with a lower test error. We emphasize that the notion of optimality under our consideration overlooks constant factors. In practice, a better personalization result could be achieved by more sophisticated algorithms.

4. As a side product, we provide a novel analysis of FedProx, a popular algorithm for personalized federated learning that constrains the learned local models to be close via ℓ_2 regularization (Li et al., 2018). In particular, we show that its IER_i's are of order $\mathcal{O}\big(\frac{1}{N/m} \wedge \frac{R}{\sqrt{N/m}} + \frac{\sqrt{m}}{N}\big)$, and a weighted average of the IER_i's satisfies a tighter $\mathcal{O}\big(\frac{1}{N/m} \wedge \frac{R}{\sqrt{N/m}} + \frac{1}{N}\big)$ bound, where $a \wedge b = \min\{a, b\}$ for two real numbers a and b.

5. On the technical side, our upper bound analysis is based on a generalized notion of algorithmic stability (Bousquet and Elisseeff, 2002), which we term *federated stability* and can be of independent interest. Briefly speaking, an algorithm $\mathcal{A}(\boldsymbol{S}) = \{\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\}$ has federated stability $\{\gamma_i\}$ if for any $i \in [m]$, the loss function evaluated at $\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})$ can only change by an additive term of $\mathcal{O}(\gamma_i)$, if we perturb S_i a little bit, while keeping the rest of datasets $\{S_{i'} : i' \neq i\}$ fixed. Similar ideas have appeared in Maurer (2005) and have been recently applied to multi-task learning (Wang et al., 2018). However, their notion of perturbation is based on the deletion of the whole client-wise dataset, whereas our notion of federated stability operates at the "record-level" and is more fine-grained. On the other hand, our construction of the lower bound is based on a generalization of Assound's lemma (Assouad, 1983) (see also Yu 1997), which enables us to handle multiple heterogeneous datasets.

1.2 Related Work

Ever since the proposal of federated learning by McMahan et al. (2017), recent years have witnessed a rapidly growing line of work that is concerned with various aspects of FedAvg and its variants (see, e.g., Khaled et al. 2019; Haddadpour and Mahdavi 2019; Li et al. 2020b; Bayoumi et al. 2020; Malinovsky et al. 2020; Li and Richtárik 2020; Woodworth et al. 2020; Yuan and Ma 2020; Zheng et al. 2021).

In the context of personalized federated learning, there have been significant algorithmic developments in recent years. While the idea of using ℓ_2 regularization to constrain the learned models to be similar has appeared in early works on multi-task learning (Evgeniou and Pontil, 2004), its applicability to personalized federated learning was only recently demonstrated by Li et al. (2018), where the FedProx algorithm was introduced. Similar regularization-based methods have been proposed and analyzed from the scope of convex optimization in Hanzely and Richtárik (2020); Dinh et al. (2020), and Hanzely et al. (2020). In particular, Hanzely et al. (2020) showed that an accelerated variant of FedProx is optimal in terms of communication complexity and the local oracle complexity. There is also a line of work using model-agnostic meta learning (Finn et al., 2017) to achieve personalization (Jiang et al., 2019; Fallah et al., 2020). Other strategies have been proposed (see, e.g., Arivazhagan et al. 2019; Li and Wang 2019; Mansour et al. 2020; Yu et al. 2020), and we refer readers to Kulkarni et al. (2020) for a comprehensive survey. We briefly remark here that all the papers mentioned above only consider the *optimization properties* of their proposed algorithms, while we focus on statistical properties of personalized federated learning.

Compared to the optimization understanding, our statistical understanding (in terms of sample complexity) of federated learning is still limited. Deng et al. (2020) proposed an algorithm for personalized federated learning with learning-theoretic guarantees. However, it is unclear how their bound scales with the heterogeneity among clients.

More generally, exploiting the information "shared among multiple learners" is a theme that constantly appears in other fields of machine learning such as multi-task learning (Caruana, 1997), meta learning (Baxter, 2000), and transfer learning (Pan and Yang, 2009), from which we borrow a lot of intuitions (see, e.g., Ben-David et al. 2006; Ben-David and Borbely 2008; Ben-David et al. 2010; Maurer et al. 2016; Cai and Wei 2019; Hanneke and Kpotufe 2019, 2020; Du et al. 2020; Tripuraneni et al. 2020a,b; Kalan et al. 2020; Shui et al. 2020; Li et al. 2020a; Zhang et al. 2020; Jose and Simeone 2021).

More related to our work, a series work by Denevi et al. (2018, 2019); Balcan et al. (2019), and Khodak et al. (2019) assumes the optimal local models lie in a small sub-parameter-space, and establishes "heterogeneity-aware" bounds on a weighted average of individualized excess risks. However, we would like to point out that they operate under the online learning setup, where the datasets are assumed to come in streams, and this is in sharp contrast to the federated learning setup, where the datasets are decentralized. Our notion of heterogeneity is also related to the hierarchical Bayesian model considered in Bai et al. (2020); Lucas et al. (2020); Konobeev et al. (2020), and Chen et al. (2020).

1.3 Paper Organization

The rest of this paper is organized as follows. In Section 2, we give an exposition of the problem setup and main assumptions. Section 3 presents our main results with proof sketches. We conclude this paper with a discussion of open problems in Section 4. For brevity, detailed proofs are deferred to the appendix.

2. Problem Setup

In this section, we detail some preliminaries to prepare the readers for our main results.

Notation. We introduce the notation we are going to use throughout this paper. For two real numbers a, b, we let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For two non-negative sequences a_n, b_n, we denote $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) if $a_n \leq Cb_n$ (resp. $a_n \geq Cb_n$) for some constant $C > 0$ when n is sufficiently large. We use $a_n \asymp b_n$ to indicate that $a_n \gtrsim b_n, a_b \lesssim b_n$ hold simultaneously. We also use $a_n = \mathcal{O}(b_n)$, whose meaning is the same as $a_n \lesssim b_n$, and $a_n = \Omega(b_n)$, whose meaning is the same as $a_n \gtrsim b_n$. For two probability distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, we use $\mathcal{D}_1 \otimes \mathcal{D}_2$ to denote their joint distribution under independence. We use $\mathcal{W}$ to denote the parameter space and $\mathcal{Z}$ to denote the sample space. Finally, we let $\mathcal{P}_{\mathcal{W}}(x) := \operatorname{argmin}_{y \in \mathcal{W}} \|x - y\|$ denote the operator that projects x onto $\mathcal{W}$ in Euclidean distance.

Evaluation Metrics. The presentation of our main results relies on how to evaluate the performance of a federated learning algorithm. To this end, we consider the following two evaluation metrics.

Definition 1 (Individualized excess risk) *Consider an algorithm $\mathcal{A}$ that outputs $\mathcal{A}(\boldsymbol{S}) = \{\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\}_{i=1}^m$. For the i-the client, its* individualized excess risk *(IER) is defined as*

$$\mathrm{IER}_i(\mathcal{A}) := \mathbb{E}_{Z_i \sim \mathcal{D}_i}[\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), Z_i) - \ell(\boldsymbol{w}_\star^{(i)}, Z_i)], \tag{5}$$

where $Z_i \sim \mathcal{D}_i$ is a fresh data point independent of $\boldsymbol{S}$.

Definition 2 (p-average excess risk) *Consider an algorithm $\mathcal{A}$ that outputs $\mathcal{A}(\boldsymbol{S}) = \{\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\}$. For a vector $\mathbf{p} = (p_1, \ldots, p_m)$ lying in the m-dimensional probability simplex (i.e., all p_i's are non-negative and they sum to one), we define the $\mathbf{p}$-average excess risk ($\mathrm{AER}_{\mathbf{p}}$) of $\mathcal{A}$ to be*

$$\mathrm{AER}_{\mathbf{p}}(\mathcal{A}) := \sum_{i \in [m]} p_i \cdot \mathrm{IER}_i(\mathcal{A}). \tag{6}$$

In words, IER measures the performance of the algorithm from the *client-wise* perspective, whereas AER evaluates the performance of the algorithm from the *system-wide* perspective.

Intuitively speaking, the weight vector $\mathbf{p}$ in (6) can be regarded as the importance weight on each client and controls “how many resources are allocated to each client”. For example, setting $p_i = 1/m$ enforces “fair allocation”, so that each client is treated uniformly, regardless of sample sizes. As another example, setting $p_i = n_i/N$ (recall that $N = \sum_{i \in [m]} n_i$ is the total sample size) means that the central server pays more attention to clients with larger sample sizes, which, to a certain extend, incentivize the clients to contribute more data.

Notably, while a uniform upper bound on all IER_i's can be carried over to the same bound on $\mathrm{AER}_{\mathbf{p}}$, a bound on the $\mathrm{AER}_{\mathbf{p}}$ alone in general does not imply a tight bound on each IER_i, other than the trivial bound $\mathrm{IER}_i \leq \mathrm{AER}_{\mathbf{p}}/p_i$. Such a subtlety is a distinguishing feature of personalized federated learning in the following sense: under homogeneity, it suffices to estimate a single shared global model, and thus $\mathrm{AER}_{\mathbf{p}}$ and all of IER_is are mathematically equivalent.

Regularity Conditions. In this paper, we restrict ourselves to bounded, smooth, and strongly convex loss functions. Such assumptions are common in the federated learning literature (see, e.g., Li et al. 2020b; Hanzely et al. 2020) and cover many unsupervised learning problems such as mean estimation in exponential families and supervised learning problems such as generalized linear models.

Assumption A (Regularity conditions) *Suppose the following conditions hold:*

(a) Compact and convex domain. *The parameter space $\mathcal{W}$ is a compact convex subset of $\mathbb{R}^d$ with diameter $D := \sup_{\boldsymbol{w},\boldsymbol{w}' \in \mathcal{D}} \|\boldsymbol{w} - \boldsymbol{w}'\| < \infty$;*

(b) Smoothness and strong convexity. *For any $i \in [m]$, the loss function $\ell(\cdot, z)$ is β-smooth for almost every z in the support of $\mathcal{D}_i$, and the i-th ERM objective $L_i(\cdot, S)$ is almost surely μ-strongly convex on the convex domain $\mathcal{W} \subseteq \mathbb{R}^d$. We also assume that there exists a universal constant $\|\ell\|_\infty$ such that $0 \le \ell(\cdot, z) \le \|\ell\|_\infty$ for almost every z in the support of $\mathcal{D}_i$;*

(c) Bounded gradient variance at optimum. *There exists a positive constant σ such that for any $i \in [m]$, we have $\mathbb{E}_{Z_i \sim \mathcal{D}_i} \|\nabla \ell(\boldsymbol{w}_\star^{(i)}, Z_i)\|^2 \le \sigma^2$.*

Heterogeneity Conditions. To quantify the level of heterogeneity among clients, we start by introducing the notion of an *average global model*. Assuming a strongly convex loss, the optimal local models (1) are uniquely defined. Thus, we can define the average global model as

$$\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})} = \sum_{i \in [m]} p_i \boldsymbol{w}_\star^{(i)}. \tag{7}$$

We remark that the average global model defined in (7) should *not* be interpreted as the "optimal global model". Rather, it is more suitable to think of $\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})}$ as a point in the parameter space, from which every local model is close to. Indeed, one can readily check that the average global model is the minimizer of $\sum_{i \in [m]} p_i \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}\|^2$ over $\boldsymbol{w} \in \mathbb{R}^d$.

We are now ready to quantify the level of client-wise heterogeneity as follows.

Assumption B (Level of heterogeneity) *There exists a positive constant R such that*

(a) either $\sum_{i \in [m]} p_i \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})}\|^2 \le R^2$,

(b) or $\|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})}\|^2 \le R^2 \ \forall i \in [m]$.

Our study of the $\mathrm{AER}_{\mathbf{p}}$ and IER_is will be based on Part (a) and (b) of Assumption B, respectively. Intuitively, the quantity R encodes one's belief on "how heterogeneous" the clients can be.

3. Main Results

3.1 Analyses of Two Baseline Algorithms

In this subsection, we characterize the performance of PureLocalTraining and FedAvg under the heterogeneity conditions imposed by Assumption B.

3.1.1 Warm Up: Uniform Stability and Analysis of PureLocalTraining

The analysis of PureLocalTraining is based on the classical notion of uniform stability, proposed by Bousquet and Elisseeff (2002).

Definition 3 (Uniform stability) *Consider an algorithm $\mathcal{A}$ that takes a single dataset $S = \{z_j\}_{j=1}^n$ of size n as input and outputs a* single model*: $\mathcal{A}(S) = \hat{\boldsymbol{w}}(S)$. We say $\mathcal{A}$ is γ-uniformly stable if for any dataset S, any $j \in [n]$, and any $z_j' \in \mathcal{Z}$, we have*

$$\|\ell(\hat{\boldsymbol{w}}(S), \cdot) - \ell(\hat{\boldsymbol{w}}(S^{\setminus j}), \cdot)\|_\infty \le \gamma,$$

where $S^{\setminus j}$ is the dataset formed by replacing z_j with z_j':

$$S^{\setminus j} = \{z_1, \dots, z_{j-1}, z_j', z_j, \dots, z_n\}.$$

The main implication of uniformly stable algorithms is that "stable algorithms do not overfit": if $\mathcal{A}$ is γ-uniformly stable, then its *generalization error* is upper bounded by a constant multiple of γ. Thus, one can dissect the analysis of $\mathcal{A}$ into two separate parts: (1) bounding its optimization error; (2) bounding its stability term.

Under our working assumptions, SGD with properly chosen step sizes is guaranteed to converge to the global minimum of (3) (see, e.g., Rakhlin et al. (2011)). Note that the bounds for the approximate minimizers only involve an extra additive term representing the optimization error, and this term will be negligible if we run SGD until convergence since our focus is sample complexity. Thus, we conduct the analysis for the global minimizer of (3). The performance of PureLocalTraining is given by the following theorem.

Theorem 4 (Performance of PureLocalTraining) *Let Assumption A(b) hold and assume $n_i \ge 4\beta/\mu$ $\forall i \in [m]$. Then the algorithm $\mathcal{A}_{\text{PLT}}$ which outputs the minimizer of* (3) *satisfies*

$$\mathbb{E}_{\boldsymbol{S}}[\text{IER}_i(\mathcal{A}_{\text{PLT}})] \lesssim \frac{\beta\|\ell\|_\infty}{\mu n_i}$$

for all $i = 1, \dots, m$.

Proof The proof is a direct consequence of standard results on uniform stability of strongly convex ERM (see, e.g., Section 5 of Shalev-Shwartz et al. (2009) and Section 13 of Shalev-Shwartz and Ben-David (2014)), which assert that under the current assumptions, the minimizer of (3) is $\mathcal{O}(\frac{\beta\|\ell\|_\infty}{\mu n_i})$-uniformly stable. We omit the details. ■

By definition, for any weight vector $\mathbf{p}$, $\text{AER}_{\mathbf{p}}$ of PureLocalTraining also admits the same upper bound as (4).

3.1.2 Federated Stability and Analysis of FedAvg

We consider the following weighted version of (2):

$$\min_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[m]} p_i L_i(\boldsymbol{w}, S_i). \tag{8}$$

The FedAvg algorithm (Algorithm 1) also seamlessly generalizes. The above optimization formulation is in fact covered by the general theory of Li et al. (2020b), where they showed that FedAvg is guaranteed to converge to the global optimum under a suitable hyperparameter choice, even in the presence of heterogeneity (but the convergence is slower). Thus, in the following discussion, we again consider the global minimizer of (8).

It turns out that a tight analysis of FedAvg requires a more fine-grained notion of uniform stability, which we present below.

Definition 5 (Federated stability) *An algorithm $\mathcal{A}$ that outputs $\mathcal{A}(\boldsymbol{S}) = \{\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\}$ has* federated stability $\{\gamma_i\}_{i=1}^m$ *if for every $\boldsymbol{S} \sim \bigotimes_i \mathcal{D}_i^{\otimes n_i}$ and for any $i \in [m], j_i \in [n_i], z'_{i,j_i} \in \mathcal{Z}$, we have*

$$\|\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), \cdot) - \ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}), \cdot)\|_\infty \le \gamma_i.$$

Above, $\boldsymbol{S}^{\backslash(i,j_i)}$ is the dataset formed by replacing $z_{j_i}^{(i)}$ in the i-the dataset with z'_{i,j_i}:

$$\begin{aligned} \boldsymbol{S}^{\backslash(i,j_i)} &= \{S_1, \dots, S_{i-1}, S_i^{\backslash j_i}, S_{i+1}, \dots, S_m\}, \\ S_i^{\backslash j_i} &= \{z_1^{(i)}, \dots, z_{j_i-1}^{(i)}, z'_{i,j_i}, z_{j_i+1}^{(i)}, \dots, z_{n_i}^{(i)}\}. \end{aligned}$$

Compared to the conventional uniform stability in Definition 3, federated stability provides a finer control by allowing distinct stability measures $\{\gamma_i\}$ for different clients. Moreover, the classical statement that "stable algorithms do not overfit" still holds, in the sense that the average (resp. individualized) generalization error can be upper bounded by $\mathcal{O}(\sum_{i\in[m]} n_i\gamma_i/N)$ (resp. $\mathcal{O}(\gamma_i)$), plus a term scaling with the level of heterogeneity R. And this again enables us to separate the analysis of $\mathcal{A}$ into two parts (namely bounding the optimization error and bounding the stability), as is the case with the conventional uniform stability.

The notion of federated stability has other implications when restricted to the FedProx algorithm, and we refer the readers to Section 3.4 for details.

We are now ready to state the theorem that characterizes the performance of FedAvg.

Theorem 6 (Performance of FedAvg) *Let Assumption A(b, c) hold and assume $n_i \ge 4\beta p_i/\mu \;\forall i \in [m]$. Suppose the* FedAvg *algorithm $\mathcal{A}_{\mathrm{FA}}$ outputs the minimizer of* (8). *Then under Assumption B(a), we have*

$$\mathbb{E}_{\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A}_{\mathrm{FA}})] \lesssim \frac{\beta\|\ell\|_\infty}{\mu} \sum_{i\in[m]} \frac{p_i^2}{n_i} + \beta R^2, \tag{9}$$

and under Assumption B(b), we have

$$\mathbb{E}_{\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A}_{\mathrm{FA}})] \lesssim \frac{\beta\sigma^2}{\mu^2} \sum_{i'\in[m]} \frac{p_{i'}^2}{n_{i'}} + \frac{\beta^3}{\mu^2} R^2. \tag{10}$$

Proof The proof of (9) is, roughly speaking, based on the fact that the global minimizer of (8) has federated stability $\gamma_i \lesssim \frac{\beta\|\ell\|_\infty p_i}{\mu n_i}$, and thus the first term in the right-hand side of (9) corresponds to the average federated stability $\sum_{i\in[m]} p_i\gamma_i$. The second term βR^2 in the right-hand side of (9) reflects the presence of heterogeneity. For Equation (10), we were not able to obtain a federated stability based proof, and our current proof is based on an adaptation of the arguments in Theorem 7 of Foster et al. (2019), which explains why the dependence on (σ, β, μ) are different (and slightly worse) compared to Equation (9). In particular, the bound (10) has inverse quadratic dependence on μ, wheres the bound (9) only has $1/\mu$ dependence. The $1/\mu$ dependence comes from the fact that the federated stability term has such dependence, and the $1/\mu^2$ dependence comes from the fact that the ℓ_2 estimation error has such dependence. We refer the readers to Appendix C.1 for details. ■

Note that both bounds in the above theorem are minimized by choosing $p_i = n_i/N$. With this choice of $\mathbf{p}$, the two bounds read

$$\mathbb{E}_{\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A}_{\mathrm{FA}})] \lesssim \frac{\beta\|\ell\|_\infty}{\mu N} + \beta R^2, \qquad \mathbb{E}_{\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A}_{\mathrm{FA}})] \lesssim \frac{\beta\sigma^2}{\mu^2 N} + \frac{\beta^3 R^2}{\mu^2}. \tag{11}$$

This makes sense, since this choice of weight corresponds to the ERM objective under complete homogeneity. This observation also suggests that ensuring "fair resource allocation" (i.e., setting $p_i = 1/m$) can lead to statistical inefficiency, especially when the sample sizes are imbalanced.

We conclude this subsection by noting that though the compactness assumption (Assumption A(a)) is not needed in Theorem 6, it is usually needed in the analysis of the optimization error of FedAvg and PureLocalTraining(see, e.g., Rakhlin et al. (2011); Li et al. (2020b)).

3.2 Lower Bounds

In this subsection, we present our construction of lower bounds, which characterize the information-theoretic limit of personalized federated learning. Throughout this section, we restrict out attention to the case where $p_i = n_i/N$ for any $i \in [m]$.

Our construction starts by considering a special class of problem instances: logistic regression. In logistic regression, given the collection of regression coefficients $\{\boldsymbol{w}_\star^{(i)}\} \subseteq \mathcal{W}$ where $\mathcal{W}$ has a diameter D, the data distributions $\mathcal{D}_i$'s are supported on $\mathbb{R}^d \times \{\pm 1\}$ and specified by a two-step procedure as follows:

1. Generate a feature vector $\boldsymbol{x}$, whose coordinates are i.i.d. copies from some distribution $\mathbb{P}_X$ on $\mathbb{R}$, which is assumed to have mean zero and is almost surely bounded by some absolute constant c_X;
2. Generate the binary label $y \in \{\pm 1\}$, which is a biased Rademacher random variable with head probability $\big(1 + \exp\{-\boldsymbol{x}^\top \boldsymbol{w}_\star^{(i)}\}\big)^{-1}$.

The loss function is naturally chosen to be the negative log-likelihood function, which takes the following form:

$$\ell(\boldsymbol{w}, z) = \ell(\boldsymbol{w}, \boldsymbol{x}, y) = \log(1 + e^{-y\boldsymbol{x}^\top \boldsymbol{w}}).$$

The following lemma says that Assumption A holds for the aforementioned logistic regression models.

Lemma 7 (Logistic regressions are valid problem instances) *The logistic regression problem described above is a class of problem instances that satisfies Assumption A with*

$\|\ell\|_\infty = c_X D\sqrt{d}$ and $\sigma^2 = \beta = c_X^2 d/4$. Moreover, if $m \lesssim (N/m)^c$ for some $c \geq 0$ and $N/m \geq Cd$ for some $C > 1$, then there exists some event E which only depends on the features $\{\boldsymbol{x}_j^{(i)} : i \in [m], j \in [n_i]\}$ and happens with probability at least $1 - e^{-\mathcal{O}(\sqrt{N/m})}$, such that on this event, the strongly convex constant in Assumption A satisfies

$$\mu \asymp \mu_0 = \left(\exp\{c_X D\sqrt{d}/2\} + \exp\{-c_X D\sqrt{d}\}\right)^{-2}. \tag{12}$$

Proof The compactness of the domain and the boundedness of the loss function hold by construction. To verify the rest parts of Assumption A, with some algebra one finds that

$$\nabla^2 \ell(\boldsymbol{w}, \boldsymbol{x}, y) = \frac{\boldsymbol{x}\boldsymbol{x}^\top \exp\{y\boldsymbol{x}^\top \boldsymbol{w}\}}{\big(1 + \exp\{y\boldsymbol{x}^\top \boldsymbol{w}\}\big)^2} \preceq \frac{1}{4}\boldsymbol{x}\boldsymbol{x}^\top, \tag{13}$$

where $\preceq$ is the Loewner order and the inequality holds because $x/(1+x)^2 = 1/(x^{-1/2} + x^{1/2})^2 \leq 1/4$ for $x > 0$. Since the population gradient has mean zero at optimum, the gradient variance at optimum can be upper bounded by the trace of the expected Hessian matrix, which, by the above display, is further upper bounded by $c_X^2 d/4$. Thus, we can take $\sigma^2 = c_X^2 d/4$ in Part (c). Another message of the above display is that we can set the smoothness constant in Part (b) to be $\beta = c_X^2 d/4$.

The only subtlety that remains is to ensure each local loss function is μ-strongly convex. Note that since $x/(1+x)^2$ is decreasing from $(0, 1)$ and is increasing from $(1, \infty)$, the right-hand side of (13) dominates $\mu_0 \boldsymbol{x}\boldsymbol{x}^\top$ in Loewner order, where μ_0 is the right-hand side of (12). Thus, the local population losses $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_i}[\ell(\cdot, \boldsymbol{x}, y)]$ are all μ_0-strongly convex.

Now, note that

$$\nabla^2 L_i(\boldsymbol{w}^{(i)}, S_i) = \frac{1}{n_i}\sum_{j\in[n_i]} \frac{\boldsymbol{x}_j^{(i)}(\boldsymbol{x}_j^{(i)})^\top \exp\{y_j^{(i)}\langle \boldsymbol{x}_j^{(i)}, \boldsymbol{w}^{(i)}\rangle\}}{\big(1 + \exp\{y_j^{(i)}\langle \boldsymbol{x}_j^{(i)}, \boldsymbol{w}^{(i)}\rangle\}\big)^2} \succeq \mu_0 \cdot \frac{1}{n_i}\sum_{j\in[n_i]} \boldsymbol{x}_j^{(i)}(\boldsymbol{x}_j^{(i)})^\top.$$

Invoking Theorem 5.39 of Vershynin (2010) along with a union bound over all clients, we conclude that for any $i \in [m]$, the minimum eigenvalue of $\sum_{j\in[n_i]} \boldsymbol{x}_j^{(i)}(\boldsymbol{x}_j^{(i)})^\top$ is lower bounded by a constant multiple of $n_i - p \gtrsim n_i$ (this is the definition of the event E) with probability at least $1 - me^{-\mathcal{O}(n_i)} \geq 1 - e^{-\mathcal{O}(\sqrt{N/m})}$, and the proof is concluded. ■

Note that in the proof of the above lemma, we have established the $\mu_0 \asymp \mu$-strong convexity of the client-wise population losses. Hence, lower bounding the excess risks reduces to lower bounding the ℓ_2 *estimation errors* $\|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2$ of the estimators $\widehat{\boldsymbol{w}}^{(i)}$ for $\boldsymbol{w}_\star^{(i)}$. Such a reduction allows us to use powerful tools from information theory.

To this end, we introduce two parameter spaces, corresponding to Part (a) and (b) of Assumption B. Recalling $\boldsymbol{w}_{\mathbf{p}}^{(\text{global})} = \sum_{i\in[m]} p_i \boldsymbol{w}_\star^{(i)}$, we define

$$\mathcal{P}_1 := \Big\{\{\boldsymbol{w}_\star^{(i)}\}_{i=1}^m \subseteq \mathcal{W} : \sum_{i\in[m]} p_i \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 \leq R^2\Big\},$$

$$\mathcal{P}_2 := \Big\{\{\boldsymbol{w}_\star^{(i)}\}_{i=1}^m \subseteq \mathcal{W} : \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 \leq R^2 \ \forall i \in [m]\Big\}.$$

Note that $\mathcal{P}_1$ and $\mathcal{P}_2$ index all possible values of $\{\boldsymbol{w}_\star^{(i)}\}$ that can arise in the logistic regression models under Assumption B (a) and (b), respectively.

With the notations introduced so far, we are ready to state the main result of this subsection.

Theorem 8 (Minimax lower bounds for estimation errors) *Consider the logistic regression model described above. Suppose $n_i \asymp n_{i'}$ for any $i \neq i' \in [m]$ and assume $p_i = n_i/N$ for any $i \in [m]$. Then we have*

$$\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\} \in \mathcal{P}_1} \frac{1}{N} \sum_{i \in [m]} n_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim \frac{d}{N/m} \wedge R^2 + \frac{d}{N}, \tag{14}$$

$$\inf_{\widehat{\boldsymbol{w}}^{(i)}} \sup_{\{\boldsymbol{w}_\star^{(i)}\} \in \mathcal{P}_2} \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim \frac{d}{n_i} \wedge R^2 + \frac{d}{N} \tag{15}$$

for all $i \in [m]$, where the infimum is taken over all possible $\widehat{\boldsymbol{w}}^{(i)}$s that are measurable functions of the data $\boldsymbol{S}$.

Proof See Appendix A. ■

Note that both lower bounds in Theorem 8 are a superposition of two terms, and they correspond to two distinct steps in the proof.

The first step in our proof is to argue that the lower bound under complete homogeneity is in fact a valid lower bound under our working assumptions, which gives the $\Omega(d/N)$ term. This is reasonable, since estimation under complete homogeneity is, in many senses, an "easier" problem. The proof of the $\Omega(d/N)$ term is based on the classical Assouad's method (Assouad, 1983).

The second step is to use a generalized version of Assouad's method that allows us to deal with multiple heterogeneous datasets. In particular, we need to carefully choose the prior distributions over the parameter space based on the level of heterogeneity, which ultimately leads to the $\Omega(\frac{d}{N/m} \wedge R^2)$ term. Recall that in the vanilla version of Assouad's method where there is only one parameter, say $\boldsymbol{w}_\star$, one can lower bounds the minimax risk by the Bayes risk, and the prior distribution is usually chosen to be $\boldsymbol{w}_\star = \delta \boldsymbol{v}$, where $\boldsymbol{v}$ follows a uniform distribution over all d-dimensional binary vectors and δ is chosen so that the resulting hypothesis testing problem has large type-I plus type-II error. In our case where there are m parameters $\{\boldsymbol{w}_\star^{(i)}\}$, we need to consider a different prior of the following form:

$$\boldsymbol{w}_\star^{(i)} = \delta_i \boldsymbol{v}^{(i)},$$

where $\boldsymbol{v}^{(i)}$ are i.i.d. samples from the uniform distribution over all d-dimensional binary vectors, and δ_i's are scalers that need to be carefully chosen to make the resulting hypothesis testing problem hard.

The following result is an immediate corollary of Theorem 8.

Corollary 9 (Minimax lower bounds for excess errors) *Assume there exist constants $C, C' > 0, c \geq 0$ such that $n_i \geq C\beta \ \forall i \in [m]$ and $m \leq C'(N/m)^c$. Moreover, assume $n_i \asymp n_{i'}$*

for any $i \neq i' \in [m]$ and $p_i = n_i/N$ for any $i \in [m]$. Then there exists an absolute constant c' such that the following two statements hold:

1. There exists a problem instance such that Assumptions A and B(a) are satisfied with probability at least $1-e^{-c'\sqrt{N/m}}$. Call this high probability event E. On this problem instance, any randomized algorithm $\mathcal{A}$ must suffer

$$\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A}) \cdot \mathbb{1}_{\mathsf{E}}] \gtrsim \mu \cdot \left(\frac{\beta}{N/m} \wedge R^2 + \frac{\beta}{N}\right); \tag{16}$$

2. For any $i \in [m]$, there exists a problem instance such that Assumptions A and B(b) are satisfied with probability at least $1 - e^{-c'\sqrt{N/m}}$. Call this high probability event E_i. On this problem instance, any randomized algorithm $\mathcal{A}$ must suffer

$$\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A}) \cdot \mathbb{1}_{\mathsf{E}_i}] \gtrsim \mu \cdot \left(\frac{\beta}{n_i} \wedge R^2 + \frac{\beta}{N}\right). \tag{17}$$

In the two displays above, the expectation is taken over the randomness in both the algorithm $\mathcal{A}$ and the sample $\boldsymbol{S}$.

Proof Along with Lemma 7 and Theorem 8, this corollary follows by the fact that the smoothness constant β is of the same order as d and the population losses are all $\mu_0 \asymp \mu$-strongly convex. ■

3.3 Implications of the Main Results

The upper bounds in Section 3.1 and the lower bounds in Section 3.2 together reveal several intriguing phenomena regarding personalized FL, which we detail in this subsection.

Focusing on the dependence on the sample sizes and assuming the client-wise samples sizes are balanced (i.e., $n_i \asymp N/m$), the heterogeneity measure R enters the lower and upper bounds in a dichotomous fashion:

- If $R^2 \lesssim m/N$, then both lower bounds become $\Omega(R^2 + 1/N)$, and this lower bound can be attained by FedAvg up to factors that do not depend on the sample sizes;
- If $R^2 \gtrsim m/N$, then both lower bounds become $\Omega(m/N)$. They agree with the minimax rate as if we were under complete heterogeneity and can be achieved by PureLocalTraining.

Now, let us consider the following naïve dichotomous strategy: if output $R^2 \leq \frac{\|\ell\|_\infty}{\mu} \cdot m/N$, then output $\mathcal{A} = \mathcal{A}_{\mathrm{FA}}$; otherwise, output $\mathcal{A} = \mathcal{A}_{\mathrm{PLT}}$. That is, we switch between the two baseline algorithms at the threshold of $R^2 \asymp m/N$. Then under the assumptions in Theorems 4 and 6, one can readily check that this dichotomous strategy satisfies the following AER guarantee:

$$\mathbb{E}_{\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A})] \lesssim \beta\left(\frac{\|\ell\|_\infty}{\mu N/m} \wedge R^2\right) + \frac{\beta\|\ell\|_\infty}{\mu} \sum_{i\in[m]} \frac{p_i^2}{n_i}. \tag{18}$$

If in addition, $n_i \asymp n_{i'}$ for any $i \neq i' \in [m]$, then it also satisfies the following IER guarantee:

$$\mathbb{E}_{\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A})] \lesssim \frac{\beta^3}{\mu^2}\left(\frac{\|\ell\|_\infty}{\mu n_i} \wedge R^2\right) + \frac{\beta\sigma^2}{\mu^2}\sum_{i' \in [m]} \frac{p_{i'}^2}{n_{i'}}. \tag{19}$$

When $p_i = n_i/N$, the two displays above simplify to

$$\mathbb{E}_{\boldsymbol{S}}[\mathrm{AER}(\mathcal{A})] \lesssim \beta\left(\frac{\|\ell\|_\infty}{\mu N/m} \wedge R^2\right) + \frac{\beta\|\ell\|_\infty}{\mu N}, \qquad \mathbb{E}_{\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A})] \lesssim \frac{\beta^3}{\mu^2}\left(\frac{\|\ell\|_\infty}{\mu n_i} \wedge R^2\right) + \frac{\beta\sigma^2}{\mu^2 N},$$

which matches the lower bound in Corollary 9 up to constant factors, provided $(\beta, \mu, \|\ell\|_\infty, \sigma)$ are all of constant order. In other words, switching between the two algorithms at the threshold of $R^2 \asymp m/N$ gives an oracle algorithm that is minimax rate optimal.

Thus, we have shown an interesting property for personalized FL on the choice of the two baseline algorithms. In particular, consider a collection of problem instances indexed by $(R, \beta, \mu, \|\ell\|, \sigma)$ using Assumptions A and B and assume $(\beta, \mu, \|\ell\|_\infty, \sigma)$ are all of constant order. Now, for a fixed value of R, *exactly one of these two algorithms is minimax optimal*, where the optimality is defined over the specified collection of problem instances and with respect to both AER and IER. Moreover, the oracle dichotomous strategy that switches between the two baseline algorithms at the threshold of $R^2 \asymp m/N$ is minimax optimal.

More implications of the theoretical results are described below.

Optimality of a dichotomous strategy. From the practical side, for supervised learning problems, such a dichotomous strategy can be implemented without prior knowledge of R if test errors can be evaluated in a distributed fashion. Indeed, we can first run both FedAvg and PureLocalTraining separately, evaluate their test errors (in a distributed fashion), and deploy the one with a lower test error. Due to the upper and lower bounds proved in Sections 3.1 and 3.2, such a strategy is guaranteed to be minimax rate optimal. As a caveat, however, one should refrain from interpreting our results as saying either of the two baseline algorithms is sufficient for practical problems. From a practical viewpoint, constants that are omitted in the minimax analysis are crucial. Even for supervised problems, a better personalization result could be achieved by more sophisticated algorithms in practice. Nevertheless, our results suggest that the two baseline algorithms can at least serve as a good starting point in the search for efficient personalized algorithms.

For unsupervised problems where the quality of a model is hard to evaluate, implementing the dichotomous strategy requires estimating an upper bound R of the level of heterogeneity. This is an important open problem, which we leave for future work.

Optimality of FedAvg *followed by local fine tuning.* Another popular baseline algorithm for personalized FL is to first run FedAvg until convergence, and then let each client run PureLocalTraining to fine tune the model. In strongly convex problems, global optima can be reached by gradient descent regardless the initialization with a suitable choice of the learning rate (see, e.g., Theorem 2.1.15 of Nesterov 2018). Thus, if each client run PureLocalTraining for long enough, the global optima for its local loss function will finally be reached. This fact tells that along the whole fine tuning trajectory, there is a point at which the model gives the worst-case optimal AER and IER, and for a fixed level of heterogeneity, this point is either at the very beginning (which is FedAvg), or at the

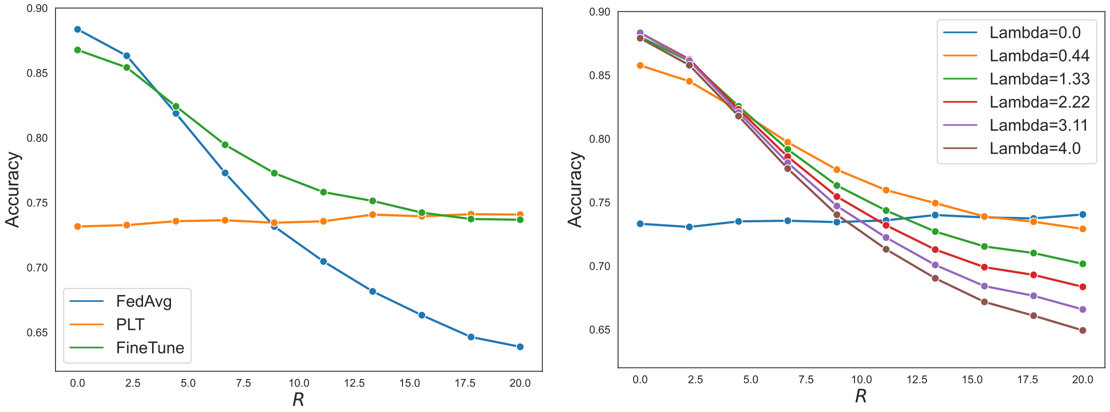

Figure 1: Average classification accuracy of FedAvg, PureLocalTraining and FedAvg followed by fine tuning (left panel) as well as FedProx with different choice of λ (right panel).

very end (which is PureLocalTraining). Although this conclusion is almost trivial from a technical point of view given our minimax results, it provides a reassuring theoretical property (of being minimax optimal) for a popular method used by practitioners.

Illustrating the minimaxity in a simulated example. We conduct a simulation on federated logistic regression to corroborate our theoretical results and the optimality of the FedAvg following by local fine tuning strategy. In the simulation, we set $m = 5$, $n_i = 100, \forall i \in [m]$ and we vary R from 0 to 20 (see Appendix D for details). In the left panel of Figure 1, we plot the test accuracy (averaged over 100 rounds of simulations) of those three methods against the value of R. One can see that the accuracy of the fine tuning strategy roughly follows the maximum of the accuracies achieved by FedAvg and PureLocalTraining, confirming our theoretical prediction that the fine tuning strategy can indeed perform as well as the best between FedAvg and PureLocalTraining.

Beyond the current heterogeneity assumption. Our minimax results are established under Assumption B, which states that all optimal local models are close to a certain "centroid" (i.e., the average global model defined in (7)). If we draw a graph of clients and connect two clients if their optimal local models are similar, then the current heterogeneity assumption gives rise to a complete graph (or a star-shaped graph if we introduce another node to represent the average global model). While such an idealized graph structure enables a clean theoretical analysis, the real world proximity patterns among clients are clearly far more sophisticated. In fact, counterexamples exist under which the minimax results does not hold in a "global" sense.

Suppose all m clients exhibit a clustering structure as follows. We have $\sqrt{m}$ clients whose optimal global models serve as cluster centroids, and those centroids are very far apart. In the neighborhood of each centroid, there are $\sqrt{m}$ clients whose optimal local models are R_c away from the centroid in the sense of Assumption B. Additionally, assume each client has equal sample sizes, so that $n_i = n$ for some n. Under this setting, the "global" heterogeneity parameter R in Assumption B is very large, so our theory would suggest choosing PureLocalTraining, which gives a rate of $\mathcal{O}(1/n)$. However, this rate is clearly suboptimal. If one can successfully cluster the m clients into $\sqrt{m}$ clusters (which is hopeful as the centroids are assumed to be far apart), then one can apply our theory to each cluster (i.e., run the dichotomous strategy for each cluster) and conclude that the rate for each cluster is $\mathcal{O}(\frac{1}{n\sqrt{m}} + \frac{1}{n} \wedge R_c^2) \ll \frac{1}{n}$ if m diverges to infinity and $R_c \ll 1/\sqrt{n}$. The foregoing discussion reveals in such a clustered setting, our theoretical results only make sense at the cluster level, but not at the global level.

The behaviors of the minimax rate for more general client proximity graphs can be even more complicated, which we leave for future work.

3.4 More Implications of Federated Stability and Analysis of FedProx

In this subsection, we are concerned with the performance guarantees for FedProx. As with our earlier analysis of FedAvg, we consider a **p**-weighted version of FedProx, whose optimization formulation is given below:

$$\min_{\substack{\boldsymbol{w}^{(\text{global})} \in \mathcal{W} \\ \{\boldsymbol{w}^{(i)}\}_{i=1}^m \subseteq \mathcal{W}}} \sum_{i \in [m]} p_i \left(L_i(\boldsymbol{w}^{(i)}, S_i) + \frac{\lambda}{2} \|\boldsymbol{w}^{(\text{global})} - \boldsymbol{w}^{(i)}\|^2 \right), \tag{20}$$

where we recall that $L_i(\boldsymbol{w}, S_i) := \sum_{j \in [n_i]} \ell(\boldsymbol{w}, z_j^{(i)})/n_i$ is the ERM objective for the i-th client. In this subsection, we let $(\tilde{\boldsymbol{w}}^{(\text{global})}, \{\tilde{\boldsymbol{w}}^{(i)}\})$ be the global minimizer of the above problem. Compared to (8), which imposes a "hard" constraint $\boldsymbol{w}^{(i)} = \boldsymbol{w}^{(\text{global})}$, and compared to (3), where there is no constraint at all, the above formulation imposes a "soft" constraint that the norm of $(\boldsymbol{w}^{(\text{global})} - \boldsymbol{w}^{(i)})$ should be small, with a hyperparameter λ controlling the strength of this constraint.

The rationale behind the optimization formulation (20) of FedProx is clear: by setting $\lambda = 0$, the optimization formulation of PureLocalTraining (3) is recovered, and as $\lambda \to \infty$, the optimization formulation of FedAvg (8) is recovered. The hope is that by varying $\lambda \in (0, \infty)$, one can interpolate between the two extremes.

Applying the idea of local SGD to (20), one obtains the FedProx algorithm[1], which we detail in Algorithm 2. We separate the whole algorithm into two stages as they has distinct interpretations: in Stage I, the central server aims to learn a good global model with the help of local clients, whereas in Stage II, each local client takes advantage of the global model to personalize. Alternatively, one can also interpret FedProx as an instance of the general framework of *model-agnostic meta learning* (Finn et al., 2017), where Stage I learns a good initialization, and Stage II trains the local models starting from this initialization.

In contrast to our analyses for FedAvg and PureLocalTraining in Section 3.1, where we largely focused on global minimizers, the analysis for FedProx will be carried out for the approximate minimizer output by Algorithm 2. The reason for this is rooted in the tradeoff between the optimization error and the generalization error. Note that given the results derived in Section 3.1, the analysis for the global minimizer $(\tilde{\boldsymbol{w}}^{(\text{global})}, \{\tilde{\boldsymbol{w}}^{(i)}\})$ becomes trivial: by setting $\lambda = 0$, we reduce the task to analyzing PureLocalTraining; by sending $\lambda \to \infty$, we reduce the task to analyzing FedAvg. Based on Theorems 4 and 6, one immediately concludes that there exists a choice of λ, such that the AER and IER of

[1] In fact, Algorithm 2 is not exactly the same as the original FedProx algorithm introduced in Li et al. (2018). But since both algorithms share the idea of imposing regularization, we still call Algorithm 2 FedProx for conceptual simplicity.

Algorithm 2: FedProx, general version

```
Input: Initial global model $\boldsymbol{w}_0^{(\mathrm{global})}$, initial local models $\{\boldsymbol{w}_0^{(i)}\}_{i=1}^m \equiv \{\boldsymbol{w}_{0,0}^{(i)}\}_{i=1}^m$,
       global rounds $T$, global batch size $B^{(\mathrm{global})}$, global step sizes $\{\eta_t^{(\mathrm{global})}\}_{t=0}^{T-1}$,
       local rounds $\{K_t\}_{t=0}^T$, local batch sizes $\{B^{(i)}\}_{i=0}^m$ local step sizes
       $\{\eta_{t,k}^{(i)} : 0 \le t \le T, 0 \le k \le K_t - 1\}$.
Output: Local models $\{\boldsymbol{w}_{T+1}^{(i)}\}_{i=1}^m$.
# Stage I: joint training
for $t = 0, 1, \ldots, T-1$ do
    Randomly sample a batch $\mathcal{C}_t \subseteq [m]$ of size $B^{(\mathrm{global})}$
    for $i \in [m]$ do
        if $i \in \mathcal{C}_t$ then $\boldsymbol{w}_{t+1}^{(i)} \leftarrow \boldsymbol{w}_t^{(i)}$
        else
            Pull $\boldsymbol{w}_t^{(\mathrm{global})}$ from the server
            $\boldsymbol{w}_{t+1}^{(i)} \leftarrow$ SoftLocalSGD$(i, \boldsymbol{w}_t^{(i)}, \boldsymbol{w}_t^{(\mathrm{global})}, K_t, B^{(i)}, \{\eta_{t,k}^{(i)}\}_{k=0}^{K_t-1})$
            Push $\boldsymbol{w}_{t+1}^{(i)}$ to the server
    $\boldsymbol{w}_{t+1}^{(\mathrm{global})} \leftarrow \boldsymbol{w}_t^{(\mathrm{global})} - \frac{\lambda m \eta_t^{(\mathrm{global})}}{B^{(\mathrm{global})}} \sum_{i \in \mathcal{C}_t} p_i(\boldsymbol{w}_t^{(\mathrm{global})} - \boldsymbol{w}_{t+1}^{(i)})$
# Stage II: final training before deployment
for $i \in [m]$ do
    Pull $\boldsymbol{w}_T^{(\mathrm{global})}$ from the server
    $\boldsymbol{w}_{T+1}^{(i)} \leftarrow$ SoftLocalSGD$(i, \boldsymbol{w}_T^{(i)}, \boldsymbol{w}_T^{(\mathrm{global})}, K_T, B^{(i)}, \{\eta_{T,k}^{(i)}\}_{k=0}^{K_T-1})$
return $\{\boldsymbol{w}_{T+1}^{(i)}\}_{i=1}^m$
# Local SGD subroutine
Function SoftLocalSGD$(i, \boldsymbol{w}^{(i)}, \boldsymbol{w}^{(\mathrm{global})}, K, B, \{\eta_k\}_{k=0}^{K-1})$
    for $k = 0, 1, \ldots, K-1$ do
        Randomly sample a batch $I \subseteq [n_i]$ of size $|I| = B$
        $\boldsymbol{w}^{(i)} \leftarrow \mathcal{P}_{\mathcal{W}}\Big[\boldsymbol{w}^{(i)} - \frac{\eta_k}{B} \sum_{j \in I} \Big(\nabla \ell(\boldsymbol{w}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}^{(i)} - \boldsymbol{w}^{(\mathrm{global})})\Big)\Big]$
    return $\boldsymbol{w}^{(i)}$
```

$\{\tilde{\boldsymbol{w}}^{(i)}\}$ satisfy the bounds in (18) and (19), respectively. However, the foregoing discussion is purely restricted to generalization error. When we set $\lambda = 0$ or send $\lambda \to \infty$, it is not known a priori whether FedProx algorithm will converge to the global minima. Worse still, the optimization error may depends on λ in a particular way so that it becomes unbounded when λ approaches zero or infinity. To the best of our knowledge, prior work only proved the optimization convergence of FedProx for the global model with a fixed value of λ, namely the convergence of $\boldsymbol{w}_T^{(\mathrm{global})}$ to $\tilde{\boldsymbol{w}}^{(\mathrm{global})}$ as the number of global communication rounds T tends to infinity (Li et al., 2018; Dinh et al., 2020). To have a theoretical understanding of the performance of FedProx, it is crucial to (1) establish the optimization convergence for both global and local models; (2) bound the generalization error; and (3) balance the optimization error and the generalization error, both of which are functions of λ. In the following, we execute the those steps with the aid of federated stability.

Implications of federated stability for FedProx. We have briefly mentioned the main implications of federated stability in Section 3.1.2: for an algorithm $\mathcal{A} = \{\widehat{\boldsymbol{w}}^{(i)}\}$ with federated stability $\{\gamma_i\}$, its average generalization error (resp. individualized generalization error) can be upper bounded by $\mathcal{O}(\sum_{i\in[m]} p_i\gamma_i)$ (resp. $\mathcal{O}(\gamma_i)$), plus a term scaling with the level of heterogeneity R. We make such a statement precise here. Let us first define the optimization error of a generic algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\mathrm{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\})$ (which tries to solve (20)) as

$$\begin{aligned}\mathcal{E}_{\mathrm{OPT}} := \sum_{i\in[m]} p_i\bigg(L_i(\widehat{\boldsymbol{w}}^{(i)}, S_i) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \widehat{\boldsymbol{w}}^{(i)}\|^2\bigg) \\ - \sum_{i\in[m]} p_i\bigg(L_i(\tilde{\boldsymbol{w}}^{(i)}, S_i) + \frac{\lambda}{2}\|\tilde{\boldsymbol{w}}^{(\mathrm{global})} - \tilde{\boldsymbol{w}}^{(i)}\|^2\bigg).\end{aligned}$$

The main implications of federated stability, when applied to the specifics of FedProx, can then be summarized in the following proposition.

Proposition 10 (Implications of federated stability restricted to FedProx) *Consider an algorithm* $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\mathrm{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\})$ *with federated uniform stability* $\{\gamma_i\}_1^m$*. Then we have*

$$\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A})] \le \mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\mathcal{E}_{\mathrm{OPT}}] + 2\sum_{i\in[m]} p_i\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\gamma_i] + \frac{\lambda}{2}\sum_{i\in[m]} p_i\|\boldsymbol{w}_{\mathrm{avg}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2, \tag{21}$$

$$\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A})] \le \frac{\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\mathcal{E}_{\mathrm{OPT}}]}{p_i} + 2\mathbb{E}_{\mathcal{A},\boldsymbol{S}}[\gamma_i] + \frac{\lambda}{2}\mathbb{E}_{\mathcal{A},\boldsymbol{S}}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2 \quad \forall i\in[m]. \tag{22}$$

Proof The proof of (21) is based on the following basic inequality for the AER:

$$\sum_{i\in[m]} p_i\bigg(L_i(\tilde{\boldsymbol{w}}^{(i)}, S_i) + \frac{\lambda}{2}\|\tilde{\boldsymbol{w}}^{(\mathrm{global})} - \tilde{\boldsymbol{w}}^{(i)}\|^2\bigg) \le \sum_{i\in[m]} p_i\bigg(L_i(\boldsymbol{w}_\star^{(i)}, S_i) + \frac{\lambda}{2}\|\boldsymbol{w}_{\mathrm{avg}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2\bigg), \tag{23}$$

whereas the proof of (22) is based on the following basic inequality for the IER: for any $s\in[m]$, we have

$$\begin{aligned}&\sum_{i\in[m]} p_i\bigg(L_i(\tilde{\boldsymbol{w}}^{(i)}, S_i) + \frac{\lambda}{2}\|\tilde{\boldsymbol{w}}^{(\mathrm{global})} - \tilde{\boldsymbol{w}}^{(i)}\|^2\bigg) \\ &\le p_sL_s(\boldsymbol{w}_\star^{(s)}, S_s) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(s)}\|^2 + \sum_{i\neq s} p_i\bigg(L_i(\widehat{\boldsymbol{w}}^{(i)}, S_i) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \widehat{\boldsymbol{w}}^{(i)}\|^2\bigg).\end{aligned} \tag{24}$$

We refer the readers to Appendix C.2 for details. ∎

Note that both bounds in Proposition 10 involve a term that scales linearly with both λ and the heterogeneity measure. In general, we expect the stability measures to scale inversely with λ, and thus opening the possibility of carefully choosing λ to balance the stability term and the heterogeneity term.

Let us observe that the heterogeneity term of (22) is slightly different than that of (21), in that it involves the estimated global model $\widehat{\boldsymbol{w}}^{(\text{global})}$. This suggests that achieving the IER guarantees might be intrinsically more difficult than achieving the AER guarantees.

In view of Proposition 10, we are left to bound the optimization error and the federated stability of FedProx. As discussed above, achieving the AER and IER guarantees requires somewhat different assumptions, as the latter involves characterizing the performance of the global model. So we split our discussion into two parts below.

Bounding the average excess error. The following theorem characterize the performance of FedProx in terms of the AER.

Theorem 11 (AER guarantees for FedProx) *Let Assumptions A and B(a) hold, and assume $n_i \geq 4\beta/\mu$ for all $i \in [m]$. Choose the weight vector $\mathbf{p}$ such that*

$$\frac{p_{\max} \sum_{i\in[m]} p_i/n_i}{\sum_{i\in[m]} p_i^2/n_i} \leq C_{\mathbf{p}} \tag{25}$$

for some constant $C_{\mathbf{p}}$, where $p_{\max} = \max_i p_i$. Consider the FedProx *algorithm, $\mathcal{A}_{\text{FP}}$, with the following hyperparameter configuration:*

1. *In the joint training stage (i.e., $0 \leq t \leq T-1$), set*

$$\begin{aligned} &\eta_{t,k}^{(i)} = \frac{1}{(\mu+\lambda)(k+1)}, \quad \eta_t^{(\text{global})} = \frac{2(\mu+\lambda)}{\lambda\mu(t+1)}, \quad K_t + 1 \geq C_1(\lambda^2 \vee 1)t, \\ &T \geq C_2\lambda(\lambda\vee 1)m\|\mathbf{p}\|^2 \cdot \Big(\big[\sum_{i\in[m]} p_i/n_i\big]^{-1} \vee \big[\lambda(\lambda\vee 1)n_{\max}^2\big]\Big); \end{aligned} \tag{26}$$

2. *In the final training stage (i.e., $t = T$), set*

$$\begin{aligned} &\eta_{T,k}^{(i)} = \frac{1}{(\mu+\lambda)(k+1)}, \\ &K_T \geq C_3(\lambda+1)^2 \cdot \Big(\big[\sum_{i\in[m]} p_i/n_i\big]^{-1} \vee \big[\lambda^2 \max_{i\in[m]}(p_i n_i)^2\big]\Big), \end{aligned} \tag{27}$$

where C_1, C_2, C_3 are constants depending only on $(\mu, \beta, \|\ell\|_\infty, D)$. Then, there exists a choice of λ such that

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}_{\text{FP}},\boldsymbol{S}}[\text{AER}_{\mathbf{p}}(\mathcal{A}_{\text{FP}})] \\ &\lesssim \Big(\frac{\mu}{1\wedge C_{\mathbf{p}}} + \frac{(1\vee C_{\mathbf{p}})\beta\|\ell\|_\infty}{\mu}\Big)\Big[\Big(R\sqrt{\sum_{i\in[m]}\frac{p_i}{n_i}}\Big) \wedge \Big(\sum_{i\in[m]}\frac{p_i}{n_i}\Big) + \sum_{i\in[m]}\frac{p_i^2}{n_i}\Big]. \end{aligned} \tag{28}$$

Proof See Appendix C.3. ■

A few remarks are in order. First, (25) essentially says that the weight $\mathbf{p}$ cannot be too imbalanced, and too much imbalance in $\mathbf{p}$ can hurt the performance in view of the multiplicative factor of $C_{\mathbf{p}}$ in our bound (28). If we set $p_i = 1/m$, then $C_{\mathbf{p}}$ is naturally of constant order; whereas if we set $p_i = n_i/N$, we have $C_{\mathbf{p}} \asymp mn_{\max}/N$, where $n_{\max} = \max_i n_i$, which calls for relative balance of the sample sizes.

We then briefly comment on the hyperparameter choice in the above theorem. The step sizes are of the form 1/(strongly convex constant × iteration counter), and such a choice is common in strongly convex stochastic optimization problems (see, e.g., Rakhlin et al. 2011; Shamir and Zhang 2013). Such a choice, along with the smoothness of the problem, is also the key for us to by-pass the need of doing any time-averaging operation, as is done in, for example, Dinh et al. (2020).

In Theorem 11, the choice of the communication rounds T and the final local training round K_T both scale polynomially with λ, which means that the optimization convergence of FedProx is slower when the data are less heterogeneous. This phenomenon happens more generally. For example, in Hanzely and Richtárik (2020), they proposed a variant of SGD that optimizes (20) with $p_i = 1/m$ in $\mathcal{O}\big(\frac{L+\lambda}{\mu}\log 1/\varepsilon\big)$-many iterations, where L is the Lipschitz constant of the loss function and ε is the desired accuracy level.

The constants C_1, C_2, C_3 in the statement of Theorem 11 can be explicitly traced in our proof. We remark that the dependence on problem-specific constants $(\mu, \beta, \|\ell\|_\infty, D)$ in our hyperparameter choice and on λ may not be tight. A tight analysis of the optimization error is interesting, but less relevant for our purpose of understanding the sample complexity. So we defer such an analysis to future work[2].

Bounding individualized excess errors. The following theorem gives the IER guarantees for FedProx.

Theorem 12 (IER guarantees for FedProx) *Let Assumptions A and B(b) hold. Moreover, assume that $n_i \asymp n_{i'}$ for any $i \neq i' \in [m]$ and $n_i \geq 4\beta/\mu$ $\forall i \in [m]$. Let the weight vector be chosen as $p_i \asymp 1/m$ $\forall i \in [m]$. Consider the* FedProx *algorithm, $\mathcal{A}_{\rm FP}$, with the following hyperparameter configuration:*

1. *In the joint training stage (i.e., $0 \leq t \leq T-1$), set $\eta_{t,k}^{(i)}, \eta_t^{(\rm global)}, K_t$ as in (26), and set*
$$T \geq C_2' \lambda(\lambda \vee 1) \max_{i\in[m]} n_i \cdot \bigg(p_i^{-1} \vee [\lambda(\lambda \vee 1) n_i]\bigg);$$
2. *In the final training stage (i.e., $t = T$), set $\eta_{T,k}^{(i)}$ as in (27), and set*
$$K_T \geq C_3'(\lambda+1)^2 \max_{i\in[m]} n_i \bigg(p_i^{-1} \vee \lambda^2 p_i^2 n_i\bigg),$$

where C_2', C_3' are constants only depending on $(\mu, \beta, \|\ell\|_\infty, D)$. Then, there exists a choice of λ such that for any $i \in [m]$, we have
$$\mathbb{E}_{\mathcal{A}_{\rm FP}, \boldsymbol{S}}[{\rm IER}_i(\mathcal{A}_{\rm FP})] \lesssim \bigg[(\mu+\mu^{-1})\bigg(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\bigg) + \mu D^2\bigg] \cdot \bigg(\frac{R}{\sqrt{n_i}} \wedge \frac{1}{n_i} + \frac{\sqrt{m}}{N}\bigg). \tag{29}$$

[2] The theories developed by Hanzely et al. (2020) can be useful for such an analysis.

Proof See Appendix C.4. ■

Compared to Theorem 11, the above theorem imposes extra assumptions that the sample sizes are relative balanced and that $p_i \asymp 1/m$, both of which are due to the fact that we need to additionally take care of the estimation error of the global model. The hyperparameter choice slightly differs from that in Theorem 11 for the same reason. In practice, when one is to use FedProx to optimize highly non-convex functions like the loss function of deep neural networks, instead of sticking to the choices made in Theorems 11 and 12, the hyperparameters are usually tuned by trial-and-error for best test performance.

Comparison with the lower bounds. In order to comment about the optimality/suboptimality of FedProx, let us restrict to the case when $p_i = n_i/N$. In this case, the bound in Theorem 11 becomes

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A}_{\mathrm{FP}})] \lesssim \left(\mu + \frac{\beta\|\ell\|_\infty}{\mu}\right) \cdot \left(\frac{1}{N/m} \wedge \frac{R}{\sqrt{N/m}} + \frac{1}{N}\right). \tag{30}$$

Recall the lower bound in (16). Focusing on the dependence of sample sizes and heterogeneity measure, we have the following three cases. If $R^2 \gtrsim m/N$, then (30) becomes $\mathcal{O}(m/N)$, which matches the lower bound. Meanwhile, if $1/mN \lesssim R^2 \lesssim m/N$, then (30) becomes $\mathcal{O}(m/N)$, whereas the lower bound reads $\Omega(R^2 + 1/N)$, and thus (30) is suboptimal unless $R^2 \asymp m/N$. Moreover, if $R^2 \lesssim 1/mN$, then (30) becomes $\mathcal{O}(1/N)$, and is minimax optimal again.

A similar trilogy holds for IER of FedProx. Comparing the upper bound in (29) and the lower bound in (17), we still have three cases as follows. If $R^2 \gtrsim m/N$, then (29) is $\mathcal{O}(1/n_i)$, which agrees with the lower bound. Meanwhile, if $1/N \lesssim R^2 \lesssim m/N$, then (29) is $\mathcal{O}(R/\sqrt{n_i})$, and is suboptimal compared to the $\Omega(R^2+1/N)$ lower bound unless $R^2 \asymp m/N$. Moreover, if $R^2 \lesssim 1/N$, then (29) is $\mathcal{O}(\sqrt{m}/N)$, and is off by a factor of order $\sqrt{m}$ compared to the $\Omega(1/N)$ lower bound.

While the bounds in Theorems 11 and 12 in general do not attain the lower bounds in Corollary 9, they are still non-trivial in the sense that they scale with the heterogeneity measure R. While there are some recent works establishing the AER guarantees for an objective similar to (20) under the online learning setup (see, e.g., Denevi et al. 2019; Balcan et al. 2019; Khodak et al. 2019), to the best of our knowledge, Theorems 11 and 12 are the first to establish *both* the AER and IER guarantees for (20) under the federated learning setup.

Curious readers may wonder if the suboptimality of the theoretical guarantees for FedProx (with non-zero λ) is a characteristic of this algorithm or if it is due to the artifact of our technical proof. To answer this question, we conduct a simulation where we apply FedProx with different λs on datasets generated by federated logistic regression (see Appendix D for details). The accuracies versus different values of R is shown in the right panel of Figure 1. As expected by our theory, the performance of FedProx with $\lambda = 0$ mimics that of PureLocalTraining, whereas the performance with $\lambda = 4$ resembles that of FedAvg. Interestingly, FedProx with $\lambda = 0.44$ bears a similar performance with the FedAvg followed by fine tuning strategy, which we know is minimax optimal. This observation supports the conjecture that optimally tuned FedProx is indeed minimax optimal, and the suboptimality of bounds from Theorems 11 and 12 are likely to be a consequence of the artifact of our theoretical analysis.

4. Discussion

This paper studies the statistical properties of personalized federated learning. Focusing on strongly-convex, smooth, and bounded empirical risk minimization problems, we have uncovered an intriguing phenomenon that given a specific level of heterogeneity, exactly one of FedAvg or PureLocalTraining is minimax optimal. In the course of proving this result, we obtained a novel analysis of FedProx and introduced a new notion of algorithmic stability termed federated stability, which is possibly of independent interest for analyzing generalization properties in the context of federated learning.

We close this paper by mentioning several open problems.

- *Dependence on problem-specific parameters.* This paper focuses on the dependence on the sample sizes, and in our bounds, the dependence on problem-specific parameters (e.g., the smoothness and strong convexity constants) may not be optimal. This can be problematic if those parameters are not of constant order, and it will be interesting to give a refined analysis that gives optimal dependence on those parameters.

- *A refined analysis of* FedProx*.* The upper bounds we develop for FedProx, as we have mentioned, do not match our minimax lower bounds. According to a simulated example, we suspect that this is an artifact of our analysis and a refined analysis of FedProx would be a welcome advance.

- *Estimation of the level of heterogeneity and development of adaptive algorithms.* For unsupervised problems where evaluation of a model is difficult, implementation of the oracle dichotomous strategy described in Section 3.3 would require estimating the level of heterogeneity R. Even for supervised problems, estimation of R would be interesting, as it allows one to decide which algorithm to choose without model training. More generally, developing adaptive algorithms that attains the lower bound without prior information of R is an important open problem.

- *Beyond the current heterogeneity assumption.* As discussed in Section 3.3, our theoretical results may not hold globally when one moves from Assumption B to more general heterogeneity assumptions. Establishing the minimax rates and designing provably optimal algorithms under those assumptions are of both theoretical and practical interest.

- *Beyond convexity.* Our analysis is heavily contingent upon the strong convexity of the loss function, which, to the best of our knowledge, is not easily generalizable to the non-convex case. Meanwhile, our notion of heterogeneity, which is based on the distance of optimal local models to the convex combination of them, may not be natural for non-convex problems. It is of interest, albeit difficult, to have a theoretical investigation of personalized federated learning for non-convex problems.

Acknowledgments

This work was supported in part by NIH through R01-GM124111 and RF1-AG063481, NSF through CAREER DMS-1847415, CCF-1763314, and CCF-1934876, and an Alfred P. Sloan Research Fellowship.

Appendices for "Minimax Estimation for Personalized Federated Learning: An Alternative between FedAvg and Local Training?"

Contents

A. Proof of Theorem 8: Lower Bounds

We start by presenting a lower bound when all $\boldsymbol{w}_\star^{(i)}$'s are the same.

Lemma 13 (Lower bound under homogeneity) *Consider the logistic regression model with $\boldsymbol{w}_\star^{(i)} = \boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}$ for any $i \in [m]$. Then*

$$\inf_{\widehat{\boldsymbol{w}}^{(\mathrm{global})}} \sup_{\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}} \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}\|^2 \gtrsim \frac{d}{N}.$$

Proof This is a classical result. See, e.g., Example 8.4 of Duchi (2019). ∎

Proof [Proof of (14)] We first give a lower bound based on the observation that the homogeneous case is in fact included in the parameter space $\mathcal{P}_1$. More explicitly, let us define $\mathcal{P}_0 = \{\{\boldsymbol{w}_\star^{(i)}\} \in \mathcal{P}_1 : \boldsymbol{w}_\star^{(i)} = \boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})} \ \forall i \in [m]\}$. By Lemma 13, we have

$$\begin{aligned}\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\} \in \mathcal{P}_1} \sum_{i \in [m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 &\geq \inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\} \in \mathcal{P}_0} \sum_{i \in [m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \\ &= \inf_{\widehat{\boldsymbol{w}}^{(\mathrm{global})}} \sup_{\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}} \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}\|^2 \\ &\gtrsim \frac{d}{N}. \end{aligned} \tag{31}$$

We now use a variant of Assouad's method (Assouad, 1983) that allows us to tackle multiple datasets. Consider the following data generating process: nature generates $\boldsymbol{V} = \{\boldsymbol{v}^{(i)} : i \in [m]\}$ i.i.d. from the uniform distribution on $\mathcal{V} = \{\pm 1\}^d$ and sets $\boldsymbol{w}_\star^{(i)} = \delta_i \boldsymbol{v}^{(i)}$ for some δ_i such that the following constraint is satisfied:

$$\sum_{i\in[m]} p_i \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 = \sum_{i\in[m]} p_i \left\| \delta_i \boldsymbol{v}^{(i)} - \sum_{s\in[m]} p_s \delta_s \boldsymbol{v}^{(s)} \right\|^2 \leq R^2. \tag{32}$$

We will specify the choice of δ_i's later. Denoting $\mathbb{E}_{\boldsymbol{X}}$ as the marginal expectation operator with respect to all the features $\{\boldsymbol{x}_j^{(i)}\}$ and $\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}$ as the conditional expectation operator with respect to $\{y_j^{(i)}\}|\{\boldsymbol{x}_j^{(i)}\}$, we can lower bound the minimax risk by the Bayes risk as follows:

$$\begin{aligned}
&\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \\
&\geq \inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \mathbb{E}_{\{\boldsymbol{v}^{(i)}\}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \delta_i \boldsymbol{v}^{(i)}\|^2 \\
&= \inf_{\{\widehat{\boldsymbol{v}}^{(i)}\}\subseteq\mathcal{V}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{V},\boldsymbol{S}} \|\delta_i \widehat{\boldsymbol{v}}^{(i)} - \delta_i \boldsymbol{v}^{(i)}\|^2 \\
&\geq \mathbb{E}_{\boldsymbol{X}} \sum_{i\in[m]} p_i \delta_i^2 \inf_{\widehat{\boldsymbol{v}}^{(i)}\in\mathcal{V}} \mathbb{E}_{\boldsymbol{V},\boldsymbol{Y}|\boldsymbol{X}} \|\widehat{\boldsymbol{v}}^{(i)} - \boldsymbol{v}^{(i)}\|^2 \\
&\geq \mathbb{E}_{\boldsymbol{X}} \sum_{i\in[m]} p_i \delta_i^2 \sum_{k\in[d]} \inf_{\widehat{\boldsymbol{v}}_k^{(i)}\in\{\pm 1\}} \mathbb{E}_{\boldsymbol{V},\boldsymbol{Y}|\boldsymbol{X}} (\widehat{\boldsymbol{v}}_k^{(i)} - \boldsymbol{v}_k^{(i)})^2 \\
&\gtrsim \mathbb{E}_{\boldsymbol{X}} \sum_{i\in[m]} p_i \delta_i^2 \sum_{k\in[d]} \inf_{\widehat{\boldsymbol{v}}_k^{(i)}\in\{\pm 1\}} \mathbb{P}_{\boldsymbol{V},\boldsymbol{Y}|\boldsymbol{X}} (\widehat{\boldsymbol{v}}_k^{(i)} \neq \boldsymbol{v}_k^{(i)}) \\
&= \frac{1}{2} \mathbb{E}_{\boldsymbol{X}} \sum_{i\in[m]} p_i \delta_i^2 \sum_{k\in[d]} \inf_{\widehat{\boldsymbol{v}}_k^{(i)}\in\{\pm 1\}} \Big(\mathbb{P}_{i,+k}(\widehat{\boldsymbol{v}}_k^{(i)} = -1) + \mathbb{P}_{i,-k}(\widehat{\boldsymbol{v}}_k^{(i)} = +1) \Big),
\end{aligned}$$

where in the last line, we have let $\mathbb{P}_{i,\pm k}(\cdot) = \mathbb{P}_{\boldsymbol{V},\boldsymbol{Y}|\boldsymbol{X}}(\cdot|\boldsymbol{v}_k^{(i)} = \pm 1)$ to denote the probability measure with respect to the randomness in $(\boldsymbol{V}, \boldsymbol{S})$ conditional on the features $\{\boldsymbol{x}_j^{(i)}\}$ as well as the realization of $\boldsymbol{v}_k^{(i)} = \pm 1$. More explicitly, we can write

$$\begin{aligned}
\mathbb{P}_{i,\pm k} &= \Bigg(\bigotimes_{s\neq i} \mathbb{P}_{\boldsymbol{v}^{(s)}} \otimes \mathbb{P}_{\{y^{(s)}\}_{j=1}^{n_s}|\boldsymbol{v}^{(s)},\{\boldsymbol{x}_j^{(s)}\}_{j=1}^{n_s}} \Bigg) \otimes \Bigg(\mathbb{P}_{\boldsymbol{v}^{(i)}|\boldsymbol{v}_k^{(i)}=\pm 1} \otimes \mathbb{P}_{\{y_j^{(i)}\}_{j=1}^{n_i}|\boldsymbol{v}^{(i)},\boldsymbol{v}_k^{(i)}=\pm 1,\{\boldsymbol{x}_j^{(i)}\}_{j=1}^{n_i}} \Bigg) \\
&= \frac{1}{2^{(m-1)d+d-1}} \sum_{\boldsymbol{V}\backslash\{\boldsymbol{v}_k^{(i)}\}} \mathbb{P}_{\boldsymbol{V},i,\pm k},
\end{aligned}$$

where the ⊗ symbol stands for taking the product of two measures and $\mathbb{P}_{\boldsymbol{V},i,\pm k}$ corresponds to the law of all the labels $\boldsymbol{Y}$ conditional on a specific realization of $\{\boldsymbol{V} : \boldsymbol{v}_k^{(i)} = \pm 1\}$ and the features $\boldsymbol{X}$. With the current notations and letting $\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}}$ be the total variation distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$, we can invoke Neyman-Pearson lemma to get

$$\begin{aligned}\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 &\gtrsim \mathbb{E}_{\boldsymbol{X}} \sum_{i\in[m]} p_i\delta_i^2 \sum_{k\in[d]} \Big(1 - \|\mathbb{P}_{i,+k}-\mathbb{P}_{i,-k}\|_{\mathrm{TV}}\Big)\\ &= d\sum_{i\in[m]} p_i\delta_i^2 - \mathbb{E}_{\boldsymbol{X}} \sum_{i\in[m]} p_i\delta_i^2 \sum_{k\in[d]} \|\mathbb{P}_{i,+k}-\mathbb{P}_{i,-k}\|_{\mathrm{TV}}.\end{aligned} \tag{33}$$

We then proceed by

$$\begin{aligned}&\sum_{i\in[m]} p_i\delta_i^2 \sum_{k\in[d]} \|\mathbb{P}_{i,+k}-\mathbb{P}_{i,-k}\|_{\mathrm{TV}}\\ &\le \sum_{i\in[m]} p_i\delta_i^2\sqrt{d}\Big(\sum_{k\in[d]} \|\mathbb{P}_{i,+k}-\mathbb{P}_{i,-k}\|_{\mathrm{TV}}^2\Big)^{1/2}\\ &= \sum_{i\in[m]} p_i\delta_i^2\sqrt{d}\Bigg(\sum_{k\in[d]} \Bigg\|\frac{1}{2^{(m-1)d+d-1}} \sum_{\boldsymbol{V}\setminus\{\boldsymbol{v}_k^{(i)}\}} \mathbb{P}_{\boldsymbol{V},i,+k} - \mathbb{P}_{\boldsymbol{V},i,-k}\Bigg\|_{\mathrm{TV}}^2\Bigg)^{1/2}\\ &= \sum_{i\in[m]} p_i\delta_i^2\sqrt{d}\Bigg(\sum_{k\in[d]} \frac{1}{2^{(m-1)d+d-1}} \sum_{\boldsymbol{V}\setminus\{\boldsymbol{v}_k^{(i)}\}} \|\mathbb{P}_{\boldsymbol{V},i,+k} - \mathbb{P}_{\boldsymbol{V},i,-k}\|_{\mathrm{TV}}^2\Bigg)^{1/2},\end{aligned}$$

where the last inequality is by convexity of the total variation distance. Note that $\mathbb{P}_{\boldsymbol{V},i,\pm k}$ is the product of biased Rademacher random variables: if we let $\mathrm{Rad}(p)$ be the ±1-valued random variable with positive probability p, we can write

$$\mathbb{P}_{\boldsymbol{V},i,\pm k} = \bigotimes_{s\in[m]} \bigotimes_{j\in[n_s]} \mathrm{Rad}\Bigg(\frac{1}{1+\exp\{-\delta_s\langle \boldsymbol{v}^{(s)}, \boldsymbol{x}_j^{(s)}\rangle\}}\Bigg), \qquad \boldsymbol{v}_k^{(i)} = \pm 1.$$

Thus, by Pinsker's inequality, we have

$$\begin{aligned}&\|\mathbb{P}_{\boldsymbol{V},i,+k} - \mathbb{P}_{\boldsymbol{V},i,-k}\|_{\mathrm{TV}}^2\\ &\le \frac{1}{2} D_{\mathrm{JS}}(\mathbb{P}_{\boldsymbol{V},i,+k}\|\mathbb{P}_{\boldsymbol{V},i,-k})\\ &= \frac{1}{2}\sum_{s\neq i}\sum_{j\in[n_s]} 0 + \frac{1}{2}\sum_{j\in[n_i]} D_{\mathrm{JS}}\Bigg[\mathrm{Rad}\Bigg(\frac{1}{1+\exp\{-\delta_i\langle \boldsymbol{v}^{(i)}, \boldsymbol{x}_j^{(i)}\rangle\}}\Bigg)\Bigg\|\mathrm{Rad}\Bigg(\frac{1}{1+\exp\{-\delta_i\langle \tilde{\boldsymbol{v}}^{(i)}, \boldsymbol{x}_j^{(i)}\rangle\}}\Bigg)\Bigg],\end{aligned}$$

where $D_{\mathrm{JS}}(\mathbb{P}\|\mathbb{Q}) = \frac{D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})+D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{2}$ is the Jensen–Shannon divergence between $\mathbb{P}$ and $\mathbb{Q}$, and $\boldsymbol{v}^{(s)}, \tilde{\boldsymbol{v}}^{(s)}$ are two $\mathcal{V}$-valued vectors that only differs in the k-th coordinate. By a standard calculation, one finds that

$$\begin{aligned}D_{\mathrm{JS}}\Bigg[\mathrm{Rad}\Bigg(\frac{1}{1+\exp\{-\delta_i\langle \boldsymbol{v}^{(i)}, \boldsymbol{x}_j^{(i)}\rangle\}}\Bigg)\Bigg\|\mathrm{Rad}\Bigg(\frac{1}{1+\exp\{-\delta_i\langle \tilde{\boldsymbol{v}}^{(i)}, \boldsymbol{x}_j^{(i)}\rangle\}}\Bigg)\Bigg] &\le \delta_i^2(\boldsymbol{v}_k^{(i)} - \tilde{\boldsymbol{v}}_k^{(i)})^2(\boldsymbol{x}_{j,k}^{(i)})^2\\ &= 4\delta_i^2(\boldsymbol{x}_{j,k}^{(i)})^2.\end{aligned}$$

This gives

$$\|\mathbb{P}_{\boldsymbol{V},i,+k} - \mathbb{P}_{\boldsymbol{V},i,-k}\|_{\mathrm{TV}}^2 \le 2\delta_i^2 \sum_{j\in[n_i]} (\boldsymbol{x}_{j,k}^{(i)})^2 \le 2\delta_i^2 c_X^2 n_i.$$

and hence

$$\sum_{i\in[m]} p_i\delta_i^2 \sum_{k\in[d]} \|\mathbb{P}_{i,+k} - \mathbb{P}_{i,-k}\|_{\mathrm{TV}} \le \sqrt{2}c_X \sum_{i\in[m]} p_i\delta_i^3 d n_i^{1/2}.$$

Plugging the above display to (33) gives

$$\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim d\bigg(\sum_{i\in[m]} p_i\delta_i^2 - \sqrt{2}c_X \sum_{i\in[m]} p_i\delta_i^3\sqrt{n_i}\bigg). \tag{34}$$

To this end, all that is left is to choose δ_i approriately so that (1) the above display is as tight as possible; (2) (32) is satisfied. We consider the following two cases:

1. Assume $R^2 \ge d\sum_{i\in[m]} p_i/n_i = dm/N$. Note that we can re-write the requirement (32) to be
$$d\sum_{i\in[m]} p_i\delta_i^2 - \|\sum_{i\in[m]} p_i\delta_i \boldsymbol{v}^{(i)}\|^2 \le R^2.$$
Under the current assumption, this requirement will be satisfied if we choose $\delta_i = c/\sqrt{n_i}$ for any $c \le 1$. Under such a choice, the right-hand side of (34) becomes $\frac{c^2dm}{N}(c - \sqrt{2}c_X)$. Thus, by setting $c = 2\sqrt{2}c_X$, we get the following lower bound:
$$\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim \frac{d}{N/m}.$$

2. Assume $R^2 \le d\sum_{i\in[m]} p_i/n_i = dm/N$. Note that if we set $\delta_i \equiv \delta = cR/\sqrt{d}$ where $c \le 1$, (32) reads
$$c^2R^2 - \|\sum_{i\in[m]} p_i\delta_i \boldsymbol{v}^{(i)}\|^2 \le R^2,$$
which trivially holds. Now, the right-hand side of (34) becomes
$$c^2R^2(1 - \sqrt{2}cc_X \sum_{i\in[m]} p_iR\sqrt{n_i}/\sqrt{d}).$$
Since $p_i = n_i/N$ and $n_i \asymp N/m$, our assumption on R gives
$$\sqrt{2}cc_X \sum_{i\in[m]} p_iR\sqrt{n_i}/\sqrt{d} \lesssim \sum_i \frac{n_i}{N} \cdot \sqrt{\frac{mn_i}{N}} = 1.$$
This means that we can choose c to be a small constant such that the following lower bound holds:
$$\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim R^2.$$

Summarizing the above two cases, we arrive at

$$\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim \frac{d}{N/m} \wedge R^2.$$

Combining the above bound with (31), we get

$$\inf_{\{\widehat{\boldsymbol{w}}^{(i)}\}} \sup_{\{\boldsymbol{w}_\star^{(i)}\}\in\mathcal{P}} \sum_{i\in[m]} p_i \mathbb{E}_{\boldsymbol{S}} \|\widehat{\boldsymbol{w}}^{(i)} - \boldsymbol{w}_\star^{(i)}\|^2 \gtrsim \frac{d}{N/m} \wedge R^2 + \frac{d}{N},$$

which is the desired result. ■

Proof [Proof of (15)] The proof is similar to the proof of (16), and we only provide a sketch here. Without loss of generality we consider the first client. By the same arguments as in the proof of (16), the left-hand side of (15) is lower bounded by a constant multiple of d/N. Now, by considering the same prior distribution on $\mathcal{P}$ as in the proof of (16), we get

$$\inf_{\bar{\boldsymbol{w}}^{(1)}} \sup_{\{\boldsymbol{w}^{(i)}\}\in\mathcal{P}} \mathbb{E}_{\boldsymbol{S}} \|\bar{\boldsymbol{w}}^{(1)} - \boldsymbol{w}_\star^{(1)}\|^2 \gtrsim d\delta_1^2(1-\delta_1\sqrt{n_1}),$$

where the δ_i's should obey the following inequality:

$$\|\delta_i \boldsymbol{v}^{(i)} - \sum_{s\in[m]} p_s \delta_s \boldsymbol{v}^{(s)}\|^2 \le R^2.$$

Choosing $\delta_i \asymp 1/\sqrt{n_i}$ when $R \ge dm/N$ and $\delta_i \asymp R/\sqrt{d}$ otherwise, we arrive at

$$\inf_{\bar{\boldsymbol{w}}^{(1)}} \sup_{\{\boldsymbol{w}^{(i)}\}\in\mathcal{P}} \mathbb{E}_{\boldsymbol{S}} \|\bar{\boldsymbol{w}}^{(1)} - \boldsymbol{w}_\star^{(1)}\|^2 \gtrsim \frac{d}{n_1} \wedge R^2,$$

and the proof is concluded. ■

B. Optimization Convergence of FedProx

This section concerns the optimization convergence of FedProx. We first introduce some notations. Let $\boldsymbol{w}_{t,k}^{(i)}$ be the output of k-th step of Algorithm 2 when the initial local model is given by $\boldsymbol{w}_t^{(i)} \equiv \boldsymbol{w}_{t,0}^{(i)} \equiv \boldsymbol{w}_{t-1,K}^{(i)}$, let $I_{t,k}^{(t)}$ be the corresponding minibatch taken, and denote the initial global model by $\boldsymbol{w}_t^{(\text{global})}$. Let $\mathcal{F}_{t,k}$ be the sigma algebra generated by the randomness by Algorithm 2 up to $\boldsymbol{w}_{t,k}^{(i)}$, namely the randomness in $\Big\{\mathcal{C}_\tau, \{I_{\tau,l}^{(i)} : i \in \mathcal{C}_\tau, 0 \le l \le K_\tau - 1\}\Big\}_{\tau=0}^{t-1}$, $\mathcal{C}_t$, and $\{I_{t,l}^{(i)} : i \in \mathcal{C}_t, 0 \le l \le k-1\}$. For notational convenience we let $\mathcal{C}_T = [m]$ (i.e., all clients are involved in local training in Stage II of Algorithm 2). Then the sequence $\{\boldsymbol{w}_{t,k}^{(i)}\}$ is adapted to the following filtration:

$$\mathcal{F}_{0,0} \subseteq \mathcal{F}_{0,1} \subseteq \cdots \subseteq \mathcal{F}_{0,K} \subseteq \mathcal{F}_{1,0} \subseteq \mathcal{F}_{1,1} \subseteq \cdots \subseteq \mathcal{F}_{1,K} \subseteq \cdots \subseteq \mathcal{F}_{T,K}.$$

We write the optimization problem (20) as

$$\min_{\boldsymbol{w}^{(\text{global})}\in\mathcal{W}} \sum_{i\in[m]} p_i F_i(\boldsymbol{w}^{(\text{global})}, S_i), \tag{35}$$

where

$$F_i(\boldsymbol{w}^{(\text{global})}, S_i) = \min_{\boldsymbol{w}^{(i)}\in\mathcal{W}} \left\{ L_i(\boldsymbol{w}^{(i)}, S_i) + \frac{\lambda}{2}\|\boldsymbol{w}^{(\text{global})} - \boldsymbol{w}^{(i)}\|^2 \right\}. \tag{36}$$

To simplify notations, we introduce the proximal opertor

$$\text{Prox}_{L_i/\lambda}(\boldsymbol{w}^{(\text{global})}) = \text{Prox}_{L_i/\lambda}(\boldsymbol{w}^{(\text{global})}, S_i) = \text{argmin}_{\boldsymbol{w}^{(i)}\in\mathcal{W}} \left\{ L_i(\boldsymbol{w}^{(i)}, S_i) + \frac{\lambda}{2}\|\boldsymbol{w}^{(\text{global})} - \boldsymbol{w}^{(i)}\|^2 \right\}. \tag{37}$$

The high-level idea of this proof is to regard $\lambda \sum_{i\in\mathcal{C}_t}(\boldsymbol{w}_t^{(\text{global})} - \boldsymbol{w}_{t+1}^{(i)})/B^{(\text{global})}$ as a biased stochastic gradient of $\frac{1}{n}\sum_{i\in[m]} F_i(\boldsymbol{w}_t^{(\text{global})}, S_i)$. This idea has appeared in various places (see, e.g., the proof of Proposition 5 in Denevi et al. (2019) and the proof of Theorem 1 in Dinh et al. (2020)). However, the implementation of this idea in our case is more complicated than the above mentioned works in that (1) we are not in an online learning setup (compared to Denevi et al. (2019)); (2) we don't need to assume all clients are training at every round (compared to Dinh et al. (2020)); and (3) we use local SGD for the inner loop (instead of assuming the inner loop can be solved with arbitrary precision as assumed in Dinh et al. (2020)), so the gradient norm depends on λ, and could in principle be arbitrarily large, which causes extra complications.

Lemma 14 (Convergence of the inner loop) *Let Assumption A(a, b) holds. Choose* $\eta_{t,k}^{(i)} = \frac{1}{(\mu+\lambda)(k+1)}$. *Then for any* $k \geq 0$, *we have*

$$\mathbb{E}\left[\|\boldsymbol{w}_{t,k}^{(i)} - \text{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})})\|^2 \;\middle|\; \mathcal{F}_{t,0}, i\in\mathcal{C}_t\right] \leq \frac{8\beta^2 D^2}{\mu^2(k+1)}.$$

Proof See Appendix B.1. ■

Lemma 15 (Convergence of the outer loop) *Let the assumptions in Lemma 14 hold. Choose* $\eta_t^{(\text{global})} = \frac{2(\mu+\lambda)}{\lambda\mu(t+1)}$ *and assume*

$$K_\tau + 1 \geq \frac{(4\tau+20)\lambda^2\beta^2 D^2}{\mu^2(\beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)} \qquad \forall 0 \leq \tau \leq t-1. \tag{38}$$

Then for any $t \geq 0$, *we have*

$$\mathbb{E}_{\mathcal{A}_{\text{FP}}}\|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 \leq \frac{12(\lambda+\mu)^2 m\|\mathbf{p}\|^2(\beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)}{\lambda^2\mu^2(t+1)}, \tag{39}$$

where the expectation is taken over the randomness in Algorithm 2.

Proof See Appendix B.2. ■

Proposition 16 (Optimization error of $\mathcal{A}_{\text{FP}}$) *Under the assumptions of Lemma 14 and 15, for any dataset $\boldsymbol{S} \sim \bigotimes_i \mathcal{D}_i^{\otimes n_i}$, we have*

$$\mathbb{E}_{\mathcal{A}_{\text{FP}}}[\mathcal{E}_{\text{OPT}}] \leq \frac{4(\beta+\lambda)\beta^2 D^2}{\mu^2(K_T+1)} + \frac{6(\lambda+\mu)^2 m\|\mathbf{p}\|^2(\beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)}{\lambda\mu^2(t+1)}$$

Proof By definition we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}_{\text{FP}}}[\mathcal{E}_{\text{OPT}}] &:= \mathbb{E}_{\mathcal{A}_{\text{FP}}}\left[\sum_{i\in[m]} p_i\left(L_i(\boldsymbol{w}_{T+1}^{(i)}, S_i) + \frac{\lambda}{2}\|\boldsymbol{w}_T^{(\text{global})} - \boldsymbol{w}_{T+1}^{(i)}\|^2\right) - \sum_{i\in[m]} p_i F_i(\tilde{\boldsymbol{w}}^{(\text{global})}, S_i)\right] \\
&\overset{\text{(a)}}{\leq} \sum_{i\in[m]} \frac{p_i(\beta+\lambda)}{2} \cdot \mathbb{E}_{\mathcal{A}_{\text{FP}}}\|\boldsymbol{w}_{T+1}^{(i)} - \operatorname{Prox}_{L_i/\lambda}(\boldsymbol{w}_T^{(\text{global})})\|^2 \\
&\quad + \mathbb{E}_{\mathcal{A}_{\text{FP}}}\left[\sum_{i\in[m]} p_i F_i(\boldsymbol{w}_T^{(\text{global})}, S_i) - \sum_{i\in[m]} p_i F_i(\tilde{\boldsymbol{w}}^{(\text{global})}, S_i)\right] \\
&\overset{\text{(b)}}{\leq} \frac{4(\beta+\lambda)\beta^2 D^2}{\mu^2(K_T+1)} + \frac{\lambda}{2}\mathbb{E}_{\mathcal{A}_{\text{FP}}}\|\boldsymbol{w}_T^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 \\
&\overset{\text{(c)}}{\leq} \frac{4(\beta+\lambda)\beta^2 D^2}{\mu^2(K_T+1)} + \frac{6(\lambda+\mu)^2 m\|\mathbf{p}\|^2(\beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)}{\lambda\mu^2(t+1)}.
\end{aligned}$$

where (a) is by smoothness of L_i, (b) is by Lemma 14 and λ-smoothness of $\sum_{i\in[m]} p_i F_i$ (which holds by Lemma 17), and (c) is by Lemma 15. ■

B.1 Proof of Lemma 14: Convergence of the Inner Loop

The proof is an adaptation of the proof of Lemma 1 in Rakhlin et al. (2011). However, we need to deal with the extra complication that the hyperparameter λ can in principle be arbitrarily large. We start by noting that

$$\begin{aligned}
&\|\boldsymbol{w}_{t,k+1}^{(i)} - \operatorname{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})})\|^2 \\
&= \left\|\mathcal{P}_{\mathcal{W}}\left[\boldsymbol{w}_{t,k}^{(i)} - \frac{\eta_{t,k}^{(i)}}{B^{(i)}} \sum_{j\in I_{t,k}^{(i)}}\left(\nabla\ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\text{global})})\right)\right] - \operatorname{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})})\right\|^2 \\
&\leq \left\|\boldsymbol{w}_{t,k}^{(i)} - \frac{\eta_{t,k}^{(i)}}{B^{(i)}} \sum_{j\in I_{t,k}^{(i)}}\left(\nabla\ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\text{global})})\right) - \operatorname{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})})\right\|^2 \\
&= \|\boldsymbol{w}_{t,k}^{(i)} - \operatorname{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})}))\|^2 + \left\|\frac{\eta_{t,k}^{(i)}}{B^{(i)}} \sum_{j\in I_{t,k}^{(i)}}\left(\nabla\ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\text{global})})\right)\right\|^2
\end{aligned}$$

$$- 2\Big\langle \boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})}), \frac{\eta_{t,k}^{(i)}}{B^{(i)}} \sum_{j \in I_{t,k}^{(i)}} \Big(\nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\mathrm{global})}) \Big) \Big\rangle,$$

where the inequality is because $\mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})}) \in \mathcal{W}$ and $\mathcal{P}_{\mathcal{W}}$ is non-expansive. Now by strong convexity and unbiasedness of the stochastic gradients, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})}), \frac{1}{B^{(i)}} \sum_{j \in I_{t,k}^{(i)}} \Big(\nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\mathrm{global})}) \Big) \Big\rangle \ \Big|\ \mathcal{F}_{t,k}, i \in \mathcal{C}_t \Big] \\
&\geq \Big(L_i(\boldsymbol{w}_{t,k}^{(i)}, S_i) + \frac{\lambda}{2} \|\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}^{(\mathrm{global})}\|^2 \Big) \\
&\quad - \Big(L_i(\mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})}), S_i) + \frac{\lambda}{2} \|\mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})}) - \boldsymbol{w}_t^{(\mathrm{global})}\|^2 \Big) \\
&\quad + \frac{1}{2}\Big(\mu_i + \frac{\lambda n}{m n_i} \Big) \|\boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2 \\
&\geq (\mu + \lambda) \|\boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2.
\end{aligned}$$

On the other hand, applying Lemma 19 gives

$$\begin{aligned}
&\mathbb{E}\Big[\Big\| \frac{\eta_{t,k}^{(i)}}{B^{(i)}} \sum_{j \in I_{t,k}^{(i)}} \Big(\nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\mathrm{global})}) \Big) \Big\|^2 \ \Big|\ \mathcal{F}_{t,k}, i \in \mathcal{C}_t \Big] \\
&= (\eta_{t,k}^{(i)})^2 \cdot \Big[\frac{n_i/B^{(i)} - 1}{n_i(n_i - 1)} \sum_{j \in [n_i]} \Big\| \nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) - \overline{\nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_{\boldsymbol{\cdot}}^{(i)})} \Big\|^2 \\
&\qquad + \Big\| \frac{1}{n_i} \sum_{j \in [n_i]} \nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)}) + \lambda(\boldsymbol{w}_{t,k}^{(i)} - \boldsymbol{w}_t^{(\mathrm{global})}) \Big\|^2 \Big] \\
&\leq 2(\eta_{t,k}^{(i)})^2 \beta^2 D^2 \cdot \frac{n_i/B^{(i)} - 1}{(n_i - 1)} + \Big(\beta + \frac{\lambda n}{m n_i} \Big)^2 \|\boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2,
\end{aligned}$$

where in the second line we let $\overline{\nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_{\boldsymbol{\cdot}}^{(i)})} := \sum_{j \in [n_i]} \nabla \ell(\boldsymbol{w}_{t,k}^{(i)}, z_j^{(i)})/n_i$, and in the last line is by the β-smoothness of $\ell(\cdot, z)$. Thus, we get

$$\begin{aligned}
&\mathbb{E}\Big[\|\boldsymbol{w}_{t,k+1}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2 \ \Big|\ \mathcal{F}_{t,k}, i \in \mathcal{C}_t \Big] \\
&\leq \Big[1 - 2\eta_{t,k}^{(i)}(\mu + \lambda) + (\eta_{t,k}^{(i)})^2 (\beta + \lambda)^2 \Big] \|\boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2 \\
&\qquad + 2(\eta_{t,k}^{(i)})^2 \beta^2 D^2 \cdot \frac{n_i/B^{(i)} - 1}{(n_i - 1)}.
\end{aligned} \tag{40}$$

We then proceed by induction. Note that if $k + 1 \leq \frac{8\beta^2}{\mu^2}$, then we have the following trivial bound:

$$\mathbb{E}\Big[\|\boldsymbol{w}_{t,k}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2 \ \Big|\ \mathcal{F}_{t,0}, i \in \mathcal{C}_t \Big] \leq D^2 \leq \frac{8\beta^2 D^2}{\mu^2(k+1)}, \tag{41}$$

where the first inequality is by $\boldsymbol{w}_{t,k}^{(i)}, \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})}) \in \mathcal{W}$ and the second inequality is by our assumption on k. Thus, it suffices to show

$$\mathbb{E}\left[\|\boldsymbol{w}_{t,k+1}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2 \;\middle|\; \mathcal{F}_{t,0}, i \in \mathcal{C}_t\right] \leq \frac{8\beta^2 D^2}{\mu^2(k+2)} \tag{42}$$

based on the inductive hypothesis (41) and $k+1 \geq 8\beta^2/\mu^2$. By the recursive relationship (40) and taking expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}_{t,k+1}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\mathrm{global})})\|^2 \;\middle|\; \mathcal{F}_{t,0}, i \in \mathcal{C}_t\right] \\
&\leq \left[1 - 2\eta_{t,k}^{(i)}(\mu+\lambda) + (\eta_{t,k}^{(i)})^2\left(\beta+\lambda\right)^2\right]\frac{8\beta^2 D^2}{k+1} + 2(\eta_{t,k}^{(i)})^2\beta^2 D^2 \cdot \frac{n_i/B^{(i)} - 1}{(n_i - 1)}.
\end{aligned}$$

Hence (42) is satisfied if

$$8\beta^2 D^2 \cdot \left[\frac{1}{k+2} - \frac{1}{k+1} + \frac{2\eta_{t,k}^{(i)}}{k+1}(\mu+\lambda) - \frac{(\eta_{t,k}^{(i)})^2}{k+1}(\beta+\lambda)^2\right] \geq 2(\eta_{t,k}^{(i)})^2\beta^2 D^2 \cdot \frac{n_i/B^{(i)} - 1}{(n_i - 1)}.$$

By our choice of $\eta_{t,k}^{(i)}$, the above display is equivalent to

$$8\beta^2 D^2 \cdot \left[-\frac{1}{(k+1)(k+2)} + \frac{2}{(k+1)^2} - \frac{1}{(k+1)^3}\left(\frac{\beta+\lambda}{\mu+\lambda}\right)^2\right] \geq \frac{2\beta^2 D^2}{(\mu+\lambda)^2(k+1)^2} \cdot \frac{n_i/B^{(i)} - 1}{n_i - 1},$$

which is further equivalent to

$$8\beta^2 D^2 \cdot \left[-\frac{k+1}{k+2} + 2 - \frac{1}{k+1}\left(\frac{\beta+\lambda}{\mu+\lambda}\right)^2\right] \geq \frac{2\beta^2 D^2}{(\mu+\lambda)^2} \cdot \frac{n_i/B^{(i)} - 1}{n_i - 1}.$$

We now claim that

$$\frac{1}{k+1}\left(\frac{\beta+\lambda}{\mu+\lambda}\right)^2 \leq \frac{1}{2}.$$

Indeed, since $k+1 \geq 8\beta^2/\mu^2$, (1) if $\lambda \leq \beta$, then the left-hand side above is less than $\frac{4\beta^2}{\mu^2(k+1)} \leq \frac{1}{2}$; and (2) if $\lambda \geq \beta$, the left-hand side above is less than $\frac{4}{k+1} \leq \frac{\mu^2}{2\beta^2} \leq \frac{1}{2}$. By the above claim, (42) would hold if

$$4\beta^2 D^2 \geq \frac{2\beta^2 D^2}{(\mu+\lambda)^2} \cdot \frac{n_i/B^{(i)} - 1}{n_i - 1}.$$

We finish the proof by noting that the right-hand side above is bounded above by $\frac{2\beta^2 D^2}{\mu^2}$.

B.2 Proof of Lemma 15: Convergence of the Outer Loop

By construction we have

$$\|\boldsymbol{w}_t^{(\mathrm{global})} - \tilde{\boldsymbol{w}}^{(\mathrm{global})}\|^2$$

$$
\begin{aligned}
&= \left\| \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i (\boldsymbol{w}_t^{(\text{global})} - \boldsymbol{w}_{t+1}^{(i)}) \right\|^2 \\
&= \|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 + \left\| \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i (\boldsymbol{w}_t^{(\text{global})} - \boldsymbol{w}_{t+1}^{(i)}) \right\|^2 \\
&\quad - 2 \left\langle \boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}, \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i (\boldsymbol{w}_t^{(\text{global})} - \boldsymbol{w}_{t+1}^{(i)}) \right\rangle \\
&\le \|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 - \underbrace{2 \left\langle \boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}, \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i \left(\boldsymbol{w}_t^{(\text{global})} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})}) \right) \right\rangle}_{\text{I}} \\
&\quad + \underbrace{2 \left\| \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i \left(\boldsymbol{w}_t^{(\text{global})} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})}) \right) \right\|^2}_{\text{II}} \\
&\quad + \underbrace{2 \left\| \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i \left(\mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})}) - \boldsymbol{w}_{t+1}^{(i)} \right) \right\|^2}_{\text{III}} \\
&\quad \underbrace{- 2 \left\langle \boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}, \frac{\lambda m \eta_t^{(\text{global})}}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i \left(\mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})}) \right) - \boldsymbol{w}_{t+1}^{(i)} \right\rangle}_{\text{IV}}.
\end{aligned}
$$

We first consider Term I. Note that $\frac{\lambda m}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} p_i (\boldsymbol{w}_t^{(\text{global})} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_t^{(\text{global})}))$ is an unbiased stochastic gradient of $\sum_i p_i F_i$, which is $\mu_F = \lambda\mu/(\lambda+\mu)$-strongly convex. Thus, we have

$$
\begin{aligned}
\mathbb{E}[\text{I} \mid \mathcal{F}_{t-1,K_{t-1}}] &= 2\eta_t^{(\text{global})} \left\langle \boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}, \sum_{i \in [m]} p_i \nabla F_i(\boldsymbol{w}_t^{(\text{global})}, S_i) \right\rangle \\
&\ge 2\eta_t^{(\text{global})} \mu_F \|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2.
\end{aligned}
$$

Now for Term II, we have

$$
\begin{aligned}
&\mathbb{E}[\text{II} \mid \mathcal{F}_{t-1,K_{t-1}}] \\
&\le 2(\eta_t^{(\text{global})})^2 \cdot \mathbb{E}\left[\left(\frac{1}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} m p_i \right)^2 \mid \mathcal{F}_{t-1,K_{t-1}} \right] \cdot \max_{i \in [m]} \|\nabla F_i(\boldsymbol{w}_t^{(\text{global})}, S_i)\|^2 \\
&\le 2(\eta_t^{(\text{global})})^2 \cdot \mathbb{E}\left[\left(\frac{1}{B^{(\text{global})}} \sum_{i \in \mathcal{C}_t} m p_i \right)^2 \mid \mathcal{F}_{t-1,K_{t-1}} \right] \cdot \max_{i \in [m]} \cdot (\beta^2 D^2 \wedge 2\lambda \|\ell\|_\infty \wedge \lambda^2 D^2) \\
&\le 2(\eta_t^{(\text{global})})^2 \cdot \left(\frac{1}{m} \sum_{i \in [m]} (m p_i - 1)^2 + 1 \right) \cdot (\beta^2 D^2 \wedge 2\lambda \|\ell\|_\infty \wedge \lambda^2 D^2) \\
&= 2(\eta_t^{(\text{global})})^2 m \|\mathbf{p}\|^2 (\beta^2 D^2 \wedge 2\lambda \|\ell\|_\infty \wedge \lambda^2 D^2),
\end{aligned}
$$

where the second line is by Lemma 18 and the third line is by Lemma 19. For Term III, we invoke Lemma 14 to get

$$\begin{aligned}\mathbb{E}[\mathrm{III} \mid \mathcal{F}_{t-1,K_{t-1}}] &\leq 2\lambda^2(\eta_t^{(\text{global})})^2 \cdot \frac{8\beta^2 D^2}{\mu^2(K_t+1)} \cdot \mathbb{E}\left[\left(\frac{1}{B^{(\text{global})}}\sum_{i\in\mathcal{C}_t} mp_i\right)^2 \mid \mathcal{F}_{t-1,K_{t-1}}\right] \\ &\leq \frac{16\lambda^2(\eta_t^{(\text{global})})^2\beta^2 D^2 m\|\mathbf{p}\|^2}{\mu^2(K_t+1)},\end{aligned}$$

where the last line is again by Lemma 19. For Term IV, we invoke Young's inequality for products to get

$$\begin{aligned}&\mathbb{E}[-\mathrm{IV} \mid \mathcal{F}_{t-1,K_{t-1}}] \\ &\leq (\eta_t^{(\text{global})}\mu_F)\|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 + (\eta_t^{(\text{global})}\mu_F)^{-1} \cdot \mathbb{E}\left[\frac{\mathrm{III}}{2} \,\middle|\, \mathcal{F}_{t-1,K_{t-1}}\right] \\ &\leq (\eta_t^{(\text{global})}\mu_F)\|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 + (\eta_t^{(\text{global})}\mu_F)^{-1} \cdot \frac{8\lambda^2(\eta_t^{(\text{global})})^2\beta^2 D^2 m\|\mathbf{p}\|^2}{\mu^2(K_t+1)}.\end{aligned}$$

Summarizing the above bounds on the four terms, we arrive at

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{w}_{t+1}^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 \,\middle|\, \mathcal{F}_{t-1,K_{t-1}}\right] \\ &\leq (1-\eta_t^{(\text{global})}\mu_F)\|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 + 2\underbrace{(\eta_t^{(\text{global})})^2 m\|\mathbf{p}\|^2(\beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)}_{\mathrm{V}} \\ &\quad + \underbrace{\frac{\lambda^2(\eta_t^{(\text{global})})^2\beta^2 D^2 m\|\mathbf{p}\|^2}{\mu^2(K_t+1)} \cdot \left(16 + \frac{8}{\delta_t^{(\text{global})}\mu_F}\right)}_{\mathrm{VI}}.\end{aligned}$$

We claim that VI ≤ V. Indeed, with our choice of $\eta_t^{(\text{global})} = \frac{2}{\mu_F(t+1)}$, with some algebra, one recognizes that this claim is equivalent to

$$\frac{20+4t}{\mu^2(K_t+1)} \leq \left(\frac{1}{\lambda^2} \wedge \frac{2\|\ell\|_\infty}{\lambda\beta^2 D^2} \wedge \frac{1}{\beta^2}\right),$$

which is exactly (38). Thus, we have

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{w}_{t+1}^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 \,\middle|\, \mathcal{F}_{t-1,K_{t-1}}\right] \\ &\leq (1-\eta_t^{(\text{global})}\mu_F)\|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 + 3\cdot\mathrm{V} \\ &= \left(1-\frac{2}{t+1}\right)\|\boldsymbol{w}_t^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 + \frac{12m\|\mathbf{p}\|^2(\beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)}{\mu_F^2(t+1)^2}.\end{aligned} \tag{43}$$

We then proceed by induction. For the base case, we invoke the strong convexity of $\sum_i p_i F_i$ and Lemma 18 to get

$$\frac{\mu_F^2}{4}\|\boldsymbol{w}_0^{(\text{global})} - \tilde{\boldsymbol{w}}^{(\text{global})}\|^2 \leq \left\|\sum_{i\in[m]} p_i \nabla F_i(\boldsymbol{w}_0^{(\text{global})}, S_i)\right\|^2 \leq \beta^2 D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2.$$

Along with the fact that $1 = (\sum_{i\in[m]} p_i)^2 \le m\|\mathbf{p}\|^2$, we conclude that (15) is true for $t = 0$. Now assume (39) hold for any $0 \le t \le \tau$. For $t = \tau + 1$, using (43) and the inductive hypothesis, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}\|\boldsymbol{w}_{\tau+1}^{(\mathrm{global})} - \tilde{\boldsymbol{w}}^{(\mathrm{global})}\|^2 &\le \left(1 - \frac{2}{\tau+1}\right)\frac{12m\|\mathbf{p}\|^2(\beta^2D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2D^2)}{(\tau+1)\mu_F^2} \\ &\quad + \frac{12m\|\mathbf{p}\|^2(\beta^2D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2D^2)}{(\tau+1)^2\mu_F^2} \\ &= \left(\frac{1}{\tau+1} - \frac{1}{(\tau+1)^2}\right)\cdot\frac{12m\|\mathbf{p}\|^2(\beta^2D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2D^2)}{\mu_F^2} \\ &\le \frac{12m\|\mathbf{p}\|^2(\beta^2D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2D^2)}{(\tau+2)\mu_F^2},\end{aligned}$$

which is the desired result.

B.3 Auxiliary lemmas

Lemma 17 (Convexity and smoothness F_i) *Under Assumption A(b), each F_i is λ-smooth and $\frac{\mu\lambda}{\mu+\lambda}$-strongly convex.*

Proof The smoothness is a standard fact about the Moreau envelope. The strongly convex constant of F_i follows from Theorem 2.2 of Lemaréchal and Sagastizábal (1997). ■

Lemma 18 (A priori gradient norm bound) *Under Assumption A(a, b), for any $w \in \mathcal{W}$ and $i \in [m]$, we have*

$$\|\nabla F_i(w, S_i)\|^2 \le \beta^2D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2D^2.$$

Proof Since $\nabla F_i(w, S_i) = \lambda(w - \mathrm{Prox}_{L_i/\lambda}(w))$, its norm is trivially bounded by λD. Now, since $\mathrm{Prox}_{L_i/\lambda}(w)$ achieves a lower objective value than w for the objective function $L_i(\cdot, S_i) + \frac{\lambda}{2}\|w - \cdot\|^2$, we have

$$\frac{\lambda}{2}\|w - \mathrm{Prox}_{L_i/\lambda}(w)\|^2 \le L_i(w, S_i) - L_i(\mathrm{Prox}_{L_i/\lambda}(w), S_i) \le \|\ell\|_\infty,$$

and hence $\|\nabla F_i(w, S_i)\|^2 \le 2\lambda\|\ell\|_\infty$. Finally, by the first-order condition, we have

$$\nabla L_i(\mathrm{Prox}_{L_i/\lambda}(w), S_i) + \lambda(\mathrm{Prox}_{L_i/\lambda}(w) - w) = 0.$$

Hence, we get $\|\nabla F_i(w, S_i)\| = \|\nabla L_i(\mathrm{Prox}_{L_i/\lambda}(w), S_i)\| \le \beta D$. ■

Lemma 19 (Variance of minibatch sampling) *Let $\mathcal{B} \subseteq [n]$ be a randomly sampled batch with batch size B and let $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$ be an arbitrary set of vectors, then*

$$\mathbb{E}_{\mathcal{B}}\|\frac{1}{B}\sum_{i\in\mathcal{B}} x_i\|^2 = \frac{n/B - 1}{n(n-1)}\sum_{i\in[n]}\|x_i - \bar{x}\|^2 + \|\bar{x}\|^2 \le \frac{1}{n}\sum_{i\in[n]}\|x_i - \bar{x}\|^2 + \|\bar{x}\|^2,$$

where $\bar{x} := \sum_{i\in[n]} x_i/n$.

Proof Since $\mathbb{E}_{\mathcal{B}} \sum_{i \in \mathcal{B}} x_i / B = \bar{x}$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{B}} \| \frac{1}{B} \sum_{i \in \mathcal{B}} x_i \|^2 &= \mathbb{E}_{\mathcal{B}} \| \frac{1}{B} \sum_{i \in \mathcal{B}} x_i - \bar{x} \|^2 + \|\bar{x}\|^2 \\
&= \frac{1}{B^2} \bigg(\sum_{i \in [n]} \mathbb{1}\{i \in \mathcal{B}\} \|x_i - \bar{x}\|^2 + 2 \sum_{i<j} \mathbb{1}\{i, j \in \mathcal{B}\} \langle x_i - \bar{x}, x_j - \bar{x} \rangle \bigg) + \|\bar{x}\|^2 \\
&= \frac{1}{B^2} \bigg(\frac{B}{n} \sum_{i \in [n]} \|x_i - \bar{x}\|^2 + \frac{2B(B-1)}{n(n-1)} \sum_{i<j} \langle x_i - \bar{x}, x_j - \bar{x} \rangle \bigg) + \|\bar{x}\|^2,
\end{aligned}
$$

where the last line is by $\mathbb{P}_{\mathcal{B}}(i \in \mathcal{B}) = B/n$ and $\mathbb{P}_{\mathcal{B}}(i, j \in \mathcal{B}) = B(B-1)n^{-1}(n-1)^{-1}$ for any $i \neq j$. Now, since $\sum_{i \in [n]} \|x_i - \bar{x}\|^2 + 2 \sum_{i<j} \langle x_i - \bar{x}, x_j - \bar{x} \rangle = 0$, we arrive at

$$
\begin{aligned}
\mathbb{E}_{\mathcal{B}} \| \frac{1}{B} \sum_{i \in \mathcal{B}} x_i \|^2 &= \frac{1}{B^2} \bigg(\frac{B}{n} - \frac{B(B-1)}{n(n-1)} \bigg) \sum_{i} \|x_i - \bar{x}\|^2 + \|x\|^2 \\
&= \frac{n/B - 1}{n(n-1)} \sum_{i \in [n]} \|x_i - \bar{x}\|^2 + \|\bar{x}\|^2,
\end{aligned}
$$

which is the desired result. ■

C. Proofs of Upper Bounds

C.1 Proof of Theorem 6

In this proof, we let $\widehat{\boldsymbol{w}}^{(\text{global})}$ be the global minimizer of (8) and we write $\boldsymbol{w}_{\text{avg}}^{(\text{global}, \mathbf{p})} \equiv \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}$ when there is no ambiguity.

Proof [Proof of (9)] We have

$$
\begin{aligned}
0 &= - \sum_{i \in [m]} p_i L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i) + \sum_{i \in [m]} p_i L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i) \\
&\leq - \sum_{i \in [m]} p_i L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i) + \sum_{i \in [m]} p_i L_i(\boldsymbol{w}_{\star}^{(1)}, S_i) \\
&= - \sum_{i \in [m]} \frac{p_i}{n_i} \sum_{j \in [n_i]} \Big(\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus (i,j)}), z_j^{(i)}) - \ell(\boldsymbol{w}_{\star}^{(1)}, z_j^{(i)}) \Big) \\
&\quad + \sum_{i \in [m]} \frac{p_i}{n_i} \sum_{j \in [n_i]} \Big(\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus (i,j)}), z_j^{(i)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_j^{(i)}) \Big),
\end{aligned}
$$

where $\boldsymbol{S}^{\setminus (i,j)}$ stands for the dataset formed by replacing $z_j^{(i)}$ by another $z'_{i,j} \sim \mathcal{D}_i$, which is independent of everything else. Taking expectation in both sides, we get

$$
0 \leq - \sum_{i \in [m]} p_i \cdot \mathbb{E}_{\boldsymbol{S}, Z_i \sim \mathcal{D}_i} [\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), Z_i) - \ell(\boldsymbol{w}_{\star}^{(1)}, Z_i)]
$$

$$
\begin{aligned}
&+ \sum_{i\in[m]} \frac{p_i}{n_i} \sum_{j\in[n_i]} \mathbb{E}_{\boldsymbol{S},z'_{i,j}}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(i,j)}), z_j^{(i)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_j^{(i)})] \\
= &- \sum_{i\in[m]} p_i \cdot \mathbb{E}_{\boldsymbol{S},Z_i\sim\mathcal{D}_i}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), Z_i) - \ell(\boldsymbol{w}_\star^{(i)}, Z_i)] \\
&- \sum_{i\in[m]} p_i \cdot \mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\boldsymbol{w}_\star^{(i)}, Z_i) - \ell(\boldsymbol{w}_\star^{(1)}, Z_i)] \\
&+ \sum_{i\in[m]} \frac{p_i}{n_i} \sum_{j\in[n_i]} \mathbb{E}_{\boldsymbol{S},z'_{i,j}}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(i,j)}), z_j^{(i)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_j^{(i)})].
\end{aligned}
$$

Noting that $\boldsymbol{w}_\star^{(i)}$ is the argmin of $\mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\cdot, Z_i)]$ and invoking the β-smoothness assumption, we get

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{S}}[\text{AER}_{\mathbf{p}}(\widehat{\boldsymbol{w}}^{(\text{global})})] \\
&\le \beta \sum_{i\in[m]} p_i \|\boldsymbol{w}_\star^{(1)} - \boldsymbol{w}_\star^{(i)}\|^2 + \sum_{i\in[m]} \frac{p_i}{n_i} \sum_{j\in[n_i]} \mathbb{E}_{\boldsymbol{S},z'_{i,j}}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(i,j)}), z_j^{(i)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_j^{(i)})] \\
&\le 2\beta \|\boldsymbol{w}_\star^{(1)} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 + 2\beta \sum_{i\in[m]} p_i \|\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\| \\
&\quad + \sum_{i\in[m]} \frac{p_i}{n_i} \sum_{j\in[n_i]} \mathbb{E}_{\boldsymbol{S},z'_{i,j}}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(i,j)}), z_j^{(i)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_j^{(i)})].
\end{aligned}
$$

Taking a weighted average, we arrive at

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{S}}[\text{AER}_{\mathbf{p}}(\widehat{\boldsymbol{w}}^{(\text{global})})] \\
&\le 4\beta R^2 + \sum_{i\in[m]} \frac{p_i}{n_i} \sum_{j\in[n_i]} \mathbb{E}_{\boldsymbol{S},z'_{i,j}}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(i,j)}), z_j^{(i)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_j^{(i)})] \qquad (44)
\end{aligned}
$$

To bound the second term in the right-hand side above, we bound the federated stability of $\widehat{\boldsymbol{w}}^{(\text{global})}$. Without loss of generality we consider the first client. By μ-strongly convexity of L_1, for any $j_1 \in [n_1]$ we have

$$
\begin{aligned}
&\frac{\mu}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2 \\
&\le \sum_{i\in[m]} p_i \Big(L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_i) - L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i) \Big) \\
&= \Big(\sum_{i\ne 1} p_i L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_i) + p_1 L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}) \Big) \\
&\quad - \Big(\sum_{i\ne 1} p_i L_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i) + p_1 L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1^{\setminus j_1}) \Big) \\
&\quad + p_1 \Big(L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1) - L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}) \Big) \\
&\quad + p_1 \Big(L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1^{\setminus j_1}) - L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1) \Big)
\end{aligned}
$$

$$
\begin{aligned}
&\le p_1\left(L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), S_1) - L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), S_1^{\backslash j_1})\right) \\
&\qquad + p_1\left(L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1^{\backslash j_1}) - L_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1)\right) \\
&= \frac{p_1}{n_1}\left(\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), z_{j_1}^{(1)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_{j_1}^{(1)})\right) \\
&\qquad + \frac{p_1}{n_1}\left(\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z'_{1,j_1}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), z'_{1,j_1})\right), \qquad (45)
\end{aligned}
$$

where the second inequality is because $\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)})$ minimizes $L_1(\cdot, S_1^{\backslash j_1}) + \sum_{i\neq 1} n_i L_i(\cdot, S_i)$. By an identical argument as in the proof of Lemma 26, we have

$$
\begin{aligned}
&\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), z_{j_1}^{(1)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_{j_1}^{(1)}) \\
&\le \sqrt{2\beta\|\ell\|_\infty}\cdot\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)})\| + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)})\|^2 \qquad (46)
\end{aligned}
$$

The same bound also holds for $\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z'_{1,j_1}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), z'_{1,j_1})$. Plugging these two bounds to (45) and rearranging terms, we get

$$
\left(\frac{\mu}{2} - \frac{\beta p_1}{n_1}\right)\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)})\| \le \frac{2\sqrt{2\beta\|\ell\|_\infty}\cdot p_1}{n_1}.
$$

Since $n_1 \ge 4\beta p_1/\mu$, we in fact have

$$
\frac{\mu}{4}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)})\| \le \frac{2\sqrt{2\beta\|\ell\|_\infty}\cdot p_1}{n_1}.
$$

Plugging the above display back to (46), we arrive at

$$
\ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), z_{j_1}^{(1)}) - \ell(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), z_{j_1}^{(1)}) \le \frac{16\beta\|\ell\|_\infty p_1}{\mu n_1}\left(1 + \frac{4\beta p_1}{\mu n_1}\right) \le \frac{32\beta\|\ell\|_\infty p_1}{\mu n_1},
$$

where the last inequality is again by $n_1 \ge 4\beta p_1/\mu$. The desired result follows by plugging the above inequality back to (44). ■

Proof [Proof of (10)] Without loss of generality we consider the first client. Since $\boldsymbol{w}_\star^{(1)}$ is the minimizer of $\mathbb{E}_{Z_1\sim\mathcal{D}_1}\ell(\cdot, Z_1)$, by β-smoothness we have

$$
\begin{aligned}
\mathbb{E}_{Z_1\sim\mathcal{D}_1}[\ell(\widehat{\boldsymbol{w}}^{(\text{global})}, Z_1) - \ell(\boldsymbol{w}_\star^{(1)}, Z_1)] &\lesssim \beta\cdot\mathbb{E}_{Z_1\sim\mathcal{D}_1}\|\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_\star^{(1)}\|^2 \\
&\lesssim \beta\cdot\mathbb{E}_{Z_1\sim\mathcal{D}_1}\|\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 + \beta R^2, \qquad (47)
\end{aligned}
$$

where the last inequality is by Part (b) of Assumption B. By optimality of $\widehat{\boldsymbol{w}}^{(\text{global})}$ and the strong convexity of L_i's, we have

$$
\left\langle \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}, S_i), \widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\right\rangle + \frac{\mu}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 \le 0.
$$

If $\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})} = 0$ then we are done. Otherwise, the above display gives

$$\begin{aligned}
&\|\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\| \\
&\le \frac{2}{\mu}\|\sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}, S_i)\| \\
&\le \frac{2}{\mu}\Big(\|\sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\| + \|\sum_{i\in[m]} p_i\big(\nabla L_i(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}, S_i) - \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\big)\|\Big) \\
&\le \frac{2}{\mu}\Big(\|\sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\| + \beta\sum_{i\in[m]} p_i\|\boldsymbol{w}_{\mathbf{p}}^{(\text{global})} - \boldsymbol{w}_\star^{(i)}\|\Big) \\
&\le \frac{2}{\mu}\Big(\|\sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\| + \beta R\Big).
\end{aligned}$$

Thus, we get

$$\|\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 \le \frac{8}{\mu^2}\Big(\|\sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\|^2 + \beta^2 R^2\Big).$$

Taking expectation with respect to the sample $\boldsymbol{S}$ at both sides, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{S}}\|\widehat{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 &\lesssim \frac{1}{\mu^2}\mathbb{E}_{\boldsymbol{S}}\Big\|\sum_{i\in[m]} p_i\big(\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i) - \mathbb{E}_{\boldsymbol{S}}[\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)]\big)\Big\|^2 + \frac{\beta^2 R^2}{\mu^2} \\
&\le \frac{1}{\mu^2}\cdot\sum_{i\in[m]}\frac{p_i^2\sigma^2}{n_i} + \frac{\beta^2 R^2}{\mu^2}.
\end{aligned}$$

Plugging the above inequality to (47) gives the desired result. ■

C.2 Proof of Proposition 10

Proof [Proof of (21)] By the definitions of the AER and $\mathcal{E}_{\text{OPT}}$, we have

$$\begin{aligned}
\text{AER}_{\mathbf{p}} = \mathcal{E}_{\text{OPT}} &+ \frac{\lambda}{2}\sum_{i\in[m]} p_i\Big(\|\tilde{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \tilde{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|^2 - \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|^2\Big) \\
&+ \sum_{i\in[m]} p_i\Big(\mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), Z_i)] - L_i(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), S_i)\Big) \\
&+ \sum_{i\in[m]} p_i\Big(L_i(\tilde{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), S_i) - \mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\boldsymbol{w}_\star^{(i)}, Z_i)]\Big).
\end{aligned}$$

By the basic inequality (23), we can bound the AER by

$\text{AER}_{\mathbf{p}}$

$$\le \mathcal{E}_{\mathrm{OPT}} + \frac{\lambda}{2}\sum_{i\in[m]} p_i \|\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2 + \sum_{i\in[m]} p_i\Big(\mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), Z_i)] - L_i(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), S_i)\Big)$$
$$+ \sum_{i\in[m]} p_i\Big(L_i(\boldsymbol{w}_\star^{(i)}, S_i) - \mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\boldsymbol{w}_\star^{(i)}, Z_i)]\Big).$$

Now, invoking federated stability, we can further bound the AER by

$$\mathrm{AER}_{\mathbf{p}} \le \mathcal{E}_{\mathrm{OPT}} + \frac{\lambda}{2}\sum_{i\in[m]} p_i \|\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2 + 2\sum_{i\in[m]} p_i\gamma_i$$
$$+ \sum_{i\in[m]} p_i \cdot \frac{1}{n_i}\sum_{j\in[n_i]} \mathbb{E}_{z'_{i,j}\sim\mathcal{D}_i}\Big[\mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j)}), Z_i)] - \ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j)}), z_j^{(i)})\Big]$$
$$+ \sum_{i\in[m]} p_i\Big(L_i(\boldsymbol{w}_\star^{(i)}, S_i) - \mathbb{E}_{Z_i\sim\mathcal{D}_i}[\ell(\boldsymbol{w}_\star^{(i)}, Z_i)]\Big),$$

where $\boldsymbol{S}^{\backslash(i,j)}$ is the dataset formed by replacing $z_j^{(i)}$ with a new sample $z'_{i,j}$, and here we are choosing $z'_{i,j}$ to be an independent sample from $\mathcal{D}_i$. Note that the last two terms of the above display have mean zero under the randomness of the algorithm $\mathcal{A}$, the dataset $\boldsymbol{S}$, and $\{z'_{i,j} : i \in [m], j \in [n_i]\}$. Thus, the desired result follows by taking expectation in both sides. ■

Proof [Proof of (22)] Without loss of generality we consider the first client. By definitions of IER_1 and $\mathcal{E}_{\mathrm{OPT}}$, we have

$$p_1 \cdot \mathrm{IER}_1 = \mathcal{E}_{\mathrm{OPT}} + \sum_{i\in[m]} p_i\Big(L_i(\tilde{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), S_i) + \frac{\lambda}{2}\|\tilde{\boldsymbol{w}}^{(\mathrm{global})}(\boldsymbol{S}) - \tilde{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|^2\Big)$$
$$- \sum_{i\in[m]} p_i\Big(L_i(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), S_i) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|^2\Big)$$
$$+ p_1\mathbb{E}_{Z_1\sim\mathcal{D}_1}[\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}), Z_1) - \ell(\boldsymbol{w}_\star^{(1)}, Z_1)].$$

Invoking the basic inequality (24), with some algebra, we arrive at

$$p_1 \cdot \mathrm{IER}_1$$
$$\le \mathcal{E}_{\mathrm{OPT}} + \frac{p_1\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})}(\boldsymbol{S}) - \boldsymbol{w}_\star^{(1)}\|^2 + p_1\Big(\mathbb{E}_{Z_1\sim\mathcal{D}_1}[\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}), Z_1)] - L_1(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}), S_1)\Big)$$
$$+ p_1\Big(L_1(\boldsymbol{w}_\star^{(1)}, S_1) - \mathbb{E}_{Z_1\sim\mathcal{D}_1}[\ell(\boldsymbol{w}_\star^{(1)}, Z_1)]\Big).$$

Now, invoking federated stability for the first client, we can bound its IER by

$$p_1 \cdot \mathrm{IER}_1 \le \mathcal{E}_{\mathrm{OPT}} + \frac{p_1\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})}(\boldsymbol{S}) - \boldsymbol{w}_\star^{(1)}\|^2 + 2p_1\gamma_1$$
$$+ \frac{p_1}{n_1}\sum_{j\in[n_1]} \mathbb{E}_{z'_{1,j}\sim\mathcal{D}_1}\Big[\mathbb{E}_{Z_1\sim\mathcal{D}_1}[\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\backslash(1,j)}), Z_1)] - \ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\backslash(1,j)}), z_j^{(1)})\Big]$$

$$+ p_1\bigg(L_1(\boldsymbol{w}_\star^{(1)}, S_1) - \mathbb{E}_{Z_1\sim\mathcal{D}_1}[\ell(\boldsymbol{w}_\star^{(1)}, Z_1)]\bigg),$$

where we recall that $\boldsymbol{S}^{\backslash(1,j)}$ is the dataset formed by replacing $z_j^{(1)}$ with a new sample $z'_{1,j}$, and here we are choosing $z'_{1,j}$ to be an independent sample from $\mathcal{D}_1$. We finish the proof by taking the expectation with respect to $\mathcal{A}, \boldsymbol{S}, \{z'_{1,j} : j \in [n_1]\}$ at both sides. ■

C.3 Proof of Theorem 11

In this proof, we let $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\mathrm{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\}$ be a generic algorithm that tries to minimize (20). For notiontional simplicity, we use $a_n \lesssim_\beta b_n$ (resp. $a_n \gtrsim_\beta b_n$) to denote that $a_n \le C_\beta b_n$ (resp. $a_n \ge C_\beta b_n$) for large n, where C_β has explicit dependence on a parameter β.

Recall that $(\tilde{\boldsymbol{w}}^{(\mathrm{global})}, \{\tilde{\boldsymbol{w}}^{(i)}\})$ is the global minimizer of (20), and recall the notations in (35)–(37). We start by bounding the federated stability of approximate minimizers of (20). We need the following definition.

Definition 20 (Approximate minimizers) *We say an algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\mathrm{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\}_1^m)$ produces an $(\varepsilon^{(\mathrm{global})}, \{\varepsilon^{(i)}\}_1^m)$-minimizer of the objective function* (20) *on the dataset $\boldsymbol{S}$ if the following two conditions hold:*

1. *there exist a positive constant $\varepsilon^{(\mathrm{global})}$ such that $\|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \tilde{\boldsymbol{w}}^{(\mathrm{global})}\| \le \varepsilon^{(\mathrm{global})}$;*
2. *for any $i \in [m]$, there exist a positive constant $\varepsilon^{(i)}$ such that $\|\widehat{\boldsymbol{w}}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\widehat{\boldsymbol{w}}^{(\mathrm{global})})\| \le \varepsilon^{(i)}$.*

The stability bound is as follows.

Proposition 21 (Federated stability of approximate minimizers) *Let Assumption A(b) holds, and consider an algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\mathrm{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\}_1^m)$ that produces an $(\varepsilon^{(\mathrm{global})}, \{\varepsilon^{(i)}\}_1^m)$-minimizer of the objective function* (20) *on the dataset $\boldsymbol{S}$. Assume in addition that*

$$n_i \ge \frac{4\beta}{\mu}, \qquad p_i\lambda \le \frac{\mu}{16} \qquad \forall i \in [m]. \tag{48}$$

Then $\mathcal{A}$ has federated stability

$$\gamma_i \le \frac{160\beta\|\ell\|_\infty}{n_i(\mu+\lambda)} + \mathrm{Err}_i,$$

where

$$\begin{aligned}\mathrm{Err}_i := {} & 2\sqrt{2\beta\|\ell\|_\infty}\bigg[4\varepsilon^{(\mathrm{global})}\bigg(\frac{\beta+\lambda}{\mu+\lambda} + \frac{3\lambda}{\mu}\bigg) + \varepsilon^{(i)}\bigg(\sqrt{\frac{\beta+\lambda}{\mu+\lambda}} + \frac{16p_i\lambda}{\mu}\bigg)\bigg] \\ & + 8\beta^2\bigg[16(\varepsilon^{(\mathrm{global})})^2\bigg(\frac{\beta+\lambda}{\mu+\lambda} + \frac{3\lambda}{\mu}\bigg)^2 + (\varepsilon^{(i)})^2\bigg(\sqrt{\frac{\beta+\lambda}{\mu+\lambda}} + \frac{16p_i\lambda}{\mu}\bigg)^2\bigg]\end{aligned}$$

is the error term due to not exactly minimizing the soft weight sharing objective (20).

Proof See Appendix C.3.1. ■

Taking the optimization error into account, we have the following result.

Proposition 22 (Federated stability of $\mathcal{A}_{\mathrm{FP}}$) *Let Assumption A(a, b) and* (48) *hold. Run $\mathcal{A}_{\mathrm{FP}}$ with hyperparameters chosen as in Lemma 14 and 15. Then, as long as*

$$T \geq C_1 \cdot \lambda^2(\lambda \vee 1)^2 m\|\mathbf{p}\|^2 n_i^2, \qquad K_T \geq C_2 \cdot \lambda^2(\lambda \vee 1)^2 p_i^2 n_i^2 \qquad \forall i \in [m], \tag{49}$$

the algorithm $\mathcal{A}_{\mathrm{FP}}$ have expected federated stability

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[\gamma_i] \leq C \cdot \frac{\beta\|\ell\|_\infty}{n_i(\mu+\lambda)},$$

where C_1, C_2 are two constants only depending on $(\mu, \beta, \|\ell\|_\infty, D)$, and C is an absolute constant.

Proof By Proposition 21, it suffices to upper bound the error term Err_i by a constant multiple of $\frac{\beta\|\ell\|_\infty}{n_i(\mu+\lambda)}$. Invoking Lemma 15, we have

$$\left(\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[\varepsilon^{(\mathrm{global})}]\right)^2 \leq \mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[(\varepsilon^{(\mathrm{global})})^2] \leq \frac{12(\lambda+\mu)^2 m\|\mathbf{p}\|^2(\beta^2D^2 \wedge 2\lambda\|\ell\|_\infty \wedge \lambda^2 D^2)}{\lambda^2\mu^2(T+1)}.$$

This gives

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[(\varepsilon^{(\mathrm{global})})^2] \lesssim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{(\lambda\vee 1)^2 m\|\mathbf{p}\|^2(1\wedge\lambda\wedge\lambda^2)}{\lambda^2(T+1)} \lesssim \frac{m\|\mathbf{p}\|^2}{T+1}, \tag{50}$$

where we recall that $a_n \lesssim_{(\mu,\beta,\|\mu\|_\infty,D)} b_n$ means $|a_n| \leq Cb_n$ for a constant C that only depending on $(\mu, \beta, \|\mu\|_\infty, D)$, and the last inequality follows from $(\lambda\vee 1)^2(1\wedge\lambda\wedge\lambda^2) \leq \lambda^2$ regardless $\lambda \geq 1$ or $\lambda \leq 1$. Meanwhile, by Lemma 14, we have

$$\left(\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[\varepsilon^{(i)})]\right)^2 \leq \mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[(\varepsilon^{(i)})^2] \leq \frac{8\beta^2D^2}{\mu^2(K_T+1)} \lesssim_{(\beta,D)} \frac{1}{K_T+1}.$$

Recalling the definition of Err_i, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[\mathrm{Err}_i] \\
&\lesssim_{(\mu,\beta,\|\mu\|_\infty,D)} \lambda\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[\varepsilon^{(\mathrm{global})}] + \lambda p_i\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[\varepsilon^{(i)}] + \lambda^2\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[(\varepsilon^{(\mathrm{global})})^2] + p_i^2\lambda^2\mathbb{E}_{\mathcal{A}_{\mathrm{FP}}}[(\varepsilon^{(i)})^2] \\
&\lesssim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{\lambda\sqrt{m}\|\mathbf{p}\|}{\sqrt{T+1}} + \frac{\lambda p_i}{\sqrt{K_T+1}} + \frac{\lambda^2 m\|\mathbf{p}\|^2}{T+1} + \frac{p_i^2\lambda^2}{K_T+1}.
\end{aligned}$$

Thus, it suffices to require

$$\begin{aligned}
\sqrt{T} &\gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \lambda\sqrt{m}\|\mathbf{p}\|n_i(\mu+\lambda), \qquad T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \lambda^2 m\|\mathbf{p}\|^2 n_i(\mu+\lambda), \\
\sqrt{K_T} &\gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \lambda p_i n_i(\mu+\lambda), \qquad K_t \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} p_i^2\lambda^2 n_i(\mu+\lambda),
\end{aligned}$$

which is equivalent to

$$T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \max\{\lambda^2 m\|\mathbf{p}\|^2 n_i^2(\lambda\vee 1)^2, \lambda^2 m\|\mathbf{p}\|^2 n_i(\lambda\vee 1)\} = \lambda^2 m\|\mathbf{p}\|^2 n_i^2(\lambda\vee 1)^2$$

$$K_T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \max\{\lambda^2 p_i^2 n_i^2(\mu \vee 1)^2, p_i^2\lambda^2 n_i(\mu\vee 1)\} = \lambda^2 p_i^2 n_i^2 (\lambda \vee 1)^2,$$

which is exactly (49). ■

Combining the above proposition with Proposition 10. we get the following result.

Proposition 23 (λ-dependent bound on the AER) *Let Assumption A(a, b) and (48) hold. Run $\mathcal{A}_{\rm FP}$ with hyperparameters chosen as in Lemma 14 and 15. Then, as long as*

$$T \geq C_1 \cdot \lambda(\lambda\vee 1)m\|\mathbf{p}\|^2 \cdot \left(\big[\sum_{s\in[m]} p_s/n_s\big]^{-1} \vee \big[\lambda(\lambda\vee 1)n_i^2\big]\right),$$
$$K_T \geq C_2 \cdot (\lambda+1)^2 \cdot \left(\big[\sum_{s\in[m]} p_s/n_s\big]^{-1} \vee \big[\lambda^2 p_i^2 n_i^2\big]\right), \tag{51}$$

for any $i\in[m]$, the algorithm $\mathcal{A}_{\rm FP}$ satisfies

$$\mathbb{E}_{\mathcal{A}_{\rm FP},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}(\mathcal{A}_{\rm FP})] \leq C\cdot \frac{\beta\|\ell\|_\infty}{\mu+\lambda}\sum_{i\in[m]}\frac{p_i}{n_i} + \frac{\lambda}{2}\sum_{i\in[m]} p_i\|\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2,$$

where C_1, C_2 are two constants only depending on $(\mu,\beta,\|\ell\|_\infty, D)$, and C is an absolute constant.

Proof In view of Propositions 10 and 22, it suffices to set T, K_T such that (1) (49) is satisfied; and (2) $\mathbb{E}_{\mathcal{A}_{\rm FP}}[\mathcal{E}_{\rm OPT}]$ is upper bounded by a constant multiple of $\frac{\beta\|\ell\|_\infty}{\mu+\lambda}\sum_{i\in[m]}\frac{p_i}{n_i}$. To achive the second goal, note that by Proposition 16, the optimization error is bounded by

$$\mathbb{E}_{\mathcal{A}_{\rm FP}}[\mathcal{E}_{\rm OPT}] \lesssim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{\lambda\vee 1}{K_T+1} + \frac{\lambda m\|\mathbf{p}\|^2}{T+1}. \tag{52}$$

Thus, it suffices to require $T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{\lambda(\lambda\vee 1)m\|\mathbf{p}\|^2}{\sum_{i\in[m]} p_i/n_i}$ and $K_T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{(\lambda\vee 1)^2}{\sum_{i\in[m]} p_i/n_i}$. This requirement, combined with (49), is exactly (51). ■

With the above proposition at hand, we are ready to give our proof of Theorem 11.

Proof [Proof of Theorem 11] We first define the following three events:

$$\mathsf{A} := \left\{R \geq \sqrt{\sum_{i\in[m]}\frac{p_i}{n_i}}\right\},\ \mathsf{B} := \left\{\frac{\sum_{i\in[m]} p_i^2/n_i}{\sqrt{\sum_{i\in[m]} p_i/n_i}} \leq R \leq \sqrt{\sum_{i\in[m]}\frac{p_i}{n_i}}\right\},\ \mathsf{C} := \left\{R \leq \frac{\sum_{i\in[m]} p_i^2/n_i}{\sqrt{\sum_{i\in[m]} p_i/n_i}}\right\}.$$

We then choose λ to be

$$\lambda = \frac{\mu}{16R^2}\sum_{i\in[m]}\frac{p_i}{n_i}\cdot \mathbb{1}_{\mathsf{A}} + \frac{\mu}{16C_{\mathbf{p}}R}\sqrt{\sum_{i\in[m]}\frac{p_i}{n_i}}\cdot\mathbb{1}_{\mathsf{B}} + \frac{\mu}{16C_{\mathbf{p}}\sum_{i\in[m]} p_i^2/n_i}\sum_{i\in[m]}\frac{p_i}{n_i}\cdot\mathbb{1}_{\mathsf{C}}.$$

We now consider the three events separately.

1. If A holds, then $p_i\lambda = \frac{p_i\mu}{16R^2}\sum_{i\in[m]}\frac{p_i}{\mu_i} \le \frac{p_i\mu}{16} \le \frac{\mu}{16}$. Thus we can invoke Proposition 23 to get

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}] \le \left(\frac{C\beta\|\ell\|_\infty}{\mu} + \frac{\mu}{32}\right)\sum_{i\in[m]}\frac{p_i}{n_i} \lesssim \text{right-hand side of (28).}$$

2. If B holds, then $p_i\lambda = \frac{p_i\mu}{16C_{\mathbf{p}}R}\sqrt{\sum_{i\in[m]}\frac{p_i}{n_i}} \le \frac{p_{\max}\mu N}{16C_{\mathbf{p}}\sum_{i\in[m]}p_i^2/n_i}\sum_{i\in[m]}\frac{p_i}{n_i} \le \frac{\mu}{16}$, where the last inequality is by the definition of $C_{\mathbf{p}}$. Hence, by Proposition 23, we have

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}] \le \left(\frac{16CC_{\mathbf{p}}\beta\|\ell\|_\infty}{\mu} + \frac{\mu}{32C_{\mathbf{p}}}\right)R\sqrt{\sum_{i\in[m]}\frac{p_i}{n_i}} \lesssim \text{right-hand side of (28).}$$

3. If C holds, then $p_i\lambda = \frac{p_i\mu}{16C_{\mathbf{p}}\sum_{i\in[m]}p_i^2/n_i}\sum_{i\in[m]}\frac{p_i}{n_i} \le \frac{\mu}{16}$, and thus Proposition 23 gives

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{AER}_{\mathbf{p}}] \le \left(\frac{16CC_{\mathbf{p}}\beta\|\ell\|_\infty}{\mu} + \frac{\mu}{32C_{\mathbf{p}}}\right)\sum_{i\in[m]}\frac{p_i^2}{n_i} \lesssim \text{right-hand side of (28).}$$

The desired result follows by combining the above three cases together. ■

C.3.1 Proof of Proposition 21: Stability of Approximate Minimizers

We first present two lemmas, from which Proposition 21 will follow.

Lemma 24 (Federated stability of approximate minimizers, Part I) *Let Assumption A(b) holds, and consider an algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\mathrm{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\}_1^m)$ that satisfies the following conditions:*

1. *there exist positive constants $\delta^{(\mathrm{global})}, \zeta^{(\mathrm{global})}$ such that*

$$\sum_{i\in[m]} p_i F_i(\widehat{\boldsymbol{w}}^{(\mathrm{global})}, S_i) \le \delta^{(\mathrm{global})} + \sum_{i\in[m]} p_i F_i(\tilde{\boldsymbol{w}}^{(\mathrm{global})}, S_i), \tag{53}$$

$$\|\sum_{i\in[m]} p_i \nabla F_i(\widehat{\boldsymbol{w}}^{(\mathrm{global})}, S_i)\| \le \zeta^{(\mathrm{global})}. \tag{54}$$

2. *for any $i\in[m]$, there exist positive constants $\{\delta^{(i)}, \zeta^{(i)}, \varepsilon^{(i)}\}_{i=1}^m$ such that*

$$L_i(\widehat{\boldsymbol{w}}^{(i)}, S_i) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\mathrm{global})} - \widehat{\boldsymbol{w}}^{(i)}\|^2 \le \delta^{(i)} + F_i(\widehat{\boldsymbol{w}}^{(\mathrm{global})}, S_i), \tag{55}$$

$$\|\nabla L_i(\widehat{\boldsymbol{w}}^{(i)}, S_i) + \lambda(\widehat{\boldsymbol{w}}^{(i)} - \widehat{\boldsymbol{w}}^{(\mathrm{global})})\| \le \zeta^{(i)}, \tag{56}$$

$$\|\widehat{\boldsymbol{w}}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\widehat{\boldsymbol{w}}^{(\mathrm{global})})\| \le \varepsilon^{(i)}. \tag{57}$$

Assume in addition that (48) *holds. Then $\mathcal{A}$ has federated stability*

$$\gamma_i \leq \frac{160\beta\|\ell\|_\infty}{n_i(\mu+\lambda)} + \sqrt{2\beta\|\ell\|_\infty} \cdot \mathcal{E}_{\lambda,i} + \beta\mathcal{E}_{\lambda,i}^2, \tag{58}$$

where

$$\mathcal{E}_{\lambda,i} := \frac{8\zeta^{(i)}}{\mu+\lambda} + \sqrt{\frac{8\delta^{(i)}}{\mu+\lambda}} + 8\mu^{-1}\left(2\zeta^{(\text{global})} + 4p_i\lambda\varepsilon^{(i)} + \sqrt{\frac{2\mu\lambda\delta^{(\text{global})}}{\mu+\lambda}}\right) \tag{59}$$

is the error term due to not exactly minimizing (20)*.*

Lemma 25 (Federated stability of approximate minimizers, Part II) *Let Assumption A(b) holds and consider an algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\text{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\}_1^m)$ that produces an $(\varepsilon^{(\text{global})}, \{\varepsilon^{(i)}\}_1^m)$-minimizer in the sense of Definition 20. Then $\mathcal{A}$ also satisfies Equations* (53)*—*(57) *with*

$$\delta^{(\text{global})} = \frac{\lambda}{2}\varepsilon^{(\text{global})}, \quad \zeta^{(\text{global})} = \lambda\varepsilon^{(\text{global})}, \quad \delta^{(i)} = \frac{\beta+\lambda}{2}\varepsilon^{(i)}, \quad \zeta^{(i)} = (\beta+\lambda)\varepsilon^{(i)}.$$

Proof These correspondences are consequences of λ-smoothness of F_i and $(\beta+\lambda)$-smoothness of $L_i(\cdot, S_i) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})} - \cdot\|^2$. We omit the details. ■

With the above two lemmas at hand, the proof of Proposition 21 is purely computational:
Proof [Proof of Proposition 21 given Lemma 24 and 25] Invoking 25, the error term $\mathcal{E}_{\lambda,i}$ defined in Equation 59 can be bounded above by

$$\begin{aligned}
\mathcal{E}_{\lambda,i} &\leq \frac{8(\beta+\lambda)}{\mu+\lambda} \cdot \varepsilon^{(\text{global})} + \sqrt{\frac{4(\beta+\lambda)}{\mu+\lambda}} \cdot \varepsilon^{(i)} + \frac{8}{\mu}\left(2\lambda\varepsilon^{(\text{global})} + 4p_i\lambda\varepsilon^{(i)} + \sqrt{\frac{\mu\lambda^2}{\mu+\lambda}}\right) \\
&= 8\varepsilon^{(\text{global})}\left(\frac{\beta+\lambda}{\mu+\lambda} + \frac{2\lambda}{\mu} + \frac{\lambda}{\sqrt{\mu(\mu+\lambda)}}\right) + 2\varepsilon^{(i)}\left(\sqrt{\frac{\beta+\lambda}{\mu+\lambda}} + \frac{16p_i\lambda}{\mu}\right) \\
&\leq 8\varepsilon^{(\text{global})}\left(\frac{\beta+\lambda}{\mu+\lambda} + \frac{3\lambda}{\mu}\right) + 2\varepsilon^{(i)}\left(\sqrt{\frac{\beta+\lambda}{\mu+\lambda}} + \frac{16p_i\lambda}{\mu}\right).
\end{aligned}$$

This gives

$$\mathcal{E}_{\lambda,i}^2 \leq 128(\varepsilon^{(\text{global})})^2\left(\frac{\beta+\lambda}{\mu+\lambda} + \frac{3\lambda}{\mu}\right)^2 + 8(\varepsilon^{(i)})^2\left(\sqrt{\frac{\beta+\lambda}{\mu+\lambda}} + \frac{16p_i\lambda}{\mu}\right)^2.$$

Plugging the above two displays to (58) gives the desired result. ■

We now present our proof of Lemma 24. We start by stating and proving several useful lemmas.

Lemma 26 (From loss stability to parameter stability) *Let Assumption A(b) holds. Then the algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\text{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\})$ has federated stability*

$$\gamma_i \leq \sqrt{2\beta\|\ell\|_\infty} \cdot \|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\| + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\|^2.$$

Proof This lemma has implicitly appeared in the proofs of many stability-based generalization bounds (see, e.g., Section 13.3.2 of Shalev-Shwartz and Ben-David (2014)), and we provide a proof for completeness. By β-smoothness, for an arbitrary $z \in \mathcal{Z}$ we have

$$\begin{aligned}
&\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), z) - \ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}), z) \\
&\le \left\langle \nabla \ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}), z), \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}) \right\rangle + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\|^2 \\
&\le \|\nabla \ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}), z)\| \cdot \|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\| + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\|^2 \\
&\le \sqrt{2\beta\Big(\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}), z) - \min_{\boldsymbol{w}^{(i)} \in \mathcal{W}} \ell(\boldsymbol{w}^{(i)}, z)\Big)} \cdot \|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\| \\
&\qquad + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\|^2 \\
&\le \sqrt{2\beta\|\ell\|_\infty} \cdot \|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\| + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)})\|^2,
\end{aligned}$$

where the last inequality follows from boundedness of ℓ. By a nearly identical argument, the above upper bound also holds for $-\ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}), z) + \ell(\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_i)}), z)$, and the desired result follows. ■

Lemma 27 (Local stability implies global stability) *Assume Assumption A(b) holds and consider an algorithm $\mathcal{A} = (\widehat{\boldsymbol{w}}^{(\text{global})}, \{\widehat{\boldsymbol{w}}^{(i)}\}_1^m)$ that satisfies Equations* (53), (54) *and* (57)*. Then for any $i \in [m], j_i \in [n_i]$, we have*

$$\begin{aligned}
&\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(i,j_i)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\| \\
&\le \frac{\lambda+\mu}{\lambda\mu}\left(2\zeta^{(\text{global})} + \sqrt{\frac{2\lambda\mu\delta^{(\text{global})}}{\lambda+\mu}} + 4p_i\lambda\varepsilon^{(i)} + 2p_i\lambda\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\backslash(i,j_1)}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|\right). \qquad (60)
\end{aligned}$$

Proof Without loss of generality we consider the first client. Let μ_F be the strongly convex constant of $\sum_i p_i F_i$, which, by Lemma 17, is equal to $\sum_i p_i \cdot \frac{\mu\lambda}{\mu+\lambda} = \lambda\mu/(\lambda+\mu)$. Now, by strong convexity, we have

$$\begin{aligned}
&\frac{\mu_F}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)})\|^2 \\
&\le \sum_{i\in[m]} p_i\Big(F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), S_i) - F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i)\Big) \\
&\qquad + \Big\langle \sum_{i\in[m]} p_i \nabla F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i), \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)} - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}))\Big\rangle \\
&\overset{(54)}{\le} \sum_{i\in[m]} p_i\Big(F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}), S_i) - F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_i)\Big) \\
&\qquad + \zeta^{(\text{global})}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\backslash(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\|
\end{aligned}$$

$$
\begin{aligned}
&= \Bigg(p_1 F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}) + \sum_{i \neq 1} p_i F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_i) \Bigg) \\
&\quad - \Bigg(p_1 F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}, S_1^{\setminus j_1}) + \sum_{i \neq 1} p_i F_i(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}, S_i) \Bigg) \\
&\quad + p_1 \Bigg(F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1) - F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}) \\
&\qquad\qquad + F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1^{\setminus j_1}) - F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1) \Bigg) \\
&\quad + \zeta^{(\text{global})} \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\| \\
&\overset{(53)}{\leq} \delta^{(\text{global})} + \zeta^{(\text{global})} \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\| \\
&\quad + p_1 \Bigg(F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1) - F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1) \\
&\qquad\qquad + F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1^{\setminus j_1}) - F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}) \Bigg).
\end{aligned}
$$

Since F_1 is λ-smooth by Lemma 17, we can proceed by

$$
\begin{aligned}
&\frac{\mu_F}{2} \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2 \\
&\leq \delta^{(\text{global})} + \zeta^{(\text{global})} \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\| + p_1 \lambda \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\|^2 \\
&\quad + p_1 \Big\langle \nabla F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1) - \nabla F_1(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}), \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) \Big\rangle.
\end{aligned}
$$

Since $\nabla F_1(\boldsymbol{w}^{(\text{global})}, S_1) = \lambda\Big(\boldsymbol{w}^{(\text{global})} - \mathrm{Prox}_{L_1/\lambda}(\boldsymbol{w}^{(\text{global})}, S_1)\Big)$, with some algebra, the right-hand side above is in fact equal to

$$
\begin{aligned}
&\delta^{(\text{global})} + \zeta^{(\text{global})} \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\| \\
&\quad + p_1 \lambda \Big\langle \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}) - \mathrm{Prox}_{L_1/\lambda}(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}), S_1), \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) \Big\rangle \\
&\quad + p_1 \lambda \Big\langle \mathrm{Prox}_{L_1/\lambda}(\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}), S_1^{\setminus j_1}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}, \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) \Big\rangle \\
&\quad + p_1 \lambda \Big\langle \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}), \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}) \Big\rangle \\
&\overset{(57)}{\leq} \delta^{(\text{global})} + (\zeta^{(\text{global})} + 2p_1\lambda) \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\| \\
&\quad + p_1 \lambda \|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\| \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\|.
\end{aligned}
$$

The above bound gives a quadratic inequality: if we let $\mathsf{s}_G := \|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})\|$ and $\mathsf{s}_1 := \|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|$, then the above bound can be written as

$$
\frac{\mu_F}{2} \cdot \mathsf{s}_G^2 - (\zeta^{(\text{global})} + 2p_1\lambda\varepsilon^{(1)} + p_1\lambda\mathsf{s}_1) \cdot \mathsf{s}_G - \delta^{(\text{global})} \leq 0.
$$

Solving this inequality gives

$$\begin{aligned}\mathsf{s}_G &\le \frac{1}{\mu_F}\cdot\left[\zeta^{(\text{global})}+2p_1\lambda\varepsilon^{(1)}+p_1\lambda\mathsf{s}_1+\sqrt{(\zeta^{(\text{global})}+2p_1\lambda\varepsilon^{(1)}+p_1\lambda\mathsf{s}_1)^2+2\mu_F\delta^{(\text{global})}}\right]\\ &\le \frac{1}{\mu_F}\left(2\zeta^{(\text{global})}+4p_1\lambda\varepsilon^{(1)}+2p_1\lambda\mathsf{s}_1+\sqrt{2\mu_F\delta^{(\text{global})}}\right),\end{aligned}$$

which is exactly (60). ■

Lemma 28 (Parameter stability) *Under the same assumptions as Proposition 21, for any $i \in [m], j_i \in [n_i]$, we have*

$$\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\setminus(i,j_i)})-\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\| \le \frac{16\sqrt{2\beta\|\ell\|_\infty}}{n_i(\mu+\lambda)}+\mathcal{E}_{\lambda,i}.$$

Proof Without loss of generality we consider the first client. Since $L_1(\cdot, S_1)+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})-\cdot\|^2$ is $(\mu+\lambda)$-strongly convex, we have

$$\begin{aligned}&\frac{1}{2}(\mu+\lambda)\|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2\\ &\le \left(L_1(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}),S_1)+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2\right)\\ &\quad-\left(L_1(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}),S_1)+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\|^2\right)\\ &\quad+\left\langle\nabla L_1(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}),S_1)+\lambda(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})),\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\right\rangle\\ &\overset{(56)}{\le}\left(L_1(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}),S_1^{\setminus j_1})+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2\right)\\ &\quad-\left(L_1(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}),S_1^{\setminus j_1})+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\|^2\right)\\ &\quad-\frac{1}{n_1}\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}),z'_{1,j_1})+\frac{1}{n_1}\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}),z^{(1)}_{j_1})+\frac{1}{n}\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}),z'_{1,j_1})-\frac{1}{n_1}\ell(\widehat{\boldsymbol{w}}^{(\boldsymbol{S})},z^{(1)}_{j_1})\\ &\quad-\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2\\ &\quad+\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\|^2-\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\|^2\\ &\quad+\zeta^{(1)}\|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\|\\ &\overset{(55)}{\le}\delta^{(1)}+\zeta^{(1)}\|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\|\\ &\quad+\lambda\left\langle\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S})-\widehat{\boldsymbol{w}}^{(\text{global})}(\boldsymbol{S}^{\setminus(1,j_1)}),\widehat{\boldsymbol{w}}^{(\boldsymbol{S})}-\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\right\rangle\\ &\quad+\frac{1}{n_1}\left(\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}),z'_{1,j_1})-\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}),z'_{1,j_1})+\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}),z^{(1)}_{j_1})-\ell(\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}),z^{(1)}_{j_1})\right)\end{aligned}$$

$$
\begin{aligned}
\leq & \delta^{(1)} + \zeta^{(1)} \|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S})\| \\
& + \frac{2}{n_1}\left(\sqrt{2\beta\|\ell\|_\infty}\|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\| + \frac{\beta}{2}\|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|^2\right) \\
& + \frac{\lambda+\mu}{\mu}\left(2\zeta^{(\text{global})} + \sqrt{\frac{2\lambda\mu\delta^{(\text{global})}}{\lambda+\mu}} + 4p_i\lambda\varepsilon^{(i)} + 2p_i\lambda\|\widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S}^{\setminus(i,j_1)}) - \widehat{\boldsymbol{w}}^{(i)}(\boldsymbol{S})\|\right) \\
& \quad \times \|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|,
\end{aligned}
$$

where the last inequality is by Lemma 26 and Lemma 27. Denoting $\mathsf{s}_1 := \|\widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}) - \widehat{\boldsymbol{w}}^{(1)}(\boldsymbol{S}^{\setminus(1,j_1)})\|$, the above inequality can be written as

$$
C_{\lambda,1}\mathsf{s}_1^2 - \left[\frac{2\sqrt{2\beta\|\ell\|_\infty}}{n_1} + \zeta^{(1)} + \frac{\lambda+\mu}{\mu}\left(2\zeta^{(\text{global})} + 4p_1\lambda\varepsilon^{(1)} + \sqrt{\frac{2\lambda\mu\delta^{(\text{global})}}{\lambda+\mu}}\right)\right]\mathsf{s}_1 - \delta^{(1)} \leq 0, \tag{61}
$$

where

$$
C_{\lambda,1} := \frac{1}{2}(\mu+\lambda) - \frac{\beta}{n_1} - \frac{2p_1\lambda(\lambda+\mu)}{\mu}.
$$

By (48), we have

$$
C_{\lambda,1} \geq \frac{\mu+\lambda}{2} - \frac{\mu}{4} - \frac{2p_1\lambda(\lambda+\mu)}{\mu} \geq \frac{\lambda+\mu}{4} - \frac{2p_1\lambda(\lambda+\mu)}{\mu} = \frac{\lambda+\mu}{4}\cdot\left(1 - \frac{8p_1\lambda}{\mu}\right) \geq \frac{\lambda+\mu}{8}.
$$

In particular, $C_{\lambda,1} > 0$, and thus we can solve the quadratic inequality (61) (similar to the proof of Lemma 27) to get

$$
\mathsf{s}_1 \leq \frac{2\sqrt{2\beta\|\ell\|_\infty}}{C_{\lambda,1}n_1} + \frac{\zeta^{(1)} + \frac{\lambda+\mu}{\mu}\left(2\zeta^{(\text{global})} + 4p_1\lambda\varepsilon^{(1)} + \sqrt{\frac{2\lambda\mu\delta^{(\text{global})}}{\lambda+\mu}}\right)}{C_{\lambda,1}} + \sqrt{\frac{\delta^{(1)}}{C_{\lambda,1}}}.
$$

Plugging in $C_{\lambda,1} \geq (\lambda+\mu)/8$ to the above inequality gives the desired result. ■

We are finally ready to present a proof of Lemma 24:

Proof [Proof of Lemma 24] Invoking Lemma 26, we have

$$
\begin{aligned}
\gamma_i &\leq \sqrt{2\beta\|\ell\|_\infty}\cdot\left(\frac{16\sqrt{2\beta\|\ell\|_\infty}}{n_i(\mu+\lambda)} + \mathcal{E}_{\lambda,i}\right) + \frac{\beta}{2}\left(\frac{16\sqrt{2\beta\|\ell\|_\infty}}{n_i(\mu+\lambda)} + \mathcal{E}_{\lambda,i}\right)^2 \\
&\leq \frac{32\beta\|\ell\|_\infty}{n_i(\mu+\lambda)}\cdot\left(1 + \frac{\beta}{n_i(\mu+\lambda)}\right) + \sqrt{2\beta\|\ell\|_\infty}\cdot\mathcal{E}_{\lambda,i} + \beta\mathcal{E}_{\lambda,i}^2,
\end{aligned}
$$

where in the last line we have used $(a+b)^2 \leq 2a^2 + 2b^2$. We finish the proof by noting that $\frac{\beta}{n_i(\mu+\lambda)} \leq \frac{\beta}{n_i\mu} \leq 4$, where the last inequality is by (48). ■

C.4 Proof of Theorem 12

Compared to the proof of Theorem 11, we need to additionally control the estimation error of the global model.

Proposition 29 (Estimation error of the global model) *Let Assumptions A(b) and B(b) hold. Then*

$$\mathbb{E}_{\boldsymbol{S}}\|\tilde{\boldsymbol{w}}^{(\mathrm{global})}-\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})}\|^2 \le \frac{48\beta^2\sigma^2}{\mu^2\lambda^2}\biggl(\sum_{i\in[m]}\frac{p_i}{\sqrt{n_i}}\biggr)^2+\frac{48\beta^2R^2}{\mu^2}+\frac{12(\mu+\lambda)^2\sigma^2}{\mu^2\lambda^2}\sum_{i\in[m]}\frac{p_i^2}{n_i}.$$

Proof See Appendix C.4.1 ■

With the above proposition, the following result is a counterpart of Proposition 23.

Proposition 30 (λ-dependent bound on the IER) *Let Assumptions A(a, b), B(b) and Equation* (48) *hold. Run $\mathcal{A}_{\mathrm{FP}}$ with hyperparameters chosen as in Lemma 14 and 15. Then, for any $i\in[m]$, as long as*

$$\begin{aligned} T &\ge C_1\lambda(\lambda\vee 1)m\|\mathbf{p}\|^2 n_i\cdot\Bigl(p_i^{-1}\vee[\lambda(\lambda\vee 1)n_i]\Bigr),\\ K_T &\ge C_2(\lambda+1)^2 n_i\Bigl(p_i^{-1}\vee\lambda^2p_i^2n_i\Bigr), \end{aligned}\tag{62}$$

the algorithm $\mathcal{A}_{\mathrm{FP}}$ satisfies both

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A}_{\mathrm{FP}})]\\ &\le\frac{C}{\lambda n_i}\biggl[\beta\|\ell\|_\infty+\frac{\sigma^2\beta^2n_i}{\mu^2}\biggl(\sum_{i\in[m]}\frac{p_i}{\sqrt{n_i}}\biggr)^2+\sigma^2n_i\sum_{i\in[m]}\frac{p_i^2}{n_i}\biggr]+C\lambda\biggl[\biggl(1+\frac{\beta^2}{\mu^2}\biggr)R^2+\frac{\sigma^2}{\mu^2}\sum_{i\in[m]}\frac{p_i^2}{n_i}\biggr)\biggr],\end{aligned}\tag{63}$$

and

$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_i(\mathcal{A}_{\mathrm{FP}})]\le C\biggl(\frac{\beta\|\ell\|_\infty}{\mu n_i}+\lambda D^2\biggr),\tag{64}$$

where C_1, C_2 are two constants only depending on $(\mu,\beta,\|\ell\|_\infty,D)$, and C is an absolute constant.

Proof Without loss of generality we consider the first client. Our assumptions allow us to invoke Propositions 10 and 22 to get

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_1]\\ &\le\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}\biggl[\frac{\mathcal{E}_{\mathrm{OPT}}}{p_1}+\frac{3\lambda}{2}(\varepsilon^{(\mathrm{global})})^2+\frac{3\lambda}{2}\|\tilde{\boldsymbol{w}}^{(\mathrm{global})}-\boldsymbol{w}_{\mathbf{p}}^{(\mathrm{global})}\|^2+\frac{3\lambda}{2}R^2+\frac{320\beta\|\ell\|_\infty}{n_1(\mu+\lambda)}\biggr].\end{aligned}$$

We first show that the expected value of $\mathcal{E}_{\mathrm{OPT}}/p_1$ and $\lambda(\varepsilon^{(\mathrm{global})})^2$ are both bounded above by a constant multiple of $\frac{\beta\|\ell\|_\infty}{n_1(\mu+\lambda)}$. Indeed, by the estimates we have established in Equations (50) and (52), it suffices to require

$$T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \lambda(\lambda\vee 1)m\|\mathbf{p}\|^2 n_1,$$

and

$$K_T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{(\lambda\vee 1)^2 n_1}{p_1}, \qquad T \gtrsim_{(\mu,\beta,\|\mu\|_\infty,D)} \frac{\lambda(\lambda\vee 1)m\|\mathbf{p}\|^2 n_1}{p_1},$$

respectively. And the above two displays, combined with (49), is exactly (62). (64) then follows from the compactness of $\mathcal{W}$. To prove (63), we invoke Proposition 29 to get

$$\begin{aligned}\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_1] &\lesssim \frac{\beta\|\ell\|_\infty}{n_1(\mu+\lambda)} + \lambda\left(1+\frac{\beta^2}{\mu^2}\right)R^2 + \frac{\beta^2\sigma^2}{\mu^2\lambda}\left(\sum_{i\in[m]}\frac{p_i}{\sqrt{n_i}}\right)^2 + \frac{(\mu+\lambda)^2\sigma^2}{\mu^2\lambda}\sum_{i\in[m]}\frac{p_i^2}{n_i} \\ &\lesssim \frac{\beta\|\ell\|_\infty}{n_1\lambda} + \lambda\left(1+\frac{\beta^2}{\mu^2}\right)R^2 + \frac{\beta^2\sigma^2}{\mu^2\lambda}\left(\sum_{i\in[m]}\frac{p_i}{\sqrt{n_i}}\right)^2 + \left(\frac{\sigma^2}{\lambda}+\frac{\lambda\sigma^2}{\mu^2}\right)\sum_{i\in[m]}\frac{p_i^2}{n_i},\end{aligned}$$

and (64) follows by rearranging terms. ■

We now present our proof of Theorem 12.

Proof [Proof of Theorem 12] Without loss of generality we consider the first client. Since all n_i's are of the same order, it suffices to show

$$\mathbb{E}[\mathrm{IER}_i(\mathcal{A}_{\mathrm{FP}})] \lesssim \left[(\mu+\mu^{-1})\left(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\right)+\mu D^2\right]\cdot\left(\frac{R}{\sqrt{N/m}}\wedge\frac{1}{N/m}+\frac{\sqrt{m}}{N}\right). \tag{65}$$

We define the following two events:

$$\mathsf{A} := \{R \geq \sqrt{m/N}\}, \qquad \mathsf{B} := \mathsf{A}^c = \{R < \sqrt{m/N}\},$$

and we set

$$\lambda = \frac{c_{\mathsf{A}}m}{D^2N}\cdot\mathbb{1}_{\mathsf{A}} + c_{\mathsf{B}}\sqrt{\frac{m}{R^2N+1}}\cdot\mathbb{1}_{\mathsf{B}},$$

where $c_{\mathsf{A}}, c_{\mathsf{B}}$ are two constants to be specified later. We consider two cases:

1. If A holds, then from (64) we have $\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_1] \lesssim \left(\frac{\beta\|\ell\|_\infty}{\mu}+c_{\mathsf{A}}\right)\cdot\frac{1}{N/m}$, provided $\lambda p_{\max} \leq \mu/16$. Note that $\lambda p_{\max} \asymp \frac{c_{\mathsf{A}}}{D^2N} \leq c_{\mathsf{A}}/D^2$. So we can choose $\lambda \asymp \mu D^2$, which gives
$$\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_1] \lesssim \left(\frac{\beta\|\ell\|_\infty}{\mu}+\mu D^2\right)\cdot\frac{1}{N/m} \leq \text{right-hand side of (65).}$$

2. If B holds, and if $\lambda p_{\max} \leq \mu/16$ holds, then from (63) we have

   $\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_1]$

$$
\begin{aligned}
&\lesssim \frac{1}{\lambda N/m}\Big(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2}{\mu^2} + \sigma^2\Big) + \lambda\Big[(1+\frac{\beta^2}{\mu^2})R^2 + \frac{\sigma^2}{\mu^2 N}\Big] \\
&\lesssim \Big(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\Big)\cdot\Big(\frac{1}{\lambda N/m} + \lambda(R^2+N^{-1})\Big) \\
&= \Big(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\Big)\cdot\Big(\frac{\sqrt{R^2N+1}}{c_{\mathsf{B}}N/\sqrt{m}} + \frac{c_{\mathsf{B}}}{N}\sqrt{m(R^2N+1)}\Big) \\
&\le \Big(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\Big)\cdot\Big(\frac{R}{c_{\mathsf{B}}\sqrt{N/m}} + \frac{1}{c_{\mathsf{B}}N/\sqrt{m}} + \frac{c_{\mathsf{B}}\sqrt{m}R}{\sqrt{N}} + \frac{c_{\mathsf{B}}\sqrt{m}}{N}\Big) \\
&= \Big(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\Big)\cdot(c_{\mathsf{B}}+c_{\mathsf{B}}^{-1})\Big(\frac{R}{\sqrt{N/m}} + \frac{\sqrt{m}}{N}\Big).
\end{aligned}
$$

Note that $p_{\max}\lambda \le c_{\mathsf{B}}p_{\max}\sqrt{m} \asymp c_{\mathsf{B}}/\sqrt{m} \le c_{\mathsf{B}}$. So to satisfy $p_{\max}\lambda \le \mu/16$, we can choose $c_{\mathsf{B}} \asymp \mu$. This gives

$$
\begin{aligned}
\mathbb{E}_{\mathcal{A}_{\mathrm{FP}},\boldsymbol{S}}[\mathrm{IER}_1] &\lesssim \Big(\beta\|\ell\|_\infty + \frac{\sigma^2\beta^2+\beta^2+\sigma^2}{\mu^2}\Big)(\mu+\mu^{-1})\cdot\Big(\frac{R}{\sqrt{N/m}} + \frac{\sqrt{m}}{N}\Big) \\
&\le \text{right-hand side of (65).}
\end{aligned}
$$

The desired result follows by combining the above two cases together. ■

C.4.1 Proof of Proposition 29: Estimation Error of the Global Model

We begin by proving a useful lemma.

Lemma 31 (Estimating $\boldsymbol{w}_\star^{(i)}$ given the knowledge of $\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}$) *Let Assumption A(b) hold. Then for any $i \in [m]$, we have*

$$
\|\boldsymbol{w}_\star^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})})\| \le \frac{2}{\mu+\lambda}\Big\|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i) + \lambda(\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})})\Big\|.
$$

Proof This follows from an adaptation of the arguments in Theorem 7 of Foster et al. (2019). By strong convexity, we have

$$
\begin{aligned}
&\Big\langle \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i) + \lambda(\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}), \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}) - \boldsymbol{w}_\star^{(i)}\Big\rangle \\
&\qquad + \frac{\mu+\lambda}{2}\|\boldsymbol{w}_\star^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})})\|^2 \\
&\le L_i(\boldsymbol{w}_\star^{(i)}, S_i) + \frac{\lambda}{2}\|\boldsymbol{w}^{(\mathrm{global})} - \boldsymbol{w}_\star^{(i)}\|^2 - L_i(\mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})}), S_i) \\
&\qquad - \frac{\lambda}{2}\|\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})})\|^2 \\
&\le 0.
\end{aligned}
$$

If $\|\boldsymbol{w}^{(i)} - \mathrm{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{P}}^{(\mathrm{global})})\| = 0$ we are done. Otherwise, Cauchy-Schwartz inequality applied to the above display gives the desired result. ■

Now, since $\sum_{i\in[m]} p_iF_i$ is $\mu_F = \mu\lambda/(\mu+\lambda)$-strongly convex, we have

$$\begin{aligned}&\left\langle \sum_{i\in[m]} p_i \nabla F_i(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}), \tilde{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})} \right\rangle + \frac{\mu_F}{2}\|\tilde{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 \\ &\leq \sum_{i\in[m]} p_iF_i(\tilde{\boldsymbol{w}}^{(\text{global})}) - \sum_{i\in[m]} p_iF_i(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}) \\ &\leq 0.\end{aligned}$$

If $\|\tilde{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\| = 0$ we are done. Otherwise, by Cauchy-Schwartz inequality, we get

$$\begin{aligned}&\|\tilde{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\| \\ &\leq \frac{2}{\mu_F}\left\| \sum_{i\in[m]} p_i \nabla F_i(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})})\right\| \\ &= \frac{2}{\mu_F}\left\| \sum_{i\in[m]} p_i \nabla L_i(\text{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}), S_i)\right\| \\ &\leq \frac{2}{\mu_F}\left\| \sum_{i\in[m]} p_i \Big(\nabla L_i(\text{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}), S_i) - \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\Big)\right\| + \frac{2}{\mu_F}\left\| \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\right\| \\ &\overset{(*)}{\leq} \frac{2\beta}{\mu_F}\sum_{i\in[m]} p_i \left\|\text{Prox}_{L_i/\lambda}(\boldsymbol{w}_{\mathbf{p}}^{(\text{global})}) - \boldsymbol{w}_\star^{(i)}\right\| + \frac{2}{\mu_F}\left\| \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\right\| \\ &\overset{(**)}{\leq} \frac{4\beta}{\mu_F(\mu+\lambda)}\sum_{i\in[m]} p_i \left\|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i) + \lambda(\boldsymbol{w}_\star^{(i)} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})})\right\| + \frac{2}{\mu_F}\left\| \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\right\| \\ &\leq \frac{4\beta}{\mu\lambda}\sum_{i\in[m]} p_i \|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\| + \frac{4\beta R}{\mu} + \frac{2(\mu+\lambda)}{\mu\lambda}\left\| \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\right\|,\end{aligned}$$

where $(*)$ is by smoothness of L_i and $(**)$ is by Lemma 31. Thus, we have

$$\begin{aligned}&\|\tilde{\boldsymbol{w}}^{(\text{global})} - \boldsymbol{w}_{\mathbf{p}}^{(\text{global})}\|^2 \\ &\leq \frac{48\beta^2}{\mu^2\lambda^2}\Big(\sum_{i\in[m]} p_i \|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\|\Big)^2 + \frac{48\beta^2R^2}{\mu^2} + \frac{12(\mu+\lambda)^2}{\mu^2\lambda^2}\left\| \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\right\|^2\end{aligned} \tag{66}$$

Note that

$$\Big(\sum_{i\in[m]} p_i \|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\|\Big)^2 = \sum_{i\in[m]} p_i^2 \|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\|^2 + \sum_{i\neq s} p_ip_s \|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\|\|\nabla L_s(\boldsymbol{w}_\star^{(s)}, S_s)\|.$$

Taking expectation at both sides, we arrive at

$$\mathbb{E}_{\boldsymbol{S}}\Big[\Big(\sum_{i\in[m]} p_i \|\nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\|\Big)^2\Big] \leq \sum_{i\in[m]} \frac{p_i^2\sigma^2}{n_i} + \sum_{i\neq s}\frac{p_ip_s\sigma_i\sigma_s}{\sqrt{n_in_s}} = \sigma^2\Big(\sum_{i\in[m]}\frac{p_i}{\sqrt{n_i}}\Big)^2.$$

Meanwhile, we have

$$\mathbb{E}_{\boldsymbol{S}}\Big\| \sum_{i\in[m]} p_i \nabla L_i(\boldsymbol{w}_\star^{(i)}, S_i)\Big\|^2 \le \sum_{i\in[m]} \frac{p_i^2\sigma^2}{n_i}.$$

The desired result follows by plugging the previous two displays to (66).

D. Details on Experiments

In each round (among 100 rounds) of simulation, we first generate $\boldsymbol{w}_\star \in \mathbb{R}^{100}$ with i.i.d. standard Gaussian entries, and we set each local model $\boldsymbol{w}_\star^{(i)} = \boldsymbol{w}_\star + R\cdot\boldsymbol{v}_i$, where $\boldsymbol{v}_i \in \mathbb{R}^{100}$ is a random unit vector that has negative correlation with $\boldsymbol{w}_\star$ and we vary R from 0 to 20. The dataset for the i-th client is then generated by a logistic regression model. We apply FedAvg (Algorithm 1), PureLocalTraining, and FedAvg followed by fine tuning, as well as FedProx (Algorithm 2) to this collection of datasets.

For FedAvg, we assume full participation (i.e., $\mathcal{C}_t = [m]$) and we set the number of communication rounds $T = 20$ and global step size $\eta_t = 0.8$. In its local training stage, we run SGD for 5 epochs with step size 0.2. For PureLocalTraining, we run SGD with step size 0.2 for $20 \cdot 5 = 100$ epochs. For the fine tuning strategy, we first run FedAvg (with the same hyperparameter as the previous case) and then for each client run SGD for 15 epochs with step size 0.2. For FedProx, we again assume full participation, and we set the number of communication rounds $T = 20$, global step size $\eta_t^{\mathrm{global}} = 0.8$, local rounds $K_t = 5$, and local step size $\eta_{t,k}^{(i)} = 0.2$. In all the experiments, the batch size is set to 16.

References

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason D Lee, Sham Kakade, Huan Wang, and Caiming Xiong. How important is the train-validation split in meta-learning? *arXiv preprint arXiv:2010.05843*, 2020.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR, 2019.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Ahmed Khaled Ragab Bayoumi, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529, 2020.

Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, and H Brendan McMahan. Towards federated learning at scale: System design. *Conference on Machine Learning and Systems*, 2019.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *arXiv preprint arXiv:1906.02903*, 2019.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Shuxiao Chen, Sifan Liu, and Zongming Ma. Global and individualized community detection in inhomogeneous multilayer networks. *arXiv preprint arXiv:2012.00933*, 2020.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In *Advances in Neural Information Processing Systems*, pages 10169–10179, 2018.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575, 2019.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Canh T Dinh, Nguyen H Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *arXiv preprint arXiv:2006.08848*, 2020.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

John Duchi. Lecture notes for statistics 311/electrical engineering 377. `http://web.stanford.edu/class/stats311/lecture-notes.pdf`, 2019. Accessed: 2020-10-03.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

Dylan J. Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization, 2019.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, pages 9871–9881, 2019.

Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*, 2020.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtarik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Sharu Theresa Jose and Osvaldo Simeone. An information-theoretic analysis of the impact of task similarity on meta-learning. *arXiv preprint arXiv:2101.08390*, 2021.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, and Rachel Cummings. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Seyed Mohammadreza Mousavi Kalan, Zalan Fabian, A Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *arXiv preprint arXiv:2006.10581*, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, pages 5917–5928, 2019.

Mikhail Konobeev, Ilja Kuzborskij, and Csaba Szepesvári. On optimality of meta-learning in fixed-design regression with weighted biased regularization. *arXiv preprint arXiv:2011.00344*, 2020.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. *arXiv preprint arXiv:2003.08673*, 2020.

Claude Lemaréchal and Claudia Sagastizábal. Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020b.

Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.

James Lucas, Mengye Ren, Irene Kameni, Toniann Pitassi, and Richard Zemel. Theoretical bounds on estimation error for meta-learning. *arXiv preprint arXiv:2010.07140*, 2020.

Grigory Malinovsky, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local sgd to local fixed point methods for federated learning. *arXiv preprint arXiv:2004.01442*, 2020.

OL Mangasarian and MV Solodov. Backpropagation convergence via deterministic nonmonotone perturbed minimization. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, pages 383–390, 1993.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(Jun):967–994, 2005.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1): 2853–2884, 2016.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Jacob Poushter. Smartphone ownership and internet usage continues to climb in emerging economies. *Pew research center*, 22(1):1–44, 2016.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79, 2013.

Changjian Shui, Qi Chen, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. Beyond $\mathcal{H}$-divergence: Domain adaptation theory with jensen-shannon divergence. *arXiv preprint arXiv:2007.15567*, 2020.

Sebastian U. Stich. Local sgd converges fast and communicates little, 2019.

Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020a.

Nilesh Tripuraneni, Michael I Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *arXiv preprint arXiv:2006.11650*, 2020b.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Weiran Wang, Jialei Wang, Mladen Kolar, and Nathan Srebro. Distributed stochastic multi-task learning with graph regularization. *arXiv preprint arXiv:1802.03830*, 2018.

Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*, 2020.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation, 2020.

Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *arXiv preprint arXiv:2006.08950*, 2020.

Hongyang R Zhang, Fan Yang, Sen Wu, Weijie J Su, and Christopher Ré. Sharp bias-variance tradeoffs of hard parameter sharing in high-dimensional linear regression. *arXiv preprint arXiv:2010.11750*, 2020.

Qinqing Zheng, Shuxiao Chen, Qi Long, and Weijie J Su. Federated f-differential privacy. *arXiv preprint arXiv:2102.11158*, 2021.