Can We Theoretically Quantify the Impacts of Local Updates on the Generalization Performance of Federated Learning?

Peizhong Ju
The Ohio State University
peizhong.ju@uky.edu

Haibo Yang
Rochester Institute of Technology
hbycis@rit.edu

Jia Liu
The Ohio State University
liu.1736@osu.edu

Yingbin Liang
The Ohio State University
liang.889@osu.edu

Ness Shroff
The Ohio State University
shroff.11@osu.edu

Auguest, 2024

Abstract

Federated Learning (FL) has gained significant popularity due to its effectiveness in training machine learning models across diverse sites without requiring direct data sharing. While various algorithms along with their optimization analyses have shown that FL with local updates is a communication-efficient distributed learning framework, the generalization performance of FL with local updates has received comparatively less attention. This lack of investigation can be attributed to the complex interplay between data heterogeneity and infrequent communication due to the local updates within the FL framework. This motivates us to investigate a fundamental question in FL: *Can we quantify the impact of data heterogeneity and local updates on the generalization performance for FL as the learning process evolves?* To this end, we conduct a comprehensive theoretical study of FL's generalization performance using a linear model as the first step, where the data heterogeneity is considered for both the stationary and online/non-stationary cases. By providing closed-form expressions of the model error, we rigorously quantify the impact of the number of the local updates (denoted as K) under three settings ($K = 1$, $K < \infty$, and $K = \infty$) and show how the generalization performance evolves with the number of rounds t. Our investigation also provides a comprehensive understanding of how different configurations (including the number of model parameters p and the number of training samples n) contribute to the overall generalization performance, thus shedding new insights (such as benign overfitting) for implementing FL over networks.

1 Introduction

Federated Learning (FL) has recently emerged as a prominent paradigm in the realm of distributed learning, facilitating the collaborative training of machine learning models among clients under the orchestration of a central server. By offering privacy preservation, scalability, and collaborative intelligence, FL holds great potential to revolutionize industries in healthcare, finance, IoT, among others [1, 2, 3, 4]. In FL, the federated averaging (FedAvg) algorithm [5] and its variants have become the prevailing approach. FedAvg leverages local computation at each client and employs a centralized parameter server to aggregate and update the model parameters. The unique feature of FedAvg is that each client runs *multiple local stochastic gradient descent (SGD) steps* between two consecutive communication rounds to reduce the communication frequency between the clients and server. In the literature, it has been shown that FedAvg-type algorithms with local updates achieve fast convergence rates while enjoying a low communication complexity. More importantly,

*This work is published in MobiHoc 2024.

the low communication complexity due to local SGD updates renders FedAvg-type algorithms ideal for deployment over wireless edge networks, where the communications links could likely be highly dynamic, stochastic, and unreliable.

However, even with the evident benefit of being communication-efficient, the impact of local updates on the **generalization performance** of FedAvg-type algorithms remains poorly understood. The lack of such theoretical understanding affects the long-term and large-scale adoption of FL. Particularly, in the FL literature, there remains a significant amount of controversy over how the FL generalization performance is affected under the intricate interplay between *data heterogeneity* and *local update steps*. Specifically, some researchers speculated that data heterogeneity results in poor generalization through empirical experiments [6, 7], while other works argued that FedAvg can generalize very well with data heterogeneity [8, 9, 10, 11]. Notably, it has been empirically demonstrated that FedAvg-type algorithms using a fine-tuned number of local update steps exhibit a better generalization performance than the parallel stochastic gradient descent (SGD) algorithm [9, 10, 11]. So far, however, there is *no* theoretical guiding principle on how to choose an appropriate number of local update steps to achieve good generalization performance in the FL literature. Given the ever-increasing importance of FL as a distributed learning mechanism over networks, a compelling open question arises:

(Q): How does the local update process, when coupled with data heterogeneity, impact the generalization performance of federated learning?

In the FL literature, there have been some initial attempts to theoretically understand the generalization performance of FL (see Section 2 for more discussions). The first line of work employs the traditional analytical tools from statistical learning, such as the "probably approximately correct" (PAC) framework. These works focus on the domain changes due to the data and system heterogeneity. For example, the works in [12] and [13] assumed that clients' data distributions are drawn from a meta-population distribution. Accordingly, two generalization gaps in FL are defined. One is the participation generalization gap, which measures the difference between the empirical and expected risk for participating clients; and the other is the non-participation generalization gap, which measures the difference in the expected risk between participating and non-participating clients. The second class of works studied the training dynamic near a manifold of minima and focused on the effect of stochastic gradient noise on generalization. For instance, the FL generalization behavior was investigated in [6] through the lens of the geometry of the loss and Hessian eigenspectrum, while the long-term FL generlization behavior was studied in [14] using the stochastic differential equation (SDE) approximation. Recently, researchers studied FL generalization under data heterogeneity through algorithmic stability [15]. Also, rate-distortion theoretic bounds on FL the generalization have been established in [16].

Despite the valuable insights on FL generalization offered by the aforementioned existing works, it is important to note that they primarily yield asymptotic results by focusing on domain changes or describing asymptotic behavior such as sufficiently large communication rounds and fine-tuned local steps. Hence, these works all fell short of providing an explicit relationship to characterize how critical factors in FL, (e.g., the number of local updates, the number of communication rounds, and data heterogeneity) affect the generalization of FL in general. To bridge this gap, as a starting point, we conduct the first theoretical study on the number of local updates on FL's generalization performance based on the recent double-descent theoretical framework for over-parameterized learning models. Our objective is *to explicitly quantify the influence of local update steps, data heterogeneity, and the total number of communication rounds on the generalization performance of FL,* all of which are particularly relevant to the deployment of FL over edge networks. We highlight our contributions as follows:

- To lay a theoretical foundation for FL generalization, we start with a linear model with Gaussian features in over-parameterized (related to benign overfitting [17, 18, 19]) and under-parameterized regimes. Specifically, in round t of FL, agent i aims to learn a model $\boldsymbol{w}$ through its own local data that follow the underlying ground truth model $\boldsymbol{y}_{(i),t} = \mathbf{X}_{(i),t}^{\top}\boldsymbol{w}_{(i),t} + \boldsymbol{\epsilon}_{(i),t}, i \in [m]$, where $\boldsymbol{w}_{(i),t}$ is the ground-truth slope. By considering different $\boldsymbol{w}_{(i),t}$, the data samples $(\mathbf{X}_i, \mathbf{y}_i)$ can simulate various patterns of data heterogeneity, including both stationary (i.e., $\boldsymbol{w}_{(i),t} = \boldsymbol{w}_{(i)}$) and online/non-stationary (i.e., $\boldsymbol{w}_{(i),t}$ is time-varying) cases. Utilizing this model allows us to efficiently disentangle the distinct influences of heterogeneous data, local update processes, and communication rounds in FL.

- Based on the aforementioned analytical model, we provide *closed-form* expressions of the generalization error of FedAvg-type algorithms in terms of the number of local update steps. Specifically, we rigorously quantify the impact of local update steps (denoted as K) under three representative regimes ($K = 1$, $K < \infty$, and $K = \infty$) and show how the generalization performance evolves with respect to the number of communication rounds t. Our results reveal some interesting insights: 1) a good pre-trained model "helps" but only to some extent; 2) the effect of noise and heterogeneity accumulates but can be limited; 3) the optimal number of local updates exists only in "some cases," hence *resolving the empirical controversy* regarding the effect of K.
- We note that, in addition to offering insights into FL's deployment over edge networks, our work is also of independent interest in learning theory. Specifically, our closed-form expressions of the FL generalization error contribute to answering, in the FL context, the fundamental question of why an over-parameterized model can generalize well. Note that over-parameterized deep neural networks (DNNs) have been widely used in machine learning (including FL), although it remains a myth why they can generalize well (also known as "benign overfitting"). In the recent literature, a promising approach toward resolving the benign overfitting question is the so-called "double-descent" theoretical framework [20, 21, 22, 23, 24, 25, 18] that starts from over-parameterized linear models. In this work, we extend such double-descent analysis into the FL regime where the distributed learning procedure is more complex than classical centralized learning due to the complications of local updates and data heterogeneity.

The rest of this paper is organized as follows. Section 2 reviews the literature to put our work in comparative perspectives. In Section 3, we introduce the over-parameterized linear model in our FL system. Section 4 presents the main generalization analysis, which is followed by the (sketched) proofs of some key results in Sections 5 and 6. The conclusion is in Section 7.

2 Related Work

1) Federated Learning: Federated Learning (FL) has emerged as a popular distributed learning framework, which harnesses the collaborative power of multiple clients to learn a shared model [26, 27, 28]. Since its inception, FL systems have demonstrated increasing prowess, effectively handling diverse forms of heterogeneity in data, network environments, and worker computing capabilities. A large number of FL algorithms, including FedAvg [29] and its various adaptations [30, 31, 32, 33, 34, 35, 36], have been proposed in the literature. However, it is worth noting that these works only provide insights into the convergence in optimization, while lacking the understanding of generalization performance for FL.

2) Generalization Performance of FL: In the literature, there have been relatively limited studies on the generalization of FL. We categorize these works into three distinct classes. The first line of work employs the traditional analytical tools from statistical learning. The work in [12] assumed that clients' data distributions are drawn from a meta-population distribution. Accordingly, they define two generalization gaps in FL: one is the participation generalization gap to measure the difference between the empirical and expected risk for participating clients, the same as the definition in classic statistical learning; the second is the non-participation generalization gap, which measures the difference of the expected risk between participating and non-participating clients. Following this two-level distribution framework, sharper bounds are provided [13]. Also, the probably approximately correct (PAC) Bayesian framework is used in [37] to investigate a tailored generalization bound for heterogeneous data in FL. Recently, some researchers studied FL generalization under data heterogeneity through algorithmic stability [15]. Meanwhile, PAC-Bayes and rate-distortion theoretic bounds on FL generalization errors have been established in [16]. Similar tools are also used to study FL generalization in [38, 39, 40, 41].

The second line of work studied the FL training dynamic near a manifold of minima and focused on the effect of stochastic gradient noise on generalization. These works used "sharpness" as a tool for characterizing generalization. For instance, the generalization behavior was investigated in [6] and [42] through the lens of the geometry of the loss and Hessian eigenspectrum, which links the model's lack of generalization capacity to the sharpness of the solution under ideal client participation. Based on sharpness, a momentum algorithm with better generalization was proposed in [43]. Also, the long-term generalization behavior of FL is studied in [14] using the stochastic differential equation (SDE) approximation, which showed that local steps could

lead to better generalization under appropriate conditions (e.g., a sufficiently small learning rate, a sufficiently large number of communication rounds, and an appropriately chosen number of local update steps).

We note that all of these existing works on FL generalization only provide asymptotic results on domain changes or describe limiting behavior, such as a large number of communication rounds under a carefully chosen number of local updates. Consequently, they all fell short of establishing a direct quantification that demonstrates how key FL factors (i.e., data heterogeneity, the number of local updates, and the communication round) affect FL generalization.

3) Benign Overfitting and Double Descent: Since our work is intimately related to the double-descent framework for resolving the "benign overfitting" mystery, it is also insightful to provide a quick overview of this research area here. As an initial step to understanding why over-parameterized DNNs generalize well (i.e., "benign overfitting") and exhibit the so-called "double-descent" phenomenon (i.e., the generalization risk descends again beyond the conventional "U-shape" curve in the over-parameterized regime), early attempts in this area started from exploring the minimum ℓ_2-norm [20, 21, 22, 23, 24] or ℓ_1-norm [25, 18] overfitted solutions of the linear models with Gaussian or Fourier features. Later studies in this area investigated the generalization performance of overfitted solutions of shallow neural network approximations. For example, researchers have considered random feature (RF) models [44], two-layer neural tangent kernel (NTK) models [45, 46, 47], and three-layer NTK models [48]. Note that all of these studies have focused only on the centralized learning settings, while our work considers the benign overfitting phenomenon in the FL settings, which are far more complex due to the multi-agent nature and unique complications due to FL, such as local updates and data heterogeneity.

3 System Model

3.1 The Ground-Truth Model, the Learning Model, and Training Samples

As a first step toward a theoretical understanding of the impacts of local updates on the FL generalization performance, we consider the general linear ground truth model which is widely used in the literature on machine learning theory (e.g., [17, 18, 19]):

$$y = \tilde{\boldsymbol{x}}^\top \tilde{\boldsymbol{w}} + \epsilon, \tag{1}$$

where $\tilde{\boldsymbol{x}} \in \mathbb{R}^s$ denotes the feature vector that consists of s true features, $\tilde{\boldsymbol{w}} \in \mathbb{R}^s$ denotes the corresponding ground-truth model parameters, and $\epsilon \in \mathbb{R}$ denote the noise in the output $y \in \mathbb{R}$.

Let p denote the number of features/parameters for the chosen learning model. In other words, a sample is in the form of $(\boldsymbol{x} \in \mathbb{R}^p, y)$. In practice, the number of features could be large (may or may not be necessary) to make sure that all true features are included. Thus, we assume that $p \geq s$ and those p features include all necessary features*. Without loss of generality, we let $\tilde{\boldsymbol{x}}$ be the first s elements of $\boldsymbol{x}$. Correspondingly, we define $\boldsymbol{w} \coloneqq \left[\begin{smallmatrix} \tilde{\boldsymbol{w}} \\ \mathbf{0} \end{smallmatrix}\right] \in \mathbb{R}^p$. Thus, Eq. (1) can be rewritten as $y = \boldsymbol{x}^\top \boldsymbol{w} + \epsilon$. We note that such a linear model is considered in many works on theoretical understanding of the double-descent phenomenon in deep learning theory [20, 21, 22, 23, 24, 25, 18]. In Section 4, we will also show that these linear models lead to insights that have been observed in practical (non-linear) FL.

Consider the FL setting with m clients, where the communication rounds are indexed by $t = 1, 2, \cdots, T$. We use $[m]$ to denote the set $\{1, 2, \cdots, m\}$, and use $[T]$ to denote the set $\{1, 2, \cdots, T\}$. We use the subscript $(\cdot)_{(i),t}$ to denote a quantity for the i-th agent at the t-th round. In the t-th communication round of FL, the i-th client uses $n_{(i),t}$ training samples. Stacking these training samples, we have the following matrix equation.

$$\boldsymbol{y}_{(i),t} = \mathbf{X}_{(i),t}^\top \boldsymbol{w}_{(i),t} + \boldsymbol{\epsilon}_{(i),t}, \tag{2}$$

where $\mathbf{X}_{(i),t} \in \mathbb{R}^{p \times n_{(i),t}}$, $\boldsymbol{w}_{(i),t} \in \mathbb{R}^p$, $\boldsymbol{y}_{(i),t} \in \mathbb{R}^{n_{(i),t}}$, and $\boldsymbol{\epsilon}_{(i),t} \in \mathbb{R}^{n_{(i),t}}$. It is worth noting that Eq. 2 is quite general, including both stationary scenarios where $\boldsymbol{w}_{(i),t} = \mathbf{w}_i$ and non-stationary scenarios with time-varying $\boldsymbol{w}_{(i),t}$ that accounts for environmental changes at the edge devices. The subscript notation $(\cdot)_{(i),t}$ in $\boldsymbol{w}_{(i),t}$ offers a more general framework to model various complications in FL, such as unbalanced

*Our result can be generalized to the case of missing features by treating the missing part as noise.

data, heterogeneity, and non-stationarity. In general FL, there exist ground-truth parameters $\boldsymbol{w}^* \in \mathbb{R}^p$ in the system, which corresponds to the target solution of FL. For example, in simple FL with balanced data, the ground truth is can be written as $\boldsymbol{w}^* = \frac{1}{m}\sum_{i\in[m]} \mathbf{w}_i$.

3.2 Data Distribution, Heterogeneity, and Non-stationarity

To analytically characterize the impact of local updates on the FL generalization performance, we need some assumptions on the distribution of the training data $\left(\mathbf{X}_{(i),t},\ \boldsymbol{y}_{(i),t}\right)_{i\in[m],t=1,2,\cdots,T}$. First, we adopt the independent Gaussian features and noise assumption, which is a common assumption in the literature (e.g., [20, 18]) for analyzing over-parameterized generalization performance. Specifically, we have the following assumption:

Assumption 1. *For any i, t, each element of $\mathbf{X}_{(i),t}$ follows i.i.d. standard Gaussian distribution, and each element of $\boldsymbol{\epsilon}_{(i),t}$ follows independent Gaussian distribution with zero mean and variance $\sigma^2_{(i),t}$.*

Assumption 1 assumes that each dataset per round is unique and freshly obtained, mirroring the conditions of an online data acquisition environment. Besides, it also serves as a realistic approximation for scenarios involving large, fixed datasets.

Since we consider linear models, the heterogeneity of the variance of $\mathbf{X}_{(i),t}$ can be normalized, i.e., it is equivalent to only consider the heterogeneity of the variance of $\boldsymbol{\epsilon}_{(i),t}$ as described in Assumption 1. Note that although $\mathbf{X}_{(i),t}$ has identical distribution among different clients, the training data are heterogeneous in $\boldsymbol{y}_{(i),t}$ because $\boldsymbol{w}_{(i),t}$ can be different and $\sigma_{(i),t}$ may have different values. In other words, $\boldsymbol{y}_{(i),t}$ and $\boldsymbol{y}_{(j),t}$ may have different distributions for different i and j in our model. To quantify the level of heterogeneity in the ground-truth $\boldsymbol{w}_{(i),t}$, we define

$$\boldsymbol{\gamma}_{(i),t} \coloneqq \boldsymbol{w}^* - \boldsymbol{w}_{(i),t}. \tag{3}$$

Intuitively, $\boldsymbol{\gamma}_{(i),t}$ describes the (small) perturbation of agent i's ground truth at the t-th round with respect to the target ground truth $\boldsymbol{w}^*$. The quantification of data heterogeneity here aligns with established research in FL, where the assumption $\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma_G^2$ is commonly used to quantify data heterogeneity [49]. In the case of a linear model, this assumption can be equivalently expressed as $\|\boldsymbol{\gamma}_{(i),t}\|^2 \leq \sigma_G^2$.

3.3 Federated Learning Process

We use mean-squared-error (MSE) as the training loss, i.e., the training loss of the parameters $\hat{\boldsymbol{w}}$ on n samples $(\mathbf{X}, \boldsymbol{y})$ is defined as:

$$L(\hat{\boldsymbol{w}}; \mathbf{X}, \boldsymbol{y}) \coloneqq \frac{1}{2n}\left\|\boldsymbol{y} - \mathbf{X}^\top \hat{\boldsymbol{w}}\right\|^2. \tag{4}$$

We consider the FedAvg algorithm [5], where a central server averages the local updates of each agent (weighted by each agent's number of samples) and then distributes the weighted averaged result to all agents as the initial point of the next local update. We use $\hat{\boldsymbol{w}}_{\text{avg},t} \in \mathbb{R}^p$ to denote the weighted average result at round t, and use $\hat{\boldsymbol{w}}_{(i),t} \in \mathbb{R}^p$ to denote the result of the local update of agent i at round t. The weighted average can be expressed as:

$$\hat{\boldsymbol{w}}_{\text{avg},t} \coloneqq \frac{\sum_{i\in[m]} n_{(i),t}\hat{\boldsymbol{w}}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}. \tag{5}$$

Let $\hat{\boldsymbol{w}}_0$ denote the initialization of the parameters (e.g., starting from a pre-trained model). For notational convenience, we define $\hat{\boldsymbol{w}}_{\text{avg},0} \coloneqq \hat{\boldsymbol{w}}_0$.

Recall that the focus of this paper is to examine the impact of local updates on FL generalization. To this end, we use a parameter $K > 0$ to denote the number of local update steps. We consider the following three regimes in terms of different K values: $K = 1$, $K < \infty$, and $K = \infty$. We use superscripts $(\cdot)^{K=1}$, $(\cdot)^{K<\infty}$, and $(\cdot)^{K=\infty}$ to these cases, respectively. For example, $\hat{\boldsymbol{w}}^{K=1}_{\text{avg},t}$ and $\hat{\boldsymbol{w}}^{K=1}_{(i),t}$ denote the values of $\hat{\boldsymbol{w}}_{\text{avg},t}$ and $\hat{\boldsymbol{w}}_{(i),t}$, respectively, when we consider the setting of $K = 1$.

3.3.1 $K = 1$ (One-Step Gradient)

The simplest algorithm in FL is to perform only one gradient step in each client's local update. Specifically, for all clients $i \in [m]$ and each round $t = 1, 2, \cdots, T$, the result of the local step (denoted by $\hat{\boldsymbol{w}}_{(i),t}^{K=1}$) can be written as:

$$\hat{\boldsymbol{w}}_{(i),t}^{K=1} := \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1} - \alpha_{(i),t} \frac{\partial L(\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}; \mathbf{X}_{(i),t}, \boldsymbol{y}_{(i),t})}{\partial \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}},$$

where $\alpha_{(i),t} > 0$ denotes client i's learning rate (i.e., step size) of the local update in round t.

3.3.2 General $K < \infty$ (Multi-Batch Local Updates)

The general case in FL is that in each round t, every client performs local updates multiple (finite) times. In the k-th update, client i uses $\tilde{n}_{(i),t}$ data $(\mathbf{X}_{(i),t,k}, \boldsymbol{y}_{(i),t,k})$ (as a batch) where $\mathbf{X}_{(i),t,k} \in \mathbb{R}^{p \times \tilde{n}_{(i),t}}$ and $\boldsymbol{y}_{(i),t,k} \in \mathbb{R}^{\tilde{n}_{(i),t}}$. In this paper, we consider the case where $\mathbf{X}_{(i),t,k}$ for all $k \in [K]$ are disjoint and their union is $\mathbf{X}_{(i),t}$. In other words, the data $\mathbf{X}_{(i),t}$ are partitioned evenly into K batches (and thus we have $K \cdot \tilde{n}_{(i),t} = n_{(i),t}$). We define $\hat{\boldsymbol{w}}_{(i),t,k}$ as the result after the k-th batch for client i in round t. Specifically, for the local update in the k-th batch ($k = 1, 2, \cdots, K$), we have

$$\hat{\boldsymbol{w}}_{(i),t,k} := \hat{\boldsymbol{w}}_{(i),t,k-1} - \alpha_{(i),t} \frac{\partial L(\hat{\boldsymbol{w}}_{(i),t,k-1}; \mathbf{X}_{(i),t,k}, \boldsymbol{y}_{(i),t,k})}{\partial \hat{\boldsymbol{w}}_{(i),t,k-1}},$$

where $\alpha_{(i),t} > 0$ denotes the learning rate. We note that $\hat{\boldsymbol{w}}_{(i),t,0} := \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1}$ and $\hat{\boldsymbol{w}}_{(i),t} := \hat{\boldsymbol{w}}_{(i),t,K}$. Also, the general case degenerates to that of Section 3.3.1 when $K = 1$.

3.3.3 $K = \infty$ (Convergence in Local Update)

In this case with $K = \infty$, we consider each client's solution that the local GD/SGD converges to[†], which is different from Sections 3.3.1 and 3.3.2 where every sample is only trained once. In the under-parameterized regime $p < n_{(i),t}$, the convergence point at each client corresponds to the solution that minimizes the local training loss, i.e.,

$$\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} := \arg\min_{\hat{\boldsymbol{w}}} L(\hat{\boldsymbol{w}}; \mathbf{X}_{(i),t}, \boldsymbol{y}_{(i),t}), \quad \text{when } p < n_{(i),t}.$$

In the over-parameterized regime $p > n_{(i),t}$, there are infinitely many solutions that make the training loss zero with probability 1, i.e., overfitted solutions. It is known in the literature that an overfitted solution corresponding to GD/SGD on a linear model in the over-parameterized regime has the smallest ℓ_2-norm of the change of parameters [52, 53]. Specifically, the convergence point of the local updates corresponds to the solution to the following optimization problem: for $t = 1, 2, \cdots, T$, when $p > n_{(i),t}$, we have

$$\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} := \arg\min_{\hat{\boldsymbol{w}}} \quad \left\| \hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty} \right\|, \tag{6}$$

$$\text{subject to} \quad \mathbf{X}_{(i),t}^{\top} \hat{\boldsymbol{w}} = \boldsymbol{y}_{(i),t}. \tag{7}$$

The constraint in Eq. (7) implies that the training loss is exactly zero (i.e., overfitted, which is also known as the interpolation regime).

3.4 Generalization Performance Metric

We then use the distance between the trained model $\hat{\boldsymbol{w}}$ and the ground truth model $\boldsymbol{w}^*$, i.e., model error, to characterize the generalization performance: $L^{\text{model}}(\hat{\boldsymbol{w}}) = \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|^2$. Such model error is equal to the expected test error in some cases.[‡] For convenience, we define

$$\boldsymbol{\Delta}_t := \boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t}, \qquad t = 0, 1, 2, \cdots, T. \tag{8}$$

[†]The difference between a very large but finite K-value and $K = \infty$ has been characterized in the literature of the convergence analysis on gradient descent, e.g., [50, 51].

[‡]We can show that the model error is equal to the expected test error for noise-free data. See Lemma 6.

Therefore, to characterize the generalization performance of FL at the end of round t, we need to quantify $\|\boldsymbol{\Delta}_t\|^2$ with respect to p, K, n, learning rates, initialization, etc. Note that $\boldsymbol{\Delta}_0$ characterizes the difference between the initial weights $\hat{\boldsymbol{w}}_0$ (which can be viewed as starting from an initial or pre-trained model) and the ideal solution $\boldsymbol{w}^*$ (thus $\boldsymbol{\Delta}_0$ is irrelevant to the configuration of K).

3.5 Extra Notations

Let $\text{seq}_i(\cdot)$ denote a sequence of numbers/vectors indexed by i. For $l = 1, 2, \cdots$, and for a real number/vector $\boldsymbol{\beta}_0$, we define a mapping $\mathcal{F}$ as follows:

$$\mathcal{F}(l, \boldsymbol{\beta}_0, \text{seq}_i(a_i), \text{seq}_i(b_i)) \coloneqq \prod_{i=1}^{l} a_i \boldsymbol{\beta}_0 + \sum_{i=1}^{l} b_i \cdot \prod_{j=i+1}^{l} a_j. \tag{9}$$

Eq. (9) corresponds to the general-term formula of $\boldsymbol{\beta}_l$ for the recurrence relation $\boldsymbol{\beta}_i = a_i \boldsymbol{\beta}_{i-1} + b_i$.

4 Main Results

In this section, we will present the closed-form expression of $\mathbb{E}\|\boldsymbol{\Delta}_t\|^2$ for all three cases of K-values. These expressions are complex since our system model considers both the non-stationarity along different rounds and the heterogeneity across different clients. To make our results more accessible, we also provide a simplified version of our results for the special case, where the system is stationary across rounds and the heterogeneity across clients are bounded. Specifically, the simple case is defined as: for all $i \in [m], t \in [T]$,

$$n_{(i),t} = n, \quad \alpha_{(i),t} = \alpha, \quad \sigma_{(i),t} = \sigma, \tag{10}$$

$$\sum\nolimits_{j\in[m]} \boldsymbol{\gamma}_{(j),t} = 0, \tag{11}$$

$$\frac{\sum_{j\in[m]} \|\boldsymbol{\gamma}_{(j),t}\|^2}{m} = \overline{\|\boldsymbol{\gamma}\|^2}, \tag{12}$$

where $\overline{\|\boldsymbol{\gamma}\|^2} \geq 0$ denotes the level of heterogeneity. Here, we consider the balanced data case with a constant learning rate and constant noise in data. The expression $\sum_{j\in[m]} \boldsymbol{\gamma}_{(j),t} = 0$ in Eq. (11) indicates that the ground-truth solution $\boldsymbol{w}^*$ is the average of the all clients' ground truth $\boldsymbol{w}_{(i),t}$, i.e., $\boldsymbol{w}^* = \frac{1}{mT}\sum_{i\in[m],t\in[T]} \boldsymbol{w}_{(i),t}$. With the above notations, we are now ready to present our main results in the following subsections. It is important to note that our general results, including Eqs. (16), (18), (26) and (27), are derived independently of the more restrictive Eqs. (10) to (12), which are only applied in simplified scenarios such as Eqs. (17), (19) and (28).

4.1 The $K = 1$ Case

We define the following short-hand notations:

$$\boldsymbol{g}_l^{K=1} \coloneqq \mathcal{F}\Big(l, \boldsymbol{\Delta}_0, \text{seq}_t\left(\frac{\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})}{\sum_{i\in[m]} n_{(i),t}}\right), \text{seq}_t\left(\frac{\sum_{i\in[m]} \alpha_{(i),t} n_{(i),t} \boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}\right)\Big), \tag{13}$$

$$H_t \coloneqq \frac{\left(\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})\right)^2 + \sum_{i\in[m]} \alpha_{(i),t}^2 n_{(i),t}(p+1)}{(\sum_{i\in[m]} n_{(i),t})^2}, \tag{14}$$

$$G_t \coloneqq \frac{\sum_{i\in[m]} \alpha_{(i),t}^2 p n_{(i),t} \sigma_{(i),t}^2}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} + \frac{\left\|\sum_{i\in[m]} \alpha_{(i),t} n_{(i),t} \boldsymbol{\gamma}_{(i),t}\right\|^2}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} + \frac{\sum_{i\in[m]} \alpha_{(i),t}^2 n_{(i),t}(p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} + \frac{2\left(\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})\right)\left(\sum_{i\in[m]} n_{(i),t}\alpha_{(i),t}\boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{g}_{t-1}\right)}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} - \frac{2\sum_{i\in[m]} \alpha_{(i),t}^2 n_{(i),t}(p+1)\boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{g}_{t-1}^{K=1}}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2}. \tag{15}$$

Theorem 1. *When $K = 1$, we have*

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=1}\right\|^2 = \mathcal{F}(t, \|\boldsymbol{\Delta}_0\|^2, seq_l(H_l), seq_l(G_l)), \forall t \in [T]. \tag{16}$$

For the simple case described by Eqs. (10) to (12), we have

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=1}\right\|^2 = H^t \|\boldsymbol{\Delta}_0\|^2 + \frac{1-H^t}{1-H} G, \tag{17}$$

where $H \coloneqq (1-\alpha)^2 + \frac{\alpha^2(p+1)}{mn}$, $G \coloneqq \frac{p\alpha^2\sigma^2}{mn} + \frac{\alpha^2(p+1)}{mn} \cdot \overline{\|\boldsymbol{\gamma}\|^2}$.

We relegate the proof of Theorem 1 to the supplemental material [54, Appendix B]. In what follows, two important insights for Theorem 1 are in order from the perspectives of model initialization effects and data heterogeneity/noise.

Insight 1) Effect of model initialization: A good initial/pre-trained model helps, but its effect attenuates as the number of communication rounds increases and it cannot address the data heterogeneity challenges. In Theorem 1, $\|\boldsymbol{\Delta}_0\|^2$ denotes the model error induced by the model initialization $\hat{\boldsymbol{w}}_0$ (cf. Eq. (8)). Theorem 1 shows that starting from a good initialization (e.g., a pre-trained model) reduces the training time required to reach a target error rate. The reason is that a good initial/pre-trained model is usually closer to the target solution $\boldsymbol{w}^*$ than a random model initialization. Thus, $\|\boldsymbol{\Delta}_0\|$ will be small and it helps to reduce the model error. This result theoretically explains previously observed experimental results that using pre-trained models as the initialization for FL accelerates the training process [55, 56]. Meanwhile, we note that the coefficient of $\|\boldsymbol{\Delta}_0\|^2$ decreases as t increases when the learning rate is relatively small.[§] This means that the effect of the pre-trained model diminishes as the number of communication rounds increases. As $t \to \infty$, the first term in Eq. (17) asymptotically goes to 0, signifying a vanishing effect of the pre-trained model. This finding is consistent with existing analyses in FL, suggesting that pre-training becomes unnecessary with a sufficiently long training [14]. In addition, Theorem 1 shows that the error induced by data noise and heterogeneity is not affected by the model initialization. This means that even a good initial/pre-trained model cannot alleviate the problems caused by heterogeneous data, which theoretically confirms prior experimental observations [55].

Insight 2) Effect of noise and heterogeneity: Errors arising from data noise and heterogeneity accumulate as the number of communication rounds increases, but eventually converge to an asymptotic limit. In Eq. (17), the coefficient of the second error term attributed to data noise and heterogeneity (G) is expressed as $\frac{1-H^t}{1-H} = 1 + H + H^2 + \cdots + H^{t-1}$. This implies that the error induced by data noise and heterogeneity accumulates as the value of t increases. Meanwhile, this error term is bounded from above and it eventually converges to $\frac{1}{1-H}G$ as $t \to \infty$. This aligns with the empirical observations that FL algorithms remain effective, despite the occurrence of model drift resulting from data heterogeneity [8, 57, 58, 59].

Experiments. We perform simulations to illustrate the influence of model initialization in FL. The experimental setup is as follows: $K = 1$, $m = 3$, $p = 200$, $n_{(i),t} = 50$, $s = 5$, $\left\|\boldsymbol{\gamma}_{(i),t}\right\| = 0.5$, and $\sigma_{(i),t} = 0.7$

[§]In Eq. (17), $H < 1$ when $\alpha_{(i),t} < \frac{2}{1+\frac{p+1}{mn}}$.

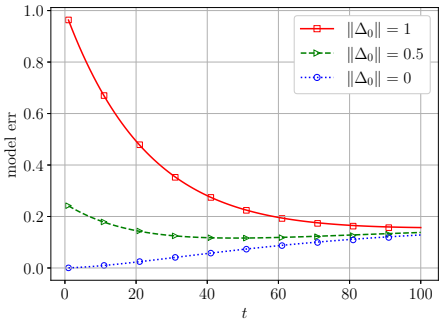

Figure 1: Experimental and analytical values of the model error w.r.t. t where $K = 1$, $m = 3$, $p = 200$, $n_{(i),t} = 50$, $s = 5$, $\left\|\boldsymbol{\gamma}_{(i),t}\right\| = 0.5$, and $\sigma_{(i),t} = 0.7$ for all i, t. Each marker point is the experimental value by averaging over 20 simulation runs. The curves are theoretical values of Theorem 1. (All markers are close to curves, which validates Theorem 1.)

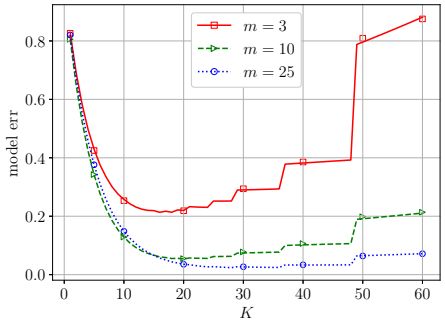

Figure 2: Experimental and theoretical values of the model error w.r.t. K where $t = 5$, $s = 5$, $p = 200$, $\|\boldsymbol{\Delta}_0\| = 1$, $n_{(i),t} = 144$, $\left\|\boldsymbol{\gamma}_{(i),t}\right\| = 0.5$, and $\sigma_{(i),t} = 0.7$ for all i, t. Each marker point is the experimental value by averaging over 20 simulation runs. The curves are theoretical values of Theorem 2. The lowest points of the three curves for cases $m = 3, 10, 25$ are located at $K = 15, 19, 27$, respectively.

for all i, t. Each marker point denotes the outcome of simulations averaged over 20 simulation trials. In Fig. 1, we plot the model error with respect to (w.r.t.) t for three different pre-trained models: $\|\boldsymbol{\Delta}_0\| = 1$ (red solid line with markers "□"), $\|\boldsymbol{\Delta}_0\| = 0.5$ (green dashed line with markers "▷"), and $\|\boldsymbol{\Delta}_0\| = 0$ (blue dotted line with markers "○"). Generally, the simulation demonstrates the tightness of our theoretical findings and confirms our two insights mentioned above. The blue curve, indicative of the smallest initial model error, initially outperforms the other two curves. However, this performance gap diminishes as time progresses. This observed trend aligns with our insights into the impact of model initialization. Conversely, as the blue curve originates from the ideal solution, its upward trend with respect to t is solely attributed to noise and heterogeneity. This observation further validates our understanding of the influence of noise and heterogeneity.

4.2 The General $K < \infty$ Case

Similar to Eqs. (14) and (15), we define $\mathcal{J}_l, \mathcal{Q}_l \in \mathbb{R}$. The expressions of $\mathcal{J}_l$ and $\mathcal{Q}_l$ only contain $n_{(i),t}$, p, $\alpha_{(i),t}$, $\boldsymbol{\gamma}_{(i),t}$, $\boldsymbol{\Delta}_0$, and the number of local steps K. The formal definitions are provided in Eqs. (64) and (65) at the beginning of supplemental material [54, Appendix C] due to space limitation.

Theorem 2. *When $K < \infty$, we have*

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2 = \mathcal{F}\left(t, \|\boldsymbol{\Delta}_0\|^2, seq_l\left(\mathcal{J}_l\right), seq_l\left(\mathcal{Q}_l\right)\right). \tag{18}$$

For the simple case described by Eqs. (10) to (12) and by further letting $\overline{\|\boldsymbol{\gamma}\|^2} = 0$, we have

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2 = \mathcal{J}^t \|\boldsymbol{\Delta}_0\|^2 + \frac{1-\mathcal{J}^t}{1-\mathcal{J}} \cdot \frac{\alpha^2 p \sigma^2}{m\tilde{n}} \cdot \frac{1-\mathcal{A}^K}{1-\mathcal{A}}, \tag{19}$$

where $\tilde{n} \coloneqq \lfloor n/K \rfloor$, $\mathcal{A} \coloneqq (1-\alpha)^2 + \frac{\alpha^2(p+1)}{\tilde{n}}$, $\mathcal{J} \coloneqq \frac{\mathcal{A}^K + (m-1)(1-\alpha)^{2K}}{m}$.

The proof for Theorem 2 is provided in the supplemental material [54, Appendix C]. Building upon the insights gained from Theorem 2, we have the following discussions concerning the impact of the local update step K.

Insight 3) Effect of the local update step number K: The optimal choice of finite K sometimes exists. In Eq. (19), the local update step number K together with several other factors simultaneously

influence two error terms. Therefore, the optimal choice of K is dependent on other configurations, such as the number of communication round t, $\|\boldsymbol{\Delta}_0\|^2$ (determined by the model initialization), and the noise denoted by σ^2. Through an analysis of how Eq. (19) evolves with K, we establish the following proposition for the optimal choice of K:

Proposition 1. *The existence of an optimal choice of K (defined by K_{opt}) for Eq. (19) in different cases are as follows:*

(1) A finite K_{opt}-value must exist when $\tilde{n}$ is fixed (i.e., n is determined by $K\tilde{n}$), α is sufficiently small[¶]*, and $t \to \infty$.*

(2) A finite K_{opt}-value does not exist (i.e., $K_{opt} = \infty$) when $\tilde{n}$ is fixed, α is sufficiently small, and $\sigma = 0$.

(3) When n is fixed (i.e., $\tilde{n}$ is determined by $\lfloor n/K \rfloor$), $t < \infty$, $\alpha \leq 0.1$, $m \geq 3$, and $\sigma = 0$, if we neglect the difference between $\lfloor n/K \rfloor$ and n/K, then

$$\frac{n}{p+1}\left(\frac{2}{\alpha}-1\right) \leq K_{opt} \leq \frac{n}{p+1}\frac{(m-2)}{\alpha^3}. \tag{20}$$

In Proposition 1, we show that the optimal and finite K-value only exists in some cases, whose value depends on other parameters in one specific problem instance. For example, the upper bound of K_{opt} in Eq. (20) indicates that **the optimal K may increase when the number of agents m increases**. This discovery offers a theoretical explanation of the experimental controversy, wherein switching to local update steps yields divergent outcomes for various tasks; some exhibit improved performance, while others do not [9, 11, 14]. Proof of Proposition 1 is provided in Section 5.

Experiments. Following a similar setting in Fig. 1, we plot the model error against the local steps K when $n_{(i),t}$ is fixed in Fig. 2. These three curves in Fig. 2 correspond to different values of m. We can see that each of the three curves in Fig. 2 has a minimum. The lowest points of the three curves for cases $m = 3, 10, 25$ are located at $K = 15, 19, 27$ (i.e., K_{opt}), respectively. This phenomenon supports our insights that the optimal K only exists "sometimes" and may increase w.r.t. m.

4.3 The $K = \infty$ Case (Convergence in Local Update)

We define the following short-hand notations:

$$\boldsymbol{g}_l^{K=\infty} \coloneqq \mathcal{F}\left(l, \boldsymbol{\Delta}_0, \text{seq}_t\left(A_t\right), \text{seq}_t\left(\boldsymbol{b}_t\right)\right), \tag{21}$$

$$A_t \coloneqq \frac{1}{\sum_{i' \in [m]} n_{(i'),t}} \sum_{i' \in [m]} n_{(i'),t}\left(1 - \frac{n_{(i'),t}}{p}\right), \tag{22}$$

$$\boldsymbol{b}_t \coloneqq \frac{1}{\sum_{i' \in [m]} n_{(i'),t}} \sum_{i' \in [m]} n_{(i'),t} \cdot \frac{n_{(i'),t}}{p} \boldsymbol{\gamma}_{(i'),t}. \tag{23}$$

$$\begin{aligned} C_t \coloneqq & \frac{\sum_{i=1}^m \left(n_{(i),t}^2\left(1 - \frac{n_{(i),t}}{p}\right)\right)}{\left(\sum_{i \in [m]} n_{(i),t}\right)^2} \\ & + \frac{\sum_{i \neq j} n_{(i),t} n_{(j),t}\left(1 - \frac{n_{(i),t}}{p}\right)\left(1 - \frac{n_{(j),t}}{p}\right)}{\left(\sum_{i \in [m]} n_{(i),t}\right)^2}, \end{aligned} \tag{24}$$

[¶]When $\alpha < \frac{2}{1+\frac{p}{\tilde{n}}}$, we have $\mathcal{A} < 1$, and thus $\mathcal{J} < \frac{1+(m-1)}{m} = 1$.

$$D_t \coloneqq \frac{\sum_{i\in[m]} \frac{n_{(i),t}^3 \sigma_{(i),t}^2}{p-n_{(i),t}-1} + \frac{n_{(i),t}^3}{p} \left\|\boldsymbol{\gamma}_{(i),t}\right\|^2}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2}$$
$$+ \frac{\sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \left(\frac{n_{(i),t}^2 n_{(j),t}^2}{p^2} \boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\gamma}_{(j),t} \right)}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} +$$
$$\frac{\sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} 2\frac{n_{(j),t}^2}{p} n_{(i),t} \left(1-\frac{n_{(i),t}}{p}\right) \boldsymbol{\gamma}_{(j),t}^\top \boldsymbol{g}_{t-1}^{K=\infty}}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2}. \tag{25}$$

Theorem 3. *In the over-parameterized (OP) regime, i.e., $p > \max n_{(i),t} + 1$, it holds that*

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2 = \mathcal{F}(t, \|\boldsymbol{\Delta}_0\|^2, seq_l\left(C_l\right), seq_l\left(D_l\right)), \forall t \in [T]. \tag{26}$$

In the under-parameterized (UP) regime, i.e., $p < \min n_{(i),t} - 1$, it holds that

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2 = \left\| \frac{\sum_{i\in[m]} n_{(i),t} \boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}} \right\|^2 + \frac{\sum_{i\in[m]} \frac{n_{(i),t}^2 p \sigma_{(i),t}^2}{n_{(i),t}-p-1}}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2}. \tag{27}$$

For the simple case described by Eqs. (10) to (12), it holds that

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2 = \begin{cases} C^t \left\|\boldsymbol{\Delta}_0\right\|^2 + \frac{1-C^t}{1-C} D & \text{if OP,} \\ \frac{p\sigma^2}{m(n-p-1)} & \text{if UP,} \end{cases} \tag{28}$$

where

$$C \coloneqq \frac{1}{m}\left(1-\frac{n}{p}\right) + \frac{m-1}{m}\left(1-\frac{n}{p}\right)^2 < 1, \tag{29}$$

$$D \coloneqq \frac{n\sigma^2}{m(p-n-1)} + \frac{n}{p}\overline{\|\boldsymbol{\gamma}\|^2}. \tag{30}$$

We provide a proof sketch of Theorem 3 in Section 6. The complete proof is in the supplemental material [54, Appendix D].

Insight 4) Benign overfitting exists in FL, and the "null risk" can be alleviated by using more communications rounds. In the over-parameterized case of Eq. (28), the term D decreases when p increases. Thus, when the term D dominates (e.g., when noise and/or heterogeneity is large, or t is large), the generalization performance of FL in this case will benefit from more parameters when overfitted. This validates the "double-descent" or benign overfitting phenomenon in the literature of the classical (single-task single-agent) linear regression (e.g. [19]). For the comparable Gaussian models we used, the expectation of the model error of such a classical (single-task single-agent) linear regression is

$$(1-\frac{n}{p})\left\|\boldsymbol{\Delta}_0\right\|^2 + \frac{n\sigma^2}{p-n-1}. \tag{31}$$

By Eq. (31) and related literature (e.g., [18]), the classical linear regression suffers from the "null risk" (i.e., converges to the initial error) when $p \to \infty$. However, for the FL result in Eq. (28), we can see that the "null risk" term $\left\|\boldsymbol{\Delta}_0\right\|^2$ can be alleviated by the coefficient C^t, which approaches zero when $t \to \infty$. In other words, for fixed n, when $p \to \infty$, as long as we let $t \to \infty$ in a faster speed (e.g., $t = p \log p$, proved in Lemma 1 in supplemental material [54, Appendix A]), then the null risk term $C^t \left\|\boldsymbol{\Delta}_0\right\|^2 \to 0$, which implies that using more communication rounds in FL (i.e., larger t) mitigates the null risk, thus "enhancing" the benefits of overfitting.

Experiments. In Fig. 3, we present a plot of model error against p in both the underparameterized regime ($p < n = 25$) and overparameterized regime ($p > n = 25$) for three cases with $t = 1$, $t = 4$, and

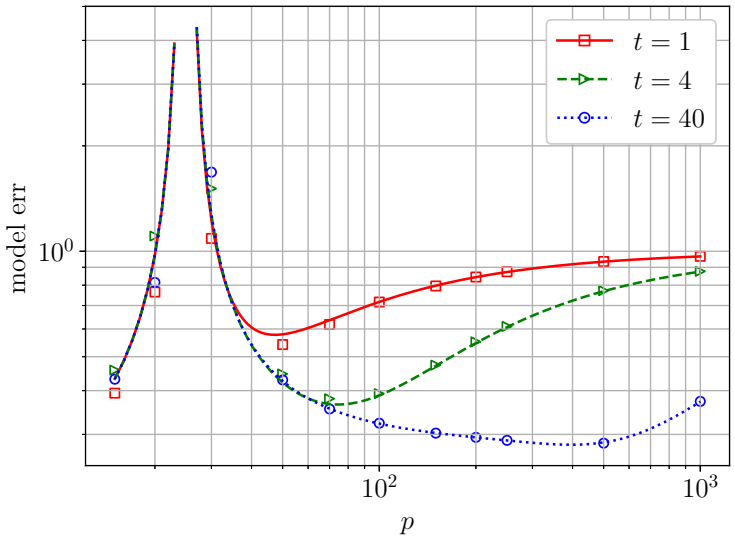

Figure 3: Experimental and theoretical values of the model error w.r.t. p where $m = 3$, $s = 5$, $n_{(i),t} = 25$, $\|\boldsymbol{\Delta}_0\| = 1$, $\left\|\boldsymbol{\gamma}_{(i),t}\right\| = 0.5$, and $\sigma_{(i),t} = 0.7$ for all i, t. Each marker is the experimental value by averaging over 20 simulation runs. The curves are drawn purely by the theoretical values from Theorem 3.

$t = 40$. The curves represent theoretical values derived from Theorem 3, while each marker signifies the average of 20 simulation trials. It is observed that all three curves exhibit a decreasing trend at the initial phase of the overparameterized regime, confirming the presence of benign overfitting. Additionally, when comparing these three cases, the curve for $t = 40$ (indicated by a blue dotted line with markers " $\circ$ ") has a more substantial and broader descent. This observation confirms our insight that a larger t-value enhances the benefits of overfitting in FL.

5 Proof Sketch of Proposition 1

(1) Since $\tilde{n}$ is fixed, then $\mathcal{A}$ does not change with K. When $t \to \infty$, the value of Eq. (19) becomes

$$\frac{1}{1-\mathcal{J}} \frac{\alpha^2 p \sigma^2}{mn} \cdot \frac{1-\mathcal{A}^K}{1-\mathcal{A}}. \tag{32}$$

The only component related to K in Eq. (32) is $\frac{1-\mathcal{A}^K}{1-\mathcal{J}}$, thus $K_{\text{opt}} = \arg\min_K \frac{1-\mathcal{A}^K}{1-\mathcal{J}}$. Notice that for any finite K, we must have

$$\mathcal{A}^K = \left(1-\alpha)^2 + \frac{\alpha^2(p+1)}{\tilde{n}}\right)^K > (1-\alpha)^{2K}.$$

Thus, we have

$$\mathcal{J} = \frac{1}{m}\mathcal{A}^K + \frac{m-1}{m}(1-\alpha)^{2K} < \mathcal{A}^K,$$

which implies that $\frac{1-\mathcal{A}^K}{1-\mathcal{J}} < 1$ for any finite K. Meanwhile, $\lim_{K\to\infty} \frac{1-\mathcal{A}^K}{1-\mathcal{J}} = 1$. Thus, K_{opt} should be finite.

(2) Since $\tilde{n}$ is fixed, then $\mathcal{A}$ does not change with K. When $\sigma = 0$, Eq. (19) becomes $\mathcal{J}^t \|\boldsymbol{\Delta}_0\|^2$. Notice that $\mathcal{J}$ is strictly monotone decreasing w.r.t. K. Therefore, $K_{\text{opt}} = \infty$.

(3) Since we use n/K to replace $\lfloor n/K \rfloor$, we have $K_{\text{opt}} = \arg\min_K f(K)$ where

$$f(K) \coloneqq \left((1-\alpha)^2 + K\frac{\alpha^2(p+1)}{n}\right)^K + (m-1)(1-\alpha)^{2K}.$$

Calculating the derivative, we have $\frac{\partial f(K)}{\partial K} =$

$$\frac{\alpha^2(p+1)}{n}\left((1-\alpha)^2+K\frac{\alpha^2(p+1)}{n}\right)^K \ln\left((1-\alpha)^2+K\frac{\alpha^2(p+1)}{n}\right) + (m-1)(1-\alpha)^{2K}\ln((1-\alpha)^2). \tag{33}$$

When $\left((1-\alpha)^2+K\frac{\alpha^2(p+1)}{n}\right)<1$, we have $\frac{\partial f(K)}{\partial K}<0$.

For any $\delta>0$, when

$$\left((1-\alpha)^2+K\frac{\alpha^2(p+1)}{n}\right)>1+\delta,$$
$$\frac{\alpha^2(p+1)}{n}(1+K\delta)\ln(1+\delta)>(m-1)\ln\frac{1}{(1-\alpha)^2},$$

we have Eq. (33) >0. (Notice that we utilize the face that $(1-\alpha)^{2K}<1$ and $(1+\delta)^K\geq 1+K\delta$.) Solving those inequalities by further letting $\ln(1+\delta)=\ln\frac{1}{(1-\alpha)^2}$, we thus have

$$\frac{n}{(p+1)}\left(\frac{2}{\alpha}-1\right)\leq K_{\text{opt}}$$
$$\leq\frac{n}{\alpha^2(p+1)}\cdot\max\left\{(2\alpha-\alpha^2)\left(1+\frac{1}{(1-\alpha)^2}\right),\ (m-2)\frac{(1-\alpha)^2}{2\alpha-\alpha^2}\right\}.$$

When $\alpha\leq 0.1$ and $m\geq 3$, we can further relax the above inequality as

$$\frac{n}{p+1}\left(\frac{2}{\alpha}-1\right)\leq K_{\text{opt}}\leq\frac{n}{p+1}\frac{(m-2)}{\alpha^3}.$$

6 Proof Sketch of Theorem 3

We provide a sketched proof of Eq. (26) here. For any $i\in[m]$, we define $\mathbf{P}_{(i),t}\in\mathbb{R}^{p\times p}$ as

$$\mathbf{P}_{(i),t}\coloneqq\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t}\right)^{-1}\mathbf{X}_{(i),t}^\top. \tag{34}$$

(We know $\mathbf{P}_{(i),t}$ is an orthogonal projection since $\mathbf{P}_{(i),t}\mathbf{P}_{(i),t}=\mathbf{P}_{(i),t}$ and $\mathbf{P}_{(i),t}^\top=\mathbf{P}_{(i),t}$.) In the overparameterized situation, after each agent trains to converge, we have

$$\begin{aligned}\hat{\boldsymbol{w}}_{(i),t}^{K=\infty}=&\mathbf{P}_{(i),t}\boldsymbol{w}_{(i),t}+(\mathbf{I}_p-\mathbf{P}_{(i),t})\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}\\&+\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}.\end{aligned} \tag{35}$$

We thus have

$$\begin{aligned}\boldsymbol{\Delta}_t^{K=\infty}=&\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{\text{avg},t}^{K=\infty}\ \ \text{(by Eq. (8))}\\=&\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}n_{(i),t}\left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}+(\mathbf{I}_p\right.\\&\left.-\mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty}-\mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top\mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}\right).\end{aligned} \tag{36}$$

Thus, we can write $\mathbb{E}_t\left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2$ into the inner product between the terms in Eq. (36). The key part of the proof is to calculate those inner product terms. The terms that involve only one agent can be calculated using the known results in the literature, e.g.,

$$\mathop{\mathbb{E}}_t\left\|\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right\|^2=\frac{n_{(i),t}}{p}\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2. \tag{37}$$

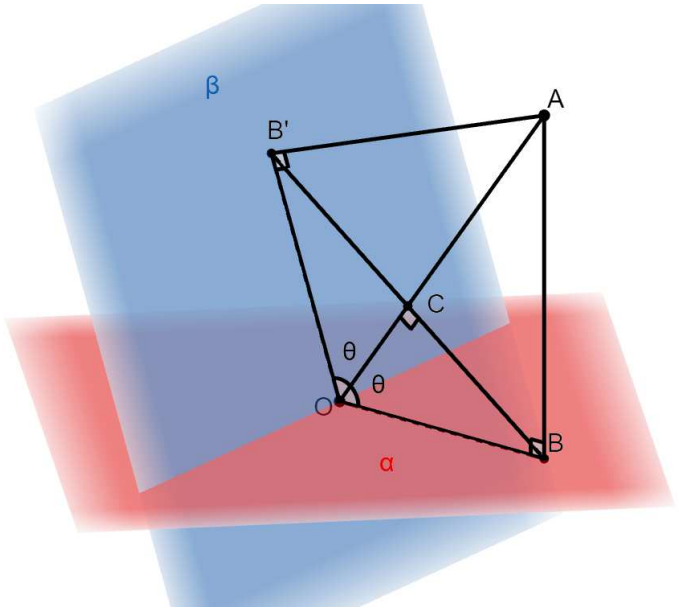

Figure 4: Geometric interpretation of Eq. (38).

The remaining terms in $\mathbb{E}_t \left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2$ involve different agents $i \neq j$, which are unique to FL and not seen in the literature. The key step is to prove

$$\mathop{\mathbb{E}}_{\mathbf{P}_{(i),t}} \left[\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right] = \frac{n_{(i),t}}{p}\boldsymbol{\Delta}_{t-1}^{K=\infty}. \tag{38}$$

We provide an intuition of Eq. (38) along with a geometric interpretation in Fig. 4 at the end of this section. By using Eq. (38), the terms involving different agents $i \neq j$ can be calculated, e.g.,

$$\mathop{\mathbb{E}}_{t}\left[\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right] = \frac{n_{(i),t}n_{(j),t}}{p^2}\boldsymbol{\gamma}_{(j),t}^\top\boldsymbol{\gamma}_{(i),t}.$$

With the above equations, we thus have

$$\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2 = C_t \cdot \mathbb{E}\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2 + D_t, \tag{39}$$

where C_t denotes the coefficient of $\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2$ and D_t denotes the remaining parts. The specific expressions of C_t and D_t are in Eqs. (24) and (25). Applying Eq. (39) recursively, we thus have Eq. (26).

Intuition of Eq. (38)*:* We use Fig. 4 to help illustrating the intuition. In Fig. 4, the vector $\overrightarrow{OA}$ denotes $\boldsymbol{\Delta}_{t-1}^{K=\infty}$, the plane α denotes the space spanned by the columns of $\mathbf{X}_{(i),t}$. Notice that $\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}$ represents result of projecting $\boldsymbol{\Delta}_{t-1}^{K=\infty}$ to the column space of $\mathbf{X}_{(i),t}$, i.e., the vector $\overrightarrow{OB}$ in Fig. 4. Therefore, in Fig. 4, calculating $\mathop{\mathbb{E}}_{\mathbf{P}_{(i),t}} \mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}$ means calculating the average of $\overrightarrow{OB}$ when the hyper-plane α rotating around the point O. Notice that $\overrightarrow{OB} = \overrightarrow{OC} + \overrightarrow{CB}$ where $\overrightarrow{OC}$ and $\overrightarrow{CB}$ are the parallel and perpendicular components of $\overrightarrow{OB}$ w.r.t. $\overrightarrow{OA}$, respectively. Because of the rotational symmetry of the hyper-plane α (due to the rotational symmetry of each column of $\mathbf{X}_{(i),t}$), we know that all the perpendicular components are cancelled out while only the parallel components remain in the averaging process. In other words, for any hyper-plane α, there exists a symmetrical (w.r.t. $\overrightarrow{OA}$) hyper-plane β with the same probability density such that the projection of $\overrightarrow{OA}$ to the hyper-plane β, named $\overrightarrow{OB'}$, has the same parallel component $\overrightarrow{OC}$ but the opposite perpendicular component $\overrightarrow{CB'} = -\overrightarrow{CB}$. Thus, we only need to calculate the average of the parallel component $\overrightarrow{OC}$, whose length equals $\cos\theta \left|\overrightarrow{OB}\right|$, where $\theta = \angle AOB$ is defined as the angle between $\boldsymbol{\Delta}_{t-1}^{K=\infty}$

and $\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}$ (i.e., the angle between $\boldsymbol{\Delta}_{t-1}^{K=\infty}$ and the hyperplane spanned by the columns of $\mathbf{X}_{(i),t}$ as

$$\theta \coloneqq \arccos \frac{\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}}{\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|}. \tag{40}$$

Also notice that $\left|\overrightarrow{OB}\right| = \cos\theta \left|\overrightarrow{OA}\right|$. Thus, the length of the parallel component equals $\left|\overrightarrow{OC}\right| = \cos^2\theta \left|\overrightarrow{OA}\right|$. Therefore, we have $\mathbb{E}\,\overrightarrow{OC} = \mathbb{E}\cos^2\theta\overrightarrow{OA} = \frac{n_{(i),t}}{p}\boldsymbol{\Delta}_{t-1}^{K=\infty}$. The last equation uses a known result in literature just as Eq. (37).

7 Conclusion

In this paper, we have precisely quantified the influence of data heterogeneity and the local update process on the generalization performance of FedAvg-type algorithms. Specifically, we undertook a thorough theoretical examination of FL's generalization performance utilizing a linear model, which yields closed-form expressions for the model error. Our analysis rigorously assesses the impact of local update steps (represented by K) across three distinct settings ($K = 1$, $K < \infty$, and $K = \infty$), elucidating how generalization performance evolves with the progression of rounds, denoted as t. Additionally, our investigation yields a comprehensive understanding of how various configurations, including the number of model parameters p, the number of training samples n, the local steps K, and the total communication round t, contribute to the overall FL generalization performance. This, in turn, unveils new insights, such as the phenomenon of benign overfitting, optimal local steps, and the impact of a good model initialization, with practical implications for the implementation of FL.

8 Acknowledgement

This work has been supported in part by NSF grants NSF AI Institute (AI-EDGE) CNS-2112471, CNS-2312836, CNS-2106933, CNS-2106932, CNS-2312836, CNS-1955535, CNS-1901057, ECCS-2113860 and 2324052, CAREER CNS-2110259, and ECCS-2331104, by Army Research Office under Grant W911NF-21-1-0244 and was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0225, by DARPA Grant D24AP00265. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[2] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

[3] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*, pages 240–254. Springer, 2020.

[4] Latif U Khan, Walid Saad, Zhu Han, Ekram Hossain, and Choong Seon Hong. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, 23(3):1759–1799, 2021.

[5] 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.

[6] Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 654–672. Springer, 2022.

[7] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[8] Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

[9] Tao Lin, Sebastian Urban Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. In *Proceedings of the 8th International Conference on Learning Representations*, 2019.

[10] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *The Journal of Machine Learning Research*, 22(1):9709–9758, 2021.

[11] Jose Javier Gonzalez Ortiz, Jonathan Frankle, Mike Rabbat, Ari Morcos, and Nicolas Ballas. Trade-offs of local sgd at scale: An empirical study. *arXiv preprint arXiv:2110.08133*, 2021.

[12] Honglin Yuan, Warren Richard Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *International Conference on Learning Representations*, 2022.

[13] Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local sgd generalize better than sgd? In *The Eleventh International Conference on Learning Representations*, 2022.

[15] Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. *arXiv preprint arXiv:2306.03824*, 2023.

[16] Milad Sefidgaran, Romain Chor, Abdellatif Zaidi, and Yijun Wan. Federated learning you may communicate less often! *arXiv preprint arXiv:2306.05862*, 2023.

[17] Sai Li, Linjun Zhang, T Tony Cai, and Hongzhe Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, pages 1–12, 2023.

[18] Peizhong Ju, Xiaojun Lin, and Jia Liu. Overfitting can be harmless for basis pursuit, but only to a degree. *Advances in Neural Information Processing Systems*, 33:7956–7967, 2020.

[19] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[20] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549, 2018.

[21] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[22] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[23] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[24] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. Harmless interpolation of noisy data in regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2299–2303. IEEE, 2019.

[25] Partha P Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*, 2019.

[26] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

[27] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

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

[29] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[30] 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.

[31] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

[32] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.

[33] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020.

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

[35] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-{iid} federated learning. In *International Conference on Learning Representations*, 2021.

[36] Haibo Yang, Xin Zhang, Prashant Khanduri, and Jia Liu. Anarchic federated learning. In *International Conference on Machine Learning*, pages 25331–25363. PMLR, 2022.

[37] Zihao Zhao, Yang Liu, Wenbo Ding, and Xiao-Ping Zhang. Federated pac-bayesian learning on non-iid data. *arXiv preprint arXiv:2309.06683*, 2023.

[38] Romain Chor, Milad Sefidgaran, and Abdellatif Zaidi. More communication does not result in smaller generalization error in federated learning. *arXiv preprint arXiv:2304.12216*, 2023.

[39] LP Barnes, Alex Dytso, and H Vincent Poor. Improved information theoretic generalization bounds for distributed and federated learning. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1465–1470. IEEE, 2022.

[40] Milad Sefidgaran, Romain Chor, and Abdellatif Zaidi. Rate-distortion theoretic bounds on generalization error for distributed learning. *Advances in Neural Information Processing Systems*, 35:19687–19702, 2022.

[41] Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, pages 4423–4434. PMLR, 2021.

[42] Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24552–24562, 2023.

[43] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning*, pages 18250–18280. PMLR, 2022.

[44] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[45] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

[46] Siddhartha Satpathi and R Srikant. The dynamics of gradient descent for overparametrized neural networks. In *Learning for Dynamics and Control*, pages 373–384. PMLR, 2021.

[47] Peizhong Ju, Xiaojun Lin, and Ness B Shroff. On the generalization power of overfitted two-layer neural tangent kernel models. *arXiv preprint arXiv:2103.05243*, 2021.

[48] Peizhong Ju, Xiaojun Lin, and Ness B Shroff. On the generalization power of the overfitted three-layer neural tangent kernel model. *arXiv preprint arXiv:2206.02047*, 2022.

[49] H Brendan McMahan et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1), 2021.

[50] Robert M Gower. Convergence theorems for gradient descent. *Lecture notes for Statistical Optimization*, 2018.

[51] Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[52] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[53] Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. *arXiv preprint arXiv:2302.05836*, 2023.

[54] Peizhong Ju, Haibo Yang, Jia Liu, Yingbin Liang, and Ness Shroff. Supplemental material. https://github.com/functionadvanced/Supp_Material/blob/main/supp.pdf, 2024. Accessed: 2024-08-25.

[55] Hong-You Chen, Cheng-Hao Tu, Ziwei Li, Han-Wei Shen, and Wei-Lun Chao. On pre-training for federated learning. *arXiv preprint arXiv:2206.11488*, 2022.

[56] John Nguyen, Jianyu Wang, Kshitiz Malik, Maziar Sanjabi, and Michael Rabbat. Where to begin? on the impact of pre-training and initialization in federated learning. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[58] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

[59] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. In *International Conference on Learning Representations*, 2020.

[60] Peizhong Ju, Yingbin Liang, and Ness Shroff. Theoretical characterization of the generalization performance of overfitted meta-learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[61] Peizhong Ju, Sen Lin, Mark S Squillante, Yingbin Liang, and Ness B Shroff. Generalization performance of transfer learning: Overparameterized and underparameterized regimes. *arXiv preprint arXiv:2306.04901*, 2023.

[62] Alberto Bernacchia. Meta-learning with negative learning rates. *arXiv preprint arXiv:2102.00940*, 2021.

Supplemental Material

We give a table to summarize the content of the supplemental material.

Section	Content
Appendix A	some useful lemmas as technical tools
Appendix B	proof of Theorem 1 for $K = 1$
Appendix C	proof of Theorem 2 for $K < \infty$
Appendix D	proof of Theorem 3 for $K = \infty$
Appendix E	a table for some important notations

Table 1: Outline of the supplemental material.

A Useful Lemmas

In this section, we provide some useful lemmas. Specifically, Lemma 1 is used to support the claim of the convergence speed in Insight 4. Lemmas 2 to 4 are some results about the Gaussian random matrices that can be found in the literature. We want to highlight Lemma 5 as part of our technical novelty, which gives the exact values of terms related to the projection formed by each agent's training inputs. Lemma 6 is used to justify the definition of model error.

Lemma 1. *Recalling the definition of C in Eq. (30), we have*

$$\lim_{t=p\ln p,\ p\to\infty} C^t = 0.$$

Proof. We have $C^t \geq 0$ and

$$\begin{aligned}
C^t &\leq \left(1-\frac{n}{p}\right)^t \quad \left(\text{since } C \leq \left(1-\frac{n}{p}\right) \text{ because } \left(1-\frac{n}{p}\right)^2 \leq \left(1-\frac{n}{p}\right)\right) \\
&= \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-t} \quad \left(\text{since } 1-\frac{n}{p} = \frac{1}{1+\frac{1}{\frac{p}{n}-1}}\right) \\
&= \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-p\ln p} \quad (\text{since } t = p\ln p) \\
&= \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-\frac{p}{n}\cdot n\cdot\ln p} \\
&\leq \left(1+\frac{1}{\frac{p}{n}-1}\right)^{-\left(\frac{p}{n}-1\right)\cdot n\cdot\ln p}.
\end{aligned}$$

Notice that

$$\lim_{p\to\infty}\left(1+\frac{1}{\frac{p}{n}-1}\right)^{-\left(\frac{p}{n}-1\right)\cdot n\cdot\ln p} = \lim_{p\to\infty} e^{-n\ln p} = 0,$$

where we use the fact that $\lim_{x\to\infty}(1+x^{-1})^x = e$. The result of this lemma thus follows by the squeeze theorem. □

The result of the following lemma can be found in the literature (e.g., [19, 60]).

Lemma 2. *Consider a random matrix $\mathbf{K} \in \mathbb{R}^{p\times n}$ where p and n are two positive integers and $p > n+1$. Each element of $\mathbf{K}$ is* i.i.d. *according to standard Gaussian distribution. For any fixed vector $\boldsymbol{a} \in \mathbb{R}^p$, we*

must have

$$\mathbb{E}\left\|\left(\mathbf{I}_p - \mathbf{K}\left(\mathbf{K}^\top\mathbf{K}\right)^{-1}\mathbf{K}^\top\right)\boldsymbol{a}\right\|^2 = \left(1-\frac{n}{p}\right)\|\boldsymbol{a}\|^2,$$
$$\mathbb{E}\left\|\mathbf{K}\left(\mathbf{K}^\top\mathbf{K}\right)^{-1}\mathbf{K}^\top\boldsymbol{a}\right\|^2 = \frac{n}{p}\|\boldsymbol{a}\|^2.$$

The following lemma can be found in Lemma 8 of [61].

Lemma 3. *Consider a random matrix $\mathbf{K} \in \mathbb{R}^{a\times b}$ where $a > b+1$. Each element of $\mathbf{K}$ is* i.i.d. *following standard Gaussian distribution $\mathcal{N}(0,1)$. Consider three Gaussian random vectors $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathbf{R}^a$ and $\boldsymbol{\beta} \in \mathbf{R}^b$ such that $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2\mathbf{I}_a)$, $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(d_1^2, d_2^2, \cdots, d_a^2))$, and $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2\mathbf{I}_b)$. Here $\mathbf{K}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are independent of each other. We then must have*

$$\mathbb{E}\left[(\mathbf{K}^\top\mathbf{K})^{-1}\right] = \frac{\mathbf{I}_b}{a-b-1}, \tag{41}$$
$$\mathbb{E}\left\|\mathbf{K}(\mathbf{K}^\top\mathbf{K})^{-1}\boldsymbol{\beta}\right\|^2 = \frac{b\sigma_\beta^2}{a-b-1}, \tag{42}$$
$$\mathbb{E}\left\|(\mathbf{K}^\top\mathbf{K})^{-1}\mathbf{K}^\top\boldsymbol{\alpha}\right\|^2 = \frac{b\sigma_\alpha^2}{a-b-1}, \tag{43}$$
$$\mathbb{E}\left\|(\mathbf{K}^\top\mathbf{K})^{-1}\mathbf{K}^\top\boldsymbol{\gamma}\right\|^2 = \frac{b\sum_{i=1}^a d_i^2}{a(a-b-1)}. \tag{44}$$

The following lemma can be found in [62] and Lemma 13 of [60].

Lemma 4. *Consider a random matrix $\mathbf{K} \in \mathbb{R}^{a\times b}$ whose each element follows* i.i.d. *standard Gaussian distribution (i.e.,* i.i.d. *$\mathcal{N}(0,1)$). We mush have*

$$\mathbb{E}[\mathbf{K}^\top\mathbf{K}] = a\mathbf{I}_b,$$
$$\mathbb{E}[\mathbf{K}\mathbf{K}^\top] = b\mathbf{I}_a,$$
$$\mathbb{E}[\mathbf{K}\mathbf{K}^\top\mathbf{K}\mathbf{K}^\top] = b(b+a+1)\mathbf{I}_a.$$

Lemma 5. *For any $i \in [m]$ and t, we must have*

$$\mathop{\mathbb{E}}_{\mathbf{P}_{(i),t}}\left[\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right] = \frac{n_{(i),t}}{p}\boldsymbol{\Delta}_{t-1}^{K=\infty}. \tag{45}$$

Consequently, when $i \neq j$, we have

$$\mathop{\mathbb{E}}_{\mathbf{P}_{(i),t},\mathbf{P}_{(j),t}}\left[\boldsymbol{\Delta}_{t-1}^{K=\infty\top}\mathbf{P}_{(i),t}\mathbf{P}_{(j),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right] = \frac{n_{(j),t}n_{(i),t}}{p^2}\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2.$$

Proof. Let $C \coloneqq \left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|$. Since we are calculating expected projection of $\boldsymbol{\Delta}_{t-1}^{K=\infty}$ onto the column space of $\mathbf{X}_{(i),t}$, by the symmetry of $\mathbf{X}_{(i),t}$, without loss of generality we let

$$\boldsymbol{\Delta}_{t-1}^{K=\infty} = C\cdot\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}. \tag{46}$$

Define

$$\tilde{\mathbf{X}}_{(i),t} \coloneqq \begin{bmatrix}-1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}\mathbf{X}_{(i),t}. \tag{47}$$

Since each element of $\mathbf{X}_{(i),t}$ follows *i.i.d.* standard Gaussian distribution, we know that $\tilde{\mathbf{X}}_{(i),t}$ and $\mathbf{X}_{(i),t}$ has identical distributioin. Thus, we have

$$\int \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} d\mu(\mathbf{X}_{(i),t}) = \int \tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})\tilde{\mathbf{X}}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} d\mu(\mathbf{X}_{(i),t}), \tag{48}$$

where $\mu(\mathbf{X}_{(i),t})$ denotes the joint probability distribution of $\mathbf{X}_{(i),t}$.

By Eq. (47), we have

$$\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t} = \mathbf{X}_{(i),t}^\top \begin{bmatrix} -1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \begin{bmatrix} -1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \mathbf{X}_{(i),t} = \mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t},$$

$\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} = [\mathbf{X}_{(i),t}]_{1,:}$, $\tilde{\mathbf{X}}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} = -[\mathbf{X}_{(i),t}]_{1,:}$ (here $[\cdot]_{1,:}$ denotes the first row of a matrix).

Thus, we have

$$\tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})^{-1}\tilde{\mathbf{X}}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} = -\tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty}. \tag{49}$$

Therefore, we have

$$\begin{aligned}
&\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} + \tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})^{-1}\tilde{\mathbf{X}}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} \\
=&(\mathbf{X}_{(i),t} - \tilde{\mathbf{X}}_{(i),t})(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} \quad \text{(by Eq. (49))} \\
=&\begin{bmatrix} 2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} \quad \text{(by Eq. (47))} \\
=&\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 2 & 0 & \cdots & 0 \end{bmatrix} \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} \\
=&2\frac{\boldsymbol{\Delta}_{t-1}^{K=\infty}}{C^2}\boldsymbol{\Delta}_{t-1}^{K=\infty\top}\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} \quad \text{(by Eq. (46))} \\
=&2\frac{\boldsymbol{\Delta}_{t-1}^{K=\infty}}{C^2}\boldsymbol{\Delta}_{t-1}^{K=\infty\top}\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty} \quad \text{(by Eq. (80))} \\
=&2\frac{\boldsymbol{\Delta}_{t-1}^{K=\infty}}{C^2}\left\|\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2 \quad \text{(since } \mathbf{P}_{(i),t}^\top \mathbf{P}_{(i),t} = \mathbf{P}_{(i),t} \text{ as } \mathbf{P}_{(i),t} \text{ is an orthogonal projection).}
\end{aligned} \tag{50}$$

Thus, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{X}_{(i),t}}[\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}] \\
=&\int \mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} d\mu(\mathbf{X}_{(i),t}) \\
=&\frac{1}{2}\int \left(\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t})^{-1}\mathbf{X}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty} + \tilde{\mathbf{X}}_{(i),t}(\tilde{\mathbf{X}}_{(i),t}^\top \tilde{\mathbf{X}}_{(i),t})\tilde{\mathbf{X}}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K=\infty}\right) d\mu(\mathbf{X}_{(i),t}) \quad \text{(by Eq. (48))} \\
=&\int \frac{\boldsymbol{\Delta}_{t-1}^{K=\infty}}{C^2}\left\|\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2 d\mu(\mathbf{X}_{(i),t}) \\
=&\frac{\boldsymbol{\Delta}_{t-1}^{K=\infty}}{C^2}\mathop{\mathbb{E}}_{\mathbf{X}_{(i),t}}\left\|\mathbf{P}_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2 \\
=&\frac{n_{(i),t}}{p}\boldsymbol{\Delta}_{t-1}^{K=\infty} \quad \text{(by Lemma 2).}
\end{aligned}$$

The result of this lemma thus follows. □

Lemma 6. *Let the noise in every test sample have zero mean and variance σ^2. For any learning result $\hat{\boldsymbol{w}}$, the mean square test error must equal to $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|^2 + \sigma^2$. Therefore, the mean squared test error for noise-free test samples equals to the model error $L^{model}(\hat{\boldsymbol{w}}) = \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|^2$.*

Proof. Considering $(\boldsymbol{x}, y)$ as a randomly generated test sample by the ground truth $y = \boldsymbol{x}^\top \boldsymbol{w}^* + \epsilon$, the mean squared error is equal to

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x},y} \left\|\boldsymbol{x}^\top \hat{\boldsymbol{w}} - y\right\| &= \mathop{\mathbb{E}}_{\boldsymbol{x},\epsilon} \left\|\boldsymbol{x}^\top \hat{\boldsymbol{w}} - (\boldsymbol{x}^\top \boldsymbol{w}^* + \epsilon)\right\|^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x},\epsilon} \left\|\boldsymbol{x}^\top (\hat{\boldsymbol{w}} - \boldsymbol{w}^*) + \epsilon\right\|^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}} \left\|\boldsymbol{x}^\top (\hat{\boldsymbol{w}} - \boldsymbol{w}^*)\right\|^2 + \mathop{\mathbb{E}}_{\epsilon} \|\epsilon\|^2 \\
&\quad \text{(since the noise } \epsilon \text{ has zero mean and is independent of other random variables)} \\
&= \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|^2 + \sigma^2 \\
&\quad \text{(notice that } \boldsymbol{x} \text{ follows standard Gaussian distribution and is independent of } \hat{\boldsymbol{w}}).
\end{aligned}$$

□

B Proof of Theorem 1

Calculating the gradient of the training loss defined at Eq. (4), we have

$$\begin{aligned}
\frac{\partial L(\hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} &= \frac{\partial (\boldsymbol{y} - \mathbf{X}^\top \hat{\boldsymbol{w}})}{\partial \hat{\boldsymbol{w}}} \cdot \frac{\partial \frac{1}{2n} \left\|\boldsymbol{y} - \mathbf{X}^\top \hat{\boldsymbol{w}}\right\|^2}{\partial (\boldsymbol{y} - \mathbf{X}^\top \hat{\boldsymbol{w}})} \quad \text{(by the chain rule)} \\
&= -\mathbf{X} \cdot \frac{1}{n} (\boldsymbol{y} - \mathbf{X}^\top \hat{\boldsymbol{w}}) \\
&= \frac{1}{n} (\mathbf{X}\mathbf{X}^\top \hat{\boldsymbol{w}} - \mathbf{X}\boldsymbol{y}).
\end{aligned}$$

When $K = 1$, with step size $\alpha_{(i),t} > 0$, we thus have

$$\begin{aligned}
\hat{\boldsymbol{w}}_{(i),t}^{K=1} &= \left(\mathbf{I}_p - \frac{\alpha_{(i),t}}{n_{(i),t}} \mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top\right) \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1} + \frac{\alpha_{(i),t}}{n_{(i),t}} \mathbf{X}_{(i),t} \boldsymbol{y}_{(i),t} \\
&= \left(\mathbf{I}_p - \frac{\alpha_{(i),t}}{n_{(i),t}} \mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top\right) \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1} + \frac{\alpha_{(i),t}}{n_{(i),t}} \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \boldsymbol{w}_{(i),t} + \boldsymbol{\epsilon}_{(i),t}\right) \quad \text{(by Eq. (2))}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\hat{\boldsymbol{w}}_{\text{avg},t}^{K=1} &= \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \hat{\boldsymbol{w}}_{(i),t}^{K=1} \\
&= \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1} + \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} \alpha_{(i),t} \left(-\mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=1} + \mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top \boldsymbol{w}_{(i),t} + \mathbf{X}_{(i),t} \boldsymbol{\epsilon}_{(i),t}\right). \quad (51)
\end{aligned}$$

By Eqs. (3) and (8), we have

$$\begin{aligned}
&\boldsymbol{\Delta}_t^{K=1} \\
&= \boldsymbol{\Delta}_{t-1}^{K=1} + \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} \alpha_{(i),t} \left(\mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top (\boldsymbol{\gamma}_{(i),t} - \boldsymbol{\Delta}_{t-1}^{K=1}) - \mathbf{X}_{(i),t} \boldsymbol{\epsilon}_{(i),t}\right) \\
&= \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} \Bigg(\underbrace{\left(n_{(i),t} \mathbf{I}_p - \alpha_{(i),t} \mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top\right) \boldsymbol{\Delta}_{t-1}^{K=1}}_{\boldsymbol{q}_{1i}} + \underbrace{\alpha_{(i),t} \mathbf{X}_{(i),t} \mathbf{X}_{(i),t}^\top \boldsymbol{\gamma}_{(i),t}}_{\boldsymbol{q}_{2i}} - \underbrace{\alpha_{(i),t} \mathbf{X}_{(i),t} \boldsymbol{\epsilon}_{(i),t}}_{\boldsymbol{q}_{3i}} \Bigg) \quad (52)
\end{aligned}$$

(since $\boldsymbol{\Delta}_{t-1}^{K=1} = \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \boldsymbol{\Delta}_{t-1}^{K=1}$).

Considering the three types of terms $\boldsymbol{q}_{1i}, \boldsymbol{q}_{2i}, \boldsymbol{q}_{3i}$ defined in Eq. (52), by Assumption 1, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t} \boldsymbol{q}_{1i} &= n_{(i),t} \left(1 - \alpha_{(i),t}\right) \boldsymbol{\Delta}_{t-1}^{K=1}, \\
\mathop{\mathbb{E}}_{t} \boldsymbol{q}_{2i} &= \alpha_{(i),t} n_{(i),t} \boldsymbol{\gamma}_{(i),t}, \\
\mathop{\mathbb{E}}_{t} \boldsymbol{q}_{3i} &= \mathbf{0}.
\end{aligned} \tag{53}$$

Notice that we use $\mathbb{E}$ to denote the expectation on all randomness and use $\mathbb{E}_t$ to denote the expectation on the randomness at the t-th round, i.e., on the randomness of $\mathbf{X}_{(i),t}$ and $\boldsymbol{\epsilon}_{(i),t}$ for all $i \in [m]$. By Eqs. (52) and (53), we thus have

$$\mathop{\mathbb{E}}_{t} \boldsymbol{\Delta}_{t}^{K=1} = \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} \left(n_{(i),t} \left(1 - \alpha_{(i),t}\right) \boldsymbol{\Delta}_{t-1}^{K=1} + \alpha_{(i),t} n_{(i),t} \boldsymbol{\gamma}_{(i),t}\right). \tag{54}$$

Applying Eq. (54) recursively and recalling Eq. (13), we thus have

$$\mathbb{E}[\boldsymbol{\Delta}_{t}^{K=1}] = \boldsymbol{g}_{t}^{K=1}. \tag{55}$$

By Assumption 1, we know that $\boldsymbol{\epsilon}_{(i),t}$ is independent of $\mathbf{X}_{(j),t}$ for all $i, j \in [m]$ and $\mathbb{E}\, \boldsymbol{\epsilon}_{(i),t} = \mathbf{0}$. Thus, we have

$$\mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{3j}] = \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{3j}] = 0.$$

Thus, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t} \left\|\boldsymbol{\Delta}_{t}^{K=1}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \left(\sum_{i\in[m]} \left(\mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{1i}\right\|^2 + \mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{2i}\right\|^2 + \mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{3i}\right\|^2 + 2 \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2i}] \right) \right. \\
& \left. + \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \left(\mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{1j}] + \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}] + \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1j}^\top \boldsymbol{q}_{2i}] + \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{2j}] \right) \right) \\
=& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \left(\sum_{i\in[m]} \left(\mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{1i}\right\|^2 + \mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{2i}\right\|^2 + \mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{3i}\right\|^2 + 2 \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2i}] \right) \right. \\
& \left. + \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \left(\mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{1j}] + 2 \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}] + \mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{2j}] \right) \right) \\
& (\text{since } \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j} + \boldsymbol{q}_{1j}^\top \boldsymbol{q}_{2i} = 2 \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}).
\end{aligned} \tag{56}$$

By Lemma 4, for any $i \in [m]$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{1i}\right\|^2 &= \left(n_{(i),t}^2 - 2\alpha_{(i),t} n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t} (n_{(i),t} + p + 1)\right) \left\|\boldsymbol{\Delta}_{t-1}^{K=1}\right\|^2 \\
&= \left(\left(1 - \alpha_{(i),t}\right)^2 n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t} (p+1)\right) \left\|\boldsymbol{\Delta}_{t-1}^{K=1}\right\|^2, \\
\mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{2i}\right\|^2 &= \alpha_{(i),t}^2 n_{(i),t} (n_{(i),t} + p + 1) \left\|\boldsymbol{\gamma}_{(i),t}\right\|^2, \\
\mathop{\mathbb{E}}_{t} \left\|\boldsymbol{q}_{3i}\right\|^2 &= \alpha_{(i),t}^2 p n_{(i),t} \sigma_{(i),t}^2, \\
\mathop{\mathbb{E}}_{t}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2i}] &= \left(\alpha_{(i),t} n_{(i),t}^2 - \alpha_{(i),t}^2 n_{(i),t} (n_{(i),t} + p + 1)\right) \boldsymbol{\Delta}_{t-1}^{K=1\top} \boldsymbol{\gamma}_{(i),t}.
\end{aligned} \tag{57}$$

Similarly, by Lemma 4, for any $i, j \in [m]$ where $i \neq j$, we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{1j}] =& n_{(i),t} n_{(j),t} \left(1-\alpha_{(i),t}\right)\left(1-\alpha_{(j),t}\right) \left\|\boldsymbol{\Delta}_{t-1}^{K=1}\right\|^2, \\
\mathbb{E}[\boldsymbol{q}_{1i}^\top \boldsymbol{q}_{2j}] =& \left(\alpha_{(j),t} n_{(i),t} n_{(j),t} - \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t}\right) {\boldsymbol{\Delta}_{t-1}^{K=1}}^\top \boldsymbol{\gamma}_{(j),t} \\
=& n_{(i),t} n_{(j),t} \alpha_{(j),t} \left(1-\alpha_{(i),t}\right) {\boldsymbol{\Delta}_{t-1}^{K=1}}^\top \boldsymbol{\gamma}_{(j),t}, \\
\mathbb{E}[\boldsymbol{q}_{2i}^\top \boldsymbol{q}_{2j}] =& \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t} \boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\gamma}_{(j),t}.
\end{aligned} \tag{58}$$

Plugging Eqs. (57) and (58) into Eq. (56), we thus have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t}[\left\|\boldsymbol{\Delta}_t^{K=1}\right\|^2] \\
=&\frac{\left\|\boldsymbol{\Delta}_{t-1}^{K=1}\right\|^2}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \left(\sum_{i\in[m]} \left((1-\alpha_{(i),t})^2 n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t}(p+1)\right) + \sum_{i\in[m]} \sum_{j\in[m]\setminus\{j\}} n_{(i),t} n_{(j),t} (1-\alpha_{(i),t})(1-\alpha_{(j),t}) \right) \\
&+ \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \alpha_{(i),t}^2 \left(p n_{(i),t} \sigma_{(i),t}^2 + n_{(i),t}(n_{(i),t}+p+1) \left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \right) \\
&+ 2\frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \left(\alpha_{(i),t} n_{(i),t}^2 - \alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1) \right) {\boldsymbol{\Delta}_{t-1}^{K=1}}^\top \boldsymbol{\gamma}_{(i),t} \\
&+ \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \left(2 n_{(i),t} n_{(j),t} \alpha_{(j),t} \left(1-\alpha_{(i),t}\right) {\boldsymbol{\Delta}_{t-1}^{K=1}}^\top \boldsymbol{\gamma}_{(j),t} + \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t} \boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\gamma}_{(j),t} \right).
\end{aligned} \tag{59}$$

Notice that

$$\begin{aligned}
&\left(\sum_{i\in[m]} \left((1-\alpha_{(i),t})^2 n_{(i),t}^2 + \alpha_{(i),t}^2 n_{(i),t}(p+1)\right) + \sum_{i\in[m]} \sum_{j\in[m]\setminus\{j\}} n_{(i),t} n_{(j),t} (1-\alpha_{(i),t})(1-\alpha_{(j),t}) \right) \\
=& \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \left(\sum_{i\in[m]} n_{(i),t} (1-\alpha_{(i),t})^2 \right)^2 + \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \alpha_{(i),t}^2 n_{(i),t}(p+1) \\
=& H_t \text{ (recalling Eq. (14))},
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1) \left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \\
&+ \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} \alpha_{(i),t}\alpha_{(j),t} n_{(i),t} n_{(j),t} \boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\gamma}_{(j),t} \\
=& \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \left\| \sum_{i\in[m]} \alpha_{(i),t} n_{(i),t} \boldsymbol{\gamma}_{(i),t} \right\|^2 + \frac{1}{\left(\sum_{i\in[m]} n_{(i),t}\right)^2} \sum_{i\in[m]} \alpha_{(i),t}^2 n_{(i),t}(p+1) \left\|\boldsymbol{\gamma}_{(i),t}\right\|^2,
\end{aligned}$$

and

$$
\begin{aligned}
&2\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\left(\alpha_{(i),t}n_{(i),t}^2-\alpha_{(i),t}^2 n_{(i),t}(n_{(i),t}+p+1)\right)\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}}\left(2n_{(i),t}n_{(j),t}\alpha_{(j),t}\left(1-\alpha_{(i),t}\right)\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(j),t}\right)\\
=&\frac{2}{(\sum_{i\in[m]} n_{(i),t})^2}\left(\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})\right)\cdot\left(\sum_{i\in[m]} n_{(i),t}\alpha_{(i),t}\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}\right)\\
&-\frac{2\sum_{i\in[m]}\alpha_{(i),t}^2 n_{(i),t}(p+1)\boldsymbol{\Delta}_{t-1}^{K=1\top}\boldsymbol{\gamma}_{(i),t}}{(\sum_{i\in[m]} n_{(i),t})^2}.
\end{aligned}
$$

Further, by Eq. (55) and recalling Eq. (15), we thus can rewrite Eq. (59) as

$$
\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=1}\right\|^2=H_t\,\mathbb{E}\left\|\boldsymbol{\Delta}_{t-1}^{K=1}\right\|^2+G_t. \tag{60}
$$

Applying Eq. (60) recursively, we thus have Eq. (16).

C Proof of Theorem 2

Define

$$
\boldsymbol{g}_l^{K<\infty}\coloneqq\mathcal{F}\left(l,\boldsymbol{\Delta}_0,\mathrm{seq}_t\left(\frac{\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})^K}{\sum_{i\in[m]} n_{(i),t}}\right),\mathrm{seq}_t\left(\frac{\sum_{i\in[m]} n_{(i),t}\left(1-(1-\alpha_{(i),t})^K\right)\boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}\right)\right) \tag{61}
$$

$$
\mathcal{A}_{(i),t}\coloneqq(1-\alpha_{(i),t})^2+\frac{\alpha_{(i),t}^2(p+1)}{\tilde{n}_{(i),t}}, \tag{62}
$$

$$
\begin{aligned}
\mathcal{B}_{(i),t,k}\coloneqq&\frac{\alpha_{(i),t}^2 p\sigma_{(i),t}^2}{\tilde{n}_{(i),t}}\\
&+\left(\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)+2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left(1-(1-\alpha_{(i),t})^{k-1}\right)\right)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2\\
&+2\left(\alpha_{(i),t}-\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)(1-\alpha_{(i),t})^{k-1}\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{g}_{t-1}^{K<\infty},
\end{aligned} \tag{63}
$$

$$
\mathcal{J}_t\coloneqq\frac{\sum_{i\in[m]} n_{(i),t}^2\mathcal{A}_{(i),t}^K}{(\sum_{i\in[m]} n_{(i),t})^2}+\frac{\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}(1-\alpha_{(i),t})^K(1-\alpha_{(j),t})^K}{(\sum_{i\in[m]} n_{(i),t})^2}, \tag{64}
$$

$$
\begin{aligned}
\mathcal{Q}_t\coloneqq&\frac{\sum_{i\in[m]} n_{(i),t}^2\sum_{k=1}^K\mathcal{B}_{(i),t,k}\mathcal{A}_{(i),t}^{K-k}}{(\sum_{i\in[m]} n_{(i),t})^2}\\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\Big(2(1-\alpha_{(i),t})^K(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(j),t}^\top\boldsymbol{g}_{t-1}^{K<\infty}\\
&+(1-(1-\alpha_{(i),t})^K)(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\Big).
\end{aligned} \tag{65}
$$

In the following, we use $\mathbb{E}_k$ to denote the expectation with respect to the randomness in the k-th batch.

We have

$$\begin{aligned}\boldsymbol{\Delta}_t^{K<\infty} =& \boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t}^{K<\infty} \\ =& \boldsymbol{w}^* - \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \hat{\boldsymbol{w}}_{(i),t} \\ =& \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}) \quad (\text{since } \boldsymbol{w}^* = \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \boldsymbol{w}^*).\end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \sum_{i\in[m]} n_{(i),t}^2 \left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}\right\|^2 \\ &+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} n_{(i),t} n_{(j),t} (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})^\top (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(j),t}).\end{aligned} \tag{66}$$

By Assumption 1, we know that at round t, different agents' data are independent with each other. Thus, we have

$$\mathop{\mathbb{E}}_t (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})^\top (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(j),t}) = \mathop{\mathbb{E}}_t (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})^\top \mathop{\mathbb{E}}_t (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(j),t}).$$

Thus, by Eq. (66), to calculate $\mathbb{E}_t \left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2$, it remains to calculate $\mathbb{E}_t \left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}\right\|^2$ and $\mathbb{E}_t(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t})$ for all $i \in [m]$. To that end, we have

$$\hat{\boldsymbol{w}}_{(i),t,k} = \left(\mathbf{I}_p - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}} \mathbf{X}_{(i),t,k} \mathbf{X}_{(i),t,k}^\top\right) \hat{\boldsymbol{w}}_{(i),t,k-1} + \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}} \mathbf{X}_{(i),t,k} (\mathbf{X}_{(i),t,k}^\top \boldsymbol{w}_{(i),t} + \boldsymbol{\epsilon}_{(i),t,k}).$$

We thus have

$$\begin{aligned}\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k} =& \left(\mathbf{I}_p - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}} \mathbf{X}_{(i),t,k} \mathbf{X}_{(i),t,k}^\top\right) (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}) + \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}} \mathbf{X}_{(i),t,k} \mathbf{X}_{(i),t,k}^\top (\boldsymbol{w}^* - \boldsymbol{w}_{(i),t}) \\ &+ \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}} \mathbf{X}_{(i),t,k} \boldsymbol{\epsilon}_{(i),t,k}.\end{aligned} \tag{67}$$

By Lemma 4 and recalling Eq. (3), we thus have

$$\mathop{\mathbb{E}}_k (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k}) = (1 - \alpha_{(i),t})(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}) + \alpha_{(i),t} \boldsymbol{\gamma}_{(i),t}. \tag{68}$$

Applying Eq. (68) recursively and recalling that $\hat{\boldsymbol{w}}_{(i),t,0} = \boldsymbol{\Delta}_{t-1}^{K<\infty}$, we thus have

$$\mathop{\mathbb{E}}_{1,2,\cdots,k} (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k}) = (1 - \alpha_{(i),t})^k \boldsymbol{\Delta}_{t-1}^{K<\infty} + \left(1 - (1 - \alpha_{(i),t})^k\right) \boldsymbol{\gamma}_{(i),t}. \tag{69}$$

By letting $k = K$ in Eq. (69) and $\hat{\boldsymbol{w}}_{(i),t,K} = \hat{\boldsymbol{w}}_{(i),t}$, we thus have

$$\mathop{\mathbb{E}}_t (\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}) = (1 - \alpha_{(i),t})^K \boldsymbol{\Delta}_{t-1}^{K<\infty} + \left(1 - (1 - \alpha_{(i),t})^K\right) \boldsymbol{\gamma}_{(i),t}. \tag{70}$$

Plugging Eq. (70) into Eq. (66), we thus have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{t}\left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2 \mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t}\right\|^2 \\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\mathop{\mathbb{E}}_{t}(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t})^\top \mathop{\mathbb{E}}_{t}(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(j),t}) \qquad (71)\\
=&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2 \mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t}\right\|^2 \\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\Big((1-\alpha_{(i),t})^K(1-\alpha_{(j),t})^K\left\|\boldsymbol{\Delta}_{t-1}^{K<\infty}\right\|^2 \\
&+(1-\alpha_{(i),t})^K(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(j),t}^\top\boldsymbol{\Delta}_{t-1}^{K<\infty}+(1-\alpha_{(j),t})^K(1-(1-\alpha_{(i),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\Delta}_{t-1}^{K<\infty} \\
&+(1-(1-\alpha_{(i),t})^K)(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\Big) \\
=&\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]} n_{(i),t}^2 \mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t}\right\|^2 \\
&+\frac{1}{(\sum_{i\in[m]} n_{(i),t})^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\Big((1-\alpha_{(i),t})^K(1-\alpha_{(j),t})^K\left\|\boldsymbol{\Delta}_{t-1}^{K<\infty}\right\|^2 \\
&+2(1-\alpha_{(i),t})^K(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(j),t}^\top\boldsymbol{\Delta}_{t-1}^{K<\infty} \\
&+(1-(1-\alpha_{(i),t})^K)(1-(1-\alpha_{(j),t})^K)\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t}\Big). \qquad (72)
\end{aligned}
$$

Notice that in Eq. (71) we use $\mathbb{E}_t(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t})^\top(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(j),t}) = \mathbb{E}_t(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t})^\top\,\mathbb{E}_t(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(j),t})$ for $i\neq j$, since $\hat{\boldsymbol{w}}_{(i),t}$ and $\hat{\boldsymbol{w}}_{(j),t}$ are independent with respect to the randomness during the local updates at round t.

By Eqs. (5) and (70), we thus have

$$
\mathbb{E}\,\boldsymbol{\Delta}_t^{K<\infty} = \frac{\sum_{i\in[m]} n_{(i),t}(1-\alpha_{(i),t})^K}{\sum_{i\in[m]} n_{(i),t}}\mathbb{E}\,\boldsymbol{\Delta}_{t-1}^{K<\infty}+\frac{\sum_{i\in[m]} n_{(i),t}\left(1-(1-\alpha_{(i),t})^K\right)\boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}. \qquad (73)
$$

Applying Eq. (73) recursively and recalling Eq. (9), we thus have

$$
\mathbb{E}[\boldsymbol{\Delta}_l^{K<\infty}] = \boldsymbol{g}_l^{K<\infty}, \qquad (74)
$$

where $\boldsymbol{g}_l^{K<\infty}$ is defined in Eq. (61).

By Eq. (67), we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{k}\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k}\right\|^2 \\
=&(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1})^\top\left(\mathbf{I}_p-2\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top+\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}^2}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\right)(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1}) \\
&+\boldsymbol{\gamma}_{(i),t}^\top\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}^2}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}\boldsymbol{\gamma}_{(i),t}+\boldsymbol{\epsilon}_{(i),t,k}^\top\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}^2}\mathbf{X}_{(i),t,k}^\top\mathbf{X}_{(i),t,k}\boldsymbol{\epsilon}_{(i),t,k} \\
&+2\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\boldsymbol{\gamma}_{(i),t}^\top\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\left(\mathbf{I}_p-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}\mathbf{X}_{(i),t,k}\mathbf{X}_{(i),t,k}^\top\right)(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1}) \\
=&\left(1-2\alpha_{(i),t}+\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left\|\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1}\right\|^2+\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \\
&+\alpha_{(i),t}^2\frac{p}{\tilde{n}_{(i),t}}\sigma_{(i),t}^2+2\alpha_{(i),t}\left(1-\frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\boldsymbol{\gamma}_{(i),t}^\top(\boldsymbol{w}^*-\hat{\boldsymbol{w}}_{(i),t,k-1}) \quad \text{(by Lemma 4)}. \qquad (75)
\end{aligned}
$$

Plugging Eq. (69) into Eq. (75), we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{1,2,\cdots,k} \left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k}\right\|^2 \\
=&\left((1-\alpha_{(i),t})^2 + \frac{\alpha_{(i),t}^2(p+1)}{\tilde{n}_{(i),t}}\right) \mathop{\mathbb{E}}_{1,2,\cdots,k-1} \left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}\right\|^2 + \frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \\
&+ \alpha_{(i),t}^2 \frac{p}{\tilde{n}_{(i),t}}\sigma_{(i),t}^2 + 2\alpha_{(i),t}\left(1 - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)(1-\alpha_{(i),t})^{k-1}\boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K<\infty} \\
&+ 2\alpha_{(i),t}\left(1 - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left(1-(1-\alpha_{(i),t})^{k-1}\right)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \\
=&\mathcal{A}_{(i),t}\,\mathbb{E}\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t,k-1}\right\|^2 + \mathcal{B}'_{(i),t,k},
\end{aligned} \tag{76}
$$

where $\mathcal{A}_{(i),t}$ is defined in Eq. (62) and

$$
\begin{aligned}
&\mathcal{B}'_{(i),t,k} \\
:=&\frac{\alpha_{(i),t}^2 p\sigma_{(i),t}^2}{\tilde{n}_{(i),t}} \\
&+ \left(\frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1) + 2\alpha_{(i),t}\left(1 - \frac{\alpha_{(i),t}}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)\left(1-(1-\alpha_{(i),t})^{k-1}\right)\right)\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 \\
&+ 2\left(\alpha_{(i),t} - \frac{\alpha_{(i),t}^2}{\tilde{n}_{(i),t}}(\tilde{n}_{(i),t}+p+1)\right)(1-\alpha_{(i),t})^{k-1}\boldsymbol{\gamma}_{(i),t}^\top \boldsymbol{\Delta}_{t-1}^{K<\infty}.
\end{aligned}
$$

We also define $\mathcal{B}_{(i),t,k}$ by replacing $\boldsymbol{\Delta}_{t-1}^{K<\infty}$ in $\mathcal{B}'_{(i),t,k}$ with $\mathcal{F}_{t-1}$, i.e., Eq. (63).

Applying Eq. (76) recursively over $k = 1, 2, \cdots, K$, we thus have

$$
\mathop{\mathbb{E}}_{t}\left\|\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{(i),t}\right\|^2 = \mathcal{A}_{(i),t}^K \left\|\boldsymbol{\Delta}_{t-1}^{K<\infty}\right\|^2 + \sum_{k=1}^{K} \mathcal{B}_{(i),t,k}\mathcal{A}_{(i),t}^{K-k}. \tag{77}
$$

Plugging Eqs. (74) and (77) into Eq. (72), we thus have

$$
\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2 = \mathcal{J}_t\,\mathbb{E}\left\|\boldsymbol{\Delta}_{t-1}^{K<\infty}\right\|^2 + \mathcal{Q}_t, \tag{78}
$$

where $\mathcal{J}_t$ is defined in Eq. (64) and $\mathcal{Q}_t$ is defined in Eq. (65).

Applying Eq. (78) recursively, we thus have Eq. (18).

D Proof of Theorem 3

Proof. In the overparameterized situation, after each agent trains to converge, we have

$$
\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} = \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\left(\boldsymbol{y}_{(i),t} - \mathbf{X}_{(i),t}^\top \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}\right) + \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}. \tag{79}
$$

For any $i \in [m]$, we define $\mathbf{P}_{(i),t} \in \mathbb{R}^{p\times p}$ as

$$
\mathbf{P}_{(i),t} \coloneqq \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\mathbf{X}_{(i),t}^\top. \tag{80}
$$

(We know $\mathbf{P}_{(i),t}$ is an orthogonal projection since $\mathbf{P}_{(i),t}\mathbf{P}_{(i),t} = \mathbf{P}_{(i),t}$ and $\mathbf{P}_{(i),t}^\top = \mathbf{P}_{(i),t}$.) By Eqs. (2), (79) and (80), we thus have

$$
\hat{\boldsymbol{w}}_{(i),t}^{K=\infty} = \mathbf{P}_{(i),t}\boldsymbol{w}_{(i),t} + (\mathbf{I}_p - \mathbf{P}_{(i),t})\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty} + \mathbf{X}_{(i),t}\left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t}\right)^{-1}\boldsymbol{\epsilon}_{(i),t}. \tag{81}
$$

We thus have

$$
\begin{aligned}
&\boldsymbol{\Delta}_t^{K=\infty} \\
=&\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t}^{K=\infty} \quad \text{(by Eq. (8))} \\
=&\boldsymbol{w}^* - \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \left(\mathbf{P}_{(i),t}\boldsymbol{w}_{(i),t} + (\mathbf{I}_p - \mathbf{P}_{(i),t})\hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty} + \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right) \\
&\text{(by Eqs. (5) and (81))} \\
=&\frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \left(\mathbf{P}_{(i),t}(\boldsymbol{w}^* - \boldsymbol{w}_{(i),t}) + (\mathbf{I}_p - \mathbf{P}_{(i),t})(\boldsymbol{w}^* - \hat{\boldsymbol{w}}_{\text{avg},t-1}^{K=\infty}) - \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right) \\
&\text{(since } \boldsymbol{w}^* = \frac{\sum_{i\in[m]} n_{(i),t}(\mathbf{P}_{(i),t} + \mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{w}^*}{\sum_{i\in[m]} n_{(i),t}}\text{)} \\
=&\frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t} \left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} + (\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty} - \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right) \\
&\text{(by Eqs. (3) and (8)).}
\end{aligned} \tag{82}
$$

For any $i, j \in [m]$, because $\boldsymbol{\epsilon}_{(j),t}$ is independent of $\boldsymbol{\Delta}_{t-1}^{K=\infty}$ and $\mathbf{X}_{(i),t}$, and also because $\boldsymbol{\epsilon}_{(j),t}$ has zero mean (by Assumption 1), we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} \right)^\top \mathbf{X}_{(j),t} \left(\mathbf{X}_{(j),t}^\top \mathbf{X}_{(j),t} \right)^{-1} \boldsymbol{\epsilon}_{(j),t} \right] \\
=&\mathbb{E}\left[((\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty})^\top \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right] \\
=&0,
\end{aligned} \tag{83}
$$

and

$$
\mathbb{E}\left[\mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right] = \mathbf{0}. \tag{84}
$$

Since $\mathbf{P}_{(i),t}(\mathbf{I}_p - \mathbf{P}_{(i),t}) = \mathbf{0}$, we have

$$
\left(\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} \right)^\top (\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty} = 0. \tag{85}
$$

Thus, by Eqs. (82), (83) and (85), we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_t \left\| \boldsymbol{\Delta}_t^{K=\infty} \right\|^2 \\
=&\frac{\sum_{i\in[m]} n_{(i),t}^2 \left(\mathbb{E}_t \left\| (\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty} \right\|^2 + \mathbb{E}_t \left\| \mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} \right\|^2 + \mathbb{E}_t \left\| \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right\|^2 \right)}{(\sum_{i\in[m]} n_{(i),t})^2} \\
&+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} n_{(i),t} n_{(j),t} \left(\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} \right. \\
&\left. + {\boldsymbol{\Delta}_{t-1}^{K=\infty}}^\top (\mathbf{I}_p - \mathbf{P}_{(j),t})(\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty} + 2\boldsymbol{\gamma}_{(j),t}^\top \mathbf{P}_{(j),t}(\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty} \right).
\end{aligned} \tag{86}
$$

For any $i \in [m]$, we have

$$
\mathop{\mathbb{E}}_t \left\| \mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t} \right\|^2 = \frac{n_{(i),t}}{p} \left\| \boldsymbol{\gamma}_{(i),t} \right\|^2 \quad \text{(by Lemma 2)}, \tag{87}
$$

$$
\mathop{\mathbb{E}}_t \left\| (\mathbf{I}_p - \mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty} \right\|^2 = \left(1 - \frac{n_{(i),t}}{p} \right) \left\| \boldsymbol{\Delta}_{t-1}^{K=\infty} \right\|^2 \quad \text{(by Lemma 2)}, \tag{88}
$$

$$
\mathop{\mathbb{E}}_t \left\| \mathbf{X}_{(i),t} \left(\mathbf{X}_{(i),t}^\top \mathbf{X}_{(i),t} \right)^{-1} \boldsymbol{\epsilon}_{(i),t} \right\|^2 = \frac{n_{(i),t}\sigma_i^2}{p - n_{(i),t} - 1} \quad \text{(by Lemma 3)}. \tag{89}
$$

For any $i, j \in [m]$ where $i \neq j$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t}\left[{\boldsymbol{\Delta}_{t-1}^{K=\infty}}^{\top}(\mathbf{I}_p-\mathbf{P}_{(j),t})(\mathbf{I}_p-\mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty}\right]\\
=&\mathop{\mathbb{E}}_{t}\left[(\mathbf{I}_p-\mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty}\right]^{\top}\mathop{\mathbb{E}}_{t}\left[(\mathbf{I}_p-\mathbf{P}_{(j),t})\boldsymbol{\Delta}_{t-1}^{K=\infty}\right]\\
&\text{(since } \mathbf{P}_{(i),t} \text{ and } \mathbf{P}_{(j),t} \text{ are independent when } i \neq j)\\
=&\left(1-\frac{n_{(i),t}}{p}\right)\left(1-\frac{n_{(j),t}}{p}\right)\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2 \text{ (by Lemma 5).}
\end{aligned} \tag{90}$$

Similarly, for $i \neq j$, we have

$$\mathop{\mathbb{E}}_{t}\left[\boldsymbol{\gamma}_{(j),t}^{\top}\mathbf{P}_{(j),t}\mathbf{P}_{(i),t}\boldsymbol{\gamma}_{(i),t}\right]=\frac{n_{(i),t}n_{(j),t}}{p^2}\boldsymbol{\gamma}_{(j),t}^{\top}\boldsymbol{\gamma}_{(i),t} \text{ (by Lemma 5),} \tag{91}$$

and

$$\mathop{\mathbb{E}}_{t}\left[\boldsymbol{\gamma}_{(j),t}^{\top}\mathbf{P}_{(j),t}(\mathbf{I}_p-\mathbf{P}_{(i),t})\boldsymbol{\Delta}_{t-1}^{K=\infty}\right]=\frac{n_{(j),t}}{p}\left(1-\frac{n_{(i),t}}{p}\right)\boldsymbol{\gamma}_{(j),t}^{\top}\boldsymbol{\Delta}_{t-1}^{K=\infty} \text{ (by Lemma 5).} \tag{92}$$

Plugging Eqs. (90) to (92) and (87) to (89) into Eq. (86), we thus have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t}\left\|\boldsymbol{\Delta}_{t}^{K=\infty}\right\|^2\\
=&\frac{\sum_{i\in[m]}n_{(i),t}^2\left(\left(1-\frac{n_{(i),t}}{p}\right)\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2+\frac{n_{(i),t}}{p}\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2+\frac{n_{(i),t}\sigma_{(i),t}^2}{p-n_{(i),t}-1}\right)}{\left(\sum_{i\in[m]}n_{(i),t}\right)^2}\\
&+\frac{1}{\left(\sum_{i\in[m]}n_{(i),t}\right)^2}\sum_{i\in[m]}\sum_{j\in[m]\setminus\{i\}}n_{(i),t}n_{(j),t}\left(\frac{n_{(i),t}n_{(j),t}}{p^2}\boldsymbol{\gamma}_{(j),t}^{\top}\boldsymbol{\gamma}_{(i),t}\right.\\
&\left.+\left(1-\frac{n_{(i),t}}{p}\right)\left(1-\frac{n_{(j),t}}{p}\right)\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2+2\frac{n_{(j),t}}{p}\left(1-\frac{n_{(i),t}}{p}\right)\boldsymbol{\gamma}_{(j),t}^{\top}\boldsymbol{\Delta}_{t-1}^{K=\infty}\right).
\end{aligned} \tag{93}$$

By Eq. (82), we also have

$$\mathop{\mathbb{E}}_{t}[\boldsymbol{\Delta}_{t}^{K=\infty}]=\frac{1}{\sum_{i\in[m]}n_{(i),t}}\sum_{i\in[m]}n_{(i),t}\left(\frac{n_{(i),t}}{p}\boldsymbol{\gamma}_{(i),t}+\left(1-\frac{n_{(i),t}}{p}\right)\boldsymbol{\Delta}_{t-1}^{K=\infty}\right). \tag{94}$$

Applying Eq. (94) recursively, we thus have

$$\mathbb{E}[\boldsymbol{\Delta}_{l}^{K=\infty}]=\boldsymbol{g}_{l}^{K=\infty}, \tag{95}$$

where $\boldsymbol{g}_{l}^{K=\infty}$ is defined in Eq. (21).

By Eqs. (93) and (95), we thus have

$$\mathbb{E}\left\|\boldsymbol{\Delta}_{t}^{K=\infty}\right\|^2=C_t\cdot\mathbb{E}\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2+D_t, \tag{96}$$

where C_t denotes the coefficient of $\left\|\boldsymbol{\Delta}_{t-1}^{K=\infty}\right\|^2$ and D_t denotes the remaining parts. The specific expressions of C_t and D_t are in Eqs. (24) and (25). Applying Eq. (96) recursively, we thus have Eq. (26).

Underparameterized situation

In the underparameterized situation, the convergence point of local steps in each round corresponds to the solution that minimizes the training loss, i.e.,

$$\begin{aligned}
\hat{\boldsymbol{w}}_{(i),t}^{K=\infty}=&(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^{\top})^{-1}\mathbf{X}_{(i),t}\boldsymbol{y}_{(i),t}\\
=&(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^{\top})^{-1}\mathbf{X}_{(i),t}(\mathbf{X}_{(i),t}^{\top}\boldsymbol{w}_{(i),t}+\boldsymbol{\epsilon}_{(i),t}) \text{ (by Eq. (2))}\\
=&\boldsymbol{w}_{(i),t}+(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^{\top})^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}.
\end{aligned}$$

Also recalling Eqs. (3) and (8), we thus have

$$\boldsymbol{\Delta}_t^{K=\infty} = \frac{1}{\sum_{i\in[m]} n_{(i),t}} \sum_{i\in[m]} n_{(i),t}(\boldsymbol{\gamma}_{(i),t} - (\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}). \tag{97}$$

For any $i, j \in [m]$, because $\boldsymbol{\epsilon}_{(j),t}$ is independent of $\mathbf{X}_{(i),t}$ and $\boldsymbol{\epsilon}_{(i),t}$, and also because $\boldsymbol{\epsilon}_{(j),t}$ has zero mean (by Assumption 1), we have

$$\mathbb{E}\left[\boldsymbol{\gamma}_{(j),t}^\top(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right] = 0 \quad \text{for all } i, j \in [m],$$
$$\mathbb{E}\left[\left(\mathbf{X}_{(j),t}\mathbf{X}_{(j),t}^\top)^{-1}\mathbf{X}_{(j),t}\boldsymbol{\epsilon}_{(j),t}\right)^\top(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right] = 0 \quad \text{for all } i \neq j.$$

Thus, by Eq. (97), we have

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2 =& \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \sum_{i\in[m]} n_{(i),t}^2 \left(\left\|\boldsymbol{\gamma}_{(i),t}\right\|^2 + \mathbb{E}\left\|(\mathbf{X}_{(i),t}\mathbf{X}_{(i),t}^\top)^{-1}\mathbf{X}_{(i),t}\boldsymbol{\epsilon}_{(i),t}\right\|^2\right) \\
&+ \frac{1}{(\sum_{i\in[m]} n_{(i),t})^2} \sum_{i\in[m]} \sum_{j\in[m]\setminus\{i\}} n_{(i),t}n_{(j),t}\boldsymbol{\gamma}_{(i),t}^\top\boldsymbol{\gamma}_{(j),t} \\
=& \left\|\frac{\sum_{i\in[m]} n_{(i),t}\boldsymbol{\gamma}_{(i),t}}{\sum_{i\in[m]} n_{(i),t}}\right\|^2 + \frac{\sum_{i\in[m]} \frac{n_{(i),t}^2 p\sigma_{(i),t}^2}{n_{(i),t}-p-1}}{(\sum_{i\in[m]} n_{(i),t})^2} \text{ (by Eq. (43) in Lemma 3).}
\end{aligned}$$

We thus have proven Eq. (27).

The result of this theorem thus follows. □

E A Table for Notations

We provide a table of some important notations used in this paper.

symbol	meaning
$n_{(i),t}$	number of training samples
$\tilde{n}_{(i),t}$	batch size
p	number of parameters
$\sigma_{(i),t}$	noise level
$\mathbf{X}_{(i),t}$	matrix for input of training samples
$\boldsymbol{y}_{(i),t}$	vector for output of training samples
$\boldsymbol{\epsilon}_{(i),t}$	vector for noise of training samples
$\hat{\boldsymbol{w}}_0$	the pre-trained parameters (initialization)
$\boldsymbol{w}^*$	the learning target
$\boldsymbol{w}_{(i),t}$	the ground-truth of agent i at round t
$\hat{\boldsymbol{w}}_{(i),t}^{K=1}, \hat{\boldsymbol{w}}_{(i),t}, \hat{\boldsymbol{w}}_{(i),t}^{K=\infty}$	the local learning result of agent i at round t
$\hat{\boldsymbol{w}}_{(i),t,k}$	learning result after k-th batch (for $K < \infty$ case)
$\hat{\boldsymbol{w}}_{\text{avg},t}^{K=1}, \hat{\boldsymbol{w}}_{\text{avg},t}^{K<\infty}, \hat{\boldsymbol{w}}_{\text{avg},t}^{K=\infty}$	the FedAvg result at round t
$\left\|\boldsymbol{\Delta}_t^{K=1}\right\|^2, \left\|\boldsymbol{\Delta}_t^{K<\infty}\right\|^2, \left\|\boldsymbol{\Delta}_t^{K=\infty}\right\|^2$	model error
$\left\|\boldsymbol{\Delta}_0\right\|^2$	initial (pre-trained) model error
$\alpha_{(i),t}$	learning rate (step size)
$\boldsymbol{\gamma}_{(i),t}$	measurement of heterogeneity

Table 2: Table for some notations.