Provable Benefits of Local Steps in Heterogeneous Federated Learning for Neural Networks: A Feature Learning Perspective

Yajie Bao [1] Michael Crawshaw [2] Mingrui Liu [2]

Abstract

Local steps are crucial for Federated Learning (FL) algorithms and have witnessed great empirical success in reducing communication costs and improving the generalization performance of deep neural networks. However, there are limited studies on the effect of local steps on heterogeneous FL. A few works investigate this problem from the optimization perspective. Woodworth et al. (2020a) showed that the iteration complexity of Local SGD, the most popular FL algorithm, is dominated by the baseline mini-batch SGD, which does not show the benefits of local steps. In addition, Levy (2023) proposed a new local update method that provably benefits over mini-batch SGD. However, in the same setting, there is still no work analyzing the effects of local steps to generalization in a heterogeneous FL setting. Motivated by our experimental findings where Local SGD learns more distinguishing features than parallel SGD, this paper studies the generalization benefits of local steps from a feature learning perspective. We propose a novel federated data model that exhibits a new form of data heterogeneity, under which we show that a convolutional neural network (CNN) trained by GD with *global* updates will miss some pattern-related features, while the network trained by GD with *local* updates can learn all features in polynomial time. Consequently, local steps help CNN generalize better in our data model. In a different parameter setting, we also prove that Local GD with *one-shot* model averaging can learn all features and generalize well in all clients. Our experimental results also confirm the benefits of local steps in improving test accuracy on real-world data.

1. Introduction

Federated learning (FL) is an efficient distributed paradigm in which local clients train a global model collaboratively by performing multiround gradient-based updates using local data (Konečný et al., 2016). Local steps in FL algorithms are initially designed to reduce the communication cost between clients and the server, while they are mysteriously useful to improve the generalization performance of deep neural networks. For example, extensive experiments in (Lin et al., 2018) showed that local steps could improve the test accuracy of models compared with parallel SGD in the homogeneous data setting. To theoretically explain this phenomenon, Gu et al. (2023b;a) investigated the dynamic of Local SGD based on the stochastic differential equation (SDE) approximation and showed that it enjoys a reduction of sharpness when choosing a large number of local steps and thus generalizes better in the test data. However, these works are restricted to the homogeneous data setting.

In the heterogeneous data setting, Woodworth et al. (2020b) proved that a baseline method mini-batch SGD that only takes global updates converges faster than Local SGD. To show the benefits of local steps in optimization, Levy (2023) proposed a new local update method and proved it requires less communication round than mini-batch SGD under the gradient dissimilarity assumption for heterogeneous data. However, these works purely focus on the optimization perspective. To the best of our knowledge, there is still no work considering the generalization effect of local steps in a heterogeneous FL regime.

Our work is motivated by an experimental finding of the CIFAR-10 task in the heterogeneous FL environment, where we can see that Local SGD learned more distinguishing features than parallel SGD in Figure 1 (e.g., one can compare the pixels inside the bounding boxes). This finding inspires us to analyze the benefits of local steps from a feature learning (Allen-Zhu & Li, 2022a;b) perspective: *How does the FL algorithm with local steps learn heterogeneous features more efficiently than that without local steps?*

To characterize the benefits of local steps in generalization, we build a new heterogeneous data model in the FL setting. Different from the traditional definition of data heterogene-

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

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

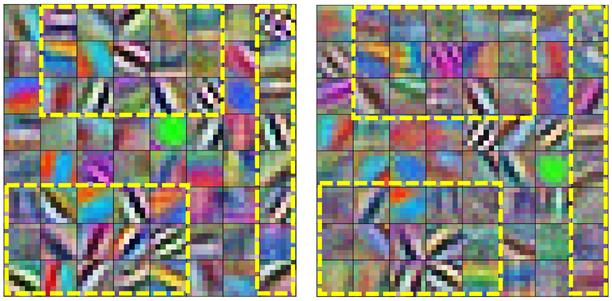

Figure 1: Weight visualization for Local SGD (left) vs Parallel SGD (right). The visualization shows the first layer convolutional weights (64 filters) for ResNet-18 trained on CIFAR-10 with heterogeneous features. We use standard data augmentation and feature heterogeneity $h = 0.25$. See Section 8 for training details.

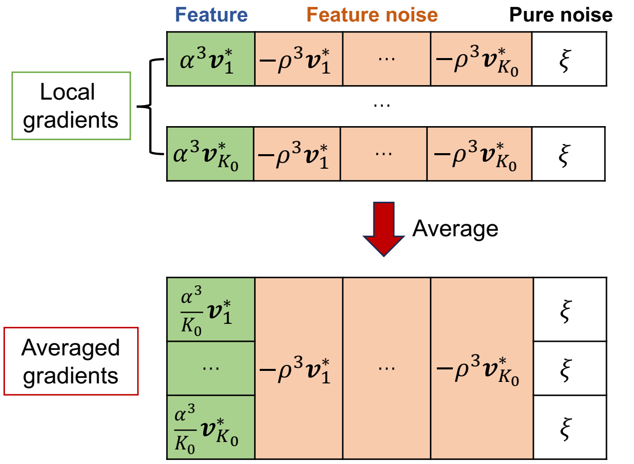
Feature
Feature noise
Pure noise
Local gradients
$\alpha^3 \boldsymbol{v}_1^*$
$-\rho^3 \boldsymbol{v}_1^*$
...
$-\rho^3 \boldsymbol{v}_{K_0}^*$
ξ
...
$\alpha^3 \boldsymbol{v}_{K_0}^*$
$-\rho^3 \boldsymbol{v}_1^*$
...
$-\rho^3 \boldsymbol{v}_{K_0}^*$
ξ
Average
Averaged gradients
$\frac{\alpha^3}{K_0} \boldsymbol{v}_1^*$
...
$\frac{\alpha^3}{K_0} \boldsymbol{v}_{K_0}^*$
$-\rho^3 \boldsymbol{v}_1^*$
...
$-\rho^3 \boldsymbol{v}_{K_0}^*$
ξ
ξ
ξ

Figure 2: The signals in local gradients for the clients containing confounding features and the signals in the globally averaged gradients, where $\alpha, \rho > 0$ are the scale of feature and feature noise. The intensity of the positive signal significantly shrinks after averaging, while the negative signal keeps the same scale.

ity in FL literature (i.e., gradient dissimilarity), our data model's heterogeneity depends solely on raw data and does not depend on the architecture of the neural network used in the learning process. The data distribution includes two key ingredients: (1) **Feature heterogeneity**. The data label in different clients is determined by different feature vectors, which means that learning the local pattern-related feature is necessary to guarantee good generalization performance in the corresponding client. The feature vectors are divided into normal features and confounding features. (2) **Adversarial heterogeneity**. The local data has a different *feature noise* structure, i.e., features of the opposite label (Allen-Zhu & Li, 2022b; Zou et al., 2021). Specifically, clients with confounding pattern-related features share the same feature noise components; the clients with normal pattern-related features have individual feature noise components.

We consider using a one-hidden-layer convolutional neural network (CNN) with a cubic activation function to learn the features of our proposed data model. Note that cubic is the smallest degree of polynomial that makes the neural network non-linear and compatible with our setting, which is also used by (Jelassi & Li, 2022). To help intuitively understand the difference between GD with global updates and local updates in the training process, we plot the signals in the local gradients and the globally averaged gradients in Figure 2. The feature noise in gradients carries a negative signal that will slow down the learning of pattern-related features. In each local client, the intensity of the negative signal can be neglected compared with that of the positive signal. Therefore, running a large number of local steps will capture the local pattern-related feature. However, averaging the gradients from clients with confounding features will shrink the relative scale of the individual positive signal. In Section 4, we formally show that gradient descent (GD) with global updates *cannot* learn the confounding features and naturally has a worse generalization performance in the clients whose data label is determined by those features. Moreover, we also prove that GD with local updates can learn all the features in polynomial time and hence enjoy a superior generalization ability.

Our contributions can be summarized as follows:

- We proposed a new heterogeneous data model in FL, which is a refined version of multi-view data from the paper (Allen-Zhu & Li, 2022b). Different from the conventional data heterogeneity definition in FL literature which uses gradient dissimilarity (Li et al., 2020b; Khaled et al., 2020; Woodworth et al., 2020a; Karimireddy et al., 2020), our data model introduces a novel *adversarial* heterogeneity regarding the feature noise structure across different local clients, which is pivotal in showing the superiority of local updates in generalization over global updates.

- Based on the proposed data model, our results formally show that the CNN model trained by GD with global updates and local updates can both attain zero training error in polynomial time with high probability. But GD with global updates can only learn normal features and prefers to fit the noise in the clients with confounding features. Finally, the network learned by global updates behaves like random guessing for the test data in these clients. On the other hand, GD with local updates successfully learns both normal and confounding features and can attain almost perfect test accuracy in all clients. The negative results for global updates rely on a careful inspection of the scale of gradients computed from the clients with confounding features since we need the negative signal to stay strong before memorizing noise. In addition, under a different parameter setting with the mild signal of feature noise, we also show that Local GD with one-shot model averaging can learn all pattern-related features and attain low test error in all clients. However, GD needs at least polynomial communication rounds to learn those features.

- We complement our theory with experimental results showing that Local SGD generalizes better than Parallel SGD on a CIFAR-10 task with heterogeneous features across clients. Both Local SGD and Parallel SGD reach almost 100% training accuracy (up to 0.1% error), but Local SGD consistently generalizes better than Parallel SGD in the presence of heterogeneous features. Further, generalization tends to improve as the number of local steps in Local SGD increases. We also evaluate two algorithms on modified CIFAR-10 data with feature noise, where Local SGD has better test accuracy than Parallel SGD across all values of the feature noise magnitude.

2. Related Work

Optimization in Federated Learning. Since the concept of FL was introduced in Konečnỳ et al. (2016); McMahan et al. (2017), there has been a series of studies on federated optimization in various settings. As the leading algorithm in FL, the convergence of FedAvg (also called Local SGD in the literature) was well studied in both convex and nonconvex problems. In the homogeneous setting, where the data distributions in all clients are the same, the convergence analysis of the Local SGD is given in (Stich, 2018; Yu et al., 2019; Khaled et al., 2020; Woodworth et al., 2020b). In the heterogeneous setting, several works (Li et al., 2020c; Glasgow et al., 2022; Woodworth et al., 2020a) showed that gradient dissimilarity will affect the convergence rate of Local SGD. To address the heterogeneity issue, several variants of Local SGD algorithms were proposed, such as FedProx (Li et al., 2020b), SCAFFOLD (Karimireddy et al., 2020). Specifically, the lower bounds in Woodworth et al. (2020a); Patel et al. (2023) showed that the accelerated mini-batch SGD is minimax optimal under some heterogeneity notations. Recently, Levy (2023) proposed a new local method in the heterogeneous setting that achieves less communication complexity than both Local SGD and mini-batch SGD. To better analyze the effectiveness of local steps, (Wang et al., 2022; Patel et al., 2023) suggested introducing a new definition of data heterogeneity. For a comprehensive survey, we refer the readers to Kairouz et al. (2019); Li et al. (2020a) and references therein. In addition, there is also a line of work focusing on the federated optimization of neural networks. Most papers (Li et al., 2021; Huang et al., 2021; Deng et al., 2022; Yue et al., 2022; Yu et al., 2022) studied federated learning algorithms on overparameterized neural networks under the neural tangent kernel (NTK) regime. Bao et al. (2023) directly analyzed the global convergence of Local SGD when training the one-layer neural network with Gaussian input beyond the NTK regime. Our theoretical results are also beyond NTK and only require a moderate level of over-parameterization.

Generalization in Deep Learning. There is a huge literature on generalization for deep learning. A classical lens for studying the generalization of a neural network is uniform convergence (Bartlett, 1998; Bartlett & Mendelson, 2002; Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018). Arora et al. (2018) improved the generalization bounds for a deep neural network using a compression approach. The second line of work studied generalization in deep learning through the lens of implicit bias. The implicit bias was originally introduced for the analysis of algorithmic regularization for linear models (Soudry et al., 2018; Gunasekar et al., 2018a; Ji & Telgarsky, 2018b), and is later widely used to analyze the generalization effect of different optimization algorithms for training neural networks (Lampinen & Ganguli, 2018; Arora et al., 2019; Gunasekar et al., 2018b; Ji & Telgarsky, 2018a; Chizat & Bach, 2020). The third line of work is connecting the generalization performance with the sharpness of the minima (Hochreiter & Schmidhuber, 1997; Neyshabur et al., 2017), and flat minima are believed to generalize well. These works consider how different algorithms find the flat minima (Keskar et al., 2016; Wu et al., 2018; Kleinberg et al., 2018; Foret et al., 2020). The most relevant work to this paper is (Gu et al., 2023b;a), which studied generalization in federated learning by investigating how FedAvg with different local steps find minimizers with different sharpness. However, their work only considers the homogeneous data setting and is not applicable to the heterogeneous data setting as considered in this paper.

Feature Learning. The feature learning is a new perspective for theoretical understanding of neural networks in recent years. In the pioneering work, Allen-Zhu & Li (2022a) developed the concept of feature learning to show the robustness of adversarial training in deep learning. Then, Allen-Zhu & Li (2022b) introduced the multi-view data model to study how the ensemble of deep neural networks and knowledge distillation improves the test accuracy of related tasks. After these two works, a line of papers studied the feature learning process for some algorithms or techniques in deep learning. Wen & Li (2021) investigated how contrastive learning learns the feature representations of neural networks. Zou et al. (2021) demonstrated the generalization gap between GD and Adam (Kingma & Ba,

2014) in training the one-hidden-layer CNN under image-like data distribution. Chen et al. (2022) studied how the Mixture-of-Experts layer (Shazeer et al., 2017) improves the performance of neural network learning. Shen et al. (2022) investigated the effect of data augmentation on the dynamic of the learning process, which helps learn rare features in the training data. Jelassi & Li (2022) showed that the momentum can improve generalization by learning the feature with a small margin. Based on a similar data model, Zou et al. (2023) showed that the Mixup training can effectively learn the rare features.Huang et al. (2023) presented a study on the generalization benefits of structural information in Graph Neural Networks. Kou et al. (2023) studied the benign overfitting problem for two-layer ReLU CNN. Huang et al. (2024) studied the generalization problem of FedAvg in the heterogeneous setting from a feature learning perspective, which shows the generalization benefits of *model averaging* compared with local training. However, to the best of our knowledge, our work is the first to show the generalization benefits of *local steps* in FL algorithms via the feature learning perspective. Moreover, our data model is very different from that in Huang et al. (2024).

3. Problem Setup

3.1. Notations

Given a positive integer n, we write $[n] = \{1, \dots, n\}$. Given an index set $\mathcal{S}$, we denote $|\mathcal{S}|$ the cardinality of $\mathcal{S}$. We use $\mathcal{N}(\mu, \Sigma)$ to denote the Gaussian distribution with mean μ and covariance Σ. We denote $\mathbf{I}_d$ the $d \times d$ identity matrix. We use asymptotic notations $\widetilde{O}, \widetilde{\Theta}, \widetilde{\Omega}$ to hide polylogarithmic factors in d. In addition, we use $\mathsf{poly}(d)$ and $\mathsf{polylog}(d)$ to hide the polynomial orders greater than 1. Throughout the paper, we use η to denote the learning rate.

3.2. Heterogeneous data distribution

Let $\{\boldsymbol{v}_k^*\}_{k\in[K]} \subseteq \mathbb{R}^d$ be the feature vectors. For simplicity, we assume $\|\boldsymbol{v}_k^*\| = 1$ for $k \in [K]$ and $\langle \boldsymbol{v}_k^*, \boldsymbol{v}_{k'}^* \rangle = 0$ if $k \neq k'$. Let $\mathcal{K}_0 \subseteq [K]$ be the indices of confounding features with size K_0. Without loss of generality, we assume $\mathcal{K}_0 = [K_0]$. Suppose there are N clients in the distributed environment. Next, we define the local data distributions.

Definition 3.1. *For the i-th client, we define its local data distribution $\mathcal{D}_i$ as:*

(1) Sample the label $y \in \{-1, 1\}$ uniformly;

(2) Generate the input $\mathbf{x}$ as a vector of P patches, i.e., $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_P) \in (\mathbb{R}^d)^P$ with $P = K_0 + 2$, where

- ***Feature patch.*** *The first patch is given by $\mathbf{x}_1 = \alpha y \cdot \mathbf{v}_i$, where $\mathbf{v}_i$ is uniformly sampled from $\{\boldsymbol{v}_k^*\}_{k=1}^K$ and $\alpha > 0$.*
- ***Feature noise patches.*** *If $\mathbf{v}_i = \boldsymbol{v}_k^*$ with $k \in \mathcal{K}_0$, those patches are given by $\mathbf{x}_p = -\rho y \cdot \boldsymbol{v}_{p-1}^*$ for $2 \le p \le K_0 + 1$. If $\mathbf{v}_i = \boldsymbol{v}_k^*$ with $k \in [K] \setminus \mathcal{K}_0$, those patches are given by $\mathbf{x}_p = -\beta y \cdot \mathbf{v}_i$ for $2 \le p \le K_0 + 1$.*
- ***Pure noise patch.*** *The last patch is given by $\mathbf{x}_P = \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \cdot \mathbf{H})$ and is independent of the label y, where $\mathbf{H} = \mathbf{I}_d - \sum_{k=1}^K \boldsymbol{v}_k^*(\boldsymbol{v}_k^*)^\top$.*

The heterogeneity in the data distribution comes from the feature patch and the components in feature noise patches. In the feature patch, each client has a pattern-relevant feature $\mathbf{v}_i$ from the vectors $\{\boldsymbol{v}_k^*\}_{k\in[K]}$, which represents the individual characteristic of the local data. In the feature noise patches, the components are determined by whether $\mathbf{v}_i$ is a confounding feature: if so, these patches carry the negative signals for all confounding features $\{\boldsymbol{v}_k^*\}_{k\in\mathcal{K}_0}$; otherwise, these patches only carry the negative signal for the pattern relevant feature $\mathbf{v}_i$. The adversarial structure of feature noise patches is crucial to show the failure of GD in learning confounding features.

To keep the analysis clean, we assume the noise patch is sampled from the orthogonal complement of the space spanned by feature vectors. A similar setting is also used in the related work (Jelassi & Li, 2022; Zou et al., 2023; 2021). We defer the detailed setting for the parameters to Definition 3.1 in Section 4.

We define the subset of clients with the feature vector $\boldsymbol{v}_k^*$ as $\mathcal{C}_k = \{i \in [N] : \mathbf{v}_i = \boldsymbol{v}_k^*\}$. We assume $|\mathcal{C}_k| = N/K$ for any $k \in [K]$. In the i-th client, we can collect a data set $\mathcal{Z}_i = \{(\mathbf{x}_{ij}, y_{ij})\}_{j\in[n]}$ generated from the local distribution $\mathcal{D}_i$ in Definition 3.1, where $\mathbf{x}_{ij} = (\mathbf{x}_{ij,1}, \dots, \mathbf{x}_{ij,P})$.

3.3. Learner model and algorithms

We consider the following one-hidden-layer convolutional neural network architecture with cubic activation function:

$$F(\boldsymbol{W}, \mathbf{x}) = \sum_{r\in[m]} \sum_{p\in[P]} \langle \boldsymbol{w}_r, \mathbf{x}_p \rangle^3, \tag{1}$$

where m is the number of hidden neurons and $\boldsymbol{W} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_m\}$ is the model weight. We denote the logistic loss function evaluated at the i-th client as

$$L(\boldsymbol{W}; \mathcal{Z}_i) = \frac{1}{n} \sum_{j\in[n]} \log\left\{1 + e^{-y_{ij}F(\boldsymbol{W}, \mathbf{x}_{ij})}\right\}.$$

At the beginning of training, we randomly initialize the CNN model (1) by independently generating the hidden weights $\{\boldsymbol{w}_r^{(0)}\}_{r\in[m]}$ from the same distribution $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$.

GD with global updates. Given the randomly initialized point $\boldsymbol{W}^{(0)}$, the global GD updates the model weight using the averaged gradient from all clients, that is

$$\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \frac{\eta}{N} \sum_{i \in [N]} \nabla L(\boldsymbol{W}^{(t)}; \mathcal{Z}_i). \tag{2}$$

GD with local updates. Given the same initial point $\boldsymbol{W}_i^{(0)} = \boldsymbol{W}^{(0)}$, the local model weights in the i-th client is updated by

$$\boldsymbol{W}_i^{(t+1)} = \boldsymbol{W}_i^{(t)} - \eta \nabla L(\boldsymbol{W}_i^{(t)}; \mathcal{Z}_i). \tag{3}$$

Let $\hat{\boldsymbol{W}}$ be the model weight of CNN in (1). The local *training error* in the i-th client is defined as

$$\frac{1}{n} \sum_{j \in [n]} \mathbb{1}\{y_{ij} F(\hat{\boldsymbol{W}}, \mathbf{x}_{ij}) < 0\}.$$

Given a test data $(\mathbf{x}, y)$ generated from some distribution $\mathcal{D}_i$, the corresponding *test error* is defined as

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_i} \left\{ yF(\hat{\boldsymbol{W}}, \mathbf{x}) < 0 \right\}.$$

4. Main Results

Before presenting the generalization results, we first give a detailed range for the parameters that appear in Section 3.

Parameter 1. *For the data model in Definition 3.1, we set $\alpha = \Theta(1)$, $\rho^3 = \frac{\alpha^3 - 1/\mathsf{poly}(d)}{K_0}$, $\beta^3 \in (0, \alpha^3/(2K_0)]$, and $\sigma_\xi = \Theta(d^{-0.51})$. For the random initialization, we set $\sigma_0 = \Theta(d^{-0.52})$. The number of confounding features satisfies $K_0 \geq \log^{\varrho}(d)$ with $\varrho > 1/2$. In addition, we set $N, n = \mathsf{polylog}(d)$ and $N > K_0$. Finally, we assume a moderate level of over-parameterization: $m = \mathsf{polylog}(d)$.*

4.1. GD with global updates generalizes badly

Our first theorem shows that the model trained by the GD with global updates fails in generalization for the clients with confounding features.

Theorem 1. *Suppose the setting in Parameter 1 holds. For GD algorithm with global updates, choosing learning rate $\eta \in (0, \widetilde{O}(1)]$, after $T = \frac{\mathsf{poly}(d)}{\eta}$* ***global iterations****,*

- *(Training error attains zero for all clients.) For any client $i \in [N]$ and data $j \in [n]$, with probability at least $1 - 1/\mathsf{poly}(d)$, it holds that*

$$y_{ij} F(\boldsymbol{W}^{(T)}, \mathbf{x}_{ij}) \geq \widetilde{\Omega}(1). \tag{4}$$

- *(Test error is almost zero for clients with normal features.) For the new data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ with $k \in [K] \setminus \mathcal{K}_0$, it holds that*

$$\mathbb{P}\left\{ yF(\boldsymbol{W}^{(T)}, \mathbf{x}) < 0 \right\} \leq \frac{1}{\mathsf{poly}(d)}.$$

- *(Test error is high for clients with confounding features.) For the new data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, it holds that*

$$\mathbb{P}\left\{ yF(\boldsymbol{W}^{(T)}, \mathbf{x}) < 0 \right\} \geq \frac{1}{2} - \frac{1}{\mathsf{polylog}(d)}. \tag{5}$$

From (4), we know that global GD can attain zero training error in all clients with high probability. However, by (5), it behaves like random guessing for test data in clients with confounding features. The reason is that global updates only fit the noise patch in the data from these clients instead of learning pattern-related features $\boldsymbol{v}_k^*$ with $k \in \mathcal{K}_0$. We defer the analysis of the training process of GD with global updates to Section 5.

Remark 4.1. *We shall provide a high-level intuition behind Theorem 1 to understand why global updates cannot learn confounding features. According to the data distribution in Definition 3.1, the data in each client $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ contains the same feature noise patches $\{-\rho \boldsymbol{v}_k^*\}_{k \in \mathcal{K}_0}$. As illustrated in Figure 2, after averaging the gradients from all clients in (2), the intensity of the negative signal of the confounding feature (say $\boldsymbol{v}_1^*$) becomes $-\rho^3 \frac{1}{N} \sum_{k \in \mathcal{K}_0} |\mathcal{C}_k| = -\rho^3 K_0 / K$. Meanwhile, the intensity of the positive signal of $\boldsymbol{v}_1^*$ shrinks to $\alpha^3 |\mathcal{C}_1| / N = \alpha^3 / K$ since the positive signal only exists in the clients $\mathcal{C}_1$. Due to the shrink of the positive signal after averaging gradients, the growth of signal intensity is significantly slower compared with the memorization of noise patches. As a consequence, GD loses the opportunity to learn the confounding features and only memorizes the noise to attain zero training error. As for the normal features $\boldsymbol{v}_k^*$ with $k \in [K] \setminus \mathcal{K}_0$, the positive and negative signals shrink at the same ratio (i.e., $1/K$), hence GD can still learn normal features.*

4.2. GD with local updates generalizes well

Our next theorem shows that GD with local updates can generalize well in all clients.

Theorem 2. *Suppose the setting in Parameter 1 holds. For GD with local updates, choosing the learning rate $\eta \in (0, \widetilde{O}(1)]$, after $T = \frac{\mathsf{poly}(d)}{\eta}$* ***local iterations*** *in each client,*

- *(Training error attains zero for all clients.) For any client $i \in [N]$ and data $j \in [n]$, with probability at least $1 - 1/\mathsf{poly}(d)$, it holds that*

$$y_{ij} F(\boldsymbol{W}_i^{(T)}, \mathbf{x}_{ij}) \geq \widetilde{\Omega}(1).$$

- *(Test error is almost zero for all clients.) For the new data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ with $k \in [K]$, it holds that*

$$\mathbb{P}\left\{ yF(\boldsymbol{W}_i^{(T)}, \mathbf{x}) < 0 \right\} \leq \frac{1}{\mathsf{poly}(d)}.$$

Theorem 2 shows that local steps can learn both normal features and confounding features after a polynomial number of local steps. Compared with the results in Theorem 1, it formally proves the benefits of local steps in generalization, which also confirms the empirical observation. We defer the analysis of local training to Section 6.

Remark 4.2. *Our results in Theorem 2 are consistent with the literature on personalized FL (Sim et al., 2019; Jiang et al., 2023; Fallah et al., 2020), which tries to find a common model by meta-learning approach and then personalizes the model on each client by running local gradient descent separately. In our theorem, the randomly initialized model can be regarded as a common model that does not bias towards any client as in (Fallah et al., 2020), and the local updates try to find individual models during the training. Personalized FL algorithms have been verified to have significantly better test performance than FedAvg on heterogeneous data, which indicates the empirical benefit of local steps.*

5. Analysis of GD with Global Updates

In this section, we provide the analysis sketch of Theorem 1. Given the model weight $\boldsymbol{W}^{(t)}$ in GD, we define its projections on feature vectors and noise patches and its local derivative during the training process:

- The signal intensity of feature learning: $\Gamma_{r,k}^{(t)} = \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{v}_k^* \rangle$ for $r \in [m]$, $k \in [K]$.
- The memorization of noise patch: $\Xi_{r,ij}^{(t)} = \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{\xi}_{ij} \rangle$ for $r \in [m]$, $i \in [N]$ and $j \in [n]$.
- The averaged derivative in each client: $\nu_i^{(t)} = \frac{1}{n} \sum_{j \in [n]} 1/(1 + e^{y_{ij} F(\boldsymbol{W}^{(t)}, \mathbf{x}_{ij})})$ for $i \in [N]$.

In Figure 3, we illustrate the growth of signal intensity of normal and confounding features and the noise memorization during the training process of GD with global updates.

Lemma 5.1 (GD learns normal features)**.** *For any $k \in [K] \in \mathcal{K}_0$, let τ_k be the first iteration that $\max_{r \in [m]} \Gamma_{r,k}^{(t)} \geq \Theta\left(\frac{1}{m^{1/3}\alpha}\right)$. It holds that $\tau_k \leq \mathrm{T}_v^0$. Meanwhile, we can also guarantee that for iterations $t \leq \mathrm{T}_v^0$,*

$$\max_{r \in [m], k' \in \mathcal{K}_0} |\Gamma_{r,k'}^{(t)}| \leq \widetilde{O}(\sigma_0),$$
$$\max_{r \in [m], i \in [N], j \in [n]} |\Xi_{r,ij}^{(t)}| \leq \widetilde{O}(\sqrt{d}\sigma_\xi \sigma_0).$$

Lemma 5.1 guarantees GD can learn all normal features before T_v^0. More importantly, the signal intensity of confounding features and the memorization of noise patches stay at the initial scale, which is plotted in the *green area* of Figure 3. Now let us see the competition of the growth between them in the following lemma.

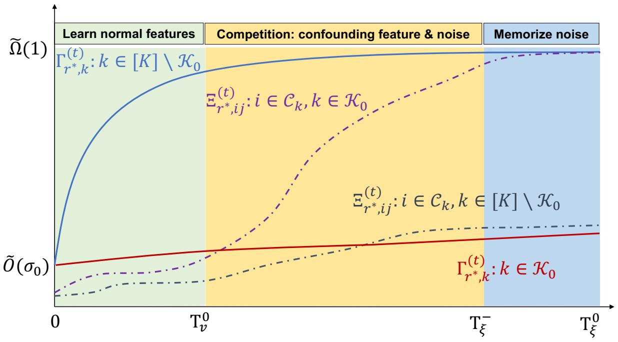

Figure 3: The growth of signal intensity and noise memorization in the training process of GD with global updates. The subscript r^* refers to the hidden weight with a maximum signal. For example, $\Gamma_{r^*,k}^{(t)} \equiv \Gamma_{r_k^*,k}^{(t)}$, where $r_k^* = \arg\max_{r \in [m]} \Gamma_{r,k}^{(0)}$.

Lemma 5.2. *Denote $\mathrm{T}_\xi^- = \Theta\left(\frac{Nn}{\eta(\sqrt{d}\sigma_\xi)^3 \sigma_0}\right)$. Given any $k \in \mathcal{K}_0$, for iterations $t \leq \mathrm{T}_\xi^-$,*

- *The signal intensity of the feature $\boldsymbol{v}_k^*$ is updated by:*
$$\Gamma_{r,k}^{(t+1)} = \Gamma_{r,k}^{(t)} + \Theta\left(\frac{\eta(\alpha^3 - K_0\rho^3)}{K}\right)\left(\Gamma_{r,k}^{(t)}\right)^2.$$
- *The memorization of noise patch $\boldsymbol{\xi}_{ij}$ for $i \in \mathcal{C}_k$ and $j \in [n]$ is updated by:*
$$y_{ij}\Xi_{r,ij}^{(t+1)} = y_{ij}\Xi_{r,ij}^{(t)} + \Theta\left(\frac{\eta d \sigma_\xi^2}{Nn}\right)\left(y_{ij}\Xi_{r,ij}^{(t)}\right)^2 + o(\sqrt{d}\sigma_\xi\sigma_0).$$

According to Parameter 1, we know $\frac{\alpha^3 - K_0\rho^3}{K} \leq O(d^{-1})$ and $\frac{d\sigma_\xi^2}{Nn} = \widetilde{O}(d^{-0.02})$. Therefore, noise memorization is much faster than learning confounding features. The tensor power method guarantees that GD first memorizes the noise patches. After that, the scale of gradients in the client with confounding features is not enough to learn these features. The competition between learning confounding features and memorizing noise is shown in the following lemma, which is also illustrated in the *yellow and blue areas* of Figure 3.

Lemma 5.3 (GD memorizes noises in clients with confounding features)**.** *Given any $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ and $j \in [n]$, let τ_{ij}^0 be the first iteration that $\max_{r \in [m]}\left(y_{ij}\Xi_{r,ij}^{(t)}\right) \geq \Theta(m^{-\frac{1}{3}})$. It holds that $\tau_{ij}^0 \leq \mathrm{T}_\xi^0 = \mathrm{T}_\xi^- + O(\log(d))$. Meanwhile, we can also guarantee that $\max_{r \in [m]} |\Gamma_{r,k}^{(t)}| \leq \widetilde{O}(\sigma_0)$ for any $t \leq \mathrm{T}_\xi^0$ and $k \in \mathcal{K}_0$.*

Lemma 5.3 guarantees the signal intensity of confounding features at iteration T_ξ^- is bounded by $\widetilde{O}(\sigma_0)$, the derivative scale after T_ξ^- is not enough to learn confounding features.

6. Analysis of GD with Local Updates

Given the local model weight $\boldsymbol{W}_i^{(t)} = \{\boldsymbol{w}_{1,i}^{(t)}, \dots, \boldsymbol{w}_{m,i}^{(t)}\}$ in the i-th client, we define the following iterates during the local training process:

- The signal intensity of feature learning: $\Gamma_{r,k,i}^{(t)} = \langle \boldsymbol{w}_{r,i}^{(t)}, \boldsymbol{v}_k^* \rangle$ for $r \in [m]$ and $k \in [K]$.
- The memorization of noise patch: $\Xi_{r,ij}^{(t)} = \langle \boldsymbol{w}_{r,i}^{(t)}, \boldsymbol{\xi}_{ij} \rangle$ for $r \in [m]$ and $j \in [n]$.
- The averaged derivative in each client: $\nu_i^{(t)} = \frac{1}{n}\sum_{j\in[n]} 1/(1+e^{y_{ij}F(\boldsymbol{W}_i^{(t)}, \mathbf{x}_{ij})})$ for $i \in [N]$.

To avoid introducing notations with complex subscripts, we use the same notations $\Xi_{r,ij}^{(t)}$ and $\nu_i^{(t)}$ those have appeared in Section 5 but were defined by global weight. In this section, they are computed by local model weights $\boldsymbol{W}_i^{(t)}$.

The next lemma presents the local update rules of the signal intensity of normal features and confounding features. In particular, we need to pay more attention to the clients with the confounding feature because it also contains the negative signal of other confounding features.

Lemma 6.1. *Denote* $\mathrm{T}_v^0 = \Theta\left(\frac{1}{\eta\alpha^3\sigma_0}\right) + O(\log d)$*. For any* $t \le \mathrm{T}_v^0$*, we have the following local update rules:*

- *If* $i \in \mathcal{C}_k$ *with* $k \in [K] \setminus \mathcal{K}_0$*,*

$$\Gamma_{r,k,i}^{(t+1)} = \Gamma_{r,k,i}^{(t)} + \Theta\left\{\eta(\alpha^3 - K_0\beta^3)\right\}\left(\Gamma_{r,k,i}^{(t)}\right)^2.$$

- *If* $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$*,*

$$\Gamma_{r,k,i}^{(t+1)} = \Gamma_{r,k,i}^{(t)} + \Theta\left\{\eta(\alpha^3 - \rho^3)\right\}\left(\Gamma_{r,k,i}^{(t)}\right)^2, \quad (6)$$

and for $k' \in \mathcal{K}_0 \setminus \{k\}$*,*

$$-\Gamma_{r,k',i}^{(t+1)} = -\Gamma_{r,k',i}^{(t)} + \Theta\left(\eta\rho^3\right)\left(\Gamma_{r,k',i}^{(t)}\right)^2. \quad (7)$$

By comparing the growth speed in (6) and (7), we know $(\alpha^3 - \rho^3)/\rho^3 = \Omega(K_0) \ge \log^{\varrho}(d)$ with $\varrho > 1/2$ according to Parameter 1. Since all clients share the same initial model, we know $\Gamma_{r,k,i}^{(0)} \equiv \Gamma_{r,k}^{(0)}$ for any $i \in [N]$. By concentration, the initial signal intensity satisfies that

$$\min_{r\in[m]} \Gamma_{r,k',i}^{(0)} \ge -O(\sigma_0\sqrt{\log d}), \quad \max_{r\in[m]} \Gamma_{r,k,i}^{(0)} \ge \Omega(\sigma_0).$$

Therefore, applying the tensor power method can guarantee that the learning of positive signals in (6) will win in the competition with the learning negative signals in (7). The consequence is stated in the following lemma.

Lemma 6.2 (Local steps learn all the features)**.** *For any* $i \in \mathcal{C}_k$ *with* $k \in [K]$*, let* $\tau_{k,i}$ *be the first iteration when* $\max_{r\in[m]} \Gamma_{r,k,i}^{(t)} > \Theta\left(\frac{1}{m^{1/3}\alpha}\right)$*. It holds that* $\tau_{k,i} \le \mathrm{T}_v^0 = \mathrm{T}_v^- + O(\log d)$*. In addition, for the client* $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$*, we also have* $\min_{r\in[m]} \Gamma_{r,k',i}^{(t)} \ge -\widetilde{O}(\sigma_0)$ *for any* $k' \in \mathcal{K}_0 \setminus \{k\}$ *and* $t \le \tau_{k,i}$*.*

Credited to local steps, each client can learn its local feature after T_v^0 local iterations. At the same time, the negative signal for confounding features stays at the initial scale. After $\tau_{k,i}$, the derivative in each local client is not large enough to pick up the negative signal. Hence, the feature noise will not affect the generalization of the local model.

7. Local GD with One-shot Model Averaging can Learn All Features

Under the data model with Parameter 1, we have shown that GD with global updates (without model averaging) can learn all features but GD with global average fails. In this section, we will compare Local GD with *one-shot* model averaging and GD with global updates under the following parameter setting.

Parameter 2. *We keep the setting identical to Parameter 1, other than choose* $\rho = 1/\mathsf{poly}(d)$*.*

The following theorem characterizes feature learning results of two algorithms under Parameter 2.

Theorem 3. *Suppose the setting in Parameter 2 holds. For GD with global updates, choosing learning rate* $\eta \in (0, \widetilde{O}(1)]$*, if the number of global iterations (communication rounds)* t *is smaller than* $\widetilde{\Omega}\left(\frac{1}{\eta\alpha^3\sigma_0}\right)$*, the global model* $\boldsymbol{W}^{(t)}$ *satisfies that for any* $k \in [K] \setminus \mathcal{K}_0$*,*

$$\max_{r\in[m]} \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{v}_k^* \rangle = \widetilde{O}(\sigma_0).$$

Theorem 4. *Suppose the setting in Parameter 2 holds. For GD with local updates, choosing learning rate* $\eta \in (0, \widetilde{O}(1)]$ *and* $T = \frac{\mathsf{poly}(d)}{\eta}$ *local iterations, the one-shot averaged model* $\bar{\boldsymbol{W}} = \frac{1}{N}\sum_{i\in[N]} \boldsymbol{W}_i^{(T)}$ *satisfies that: (1) for any* $k \in [K]$, $\max_{r\in[m]}\langle \bar{\boldsymbol{w}}_r, \boldsymbol{v}_k^* \rangle = \widetilde{\Omega}(1)$*; (2) and for a new data* $(\mathbf{x}, y) \sim \mathcal{D}_i$ *with any* $i \in \mathcal{C}_k$ *and any* $k \in [K]$*,*

$$\mathbb{P}\left\{yF(\bar{\boldsymbol{W}}, \mathbf{x}) < 0\right\} \le \frac{1}{\mathsf{poly}(d)}.$$

Under Parameter 2, Theorem 3 shows that GD with global updates cannot learn all pattern-related features if the number of communication rounds is smaller than $\mathsf{poly}(d)/\eta$. Meanwhile, Theorem 4 shows that Local GD with one-shot model averaging can learn all features and consequently generalize better than GD with global updates. Therefore, we can still show the generalization benefits of local steps after model averaging.

Algorithm	CIFAR-10 (with augmentation) $h = 0.25$	 $h = 0.5$	CIFAR-10 (no augmentation) $h = 0.25$	 $h = 0.5$
Parallel SGD	90.17 ± 0.19	90.17 ± 0.19	77.73 ± 0.20	77.73 ± 0.20
Local SGD ($I = 8$)	91.01 ± 0.17	90.71 ± 0.25	80.35 ± 0.14	80.45 ± 0.66
Local SGD ($I = 16$)	$\mathbf{91.21 \pm 0.25}$	90.84 ± 0.07	80.64 ± 0.12	80.77 ± 0.30
Local SGD ($I = 32$)	91.19 ± 0.22	$\mathbf{91.08 \pm 0.25}$	$\mathbf{80.86 \pm 0.17}$	$\mathbf{81.27 \pm 0.36}$

Table 1: Average test accuracy over three trials, with and without data augmentation and varying $h \in \{0.25, 0.5\}$. The error is the distance from the average to the max/min across three runs. All runs achieved at least 99.9% training accuracy. Note that Parallel SGD is unaffected by h, since it does not utilize local steps.

8. Experiments

To complement our theoretical results, we provide experiments to empirically verify the generalization benefit of local steps when training on real-world data. We train ResNet-18 models for image classification on a modified CIFAR-10 task, in which the dataset is split across clients to create heterogeneous features across clients. We compare the generalization performance of Local SGD with a different number of local steps against Parallel SGD (no local updates), and find that local steps tend to increase test accuracy, which is consistent with our theory. The code is available at https://github.com/MingruiLiu-ML-Lab/Provable-Benefit-Local-Steps-Feature-Learning.

8.1. Heterogeneous CIFAR-10 task

Previous works in heterogeneous FL (Karimireddy et al., 2020) use a non-IID data partitioning protocol to split a centralized dataset into local datasets that have a different distribution of labels. In our experiments, we use a modified protocol that creates *heterogeneous features* while enforcing that all clients have the same label distribution, in order to isolate the effect of heterogeneous features.

We first split the dataset $D = \{(x_i, y_i)\}_{i=1}^n$ into two pieces based on the parity of the label, so that for $j \in \{0, 1\}$ each piece is $D_j = \{(x, y) \in D \mid y \pmod 2 = j\}$. We then apply the non-IID partitioning protocol from (Karimireddy et al., 2020) to split each D_j into N partitions, allocating samples to each partition according to the 10-way label. The local dataset for client i is comprised of the i-th partition of D_0 together with the i-th partition of D_1. Finally, we replace the target of each sample from a 10-way label $y \in \{0, \dots 9\}$ to a binary label $y \pmod 2$. As a result, each local dataset has the same 50-50 distribution of binary target labels, but has a different distribution of images according to their original 10-way label.

The non-IID partitioning protocol from (Karimireddy et al., 2020) has a parameter $h \in [0, 1]$ that controls the level of heterogeneity. 0 corresponds to IID partitioning, and 1 induces maximal heterogeneity offered by the protocol. In our experiments, we evaluate two heterogeneity settings of $h = 0.25$ and $h = 0.5$.

Training Settings. We evaluate Local SGD with varying communication intervals ($I \in \{8, 16, 32\}$) and Parallel SGD, which is equivalent to Local SGD with $I = 1$. In all settings, we use a ResNet-18 architecture and the cross-entropy loss function. We set $N = 8$ clients and use a batch size of 64 for each client, so Parallel SGD is equivalent to SGD with batch size $64N = 512$.

We run training separately with and without data augmentation (random flip, random crop). Without data augmentation, we use a learning rate $\eta = 0.01$ and train for 16k update steps (≈ 170 epochs). With data augmentation, we set $\eta = 0.03$ and train for 65k steps (≈ 670 epochs). We use a number of steps sufficiently long to ensure that every training run reaches at least 99.9% training accuracy. In all settings, we decay the learning rate by a factor of 2 two times: once after $1/2$ of the total steps and once after $3/4$ of the total steps. The training was implemented in PyTorch and executed on a cluster of 8 NVIDIA A6000 GPUs.

Results. Table 1 shows the average test accuracy achieved by each algorithm across the four settings. Despite achieving the same training accuracy, Local SGD with any number of local steps reaches a higher test accuracy than Parallel SGD in all four settings. Across multiple random seeds, the worst test accuracy from Local SGD is better than the best test accuracy from Parallel SGD, indicating that the generalization boost of local steps is a consistent phenomenon.

For three of the four settings, the test accuracy of Local SGD improves as the number of local steps I increases, further suggesting that more local steps are better for generalization. This is a significant practical benefit since the cost of communication between clients is significantly reduced as I increases (assuming the number of iterations remains fixed). In practice, local steps provide a win-win situation: increased generalization and reduced communication.

8.2. CIFAR-10 with feature noise

To evaluate Local SGD under real-world data with feature noise, we train with modified CIFAR-10 data that explicitly includes feature noise similar to our theoretical framework from Definition 3.1. The modified data is constructed in the following way:

1. Let $\rho > 0$ be the feature noise magnitude.
2. Partition the dataset into N client datasets according to the heterogeneity protocol outlined in the main paper. Denote the client datasets as $D_1, \ldots, D_n$.
3. For each client i, sample $(\tilde{\mathbf{x}}_i, \tilde{y}_i)$ uniformly at random from $\cup_{j \neq i} D_j$. Then, for each data sample $(\mathbf{x}, y) \in D_i$, define a modified data sample $\mathbf{x}' = \mathbf{x} - \rho \tilde{\mathbf{x}}_i$. Let D_i' be the set of data samples $(\mathbf{x}', y)$ modified from samples $(\mathbf{x}, y) \in D_i$.

The dataset D_i' then consists of images from D_i, with an additional negative signal corresponding to an image $\tilde{\mathbf{x}}_i$ from a different client, and ρ is the magnitude of the feature noise. This construction emulates the negative signal of the feature noise patches from Definition 3.1, while still using real-world data.

Training Settings. After constructing the feature noise datasets D_i', we run training using the same experimental setup as in the CIFAR-10 experiments of the main body. We evaluate the setting of $h = 0.5$ with data augmentation. We also set $I = 128$ for Local SGD, which is an intermediate value of I between experiments of the main paper and the large I experiments of Section F. To understand the effect of the feature noise, we allow the feature noise magnitude ρ to vary over $\{0.03125, 0.0625, 0.125, 0.25\}$. Lastly, we run for 24k update steps (compared to 16k from the main paper) to ensure that all training runs reach near 100% training accuracy. Each setup was evaluated over three trials with different random seeds. Examples of the modified images are shown in Figure 4. Each image contains slightly more noise than the previous row, but overall the images retain their original signal.

Results. The testing accuracies of Parallel SGD and Local SGD are shown in Figure 5 for varying values of ρ, averaged over the three trials with different random seeds. The error bars range from minimum to the maximum test accuracy over the three trials. Note that all training runs each at least 99.9% training accuracy.

Local SGD has better test accuracy than Parallel SGD for all values of the feature noise magnitude ρ. Further, the gap between the two algorithms increases as ρ increases, showing that Local SGD can handle feature noise better than Parallel SGD. This experimental result confirms our theory, and shows that local steps can improve generalization in the presence of feature noise, even with real world data.

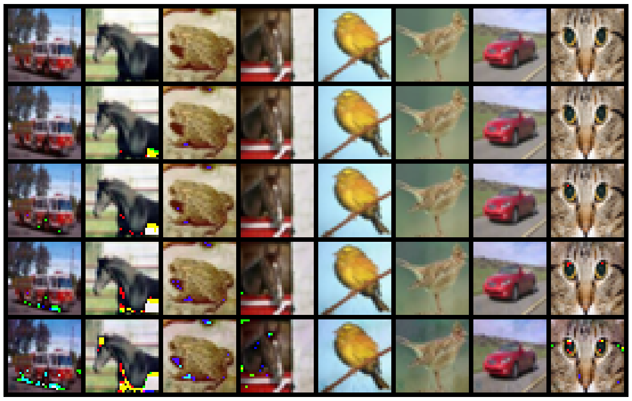

Figure 4: Modified CIFAR-10 images, where feature noise has been added. Each row shows a different value of the feature noise magnitude ρ, ranging over $\{0.0, 0.03125, 0.0625, 0.125, 0.25\}$.

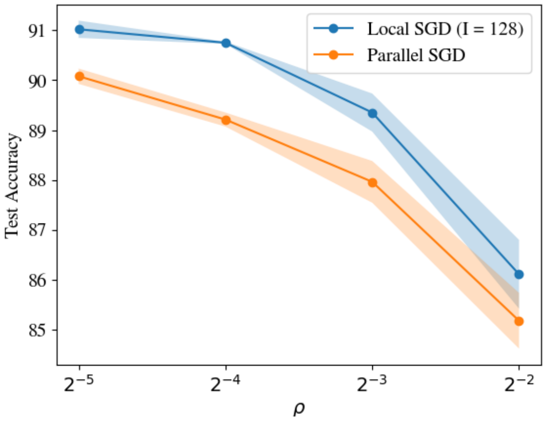

Figure 5: Test accuracy for Local SGD and Parallel SGD under different strengths of feature noise, with three trials. Local SGD always has better test accuracy.

9. Discussion

This paper explores the effectiveness of local steps of the gradient-based method in learning pattern-related features under heterogeneous FL. Due to the new adversarial form of heterogeneity in our synthetic data model, we formally prove the generalization superiority of GD with local updates over global updates. We also prove Local GD with the large number of local steps and one-shot model averaging can generalize well in all clients under mild magnitude of feature noises. In our future work, we aim to provide a more refined analysis to study the feature learning process of the vanilla Local SGD algorithm and its variants.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This work is partially supported by the Institute for Digital Innovation fellowship from George Mason University, a ORIEI seed funding from George Mason University, and a Cisco Faculty Research Award. Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu). The work of Yajie Bao was done when he was virtually visiting Mingrui Liu's research group in the Department of Computer Science at George Mason University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022a.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2022b.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Bao, Y., Shehu, A., and Liu, M. Global convergence analysis of local SGD for two-layer neural network without overparameterization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Bartlett, P. Thesamplecomplexityofp atternclassification withneuralnetworks: Thesizeoftheweightsismo reimportantthan thesizeofthenetwork. *IEEETrans. Inf. Theory*, 44 (2), 1998.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Deng, Y., Kamani, M. M., and Mahdavi, M. Local sgd optimizes overparameterized neural networks in polynomial time. In *International Conference on Artificial Intelligence and Statistics*, pp. 6840–6861. PMLR, 2022.

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

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Glasgow, M. R., Yuan, H., and Ma, T. Sharp bounds for federated averaging (local sgd) and continuous perspective. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9050–9090. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/glasgow22a.html.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Gu, X., Lyu, K., Arora, S., Zhang, J., and Huang, L. A quadratic synchronization rule for distributed deep learning, 2023a.

Gu, X., Lyu, K., Huang, L., and Arora, S. Why (and when) does local SGD generalize better than SGD? In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=svCcui6Drl.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018a.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018b.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Huang, B., Li, X., Song, Z., and Yang, X. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, pp. 4423–4434. PMLR, 2021.

Huang, W., Cao, Y., Wang, H., Cao, X., and Suzuki, T. Graph neural networks provably benefit from structural information: A feature learning perspective, 2023.

Huang, W., Shi, Y., Cai, Z., and Suzuki, T. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EcetCr4trp.

Jelassi, S. and Li, Y. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pp. 9965–10040. PMLR, 2022.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018a.

Ji, Z. and Telgarsky, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018b.

Jiang, Y., Konečný, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning, 2023.

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

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

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kleinberg, R., Li, Y., and Yuan, Y. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign overfitting for two-layer relu convolutional neural networks, 2023.

Lampinen, A. K. and Ganguli, S. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*, 2018.

Levy, K. Y. Slowcal-sgd: Slow query points improve local-sgd for stochastic convex optimization. *arXiv preprint arXiv:2304.04169*, 2023.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

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

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020c. URL https://openreview.net/forum?id=HJxNAnVtDS.

Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

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

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Patel, K. K., Glasgow, M., Wang, L., Joshi, N., and Srebro, N. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023. URL https://openreview.net/forum?id=vhS68bKv7x.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

Shen, R., Bubeck, S., and Gunasekar, S. Data augmentation as feature manipulation. In *International conference on machine learning*, pp. 19773–19808. PMLR, 2022.

Sim, K. C., Zadrazil, P., and Beaufays, F. An investigation into on-device personalization of end-to-end automatic speech recognition models, 2019.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Stich, S. U. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., and Zhang, T. On the unreasonable effectiveness of federated averaging with heterogeneous data, 2022.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

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

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

Wu, L., Ma, C., et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Yu, Y., Wei, A., Karimireddy, S. P., Ma, Y., and Jordan, M. Tct: Convexifying federated learning using bootstrapped neural tangent kernels. *Advances in Neural Information Processing Systems*, 35:30882–30897, 2022.

Yue, K., Jin, R., Pilgrim, R., Wong, C.-W., Baron, D., and Dai, H. Neural tangent kernel empowered federated learning. In *International Conference on Machine Learning*, pp. 25783–25803. PMLR, 2022.

Zou, D., Cao, Y., Li, Y., and Gu, Q. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

Zou, D., Cao, Y., Li, Y., and Gu, Q. The benefits of mixup for feature learning. *arXiv preprint arXiv:2303.08433*, 2023.

A. Preliminaries

A.1. The range of parameters

We restate the setting in Parameter 1 for our proofs:

- $\sigma_\xi = \Theta(d^{-0.51})$, $\sigma_0 = \Theta(d^{-0.52})$.
- $N, n, m = \mathsf{polylog}(d)$.
- $\Omega(\log^{\varrho}(d)) \le K_0 \le N$ with $\varrho > 1/2$.
- $\alpha = \Theta(1)$.
- $\rho^3 = \frac{\alpha^3 - 1/\mathsf{poly}(d)}{K_0}$.
- $\beta^3 \le \alpha^3/(2K_0)$.

A.2. Random initialization

Lemma A.1. *Under the Gaussian initialization, with probability* $1 - 1/\mathsf{poly}(d)$*, we have*

- *Given any* $k \in [K]$, $\max_{r\in[m]} \Gamma_{r,k}^{(0)} > \Omega(\sigma_0)$*. In addition,* $\max_{r\in[m],k\in[K]} |\Gamma_{r,k}^{(0)}| \le O\left(\sigma_0\sqrt{\log d}\right)$.
- *Given any* $i \in [N]$ *and* $j \in [n]$, $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(0)} > \Omega(\sqrt{d}\sigma_\xi\sigma_0)$*. In addition,* $\max_{r\in[m],i\in[N],j\in[n]} |\Xi_{r,ij}^{(0)}| \le O\left(\sigma_\xi\sigma_0\sqrt{d\log d}\right)$.

Proof. According to the random initialization, we know $\{\boldsymbol{w}_r^{(0)}\}_{r\in[m]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2\mathbf{I}_d)$. It follows that

- For each $k \in [K]$, $\{\Gamma_{r,k}^{(0)} = \langle \boldsymbol{w}_r^{(0)}, \boldsymbol{v}_k^*\rangle\}_{r\in[m]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2)$.
- Fixing $\boldsymbol{\xi}_{ij}$ and y_{ij}, $\{y_{ij}\Xi_{r,ij}^{(0)} = \langle \boldsymbol{w}_r^{(0)}, y_{ij}\boldsymbol{\xi}_{ij}\rangle\}_{r\in[m]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2\|\boldsymbol{\xi}_{ij}\|^2)$.

Let $\Phi(\cdot)$ be the cumulative distribution function of $\mathcal{N}(0,1)$. For any positive constant $c = O(1)$, we have

$$\mathbb{P}\left(\Gamma_{r,k}^{(0)} > c\sigma_0\right) = \mathbb{P}\left(\Gamma_{r,k}^{(0)}/\sigma_0 > c\right) = 1 - \Phi(c).$$

It follows that

$$\mathbb{P}\left(\max_{r\in[m]} \Gamma_{r,k}^{(0)} > c\sigma_0\right) = \mathbb{P}\left(\cup_{r=1}^m \{\langle \boldsymbol{w}_r^{(0)}, \boldsymbol{v}_k^*\rangle > c\sigma_0\}\right) = 1 - [1-\Phi(c)]^m \ge 1 - 1/\mathsf{poly}(d).$$

Similarly, we also have

$$\mathbb{P}\left(\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(0)} > c\sigma_0\|\boldsymbol{\xi}_{ij}\| \mid \boldsymbol{\xi}_{ij}\right) = 1 - 1/\mathsf{poly}(d). \tag{A.1}$$

By the distribution of $\boldsymbol{\xi}_{ij}$ in Definition 3.1, we can write $\boldsymbol{\xi}_{ij} = (\mathbf{I}_d - \sum_{k\in\mathcal{K}_0} \boldsymbol{v}_k^*(\boldsymbol{v}_k^*)^\top)\boldsymbol{\xi}_d$ for some $\boldsymbol{\xi}_d \sim \mathcal{N}(0, \sigma_\xi^2\mathbf{I}_d)$. Since $(\boldsymbol{v}_k^*)^\top\boldsymbol{\xi}_d \sim \mathcal{N}(0, \sigma_\xi^2)$, then we have

$$\begin{aligned}
\|\boldsymbol{\xi}_{ij}\| = \left\|\boldsymbol{\xi}_d - \sum_{k\in\mathcal{K}_0} \boldsymbol{v}_k^*(\boldsymbol{v}_k^*)^\top\boldsymbol{\xi}_d\right\| &\ge \|\boldsymbol{\xi}_d\| - \left\|\sum_{k\in\mathcal{K}_0} \boldsymbol{v}_k^*(\boldsymbol{v}_k^*)^\top\boldsymbol{\xi}_d\right\| \\
&\ge \Theta(d\sigma_\xi) - \sum_{k\in\mathcal{K}_0} \|\boldsymbol{v}_k^*\| \cdot |(\boldsymbol{v}_k^*)^\top\boldsymbol{\xi}_d| \\
&\ge \Theta(d\sigma_\xi) - O\left(K_0\sigma_\xi\sqrt{\log d}\right) \\
&= \Theta(d\sigma_\xi),
\end{aligned}$$

where we used Lemma E.1 and $K_0 = \mathsf{polylog}(d)$. Together with (A.1), we can guarantee that $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(0)} > \Omega(\sqrt{d}\sigma_\xi\sigma_0)$ with probability $1 - 1/\mathsf{poly}(d)$. □

A.3. A threshold for vanishing derivative

Recall the local loss

$$L(\boldsymbol{W}; \mathcal{Z}_i) = \frac{1}{n} \sum_{j \in [n]} \log \left(1 + e^{-y_{ij} F(\boldsymbol{W}, \mathbf{x}_{ij})}\right).$$

We compute its gradient w.r.t. the hidden weight $\boldsymbol{w}_r$ by

$$\begin{aligned} \nabla_{\boldsymbol{w}_r} L(\boldsymbol{W}; \mathcal{Z}_i) &= -\frac{1}{n} \sum_{j \in [n]} y_{ij} \ell_{ij}(\boldsymbol{W}) \nabla_{\boldsymbol{w}_r} F(\boldsymbol{W}, \mathbf{x}_{ij}) \\ &= -\frac{1}{n} \sum_{j \in [n]} y_{ij} \ell_{ij}(\boldsymbol{W}) \sum_{p \in [P]} 3 \langle \boldsymbol{w}_r, \mathbf{x}_{ij,p} \rangle^2 \cdot \mathbf{x}_{ij,p}, \end{aligned} \tag{A.2}$$

where $\ell_{ij}(\boldsymbol{W}) = 1/\left(1 + e^{y_{ij} F(\boldsymbol{W}, \mathbf{x}_{ij})}\right)$. Given the total iterations $T = \frac{\mathsf{poly}(d)}{\eta}$, we define a threshold $\vartheta = \log(\widetilde{\Omega}(T)) = \widetilde{\Omega}(1)$ such that

$$\sum_{t=0}^{T} \frac{1}{1 + \exp(\vartheta)} \leq \widetilde{O}(1). \tag{A.3}$$

Hence we can guarantee that $\sum_{s=t_0}^{T} \ell_{ij}^{(t)} \leq \widetilde{O}(1)$ if $y_{ij} F(\boldsymbol{W}^{(s)}, \mathbf{x}_{ij}) \geq \vartheta$ holds for $s \geq t_0$.

B. GD with global updates under Parameter 1

Given the iterate $\boldsymbol{W}^{(t)}$ in GD, define the following iterates during the training process:

- Signal intensity of feature learning: $\Gamma_{r,k}^{(t)} = \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{v}_k^* \rangle$ for $r \in [m]$ and $k \in [K]$.
- Noise memorization: $\Xi_{r,ij}^{(t)} = \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{\xi}_{ij} \rangle$ for $r \in [m]$, $i \in [N]$ and $j \in [n]$.
- Derivative: $\ell_{ij}^{(t)} = \ell_{ij}(\boldsymbol{W}^{(t)}) = 1/\left(1 + e^{y_{ij} F(\boldsymbol{W}^{(t)}, \mathbf{x}_{ij})}\right)$ for $i \in [N]$ and $j \in [n]$.
- Derivative in each client: $\nu_i^{(t)} = \frac{1}{n} \sum_{j \in [n]} \ell_{ij}^{(t)}$ for $i \in [N]$.
- Maximum signal intensity: $\Gamma_{r^*,k}^{(t)} \equiv \Gamma_{r_k^*,k}^{(t)}$, where $r_k^* = \arg\max_{r \in [m]} \Gamma_{r,k}^{(0)}$.
- Maximum noise memorization: $\Xi_{r^*,ij}^{(t)} \equiv \Xi_{r_{ij}^*,ij}^{(t)}$, where $r_{ij}^* = \arg\max_{r \in [m]} y_{ij} \Xi_{r,ij}^{(0)}$.

B.1. Update rules of signal intensity and noise memorization

Using the data distribution and gradients in (A.2), for the signal intensity of $\boldsymbol{v}_k^*$, we have the following update rules:

- If $k \in \mathcal{K}_0$,

$$\begin{aligned} \Gamma_{r,k}^{(t+1)} &= \Gamma_{r,k}^{(t)} - \frac{\eta}{N} \sum_{i \in [N]} \langle \nabla_{\boldsymbol{w}_r} L(\boldsymbol{W}^{(t)}; \mathcal{D}_i), \boldsymbol{v}_k^* \rangle \\ &= \Gamma_{r,k}^{(t)} + \frac{\eta}{N} \sum_{i \in [N]} \frac{1}{n} \sum_{j \in [n]} y_{ij} \ell_{ij}^{(t)} \sum_{p \in [P]} 3 \langle \boldsymbol{w}_r^{(t)}, \mathbf{x}_{ij,p} \rangle^2 \langle \mathbf{x}_{ij,p}, \boldsymbol{v}_k^* \rangle \\ &= \Gamma_{r,k}^{(t)} + \frac{3\eta}{Nn} \sum_{i \in [N]} \left\{ \mathbb{1}_{\{i \in \mathcal{C}_k\}} \alpha^3 \sum_{j \in [n]} \ell_{ij}^{(t)} \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{v}_k^* \rangle^2 - \mathbb{1}_{\{i \in \cup_{k' \in \mathcal{K}_0} \mathcal{C}_{k'}\}} \rho^3 \sum_{j \in [n]} \ell_{ij}^{(t)} \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{v}_k^* \rangle^2 \right\} \end{aligned}$$

$$
\begin{aligned}
&= \Gamma_{r,k}^{(t)} + \frac{3\eta}{Nn}\left[\alpha^3 \sum_{i\in\mathcal{C}_k}\sum_{j\in[n]} \ell_{ij}^{(t)} - \rho^3 \sum_{k'\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_{k'}}\sum_{j\in[n]} \ell_{ij}^{(t)}\right]\left(\Gamma_{r,k}^{(t)}\right)^2 \\
&= \Gamma_{r,k}^{(t)} + \frac{3\eta}{N}\left[\alpha^3 \sum_{i\in\mathcal{C}_k} \nu_i^{(t)} - \rho^3 \sum_{k'\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_{k'}} \nu_i^{(t)}\right]\left(\Gamma_{r,k}^{(t)}\right)^2. \qquad \text{(B.1)}
\end{aligned}
$$

- If $k \in [K] \setminus \mathcal{K}_0$,

$$
\begin{aligned}
\Gamma_{r,k}^{(t+1)} &= \Gamma_{r,k}^{(t)} - \frac{\eta}{N}\sum_{i\in[N]} \langle \nabla_{\boldsymbol{w}_r} L(\boldsymbol{W}^{(t)}; \mathcal{D}_i), \boldsymbol{v}_k^* \rangle \\
&= \Gamma_{r,k}^{(t)} + \frac{\eta}{N}\sum_{i\in[N]} \left\{ \mathbb{1}_{\{i\in\mathcal{C}_k\}} (\alpha^3 - K_0\beta^3)\frac{1}{n}\sum_{j\in[n]} \ell_{ij}^{(t)} 3\langle \boldsymbol{w}_r^{(t)}, \boldsymbol{v}_k^* \rangle^2 \right\} \\
&= \Gamma_{r,k}^{(t)} + \frac{3\eta}{Nn}\left[(\alpha^3 - K_0\beta^3)\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]} \ell_{ij}^{(t)}\right]\left(\Gamma_{r,k}^{(t)}\right)^2 \\
&= \Gamma_{r,k}^{(t)} + \left[\frac{3\eta(\alpha^3 - K_0\beta^3)}{N}\sum_{i\in\mathcal{C}_k} \nu_i^{(t)}\right]\left(\Gamma_{r,k}^{(t)}\right)^2. \qquad \text{(B.2)}
\end{aligned}
$$

In fact, we also used the facts $\langle \boldsymbol{v}_k^*, \boldsymbol{v}_1^* \rangle = 0$ if $k \neq 1$ and $\langle \boldsymbol{v}_k^*, \boldsymbol{\xi}_{ij} \rangle = 0$ almost surely due to $\mathbf{H}\boldsymbol{v}_k^* = \mathbf{0}$.

The noise memorization of the patch $\boldsymbol{\xi}_{ij}$ is given by:

- If $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$,

$$
\begin{aligned}
\Xi_{r,ij}^{(t+1)} &= \Xi_{r,ij}^{(t)} - \frac{\eta}{N}\sum_{i'\in[N]} \langle \nabla_{\boldsymbol{w}_r} L(\boldsymbol{W}^{(t)}; \mathcal{D}_{i'}), \boldsymbol{\xi}_{ij} \rangle \\
&= \Xi_{r,ij}^{(t)} + \frac{\eta}{N}\sum_{i'\in[N]} \frac{1}{n}\sum_{j'\in[n]} y_{i'j'} \ell_{i'j'}^{(t)} 3\langle \boldsymbol{w}_r^{(t)}, \mathbf{x}_{i'j'} \rangle^2 \langle \mathbf{x}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle \\
&= \Xi_{r,ij}^{(t)} + \eta\left[\frac{3}{Nn} y_{ij}\ell_{ij}^{(t)} \|\boldsymbol{\xi}_{ij}\|^2\right]\left(\Xi_{r,ij}^{(t)}\right)^2 \\
&\quad + \frac{3\eta}{Nn}\sum_{j'\in[n]\setminus\{j\}} y_{ij'}\ell_{ij'}^{(t)}\left(\Xi_{r,ij'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{ij'}, \boldsymbol{\xi}_{ij} \rangle \\
&\quad + \frac{3\eta}{Nn}\sum_{i'\in\mathcal{C}_k\setminus\{i\}}\sum_{j'\in[n]} y_{i'j'}\ell_{i'j'}^{(t)}\left(\Xi_{r,i'j'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle \\
&\quad + \frac{3\eta}{Nn}\sum_{k'\in\mathcal{K}_0\setminus\{k\}}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]} y_{i'j'}\ell_{i'j'}^{(t)}\left(\Xi_{r,i'j'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle \\
&\quad + \frac{3\eta}{Nn}\sum_{k'\in[K]\setminus\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]} y_{i'j'}\ell_{i'j'}^{(t-1)}\left(\Xi_{r,i'j'}^{(t-1)}\right)^2 \langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle. \qquad \text{(B.3)}
\end{aligned}
$$

- If $i \in \mathcal{C}_k$ with $k \in [K] \setminus \mathcal{K}_0$,

$$
\begin{aligned}
\Xi_{r,ij}^{(t+1)} &= \Xi_{r,ij}^{(t)} - \frac{\eta}{N}\sum_{i'\in[N]} \langle \nabla_{\boldsymbol{w}_r} L(\boldsymbol{W}^{(t)}; \mathcal{D}_{i'}), \boldsymbol{\xi}_{ij} \rangle \\
&= \Xi_{r,ij}^{(t)} + \frac{\eta}{N}\sum_{i'\in[N]} \frac{1}{n}\sum_{j'\in[n]} y_{i'j'} \ell_{i'j'}^{(t)} 3\langle \boldsymbol{w}_r^{(t)}, \mathbf{x}_{i'j'} \rangle^2 \langle \mathbf{x}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle \\
&= \Xi_{r,ij}^{(t)} + \eta\left[\frac{3}{Nn} y_{ij}\ell_{ij}^{(t)} \|\boldsymbol{\xi}_{ij}\|^2\right]\left(\Xi_{r,ij}^{(t)}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{3\eta}{Nn} \sum_{j' \in [n]\setminus\{j\}} y_{ij'} \ell_{ij'}^{(t)} \left(\Xi_{r,ij'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{ij'}, \boldsymbol{\xi}_{ij} \rangle \\
&+ \frac{3\eta}{Nn} \sum_{i' \in \mathcal{C}_k \setminus\{i\}} \sum_{j' \in [n]} y_{i'j'} \ell_{i'j'}^{(t)} \left(\Xi_{r,i'j'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle \\
&+ \frac{3\eta}{Nn} \sum_{k' \in [K]\setminus\mathcal{K}_0\setminus\{k\}} \sum_{i' \in \mathcal{C}_{k'}} \sum_{j' \in [n]} y_{i'j'} \ell_{i'j'}^{(t-1)} \left(\Xi_{r,i'j'}^{(t-1)}\right)^2 \langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle \\
&+ \frac{3\eta}{Nn} \sum_{k' \in \mathcal{K}_0} \sum_{i' \in \mathcal{C}_{k'}} \sum_{j' \in [n]} y_{i'j'} \ell_{i'j'}^{(t)} \left(\Xi_{r,i'j'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij} \rangle. \qquad \text{(B.4)}
\end{aligned}
$$

B.2. Induction hypothesis in GD

Induction hypothesis B.1 (The scale of noise memorization)**.** *Throughout the training process of GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*,*

(a) For any $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$ *and* $j \in [n]$*, we maintain*

$$\max_{r \in [m], j \in [n]} |\Xi_{r,ij}^{(t)}| \le \widetilde{O}(1) \quad \textit{for any } t \ge 0.$$

(b) For any $i \in \mathcal{C}_k$ *with* $k \in [K] \setminus \mathcal{K}_0$ *and* $j \in [n]$*, we maintain*

$$\max_{r \in [m], j \in [n]} |\Xi_{r,ij}^{(t)}| \le \widetilde{O}(d^{-1/2}) \quad \textit{for any } t \ge 0.$$

Induction hypothesis B.2 (The scale of noisy feature's signal)**.** *Throughout the training process of GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*, for any* $k \in \mathcal{K}_0$ *and* $r \in [m]$*, we maintain:*

$$\max_{r \in [m]} |\Gamma_{r,k}^{(t)}| \le \widetilde{O}(\sigma_0) \quad \textit{for any } t \ge 0.$$

Lemma B.1. *Given* $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$ *and* $j \in [n]$*, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$\min_{r \in [m]} y_{ij} \Xi_{r,ij}^{(t)} \ge -\widetilde{O}(d^{-1/2}) \quad \textit{for any} \quad t \ge 0. \qquad \text{(B.5)}$$

B.3. Proof of Theorem 1

The following lemma characterizes the growth of normal feature's signal intensity in the early stage of GD, where the scale of gradients is relatively large.

Lemma B.2. *Denote* $\mathrm{T}_v^0 = \Theta\left(\frac{K}{\eta \alpha^3 \sigma_0}\right) + O(\log d)$*. For any* $t \le \mathrm{T}_v^0$*, the signal intensity of* $\boldsymbol{v}_k^*$ *with* $k \in [K] \setminus \mathcal{K}_0$ *is updated by*

$$\Gamma_{r,k}^{(t+1)} = \Gamma_{r,k}^{(t)} + \Theta\left(\frac{\eta(\alpha^3 - K_0\beta^3)}{K}\right)\left(\Gamma_{r,k}^{(t)}\right)^2. \qquad \text{(B.6)}$$

Proof. For $t \le \mathrm{T}_v^0$, it follows from Lemmas B.6 and B.7 that $\max_{r,i,j} |\Xi_{r,ij}^{(t)}| \le \widetilde{O}(\sigma_0 \sigma_\xi \sqrt{d})$. Hence we have for $t \le \min\{\tau_k, \mathrm{T}_v^0\}$,

$$
\begin{aligned}
\ell_{ij}^{(t)} &= \frac{1}{1 + \exp\left\{\sum_{r=1}^m \left[(\alpha^3 - K_0\beta^3)[\Gamma_{r,k}^{(t)}]^3 + y_{ij}(\Xi_{r,ij}^{(t)})^3\right]\right\}} \\
&\ge \frac{1}{1 + \exp\left\{\sum_{r=1}^m (\alpha^3 - K_0\beta^3)[\Gamma_{r,k}^{(t)}]^3 + \widetilde{O}(m\sigma_0^3\sigma_\xi^3 d^{3/2})\right\}}
\end{aligned}
$$

$$\begin{aligned} &\geq \frac{1}{1+\exp\left\{1/2+\widetilde{O}(m\sigma_0^3\sigma_\xi^3 d^{3/2})\right\}} \\ &= \Theta(1). \end{aligned} \tag{B.7}$$

According to the assumption $|\mathcal{C}_k| = N/K$ for any $k \in [K]$, we can simplify (B.1) as

$$\Gamma_{r,k}^{(t+1)} = \Gamma_{r,k}^{(t)} + \Theta\left(\frac{\eta\alpha^3}{K}\right)\left(\Gamma_{r,k}^{(t)}\right)^2.$$

This completes the proof. □

B.3.1. Proof of Lemma 5.1

Proof. For any $k \in [K] \setminus \mathcal{K}_0$ and $t \leq \tau_k$, by Lemma B.2, we have

$$\begin{aligned} \Gamma_{r,k}^{(t+1)} &= \Gamma_{r,k}^{(t)} + \eta\cdot\Theta\left(\frac{\alpha^3 - K_0\beta^3}{K}\right)\left(\Gamma_{r,k}^{(t)}\right)^2 \\ &= \Gamma_{r,k}^{(t)} + \eta\cdot\Theta\left(\frac{\alpha^3}{K}\right)\left(\Gamma_{r,k}^{(t)}\right)^2. \end{aligned}$$

From Lemma A.1, we know $\Gamma_{r^*,k}^{(0)} = \Omega(\sigma_0)$ holds with probability $1 - 1/\mathsf{poly}(d)$. Applying Lemma E.4 with $h = H = \eta\cdot\Theta\left(\frac{\alpha^3}{K}\right)$, we can guarantee $\Gamma_{r^*,k}^{(t)} \geq \Theta\left(\frac{1}{m^{1/3}\alpha}\right)$ for any $t \geq \tau_k^0$ where

$$\tau_k \leq \frac{3}{h\Gamma_{r^*,k}^{(0)}} + \frac{8H}{h}\left\lceil\frac{\log(v/z^{(0)})}{\log(2)}\right\rceil \leq \frac{K}{\eta\alpha^3\sigma_0} + O(\log d) \leq \mathrm{T}_v^0.$$

The other two conclusions follows from Lemmas B.6 and B.7. □

B.3.2. Proof of Lemma 5.2

Proof. The conclusion follows from Lemma B.9. □

B.3.3. Proof of Lemma 5.3

Proof. According to the definition of τ_{ij}^0, we know $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(s)} \leq \Theta(1/m^{1/3})$ for $s \leq \tau_{ij}^0$. In addition, by Lemma B.6, we also know $\tau_{ij}^0 > \mathrm{T}_v^-$. Given $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ and $j \in [n]$, it holds for any $s \leq \tau_{ij}^0$ that

$$\begin{aligned} \ell_{ij}^{(s)} &= \frac{1}{1+\exp\left\{\sum_{r=1}^m\left[\alpha^3\left(\Gamma_{r,k}^{(s)}\right)^3 - \rho^3\sum_{k'\in\mathcal{K}_0}(\Gamma_{r,k'}^{(s)})^3 + (y_{ij}\Xi_{r,ij}^{(s)})^3\right]\right\}} \\ &\geq \frac{1}{1+\exp\left\{m\Theta(1/m) + \widetilde{O}\left(\alpha^3 m\sigma_0^3\right) + \widetilde{O}\left(\rho^3 mK_0\sigma_0^3\right)\right\}} \\ &\geq \frac{1}{1+\exp\left\{\Theta(1)+o(1)\right\}} \\ &= \Theta(1), \end{aligned} \tag{B.8}$$

where we used Induction hypothesis B.2. Let $\tau_{r,ij}^0$ be the first iteration $y_{ij}\Xi_{r,ij}^{(t)}$ reaches $\Theta(m^{-\frac{1}{3}})$. From Lemma B.9, we know $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(t)} \leq \Theta((md)^{-\frac{1}{3}})$ holds for any $t \leq \mathrm{T}_\xi^- = \Theta\left(\frac{1}{\eta}\frac{Nn}{(\sqrt{d}\sigma_\xi)^3\sigma_0}\right)$.

Let $\tau_{r^*,ij}^-$ be the first iteration $y_{ij}\Xi_{r^*,ij}^{(t)} \geq \Theta((md)^{-\frac{1}{3}})$, then $\tau_{r^*,ij}^- > \mathrm{T}_\xi^-$. After enrolling (B.3) with $r = r_{ij}^*$, for any $\tau_{r^*,ij}^- \leq t \leq \min\{\tau_{ij}^0, \mathrm{T}_\xi^0\}$, we can get

$$y_{ij}\Xi_{r^*,ij}^{(t)} \overset{(i)}{=} y_{ij}\Xi_{r^*,ij}^{(\tau_{r^*,ij}^-)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right)\sum_{s=\tau_{r^*,ij}^-}^{t-1}\left(\Xi_{r^*,ij}^{(s)}\right)^2$$

$$
\begin{aligned}
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right) \sum_{j'\in[n]\backslash\{j\}} \sum_{s=\tau_{r^*,ij}^-}^{t-1} \ell_{ij'}^{(s)} \left(\Xi_{r^*ij'}^{(s)}\right)^2 \\
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right) \sum_{i'\in\mathcal{C}_k\backslash\{i\}} \sum_{j'\in[n]} \sum_{s=\tau_{r^*,ij}^-}^{t-1} \ell_{i'j'}^{(s)} \left(\Xi_{r^*i'j'}^{(s)}\right)^2 \\
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right) \sum_{k'\in\mathcal{K}_0\backslash\{k\}} \sum_{i'\in\mathcal{C}_{k'}} \sum_{j'\in[n]} \sum_{s=\tau_{r^*,ij}^-}^{t-1} \ell_{i'j'}^{(s)} \left(\Xi_{r^*i'j'}^{(s)}\right)^2 \\
&\pm \widetilde{O}\left(\frac{\eta d^{3/2}\sigma_\xi^4\sigma_0^2}{Nn}\right) \sum_{k'\in[K]\backslash\mathcal{K}_0} \sum_{i'\in\mathcal{C}_{k'}} \left(\sum_{s=\tau_{r^*,ij}^-}^{\mathrm{T}_v^0-1} \nu_{i'j'}^{(s)} + \sum_{s=\mathrm{T}_v^0}^{t-1} \nu_{i'j'}^{(s)}\right) \\
\overset{(ii)}{=}\ & y_{ij}\Xi_{r^*ij}^{(\tau_{r^*,ij}^-)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right) \sum_{s=\tau_{r^*,ij}^-}^{t-1} \left(\Xi_{r^*ij}^{(s)}\right)^2 \pm \widetilde{O}\left(\frac{\sqrt{d}\sigma_\xi^2}{d\sigma_\xi^2}\right) \pm \widetilde{O}\left(d^{3/2}\sigma_\xi^4\sigma_0^2\right) \\
\overset{(iii)}{=}\ & y_{ij}\Xi_{r^*ij}^{(\mathrm{T}_\xi^-)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right) \sum_{s=\tau_{r^*,ij}^-}^{t-1} \left(\Xi_{r^*ij}^{(s)}\right)^2 \pm o(m^{-\frac{1}{3}}d^{-1/3}),
\end{aligned}
\tag{B.9}
$$

where (i) holds due to (B.8) and hypothesis (b); (ii) follows from Lemmas B.8 and B.10; and (iii) holds since $\sigma_\xi = \Theta(d^{-0.51})$ and $\sigma_0 = \Theta(d^{-0.52})$. Let $A = \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right)$, $C = o(\sqrt{d}\sigma_\xi\sigma_0)$ and $v = \Theta(m^{-\frac{1}{3}})$, applying the tensor power's method in (E.3) of Lemma E.6 guarantees that,

$$
\begin{aligned}
\tau_{r^*,ij}^0 &\leq \tau_{r^*,ij}^- + \frac{21}{Ay_{ij}\Xi_{r^*,ij}^{(\tau_{r^*,ij}^-)}} + 8\left\lceil \frac{\log(v/[y_{ij}\Xi_{r^*,ij}^{(\tau_{r^*,ij}^-)}])}{\log(2)} \right\rceil \\
&\leq \Theta\left(\frac{1}{\eta}\frac{Nn}{(\sqrt{d}\sigma_\xi)^3\sigma_0}\right) + \Theta\left(\frac{1}{\eta}\frac{Nnm^{1/3}}{d^{2/3}\sigma_\xi^2}\right) + O(\log d) \\
&\leq O\left(\frac{1}{\eta}\frac{Nn}{(\sqrt{d}\sigma_\xi)^3\sigma_0} + \log d\right).
\end{aligned}
\tag{B.10}
$$

By the definition of τ_{ij}^0, we know $\tau_{ij}^0 \leq \tau_{r^*,ij}^0 \leq \mathrm{T}_\xi^0$. □

B.3.4. PROOF OF LEMMA B.1

Proof. According to Lemma A.1, we know $\min_{r\in[m]} y_{ij}\Xi_{r,ij}^{(0)} \geq -\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)$ holds for any $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ and $j \in [n]$. By Lemma B.9, we know $\ell_{ij}^{(s)} \geq \frac{1}{2} - O(d^{-1})$ for any $s \leq \mathrm{T}_\xi^-$. Similar to (B.39), we can obtain that for any $t \leq \mathrm{T}_\xi^-$,

$$
\begin{aligned}
y_{ij}\Xi_{r,ij}^{(t)} &= y_{ij}\Xi_{r,ij}^{(0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right)\sum_{s=0}^{t}\left(\Xi_{r,ij}^{(s)}\right)^2 \pm o(\sqrt{d}\sigma_\xi\sigma_0) \\
&\geq -\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) - o(\sqrt{d}\sigma_\xi\sigma_0) \\
&= -\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0).
\end{aligned}
\tag{B.11}
$$

For $t > \mathrm{T}_\xi^-$, we also have

$$
y_{ij}\Xi_{r,ij}^{(t)} \geq y_{ij}\Xi_{r,ij}^{(\mathrm{T}_\xi^-)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right)\sum_{s=\mathrm{T}_\xi^-}^{t}\left(\Xi_{r,ij}^{(s)}\right)^2 - \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_\xi^-}^{t}\ell_{i'j'}^{(s)}\left(\Xi_{r,i'j'}^{(s)}\right)^2
$$

$$
\begin{aligned}
&- \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{N}\right) \sum_{k'\in[K]\setminus\mathcal{K}_0} \sum_{i'\in\mathcal{C}_{k'}} \sum_{s=\mathrm{T}_v^0}^{t} \nu_{i'}^{(s)} \left(\Xi_{r,i'j'}^{(s)}\right)^2 \\
&\geq -\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) - \widetilde{O}\left(\frac{1}{\sqrt{d}}\right) - \widetilde{O}\left(\frac{\sigma_\xi^2}{\sqrt{d}}\right) \\
&\geq -\widetilde{O}\left(\frac{1}{\sqrt{d}}\right),
\end{aligned}
$$

where the second inequality follows from Lemma B.8, Lemma B.10 and Induction hypothesis B.1 (b). Then we have finished the proof. □

B.3.5. Proof of Theorem 1

The following lemma gives the upper bound of the summation of the averaged derivative over the clients with confounding features till the end of training. Since Lemma 5.3 guarantees the signal intensity of confounding features at iteration T_ξ^- is bounded by $\widetilde{O}(\sigma_0)$, the derivative scale after T_ξ^- is not enough to learn confounding features.

Lemma B.3. *Given any $k \in \mathcal{K}_0$, in the training process of GD, for any $\mathrm{T}_\xi^- < t \leq T$, we maintain:*

$$\frac{\eta}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\mathrm{T}_\xi^-}^{t} \nu_i^{(s)} \leq \widetilde{O}\left(\frac{1}{d^{2/3}\sigma_\xi^2}\right).$$

Lemma B.4. *Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2\mathbf{H})$ be a noise vector that is independent of the training data. During the training process of GD, we maintain*

$$\max_{r\in[m]} |\langle \boldsymbol{w}_r^{(t)}, \boldsymbol{\xi}\rangle| \leq \widetilde{O}\left(d^{-\frac{1}{2}}\right) \quad \textit{for any } t \geq 0. \tag{B.12}$$

Proof. Denote $\Xi_r^{(t)} = \langle \boldsymbol{w}_r^{(t)}, \boldsymbol{\xi}\rangle$. Notice that

$$
\begin{aligned}
\Xi_r^{(t+1)} &= \Xi_r^{(t)} + \frac{\eta}{N}\sum_{i\in[N]}\frac{1}{n}\sum_{j\in[n]} y_{ij}\ell_{ij}^{(t)} \sum_{p\in[P]} 3\langle \boldsymbol{w}_r^{(t)}, \mathbf{x}_{ij,p}\rangle^2 \langle \mathbf{x}_{ij,p}, \boldsymbol{\xi}\rangle \\
&= \Xi_r^{(t)} + \frac{3\eta}{Nn}\sum_{i\in[N]}\sum_{j\in[n]} y_{ij}\ell_{ij}^{(t)} \left(\Xi_{r,ij}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{ij}, \boldsymbol{\xi}\rangle \\
&= \Xi_r^{(0)} \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{i\in[N]}\sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 \\
&= \Xi_r^{(0)} \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 \\
&\quad \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k\in[K]\setminus\mathcal{K}_0}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2.
\end{aligned}
\tag{B.13}
$$

We first consider $k \in [K] \setminus \mathcal{K}_0$. If $t \leq \mathrm{T}_v^0$, by Lemma B.6, we have

$$
\begin{aligned}
\frac{\eta}{Nn}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 &\leq \frac{\eta}{Nn}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=0}^{\mathrm{T}_v^0} \left(\Xi_{r,ij}^{(s)}\right)^2 \\
&\leq \widetilde{O}\left(\eta\mathrm{T}_v^0 d\sigma_\xi^2\sigma_0^2\right) = \widetilde{O}(d\sigma_\xi^2\sigma_0).
\end{aligned}
\tag{B.14}
$$

If $t > \mathrm{T}_v^0$, Lemma B.8 guarantees that

$$\frac{\eta}{Nn}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 \leq \widetilde{O}\left(d\sigma_\xi^2\sigma_0^2\right)\cdot\frac{\eta}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\mathrm{T}_v^0}^{t} \nu_i^{(s)} \leq \widetilde{O}\left(d\sigma_\xi^2\sigma_0^2\right). \tag{B.15}$$

For any $k \in \mathcal{K}_0$, if $t \le \mathrm{T}_\xi^-$, Lemma B.9 implies that

$$\begin{aligned}\frac{\eta}{Nn}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=0}^{t}\ell_{ij}^{(s)}\left(\Xi_{r,ij}^{(s)}\right)^2 &\le \frac{\eta}{Nn}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=0}^{\mathrm{T}_\xi^-}\left(\Xi_{r,ij}^{(s)}\right)^2\\ &\le \widetilde{O}\left(\eta\mathrm{T}_\xi^-\cdot d^{-\frac{2}{3}}\right)=\widetilde{O}\left(\frac{d^{-\frac{2}{3}}}{(\sqrt{d}\sigma_\xi)^3\sigma_0}\right).\end{aligned} \tag{B.16}$$

If $t > \mathrm{T}_\xi^-$, Lemma B.3 guarantees that

$$\frac{\eta}{Nn}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=\mathrm{T}_\xi}^{t}\ell_{ij}^{(s)}\left(\Xi_{r,ij}^{(s)}\right)^2 \le \widetilde{O}\left(\frac{1}{d\sigma_\xi^2}\right). \tag{B.17}$$

Now plugging (B.14)-(B.17) to (B.13), we can obtain

$$|\Xi_r^{(t+1)}| \le |\Xi_r^{(0)}| + \widetilde{O}\left(d^{3/2}\sigma_\xi^4\sigma_0\right)+\widetilde{O}\left(\frac{d^{-\frac{2}{3}}}{d\sigma_\xi\sigma_0}\right)+\widetilde{O}\left(d^{-\frac{1}{2}}\right)=\widetilde{O}\left(d^{-\frac{1}{2}}\right). \tag{B.18}$$

□

Proof of Theorem 1. For any $k \in [K]\setminus\mathcal{K}_0$, according to Lemma 5.1, we know $\max_{r\in[m]}\Gamma_{r,k}^{(T)} \ge \Theta(m^{-\frac{1}{3}})$. By Lemma B.5, it holds that $\min_{r\in[m]}\Gamma_{r,k}^{(T)} \ge -\widetilde{O}(\sigma_0)$. In addition, Induction hypothesis B.1 (b) also guarantees that $\max_{r\in[m]}|\Xi_{r,ij}^{(T)}| \le \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)$ if $i \in \mathcal{C}_k$ with $k \in [K]\setminus\mathcal{K}_0$. For each local data $(\mathbf{x}_{ij}, y_{ij})$ with $i \in \mathcal{C}_k$ with $k \in [K]\setminus\mathcal{K}_0$ and $j \in [n]$, we have

$$\begin{aligned} y_{ij}F(\boldsymbol{W}^{(T)},\mathbf{x}_{ij}) &= y_{ij}\sum_{r\in[m]}\sum_{p\in[P]}\langle \boldsymbol{w}_r^{(T)},\mathbf{x}_{ij,p}\rangle^3\\ &= \sum_{r\in[m]}\left[(\alpha^3-K_0\rho^3)\left(\Gamma_{r,k}^{(T)}\right)^3+\left(y_{ij}\Xi_{r,ij}^{(T)}\right)^3\right]\\ &\ge (\alpha^3-K_0\rho^3)\cdot\Theta(m^{-\frac{1}{3}})-(\alpha^3-K_0\rho^3)\cdot\widetilde{O}(m\sigma_0^3)-\widetilde{O}(md^{3/2}\sigma_\xi^3\sigma_0^3)\\ &= \widetilde{\Omega}(1)-\widetilde{O}(d^{-1.47})-\widetilde{O}(d^{-1.5})\\ &= \widetilde{\Omega}(1), \end{aligned} \tag{B.19}$$

where we used the setting $m = \mathsf{polylog}(d)$, $\sigma_0 = d^{-0.49}$ and $\sigma_\xi = d^{-0.51}$. For the new test data $(\mathbf{x}, y) \sim \mathcal{D}_i$, with the probability at least $1 - 1/\mathsf{poly}(d)$, we have

$$\begin{aligned} yF(\boldsymbol{W}^{(T)},\mathbf{x}) &= y\sum_{r\in[m]}\sum_{p\in[P]}\langle \boldsymbol{w}_r^{(T)},\mathbf{x}_p\rangle^3\\ &= \sum_{r\in[m]}\left[(\alpha^3-K_0\rho^3)\left(\Gamma_{r,k}^{(T)}\right)^3+y\langle\boldsymbol{w}_r^{(T)},\boldsymbol{\xi}\rangle^3\right]\\ &\ge \widetilde{\Omega}(1)-\widetilde{O}(md^{3/2}\sigma_\xi^3\sigma_0^3)-\widetilde{O}\left(md^{-3/2}\right)\\ &\ge \widetilde{\Omega}(1), \end{aligned} \tag{B.20}$$

where we used Lemma B.4.

For any $k \in \mathcal{K}_0$, we know $\max_{r\in[m]}|\Gamma_{r,k}^{(T)}| \le \widetilde{O}(\sigma_0)$ by Induction hypothesis B.2. If $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, Lemma 5.3 ensures that $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(T)} \ge \Theta(m^{-\frac{1}{3}})$ for any $j \in [n]$. For each local data $(\mathbf{x}_{ij}, y_{ij})$ with $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ and $j \in [n]$, we have

$$y_{ij}F(\boldsymbol{W}^{(T)},\mathbf{x}_{ij}) = y_{ij}\sum_{r\in[m]}\sum_{p\in[P]}\langle \boldsymbol{w}_r^{(T)},\mathbf{x}_{ij,p}\rangle^3$$

$$\begin{aligned}
&= \sum_{r\in[m]} \left[\alpha^3 \left(\Gamma_{r,k}^{(T)}\right)^3 - \rho^3 \sum_{k'\in\mathcal{K}_0} \left(\Gamma_{r,k'}^{(T)}\right)^3 + \left(y_{ij}\Xi_{r,ij}^{(T)}\right)^3\right] \\
&\geq -\alpha^3 \cdot \widetilde{O}(m\sigma_0^3) - \rho^3 \cdot \widetilde{O}(mK_0\sigma_0^3) + \Theta(m^{-\frac{1}{3}}) \\
&= \widetilde{\Omega}(1). \qquad\qquad (B.21)
\end{aligned}$$

For the new test data $(\mathbf{x}, y) \sim \mathcal{D}_i$, with probability at least $1 - 1/\mathsf{poly}(d)$, we have

$$\begin{aligned}
yF(\boldsymbol{W}^{(T)}, \mathbf{x}) &= y \sum_{r\in[m]} \sum_{p\in[P]} \langle \boldsymbol{w}_r^{(T)}, \mathbf{x}_p\rangle^3 \\
&= \sum_{r\in[m]} \left[\alpha^3 \left(\Gamma_{r,k}^{(T)}\right)^3 - \rho^3 \sum_{k'\in\mathcal{K}_0} \left(\Gamma_{r,k'}^{(T)}\right)^3 + y\langle \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}\rangle^3\right] \\
&\leq \sum_{r\in[m]} y\langle \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}\rangle^3 + \alpha^3 \cdot \widetilde{O}(m\sigma_0^3) + \rho^3 \cdot \widetilde{O}(mK_0\sigma_0^3) \\
&= \sum_{r\in[m]} y\langle \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}\rangle^3 + \widetilde{O}(\sigma_0^3). \qquad\qquad (B.22)
\end{aligned}$$

Denote $\mathbf{P}_v = \sum_{k\in[K]} \boldsymbol{v}_k^*(\boldsymbol{v}_k^*)^\top$ and $\mathbf{P}_v^\perp = \mathbf{I}_d - \mathbf{P}_v$. Since $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{P}_v^\perp)$, there exists some $\boldsymbol{\xi}_d \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_d)$ such that $\boldsymbol{\xi} = \mathbf{P}_v^\perp \boldsymbol{\xi}_d$. Now we write $\boldsymbol{w}_r^{(T)} = \mathbf{P}_v \boldsymbol{w}_r^{(T)} + \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}$. Recalling the definition $\Xi_{r,ij}^{(T)} = \langle \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}_{ij}\rangle = \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}_{ij}\rangle$. For each local data $(\mathbf{x}_{ij}, y_{ij})$ with $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ and $j \in [n]$, we know

$$\Theta(m^{-\frac{1}{3}}) \leq \max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(T)} = \max_{r\in[m]} \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, y_{ij}\boldsymbol{\xi}_{ij}\rangle.$$

Let $r^* = \arg\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(T)}$. Due to the fact $\sqrt{d}\sigma_\xi = \Theta(d^{-0.01})$, using Induction hypothesis B.1, we have

$$\begin{aligned}
\sum_{r\in[m]} \left\langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, \frac{y_{ij}\boldsymbol{\xi}_{ij}}{\|\boldsymbol{\xi}_{ij}\|}\right\rangle^3 &\geq \frac{1}{\|\boldsymbol{\xi}_{ij}\|^3}\left[\left(y_{ij}\Xi_{r^*,ij}^{(T)}\right)^3 - \sum_{r\neq r^*}\left(y_{ij}\Xi_{r,ij}^{(T)}\right)^3\right] \\
&\geq \widetilde{\Omega}\left(\frac{1}{d^{3/2}\sigma_\xi^3}\right)\left[\Theta(m^{-1}) - \widetilde{O}\left(m(\sqrt{d}\sigma_\xi\sigma_0)^3\right)\right] \\
&= \widetilde{\Omega}\left(\frac{1}{d^{3/2}\sigma_\xi^3}\right) \geq 1. \qquad\qquad (B.23)
\end{aligned}$$

Since the model $\boldsymbol{W}^{(T)}$ and the test label y are independent of the noise $\boldsymbol{\xi}_d$, then we know the distribution of $\sum_{r\in[m]} y\langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}_d\rangle^3$ is symmetric given $\boldsymbol{W}^{(T)}$ and $y\boldsymbol{\xi}_d \sim \mathcal{N}(0, \sigma_\xi^2\mathbf{I}_d)$ given $y \in \{-1, +1\}$. Now applying Lemma E.3 with $\boldsymbol{w}_r = \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}$ and $\boldsymbol{u} = y_{ij}\boldsymbol{\xi}_{ij}/\|\boldsymbol{\xi}_{ij}\|$, we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\xi}_d}\left(\sum_{r\in[m]} \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, y\boldsymbol{\xi}_d\rangle^3 < -\epsilon\sigma_\xi^3\right) &\geq \frac{1}{2} - \mathbb{P}_{\boldsymbol{\xi}_d}\left(\left|\sum_{r=1}^m \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, y\boldsymbol{\xi}_d\rangle^3\right| \leq \epsilon\sigma_\xi^3\left|\sum_{r=1}^m \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, \boldsymbol{u}\rangle^3\right|\right) \\
&\geq \frac{1}{2} - O\left(\epsilon^{1/3}\right).
\end{aligned}$$

Taking $\epsilon = 1/\mathsf{polylog}(d)$, together with (B.23), we can guarantee that

$$\mathbb{P}\left(\sum_{r\in[m]} \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, y\boldsymbol{\xi}_d\rangle^3 < -\widetilde{\Omega}(\sigma_\xi^3)\right) \geq \frac{1}{2} - \frac{1}{\mathsf{polylog}(d)}. \qquad (B.24)$$

Since $\sum_{r\in[m]} y\langle \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}\rangle^3 = \sum_{r\in[m]} \langle \mathbf{P}_v^\perp \boldsymbol{w}_r^{(T)}, y\boldsymbol{\xi}_d\rangle^3$, under the same event in (B.24), we have

$$yF(\boldsymbol{W}^{(T)}, \mathbf{x}) \leq \sum_{r\in[m]} y\langle \boldsymbol{w}_r^{(T)}, \boldsymbol{\xi}\rangle^3 + \widetilde{O}(\sigma_0^3)$$

$$\begin{aligned}&\le -\widetilde{\Omega}(\sigma_\xi^3) + \widetilde{O}(\sigma_0^3)\\&< 0,\end{aligned}$$

where we used the setting $\sigma_0 = \Theta(d^{-0.52})$ and $\sigma_\xi = \Theta(d^{-0.51})$. □

B.4. Proof of Induction hypotheses in GD

Lemma B.5. *Given $k \in [K] \setminus \mathcal{K}_0$. In the training process of GD, with probability at least $1 - 1/\mathsf{poly}(d)$, we maintain:*

$$\Gamma_{r,k}^{(t)} \ge -\widetilde{O}(\sigma_0) \quad \textit{for any } t \ge 0. \tag{B.25}$$

Proof. From Lemma A.1, with probability $1 - 1/\mathsf{poly}(d)$, it holds that $\Gamma_{r,k}^{(0)} \ge -\widetilde{O}(\sigma_0)$ holds for any $r \in [m]$. Observing the update rule in (B.2) for $k \in [K] \setminus \mathcal{K}_0$, we have

$$\begin{aligned}\Gamma_{r,k}^{(t+1)} &= \Gamma_{r,k}^{(t)} + \frac{3\eta(\alpha^3 - \beta^3)}{N} \sum_{i \in \mathcal{C}_k} \nu_i^{(t)} \cdot [\Gamma_{r,k}^{(t)}]^2\\&\ge \Gamma_{r,k}^{(t)} \ge \cdots \ge \Gamma_{r,k}^{(0)} \ge -\widetilde{O}(\sigma_0).\end{aligned}$$

This proves the lower bound in (B.25). □

B.4.1. Proof of Induction hypothesis B.1

Let $\mathrm{T}_v^- = \Theta\left(\frac{K}{\eta\sigma_0\alpha^3}\right)$. In this subsection, we will prove Induction hypothesis B.1 in two stages: $[0, \mathrm{T}_v^-]$ and $[\mathrm{T}_v^-, T]$. The first stage is proved by Lemma B.6. The second stage is proved by Lemma B.11.

Lemma B.6. *In the training process of GD, with probability at least $1 - 1/\mathsf{poly}(d)$, we maintain:*

$$\max_{r \in [m], i \in [N], j \in [n]} |\Xi_{r,ij}^{(t)}| \le \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) \quad \textit{for any} \quad t \le \mathrm{T}_v^-. \tag{B.26}$$

Proof. Due to the initialization and the concentration, with probability at least $1 - 1/\mathsf{poly}(d)$, we can guarantee that

$$\Xi_r^{(0)} = \max_{i,j} |\Xi_{r,ij}^{(0)}| = \max_{i,j} |\langle \boldsymbol{w}_r^{(0)}, \boldsymbol{\xi}_{ij}\rangle| \le \widetilde{O}(\sqrt{d}\sigma_0\sigma_\xi).$$

Now we assume $\Xi_r^{(s)} \le \widetilde{O}(\sqrt{d}\sigma_0\sigma_\xi)$ holds for any $s \le t$, and verify the upper bound at iteration $t+1$. Denote $\Xi_r^{(t)} = \max_{i,j} |\Xi_{r,ij}^{(t)}|$. From the update rules in (B.3) and (B.4), we have

$$\begin{aligned}\Xi_r^{(t+1)} &\le \max_{i,j}\left\{|\Xi_{r,ij}^{(t)}| + \frac{3\eta}{Nn}\ell_{ij}^{(t)}\|\boldsymbol{\xi}_{ij}\|^2\left(\Xi_{r,ij}^{(t)}\right)^2\right\}\\&\quad + \frac{3\eta}{Nn}\max_{i,j,i',j'}\left\{\sum_{(i',j') \ne (i,j)} \ell_{i'j'}^{(t)}\left(\Xi_{r,i'j'}^{(t)}\right)^2 |\langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij}\rangle|\right\}\\&\overset{(i)}{\le} \Xi_r^{(t)} + \eta|\Xi_r^{(t)}|^2 \cdot \widetilde{O}\left(\frac{d\sigma_\xi^2}{Nn}\right) + \eta|\Xi_r^{(t)}|^2 \cdot \widetilde{O}\left(P\sqrt{d}\sigma_\xi^2\right)\\&\overset{(ii)}{\le} \Xi_r^{(t)} + \eta \cdot \widetilde{O}\left(\frac{d^2\sigma_\xi^4\sigma_0^2}{Nn} + Pd^{3/2}\sigma_\xi^4\sigma_0^2\right)\\&\le \Xi_r^{(0)} + \eta(t+1) \cdot \widetilde{O}\left(\frac{d^2\sigma_\xi^4\sigma_0^2}{Nn} + Pd^{3/2}\sigma_\xi^4\sigma_0^2\right)\\&\overset{(iii)}{\le} \widetilde{O}(\sqrt{d}\sigma_0\sigma_\xi) + \widetilde{O}\left(\frac{Kd^2\sigma_\xi^4\sigma_0}{Nn\alpha^3} + \frac{PKd\sigma_\xi^4\sigma_0}{\alpha^3}\right)\end{aligned}$$

$$
\begin{aligned}
&\overset{(iv)}{=} \widetilde{O}(\sqrt{d}\sigma_0\sigma_\xi) \cdot \Big[1 + \widetilde{O}(d^{-0.03}) + \widetilde{O}\left(d^{-1.53}\right)\Big] \\
&\leq \widetilde{O}(\sqrt{d}\sigma_0\sigma_\xi),
\end{aligned}
$$

where (i) holds due to the concentration $\|\boldsymbol{\xi}_{ij}\| \leq \widetilde{O}(\sqrt{d}\sigma_\xi)$ and $|\langle \boldsymbol{\xi}_{i'j'}, \boldsymbol{\xi}_{ij}\rangle| \leq \widetilde{O}(\sqrt{d}\sigma_\xi^2)$; (ii) follows from the induction hypothesis; (iii) holds due to $t+1 \leq \mathrm{T}_v^-$; and (iv) is true because the setting $\sigma_\xi = \Theta(d^{-0.51})$ and $\alpha = \Theta(1)$. □

Lemma B.7. *In the training process of GD, given* $k \in \mathcal{K}_0$*, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$
\max_{r\in[m]} |\Gamma^{(t)}_{r,k}| \leq \widetilde{O}(\sigma_0) \quad \textit{for any} \quad t \leq \mathrm{T}_v^-.
$$

Proof. Now we suppose $\max_{r\in[m],k\in\mathcal{K}_0} |\Gamma^{(s)}_{r,k}| \leq \widetilde{O}(\sigma_0)$ holds for iterations $s \leq t \leq \mathrm{T}_v^- - 1$. By Lemma B.6, we know $\max_{r,i,j} |\Xi^{(s)}_{r,ij}| \leq \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)$ for any $s \leq t$. It means that for any $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, $j \in [n]$ and $s \leq t$,

$$
\begin{aligned}
\ell^{(s)}_{ij} &= \frac{1}{1 + \exp\Big\{\sum_{r=1}^m \Big[\alpha^3 (\Gamma^{(s)}_{r,k})^3 - \rho^3 \sum_{k'\in\mathcal{K}_0} (\Gamma^{(s)}_{r,k'})^3 + y_{ij}(\Xi^{(s)}_{r,ij})^3\Big]\Big\}} \\
&= \frac{1}{1 + \exp\Big\{\pm\widetilde{O}\left(\alpha^3 m\sigma_0^3\right) \pm \widetilde{O}\left(\rho^3 m K_0 \sigma_0^3\right) \pm \widetilde{O}\Big(m(\sqrt{d}\sigma_\xi)^3\sigma_0^3\Big)\Big\}} \\
&= \frac{1}{1 + \exp\Big\{\pm\widetilde{O}(\sigma_0^3)\Big\}} \\
&= \frac{1}{2} + \frac{\exp\Big\{\pm\widetilde{O}(\sigma_0^3)\Big\} - 1}{2(1 + \exp\{\pm\widetilde{O}(\sigma_0^3)\})} \\
&= \frac{1}{2} \pm \widetilde{O}(\sigma_0^3). \qquad \text{(B.27)}
\end{aligned}
$$

Using the recursion (B.1), for any $t+1 \leq \mathrm{T}_v^-$ we have

$$
\begin{aligned}
\Gamma^{(t+1)}_{r,k} &\leq \Gamma^{(0)}_{r,k} + \left[\frac{3\eta\alpha^3}{K} - \frac{3\eta K_0 \rho^3}{K}\right] \cdot \sum_{s=0}^{t} \left(\Gamma^{(s)}_{r,k}\right)^2 + \widetilde{O}(\sigma_0^3) \cdot \frac{3\eta\alpha^3}{2N} \sum_{i\in\mathcal{C}_k} \sum_{s=0}^{t} \left(\Gamma^{(s)}_{r,k}\right)^2 \\
&\leq \widetilde{O}(\sigma_0) + \left[\frac{3\eta\alpha^3}{K} - \frac{3\eta K_0 \rho^3}{K}\right] \cdot \widetilde{O}\left(\mathrm{T}_v^- \sigma_0^2\right) + \widetilde{O}(\sigma_0^3) \cdot \frac{3\eta\alpha^3}{K} \cdot \widetilde{O}\left(\mathrm{T}_v^- \sigma_0^2\right) \\
&\leq \widetilde{O}(\sigma_0) + \frac{\alpha^3 - K_0\rho^3}{K} \cdot \widetilde{O}\left(\frac{K\sigma_0}{\alpha^3}\right) + \widetilde{O}\left(\sigma_0^4\right) \\
&= \widetilde{O}(\sigma_0) + \widetilde{O}\left(\frac{K\sigma_0}{d\alpha^3}\right) + \widetilde{O}\left(\sigma_0^4\right) \\
&= \widetilde{O}(\sigma_0), \qquad \text{(B.28)}
\end{aligned}
$$

where we used the inductive hypothesis and the setting $\rho = \left[\frac{\alpha^3 - \Theta(K/d)}{K_0}\right]^{1/3}$. For the lower bound at time $t+1$, we have

$$
\begin{aligned}
\Gamma^{t+1}_{r,k} &\geq \Gamma^{(0)}_{r,k} + \left[\frac{3\eta\alpha^3}{K} - \frac{3\eta K_0 \rho^3}{K}\right] \cdot \sum_{s=0}^{t} \left(\Gamma^{(s)}_{r,k}\right)^2 - \widetilde{O}(\sigma_0^3) \cdot \frac{3\eta\rho^3}{2N} \sum_{k'\in\mathcal{K}_0} \sum_{i\in\mathcal{C}_{k'}} \sum_{s=0}^{t} \left(\Gamma^{(s)}_{r,k}\right)^2 \\
&\geq -\widetilde{O}(\sigma_0) - \widetilde{O}(\sigma_0^3) \cdot \frac{3K_0\eta\rho^3}{K} \cdot \widetilde{O}\left(\mathrm{T}_v^- \sigma_0^2\right) \\
&\geq -\widetilde{O}(\sigma_0). \qquad \text{(B.29)}
\end{aligned}
$$

Consequently, we can conclude that $\max_{r\in[m],k\in\mathcal{K}_0} |\Gamma^{(t)}_{r,k}| \leq \widetilde{O}(\sigma_0)$ holds for any $t \leq \mathrm{T}_v^-$. □

Lemma B.8. *Given any $k \in [K] \setminus \mathcal{K}_0$. If Induction hypothesis B.1 holds for any iterations $s \le t$, with probability at least $1 - 1/\mathsf{poly}(d)$, we have*

$$\frac{\eta}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\tau_k}^{t}\nu_i^{(s)} \le \widetilde{O}\left(1\right) \quad \textit{for any } t \ge \mathrm{T}_v^0. \tag{B.30}$$

Proof. Let $\tau^0_{r^*,k}$ be the first iteration $\Gamma^{(t)}_{r^*,k}$ reaches $\Theta\left(\frac{1}{m^{1/3}\alpha}\right)$ for $k \in [K]\setminus\mathcal{K}_0$. It follows from Lemma 5.1 that $\tau^0_{r^*,k} \le \mathrm{T}_v^-$ for any $k \in [K] \setminus \mathcal{K}_0$. Let $\tau^+_{r^*,k}$ be the first iteration $\Gamma^{(t)}_{r^*,k}$ reaches $\Theta\left(\vartheta^{1/3}/\alpha\right)$, where ϑ is defined in (A.3). Applying Lemma E.8 with $z^{(0)} = \Gamma^{(\tau^0_{r^*,k})}_{r^*,k}$, $h = H = 3\eta\alpha^3$ and $a^{(t)} = \frac{1}{N}\sum_{i\in\mathcal{C}_k}\nu_i^{(t)} \le \frac{1}{K}$, we have

$$\frac{\eta}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\mathrm{T}_v^0}^{\tau^+_{r^*,k}}\nu_i^{(s)} \le \frac{\Theta\left(m^{1/3}\alpha\right)}{\alpha^3} + O\left(\frac{\eta\log[\vartheta^{1/3}/\alpha]}{K}\right) \le \widetilde{O}(1). \tag{B.31}$$

Notice that $\Gamma^{(s)}_{r^*,k} \ge \Theta\left(\frac{\vartheta^{1/3}}{m^{1/3}\alpha}\right)$ for any $s \ge \tau^+_{r^*,k}$ since $\Gamma^{(t)}_{r,k}$ is increasing by observing (B.2). It means for any $i \in \mathcal{C}_k$ with $k \in [K] \setminus \mathcal{K}_0$ and $j \in [n]$,

$$\begin{aligned}
\ell_{ij}^{(s)} &= \frac{1}{1+\exp\left\{\sum_{r=1}^m\left[\left(\alpha^3 - K_0\beta^3\right)\left(\Gamma^{(s)}_{r,k}\right)^3 + [y_{ij}\Xi^{(s)}_{r,ij}]^3\right]\right\}}\\
&= \frac{1}{1+\exp\left\{\left(\alpha^3 - K_0\beta^3\right)\left[\left(\Gamma^{(s)}_{r^*,k}\right)^3 + \sum_{r\neq r^*}\left(\Gamma^{(s)}_{r,k}\right)^3\right] - \sum_{r=1}^m|\Xi^{(s)}_{r,ij}|^3\right\}}\\
&\le \frac{1}{1+\exp\left\{\left(\alpha^3 - K_0\beta^3\right)\left(\Gamma^{(s)}_{r^*,k}\right)^3 - \widetilde{O}(m\sigma_0^3) - \widetilde{O}(md^{3/2}\sigma_\xi^3\sigma_0^3)\right\}}\\
&\le \frac{1}{1+\exp\left\{\alpha^3/2(\Gamma^{(s)}_{r^*,k})^3 - o(1)\right\}}\\
&\le \frac{\Theta(1)}{1+e^{\vartheta}},
\end{aligned} \tag{B.32}$$

where we also used Induction hypothesis B.1 and the lower bound of Lemma B.5 in the first inequality. By the definition of ϑ in (A.3), for any $i \in \mathcal{C}_k$ with $k \in [K] \setminus \mathcal{K}_0$ we have

$$\frac{\eta}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\mathrm{T}_v^0}^{t}\nu_i^{(s)} \le \Theta\left(\frac{\eta}{K}\right)\cdot\sum_{s=0}^{T}\frac{1}{1+e^{\vartheta}} \le \widetilde{O}\left(\frac{\eta}{K}\right) = \widetilde{O}(1). \tag{B.33}$$

Combining (B.31) and (B.33), we can finish the proof. □

Lemma B.9. *Let $\mathrm{T}_\xi^- = \Theta\left(\frac{Nn}{\eta(\sqrt{d}\sigma_\xi)^3\sigma_0}\right)$. In the training process of GD, with probability at least $1 - 1/\mathsf{poly}(d)$,*

- *For any $k \in \mathcal{K}_0$, we maintain:*

$$\max_{r\in[m],i\in\mathcal{C}_k,j\in[n]} y_{ij}\Xi^{(t)}_{r,ij} \le (md)^{-1/3} \quad \textit{for any } t \le \mathrm{T}_\xi^-. \tag{B.34}$$

- *For any $k \in \mathcal{K}_0$, we maintain:*

$$\max_{r\in[m]}|\Gamma^{(t)}_{r,k}| \le \widetilde{O}(\sigma_0) \quad \textit{for any } t \le \mathrm{T}_\xi^-. \tag{B.35}$$

- *For any $k \in [K] \setminus \mathcal{K}_0$, we maintain:*

$$\max_{r\in[m], i\in\mathcal{C}_k, j\in[n]} |\Xi_{r,ij}^{(t)}| \le \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) \quad \textit{for any } t \le \mathrm{T}_\xi^- . \tag{B.36}$$

Proof of Lemma B.9. Given any $k \in \mathcal{K}_0$, let $\mathrm{T}_{\xi_{ij}}^- = \Theta\left(\frac{Nn}{\eta M_{ij}}\right)$ where $M_{ij} = \left(y_{ij}\Xi_{r^*,ij}^{(0)}\|\boldsymbol{\xi}_{ij}\|^2\right)^{-1}$ for $i \in \mathcal{C}_k$ and $j \in [n]$. We first assume (B.34), (B.35) and (B.36) hold for iterations $s \le t < \mathrm{T}_\xi^- - 1$. Then we know for any $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, $j \in [n]$ and $s \le t$,

$$\begin{aligned}
\ell_{ij}^{(s)} &= \frac{1}{1+\exp\left\{\sum_{r=1}^m \left[\alpha^3\left(\Gamma_{r,k}^{(s)}\right)^3 - \rho^3\sum_{k'\in\mathcal{K}_0}(\Gamma_{r,k'}^{(s)})^3 + (y_{ij}\Xi_{r,ij}^{(s)})^3\right]\right\}} \\
&\ge \frac{1}{1+\exp\left\{d^{-1} + \widetilde{O}\left(\alpha^3 m\sigma_0^3\right) + \widetilde{O}\left(\rho^3 m K_0\sigma_0^3\right)\right\}} \\
&\ge \frac{1}{1+\exp\left\{2d^{-1}\right\}} \\
&= \frac{1}{2} - \frac{e^{2d^{-1}}-1}{2(1+e^{2d^{-1}})} \\
&= \frac{1}{2} - O(d^{-1}).
\end{aligned} \tag{B.37}$$

From the recursion (B.1), we have

$$\begin{aligned}
\Gamma_{r,k}^{(t+1)} &= \Gamma_{r,k}^{(0)} + \left[\frac{3\eta\alpha^3}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=0}^t \nu_i^{(s)}\left(\Gamma_{r,k}^{(s)}\right)^2 - \frac{3\eta\rho^3}{N}\sum_{k'\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_{k'}}\sum_{s=0}^t \nu_i^{(s)}\left(\Gamma_{r,k}^{(s)}\right)^2\right] \\
&\overset{(i)}{\le} \Gamma_{r,k}^{(0)} + \left[\frac{3\eta\alpha^3}{2K} - \frac{3K_0\eta\rho^3}{2K}\right]\cdot\sum_{s=0}^t\left(\Gamma_{r,k}^{(s)}\right)^2 + \widetilde{O}\left(\frac{\eta}{Kd}\right)\sum_{s=0}^t\left(\Gamma_{r,k}^{(s)}\right)^2 \\
&\overset{(ii)}{=} \Gamma_{r,k}^{(0)} + \Theta\left(\frac{\eta}{d}\right)\cdot\sum_{s=0}^t\left(\Gamma_{r,k}^{(s)}\right)^2 + \widetilde{O}\left(\frac{\eta}{Kd}\right)\cdot\sum_{s=0}^t\left(\Gamma_{r,k}^{(s)}\right)^2 \\
&\le \widetilde{O}(\sigma_0) + \left[\Theta\left(\frac{\eta}{d}\right) + \widetilde{O}\left(\frac{\eta\alpha^3}{Kd}\right)\right]\cdot\widetilde{O}\left(\mathrm{T}_\xi^-\sigma_0^2\right) \\
&= \widetilde{O}(\sigma_0) + \widetilde{O}\left(\frac{\sigma_0}{d}\frac{Nn}{d^{3/2}\sigma_\xi^3}\right) \\
&= \widetilde{O}(\sigma_0)\cdot\left[1+\widetilde{O}\left(d^{-0.97}\right)\right] \\
&\le \widetilde{O}(\sigma_0),
\end{aligned} \tag{B.38}$$

where (i) holds due to (B.37) and $\ell_{ij}^{(s)} \le 1$; and (ii) follows from the setting $\rho^3 = \frac{\alpha^3 - 1/\mathsf{poly}(d)}{K_0}$. Then we have verified (B.35) at time $t+1$.

Recall that $y_{ij}\Xi_{r^*,ij}^{(0)} = \max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(0)} > \Omega(\sqrt{d}\sigma_\xi\sigma_0)$. Using the recursion (B.3), we have

$$\begin{aligned}
y_{ij}\Xi_{r^*,ij}^{(t+1)} &\overset{(i)}{=} y_{ij}\Xi_{r^*,ij}^{(0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right)\sum_{s=0}^t\left(\Xi_{r^*,ij}^{(s)}\right)^2 \\
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{j'\in[n]\setminus\{j\}}\sum_{s=0}^t \ell_{ij'}^{(s)}\left(\Xi_{r^*,ij'}^{(s)}\right)^2 \\
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{i'\in\mathcal{C}_k\setminus\{i\}}\sum_{j'\in[n]}\sum_{s=0}^t \ell_{i'j'}^{(s)}\left(\Xi_{r^*,i'j'}^{(s)}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right) \sum_{k' \in \mathcal{K}_0 \setminus \{k\}} \sum_{i' \in \mathcal{C}_{k'}} \sum_{j' \in [n]} \sum_{s=0}^{t} \ell_{i'j'}^{(s)} \left(\Xi_{r^*,i'j'}^{(s)}\right)^2 \\
&\pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right) \sum_{k' \in [K] \setminus \mathcal{K}_0} \sum_{i' \in \mathcal{C}_{k'}} \sum_{s=0}^{t} \nu_{i'}^{(s)} \left(\Xi_{r^*,i'j'}^{(s)}\right)^2 \\
\overset{(ii)}{=}\ & y_{ij}\Xi_{r^*,ij}^{(0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right) \sum_{s=0}^{t} \left(\Xi_{r^*,ij}^{(s)}\right)^2 \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn} \cdot \frac{NK_0 n}{K} \frac{1}{(md)^{\frac{2}{3}}} \cdot \mathrm{T}_\xi^-\right) \pm \widetilde{O}\left(d^{3/2}\sigma_\xi^4\sigma_0^2\right) \\
=\ & y_{ij}\Xi_{r^*,ij}^{(0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right) \sum_{s=0}^{t} \left(\Xi_{r^*,ij}^{(s)}\right)^2 \pm \widetilde{O}\left(\frac{1}{d^{\frac{2}{3}}} \frac{1}{d\sigma_\xi\sigma_0}\right) \pm \widetilde{O}\left(d^{3/2}\sigma_\xi^4\sigma_0^2\right) \\
\overset{(iii)}{=}\ & y_{ij}\Xi_{r^*,ij}^{(0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right) \sum_{s=0}^{t} \left(\Xi_{r^*,ij}^{(s)}\right)^2 \pm \widetilde{O}\left(d^{-\frac{2}{3}+0.03}\right) \pm \widetilde{O}\left(d^{-1.58}\right) \\
\overset{(iv)}{=}\ & y_{ij}\Xi_{r^*,ij}^{(0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right) \sum_{s=0}^{t} \left(\Xi_{r^*,ij}^{(s)}\right)^2 \pm o(\sqrt{d}\sigma_\xi\sigma_0),
\end{aligned} \tag{B.39}
$$

where (i) holds due to (B.37) and hypothesis (B.36); (ii) follows from Lemmas B.8 and hypothesis (B.34) for $s \leq t$; (iii) and (iv) holds since $\sigma_\xi = \Theta(d^{-0.51})$ and $\sigma_0 = \Theta(d^{-0.52})$ such that $\sqrt{d}\sigma_\xi\sigma_0 = \Theta(d^{-0.53})$.

Then (E.4) in Lemma E.7 guarantees $y_{ij}\Xi_{r^*,ij}^{(t)} \leq (md)^{-1/3}$ for any $t \leq \mathrm{T}_{\xi_{ij}}^-$ since $(md)^{-1/3} > 2y_{ij}\Xi_{r^*,ij}^{(0)}$. By concentration, we also know $\max_{ij} \mathrm{T}_{\xi_{ij}}^- + 1 \leq \mathrm{T}_\xi^-$, which means $y_{ij}\Xi_{r^*,ij}^{(t+1)} \leq (md)^{-1/3}$. In addition, for any other $r \in [m]$, we also have the following upper bounds

$$
y_{ij}\Xi_{r,ij}^{(t+1)} \leq y_{ij}\Xi_{r,ij}^{(0)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right) \sum_{s=0}^{t} \left(y_{ij}\Xi_{r,ij}^{(s)}\right)^2 + o(\sqrt{d}\sigma_\xi\sigma_0). \tag{B.40}
$$

Then we have

- For $r \in [m]$ such that $y_{ij}\Xi_{r^*,ij}^{(0)}/2 < y_{ij}\Xi_{r,ij}^{(0)} < y_{ij}\Xi_{r^*,ij}^{(0)}$, the (E.4) in Lemma E.7 ensures that $y_{ij}\Xi_{r,ij}^{(t+1)} \leq (md)^{-\frac{1}{3}}$ because $t+1 \leq \mathrm{T}_\xi^-$ and $2y_{ij}\Xi_{r,ij}^{(0)} = \Omega(\sqrt{d}\sigma_\xi\sigma_0) < (md)^{-\frac{1}{3}}$.
- For $r \in [m]$ such that $0 < y_{ij}\Xi_{r,ij}^{(0)} \leq y_{ij}\Xi_{r^*,ij}^{(0)}/2$, applying Lemma E.11 to the competition between (B.40) and (B.39) ensures that $y_{ij}\Xi_{r,ij}^{(t+1)} \leq y_{ij}\Xi_{r^*,ij}^{(t+1)} \leq (md)^{-\frac{1}{3}}$ for any $t \leq \mathrm{T}_\xi^-$.
- For $r \in [m]$ such that $-\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) \leq y_{ij}\Xi_{r,ij}^{(0)} \leq 0$, we know it cannot reach $(md)^{-\frac{1}{3}}$ before T_ξ^- since it needs to exceed $\Theta(\sqrt{d}\sigma_\xi\sigma_0)$ firstly. In fact, if $-\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) < y_{ij}\Xi_{r,ij}^{(s)} < \Theta(\sqrt{d}\sigma_\xi\sigma_0)$, then

$$
\begin{aligned}
y_{ij}\Xi_{r,ij}^{(s+1)} &\leq y_{ij}\Xi_{r,ij}^{(0)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right) \sum_{l=0}^{s} \left(\Xi_{r,ij}^{(l)}\right)^2 + o(\sqrt{d}\sigma_\xi\sigma_0) \\
&\leq y_{ij}\Xi_{r,ij}^{(s)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right) \left(\Xi_{r,ij}^{(s)}\right)^2 + o(\sqrt{d}\sigma_\xi\sigma_0) \\
&\leq \Theta(\sqrt{d}\sigma_\xi\sigma_0) + \widetilde{O}\left(\frac{\eta d^2\sigma_\xi^4\sigma_0^2}{Nn}\right) + o(\sqrt{d}\sigma_\xi\sigma_0) \\
&\leq \Theta(\sqrt{d}\sigma_\xi\sigma_0).
\end{aligned}
$$

Up to now, we have verified (B.34) at time $t+1$.

For (B.36) at time $t+1$, by update rule (B.4), we have

$$
|\Xi_{r,ij}^{(t+1)}| \leq |\Xi_{r,ij}^{(\mathrm{T}_v^0)}| + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right) \sum_{s=\mathrm{T}_v^0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right) \sum_{j' \in [n] \setminus \{j\}} \sum_{s=\mathrm{T}_v^0}^{t} \ell_{ij'}^{(s)} \left(\Xi_{r,ij'}^{(s)}\right)^2
$$

$$
\begin{aligned}
&+\widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{i'\in\mathcal{C}_k\setminus\{i\}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t}\ell_{i'j'}^{(s)}\left(\Xi_{r,i'j'}^{(s)}\right)^2\\
&+\widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in[K]\setminus\mathcal{K}_0\setminus\{k\}}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t}\ell_{i'j'}^{(s)}\left(\Xi_{r,i'j'}^{(s)}\right)^2\\
&+\widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t}\ell_{i'j'}^{(s)}\left(\Xi_{r,i'j'}^{(s)}\right)^2\\
&\overset{(i)}{\le}\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)+\widetilde{O}\left(\frac{\eta d^2\sigma_\xi^4\sigma_0^2}{Nn}\cdot\mathrm{T}_\xi^-\right)+\widetilde{O}\left(\eta d^{3/2}\sigma_\xi^4\sigma_0^2\cdot\mathrm{T}_\xi^-\right)+\widetilde{O}\left(\eta\sqrt{d}\sigma_\xi^2\cdot(md)^{-2/3}\cdot\mathrm{T}_\xi^-\right)\\
&\overset{(ii)}{\le}\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)+\widetilde{O}\left(\sqrt{d}\sigma_\xi\sigma_0\right)+\widetilde{O}\left(\frac{1}{d^{5/3}\sigma_\xi\sigma_0}\right)\\
&\overset{(iii)}{\le}\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0),
\end{aligned}
\tag{B.41}
$$

where (i) follows from the hypothesis (B.36) and (B.34); (ii) holds due to the definition of T_ξ; and (iii) is true because $\sqrt{d}\sigma_\xi\sigma_0=\Theta(d^{-0.53})$ and $(d^{5/3}\sigma_\xi\sigma_0)^{-1}=\Theta(d^{-\frac{5}{3}+1.03})\ll\Theta(d^{-0.57})$. Then we have verified (B.34) at time $t+1$. □

Lemma B.10. *Suppose Induction hypothesis B.1 holds for iterations $s\le t$, with probability at least $1-1/\mathsf{poly}(d)$, we maintain:*

$$
\frac{\eta}{Nn}\sum_{k\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=\mathrm{T}_\xi^-}^{t}\ell_{ij}^{(s)}\left[y_{ij}\Xi_{r,ij}^{(s)}\right]^2\le\widetilde{O}\left(\frac{1}{d\sigma_\xi^2}\right)\quad\textit{for any } t>\mathrm{T}_\xi^-. \tag{B.42}
$$

Proof. If $t>\mathrm{T}_\xi^-$, by the update rule (B.3) and the Induction hypothesis B.1 for any $0\le s\le t$, we have

$$
\begin{aligned}
y_{ij}\Xi_{r,ij}^{(t)}\ge y_{ij}\Xi_{r,ij}^{(\mathrm{T}_\xi^-)}&+\Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right)\cdot\sum_{s=\mathrm{T}_\xi^-}^{t-1}\ell_{ij}^{(s)}\left(\Xi_{r,ij}^{(s)}\right)^2-\widetilde{O}\left(\eta d^{3/2}\sigma_\xi^4\sigma_0^2\right)\\
&-\widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\cdot\sum_{k'\in\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_\xi^-}^{t-1}\ell_{i'j'}^{(s)}\left(\Xi_{r,i'j'}^{(s)}\right)^2,
\end{aligned}
\tag{B.43}
$$

where we also used (B.30) in Lemma B.8 since $\mathrm{T}_\xi^-\ge\tau_k$ for $k\in[K]\setminus\mathcal{K}_0$ by Lemma 5.1. Taking summation on both sides of (B.43) over $k\in\mathcal{K}_0, i\in\mathcal{C}_k, j\in[n]$, we can get

$$
\begin{aligned}
&\left[\Theta\left(d\sigma_\xi^2\right)-\widetilde{O}\left(\frac{K_0Nn\sqrt{d}\sigma_\xi^2}{K}\right)\right]\frac{\eta}{Nn}\sum_{k\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=\mathrm{T}_\xi^-}^{t-1}\ell_{ij}^{(s)}\left(\Xi_{r,ij}^{(s)}\right)^2\\
&\le\sum_{k\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\left[y_{ij}\Xi_{r,ij}^{(t)}-y_{ij}\Xi_{r,ij}^{(\mathrm{T}_\xi^-)}\right]+\widetilde{O}\left(\eta d^{3/2}\sigma_\xi^4\sigma_0^2\right).
\end{aligned}
$$

Since $K,P,N,n=\mathsf{polylog}(d)$, the inequality above implies

$$
\frac{\eta}{Nn}\sum_{k\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=\mathrm{T}_\xi^-}^{t}\ell_{ij}^{(s)}\left[y_{ij}\Xi_{r,ij}^{(s)}\right]^2\le\Theta\left(\frac{K_0Nn}{K}\right)\cdot\widetilde{O}\left(\frac{1}{d\sigma_\xi^2}\right)=\widetilde{O}\left(\frac{1}{d\sigma_\xi^2}\right), \tag{B.44}
$$

where we used the hypothesis $\max_{r,i\in\mathcal{C}_k,j}|\Xi_{r,ij}^{(t)}|\le\widetilde{O}(1)$. □

Lemma B.11. *With probability at least $1-1/\mathsf{poly}(d)$, we have:*

- *If* $k \in [K] \setminus \mathcal{K}_0$,

$$\max_{r\in[m], i\in\mathcal{C}_k, j\in[n]} |\Xi^{(t)}_{r,ij}| \le \widetilde{O}(d^{-1/2}) \quad \textit{for any } t > \mathrm{T}_v^0. \tag{B.45}$$

- *If* $k \in \mathcal{K}_0$,

$$\max_{r\in[m], i\in\mathcal{C}_k, j\in[n]} |\Xi^{(t)}_{r,ij}| \le \widetilde{O}(1) \quad \textit{for any } t > \mathrm{T}_v^0. \tag{B.46}$$

Proof. Now we suppose (B.45) and (B.46) hold for any $\mathrm{T}_v^0 < s \le t$. For $i \in \mathcal{C}_k$ with $k \in [K] \setminus \mathcal{K}_0$, using (B.4) gives

$$\begin{aligned}
|\Xi^{(t+1)}_{r,ij}| \le |\Xi^{(\mathrm{T}_v^0)}_{r,ij}| &+ \eta\Theta\left(\frac{d\sigma_\xi^2}{Nn}\right)\sum_{s=\mathrm{T}_v^0}^{t} \ell_{ij}^{(s)}\left(\Xi^{(s)}_{r,ij}\right)^2 + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{j'\in[n]\setminus\{j\}}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{ij'}^{(s)}\left(\Xi^{(s)}_{r,ij'}\right)^2 \\
&+ \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{i'\in\mathcal{C}_k\setminus\{i\}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{i'j'}^{(s)}\left(\Xi^{(s)}_{r,i'j'}\right)^2 \\
&+ \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in[K]\setminus\mathcal{K}_0\setminus\{k\}}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{i'j'}^{(s)}\left(\Xi^{(s)}_{r,i'j'}\right)^2 \\
&+ \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{i'j'}^{(s)}\left(\Xi^{(s)}_{r,i'j'}\right)^2 \\
&\overset{(i)}{\le} \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) + \widetilde{O}\left(\frac{\eta\sigma_\xi^2}{Nn}\right)\sum_{s=\mathrm{T}_v^0}^{t}\ell_{ij}^{(s)} + \widetilde{O}\left(\frac{\eta\sigma_\xi^2}{N\sqrt{d}}\right)\sum_{k'\in[K]\setminus\mathcal{K}_0}\sum_{i\in\mathcal{C}_{k'}}\nu_i^{(s)} \\
&\quad + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{i'j'}^{(s)}\left(\Xi^{(s)}_{r,i'j'}\right)^2 \\
&\overset{(ii)}{\le} \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) + \widetilde{O}\left(\sigma_\xi^2\right) + \widetilde{O}\left(\sqrt{d}\sigma_\xi^2\right) + \widetilde{O}\left(\sqrt{d}\sigma_\xi^2 \cdot \frac{1}{d\sigma_\xi^2}\right) \\
&\overset{(iii)}{\le} \widetilde{O}(d^{-1/2}),
\end{aligned} \tag{B.47}$$

where (i) holds due to the induction hypotheses; (ii) follows from Lemmas B.8 and B.10; and (iii) holds due to the setting $\sigma_\xi = \Theta(d^{-0.51})$ and $\sigma_0 = \Theta(d^{-0.52})$. Lemma 5.1 guarantees $\tau_k \le T_v^0$, which means that (B.47) can imply the conclusion (B.45). In fact, the proof of Lemma 5.1 depends only on Lemma B.6, which is independent of (B.45) and (B.46).

Notice that we have proved (B.46) holds for $\mathrm{T}_v^0 \le t \le \mathrm{T}_\xi^-$. Next we verify (B.46) at time $t+1 \ge \mathrm{T}_\xi^-$. For $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, by the update rule (B.3), we have

$$\begin{aligned}
|\Xi^{(t+1)}_{r,ij}| \le |\Xi^{(\mathrm{T}_\xi^-)}_{r,ij}| &+ \widetilde{O}\left(\frac{\eta d\sigma_\xi^2}{Nn}\right)\sum_{s=0}^{t}\ell_{ij}^{(s)}[\Xi^{(s)}_{r,ij}]^2 + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\mathrm{T}_\xi^-}^{t}\ell_{i'j'}^{(s)}\left(\Xi^{(s)}_{r,i'j'}\right)^2 \\
&+ \widetilde{O}\left(\frac{\eta\sigma_\xi^2}{N\sqrt{d}}\right)\sum_{k'\in[K]\setminus\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{s=\mathrm{T}_v^-}^{t}\nu_{i'}^{(s)} \\
&\le \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) + \widetilde{O}\left(d\sigma_\xi^2 \cdot \frac{1}{d\sigma_\xi^2}\right) + \widetilde{O}\left(\sqrt{d}\sigma_\xi^2 \cdot \frac{1}{d\sigma_\xi^2}\right) + \widetilde{O}\left(\frac{\sigma_\xi^2}{\sqrt{d}}\right) \\
&= \widetilde{O}(1),
\end{aligned} \tag{B.48}$$

where we used Lemma B.10 and Lemma B.8. Combining (B.47) and (B.48), we can finish the proof. □

B.4.2. Proof of Lemma B.3

Proof of Lemma B.3. Given $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, we define three time steps:

- τ_{ij}^-: the first iteration $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(t)} \geq \Theta((dm)^{-\frac{1}{3}})$.
- τ_{ij}^0: the first iteration $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(t)} \geq \Theta(m^{-\frac{1}{3}})$.
- τ_{ij}^+: the first iteration $\max_{r\in[m]} y_{ij}\Xi_{r,ij}^{(t)} \geq \Theta(\vartheta^{1/3})$.

Applying Lemma E.6 to (B.39), we know

$$\tau_{ij}^- \leq \Theta\left(\frac{Nn}{\eta(\sqrt{d}\sigma_\xi)^3\sigma_0}\right) + \left\lceil \frac{\log\left(\frac{1}{d^{5/6}m^{1/3}\sigma_\xi\sigma_0}\right)}{\log(2)}\right\rceil \leq \mathrm{T}_\xi^- + O\left(\log d\right),$$

which implies that

$$\sum_{s=\mathrm{T}_\xi^-}^{\tau_{ij}^-} \ell_{ij}^{(s)} \leq O\left(\log d\right). \tag{B.49}$$

Notice that $\ell_{ij}^{(t)} = \Theta(1)$ for $t \leq \tau_{ij}^0$. Now regarding $y_{ij}\Xi_{r^*,ij}^{(\tau_{ij}^-)}$ as the initial point, and applying Lemma E.6 to (B.39) again, we know

$$\sum_{s=\tau_{ij}^-}^{\tau_{ij}^0} \ell_{ij}^{(s)} \leq \tau_{ij}^0 - \tau_{ij}^- \leq \Theta\left(\frac{Nn}{\eta d\sigma_\xi^2}\cdot(md)^{1/3}\right) + O(\log d) \leq \widetilde{O}\left(\frac{1}{d^{2/3}\sigma_\xi^2}\right). \tag{B.50}$$

For $t \geq \tau_{ij}^0$, by (B.3), the following relation still holds

$$\begin{aligned}
y_{ij}\Xi_{r^*,ij}^{(t)} &= y_{ij}\Xi_{r^*,ij}^{(\tau_{ij}^0)} + \Theta\left(\frac{\eta\|\boldsymbol{\xi}_{ij}\|^2}{Nn}\right)\sum_{s=\tau_{ij}^0}^{t}\ell_{ij}^{(s)}\left(\Xi_{r^*,ij}^{(s)}\right)^2 \\
&\quad \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{j'\in[n]\setminus\{j\}}\sum_{s=\tau_{ij}^0}^{t}\ell_{ij'}^{(s)}\left(\Xi_{r^*,ij'}^{(s)}\right)^2 \\
&\quad \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{i'\in\mathcal{C}_k\setminus\{i\}}\sum_{j'\in[n]}\sum_{s=\tau_{ij}^0}^{t}\ell_{i'j'}^{(s)}\left(\Xi_{r^*,i'j'}^{(s)}\right)^2 \\
&\quad \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in\mathcal{K}_0\setminus\{k\}}\sum_{i'\in\mathcal{C}_{k'}}\sum_{j'\in[n]}\sum_{s=\tau_{ij}^0}^{t}\ell_{i'j'}^{(s)}\left(\Xi_{r^*,i'j'}^{(s)}\right)^2 \\
&\quad \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{Nn}\right)\sum_{k'\in[K]\setminus\mathcal{K}_0}\sum_{i'\in\mathcal{C}_{k'}}\sum_{s=\tau_{ij}^0}^{t}\nu_{i'}^{(s)}\left(\Xi_{r^*,i'j'}^{(s)}\right)^2 \\
&= y_{ij}\Xi_{r^*,ij}^{(\tau_{ij}^0)} + \Theta\left(\frac{\eta d\sigma_\xi^2}{Nn}\right)\sum_{s=\tau_{ij}^0}^{t-1}\ell_{ij}^{(s)}\left(\Xi_{r^*,ij}^{(s)}\right)^2 \pm o(m^{-\frac{1}{3}}).
\end{aligned}$$

Now regarding $y_{ij}\Xi_{r^*,ij}^{(\tau_{ij}^0)}$ as the initial point, and applying (E.11) in Lemma E.9, we have

$$\sum_{s=\tau_{ij}^0}^{\tau_{ij}^+} \ell_{ij}^{(s)} \leq \Theta\left(\frac{Nn}{\eta d\sigma_\xi^2}\cdot m^{1/3}\right) + O(\log d) = \widetilde{O}\left(\frac{1}{d\sigma_\xi^2}\right). \tag{B.51}$$

By the definition of ϑ in (A.3), we know

$$\sum_{s=\tau_{ij}^{+}}^{T} \ell_{ij}^{(s)} \leq \widetilde{O}(1). \tag{B.52}$$

Combining (B.49)–(B.52), we can prove the conclusion. □

B.4.3. Proof of Induction hypothesis B.2

The Induction hypothesis B.2 can be proved via Lemma B.9 and Lemma B.13. Since Induction hypothesis B.1 has been proved in Section B.4.1, we can use its conclusions in the following proof.

Lemma B.12. *Given $k \in [K] \setminus \mathcal{K}_0$. In the training process of GD, with probability at least $1 - 1/\mathsf{poly}(d)$, we have $\Gamma_{r,k}^{(t)} \leq \widetilde{O}(1)$ for any $t \geq 0$.*

Proof. Denote τ_k^{+} the first iteration $\max_{r\in[m]} \Gamma_{r,k}^{(t)} \geq \Theta(\vartheta^{1/3}) = \widetilde{O}(1)$, where ϑ is defined in (A.3). For any $t \geq \tau_k^{+}$, we know $\max_{r\in[m]} \Gamma_{r,k}^{(t)} \geq \Theta(\vartheta^{1/3})$ due to its monotonicity. For any $i \in \mathcal{C}_k$, $j \in [n]$ and $\tau_k^{+} \leq s \leq t$, it follows that

$$\begin{aligned}
\ell_{ij}^{(s)} &= \frac{1}{1+\exp\left\{\sum_{r=1}^{m}\left[(\alpha^3 - K_0\beta^3)\left(\Gamma_{r,k}^{(s)}\right)^3 + y_{ij}[\Xi_{r,ij}^{(s)}]^3\right]\right\}} \\
&\leq \frac{1}{1+\exp\left\{(\alpha^3 - K_0\beta^3)\vartheta - (m-1)(\alpha^3 - K_0\beta^3)\widetilde{O}(\sigma_0^3) - \widetilde{O}(md^{-3/2})\right\}} \\
&\leq \frac{\Theta(1)}{1+e^{\vartheta}},
\end{aligned}$$

where we used Induction hypothesis B.1 (b) and Lemma B.5. Now suppose $\Gamma_{r,k}^{(s)} \leq \widetilde{O}(1)$ for any $\tau_k^{+} \leq s \leq t$. By the definition of ϑ, we know

$$\frac{1}{N}\sum_{i\in\mathcal{C}_k}\sum_{j\in[n]}\sum_{s=\tau_k^{+}}^{t} \ell_{ij}^{(s)} \leq \widetilde{O}(1),$$

which implies that

$$\Gamma_{r,k}^{(t+1)} = \Gamma_{r,k}^{(\tau_k^{+})} + \frac{3\eta(\alpha^3 - K_0\beta^3)}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\tau_k^{+}}^{t} \nu_i^{(s)} \cdot \left(\Gamma_{r,k}^{(s)}\right)^2 \leq \widetilde{O}(1).$$

Then the proof is completed by induction. □

Lemma B.13. *Given any $k \in \mathcal{K}_0$, in the training process of GD, with probability at least $1 - 1/\mathsf{poly}(d)$, we maintain:*

$$\max_{r\in[m]} |\Gamma_{r,k}^{(t)}| \leq \widetilde{O}(\sigma_0) \quad \text{for any } t \geq \mathrm{T}_{\xi}^{-}. \tag{B.53}$$

Proof. In Lemma B.9, we have proved $\max_{r\in[m]} |\Gamma_{r,k}^{(t)}| \leq \widetilde{O}(\sigma_0)$ for any $t \leq \mathrm{T}_{\xi}^{-}$. Now suppose it holds for $\mathrm{T}_{\xi}^{-} \leq s \leq t$, at time $t+1$, we have

$$\begin{aligned}
|\Gamma_{r,k}^{(t+1)}| &\leq |\Gamma_{r,k}^{(\mathrm{T}_{\xi}^{-})}| + \left|\alpha^3\sum_{i\in\mathcal{C}_k}\sum_{s=\mathrm{T}_{\xi}^{-}}^{t} \nu_i^{(s)}\left(\Gamma_{r,k}^{(s)}\right)^2 - \rho^3\sum_{k'\in\mathcal{K}_0}\sum_{i\in\mathcal{C}_{k'}}\sum_{s=\mathrm{T}_{\xi}^{-}}^{t} \nu_i^{(s)}\left(\Gamma_{r,k}^{(s)}\right)^2\right| \\
&\overset{(i)}{\leq} |\Gamma_{r,k}^{(\mathrm{T}_{\xi}^{-})}| + \frac{3\eta\alpha^3}{N}\sum_{i\in\mathcal{C}_k}\sum_{s=\mathrm{T}_{\xi}^{-}}^{t} \nu_i^{(s)}\left(\Gamma_{r,k}^{(s)}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\le \widetilde{O}(\sigma_0) + \widetilde{O}\left(\alpha^3\sigma_0^2\right) \cdot \frac{\eta}{N} \sum_{i \in \mathcal{C}_k} \sum_{s=\mathrm{T}_\xi^-}^{t} \nu_i^{(s)} \\
&\overset{(ii)}{\le} \widetilde{O}(\sigma_0) + \widetilde{O}\left(\frac{\sigma_0^2}{d^{2/3}\sigma_\xi^2}\right) \\
&\overset{(iii)}{\le} \widetilde{O}(\sigma_0) \cdot \left[1 + \widetilde{O}\left(d^{-0.1}\right)\right],
\end{aligned}
$$

where (i) holds due to the induction hypothesis; (ii) follows from Lemma B.3; and (iii) holds due to the setting $\sigma_\xi = \Theta(d^{-0.51})$, $\sigma_0 = \Theta(d^{-0.52})$ and $\frac{2}{3} - 1.02 + 0.52 > 0.1$. □

C. GD with local updates under Parameter 1

Given the local weights $\{\boldsymbol{W}_i^{(t)}\}_{i \in [N]}$ in Local GD, we define the following iterates during the training process[1]:

- Signal intensity of $\boldsymbol{v}_k^*$ in the i-th client: $\Gamma_{r,k,i}^{(t)} = \langle \boldsymbol{w}_{r,i}^{(t)}, \boldsymbol{v}_k^* \rangle$ for $r \in [m]$ and $k \in [K]$.
- Noise memorization: $\Xi_{r,ij}^{(t)} = \langle \boldsymbol{w}_{r,i}^{(t)}, \boldsymbol{\xi}_{ij} \rangle$ for any $r \in [m]$, $i \in [N]$ and $j \in [n]$.
- Derivative: $\ell_{ij}^{(t)} = \ell_{ij}(\boldsymbol{W}_i^{(t)}) = 1/\left(1 + e^{y_{ij} F(\boldsymbol{W}_i^{(t)}, \mathbf{x}_{ij})}\right)$ for any $i \in [N]$ and $j \in [n]$.
- Derivative in each client: $\nu_i^{(t)} = \frac{1}{n} \sum_{j \in [n]} \ell_{ij}^{(t)}$ for $i \in [N]$.

C.1. Update rules of signal intensity and noise memorization

Update rules for the feature's signal in each client are

- If $k \in [K] \setminus \mathcal{K}_0$ and $i \in \mathcal{C}_k$, we have

$$
\begin{aligned}
\Gamma_{r,k,i}^{(t+1)} &= \Gamma_{r,k,i}^{(t)} + \frac{\eta}{n} \sum_{j \in [n]} y_{ij} \ell_{ij}(\boldsymbol{W}) \nabla_{\boldsymbol{w}_r} F(\boldsymbol{W}_i^{(t)}, \mathbf{x}_{ij}) \\
&= \Gamma_{r,k,i}^{(t)} + \frac{3\eta}{n} \sum_{j \in [n]} (\alpha^3 - K_0\beta^3) \ell_{ij}^{(t)} [\Gamma_{r,k,i}^{(t)}]^2 \\
&= \Gamma_{r,k,i}^{(t)} + 3\eta(\alpha^3 - K_0\beta^3) \nu_i^{(t)} \cdot [\Gamma_{r,k,i}^{(t)}]^2. \qquad \text{(C.1)}
\end{aligned}
$$

- If $k \in \mathcal{K}_0$ and $i \in \mathcal{C}_k$, we have

$$
\Gamma_{r,k,i}^{(t+1)} = \Gamma_{r,k,i}^{(t)} + 3\eta(\alpha^3 - \rho^3) \nu_i^{(t)} \cdot [\Gamma_{r,k,i}^{(t)}]^2. \qquad \text{(C.2)}
$$

- If $k \in \mathcal{K}_0$ and $i \in \mathcal{C}_k$, for $k' \in \mathcal{K}_0 \setminus \{k\}$, we have

$$
\Gamma_{r,k',i}^{(t+1)} = \Gamma_{r,k',i}^{(t)} - 3\eta\rho^3 \nu_i^{(t)} \cdot [\Gamma_{r,k',i}^{(t)}]^2. \qquad \text{(C.3)}
$$

Update rules for the noise's memorization: given any $i \in [N]$ and $j \in [n]$,

$$
\begin{aligned}
\Xi_{r,ij}^{(t+1)} &= \Xi_{r,ij}^{(t)} + \frac{1}{n} \sum_{j' \in [n]} y_{ij'} \ell_{ij'}^{(t)} \sum_{p \in [P]} 3 \langle \boldsymbol{w}_r^{(t)}, \mathbf{x}_{ij',p} \rangle^2 \langle \mathbf{x}_{ij',p}, \boldsymbol{\xi}_{ij} \rangle \\
&= \Xi_{r,ij}^{(t)} + \frac{3\eta \|\boldsymbol{\xi}_{ij}\|^2}{n} y_{ij} \ell_{ij}^{(t)} \left(\Xi_{r,ij}^{(t)}\right)^2 + \frac{3\eta}{n} \sum_{j' \in [n] \setminus \{j\}} y_{ij'} \ell_{ij'}^{(t)} \left(\Xi_{r,ij'}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{ij'}, \boldsymbol{\xi}_{ij} \rangle. \qquad \text{(C.4)}
\end{aligned}
$$

[1] To avoid new notations, we still use $\Xi_{r,ij}^{(t)}$, $\ell_{ij}^{(t)}$ and $\nu_i^{(t)}$ in this subsection, but they are different to those in the analysis of GD.

Induction hypothesis C.1. *Throughout the training process of Local GD, with probability* $1 - 1/\mathsf{poly}(d)$*,*

(a) For any $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$*, we maintain*

$$\max_{r\in[m],k'\in\mathcal{K}_0\backslash\{k\}} |\Gamma_{r,k',i}^{(t)}| \leq \widetilde{O}(\sigma_0) \quad \textit{for any } t \geq 0.$$

(b) For any $i \in [N]$ *and* $j \in [n]$*, we maintain*

$$\max_{r\in[m]} |\Xi_{r,ij}^{(t)}| \leq \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) \quad \textit{for any } t \geq 0.$$

C.2. Proof of Theorem 2

C.2.1. Proof of Lemma 6.1

Proof. The results can be easily obtained by applying Lemmas C.3 and C.4 to (C.1), (C.2) and (C.3). □

C.2.2. Proof of Lemma 6.2

Proof. Notice that, given the same initial point $\boldsymbol{W}_i^{(0)} = \boldsymbol{W}^{(0)}$, the local iterates at $t = 0$ still satisfy Lemma A.1. Specially, with probability at least $1 - 1/\mathsf{poly}(d)$, we have:

- For any $k \in [K]$ and $i \in \mathcal{C}_k$, $\max_{r\in[m]} \Gamma_{r,k,i}^{(0)} \geq \Omega(\sigma_0)$ and $\max_{r\in[m]} |\Gamma_{r,k,i}^{(0)}| \leq O\left(\sigma_0\sqrt{\log d}\right)$.
- For any $k \in [K]$, $i \in \mathcal{C}_k$ and $j \in [n]$, $\max_{r\in[m]} |\Xi_{r,ij}^{(0)}| \leq O\left(\sigma_\xi\sigma_0\sqrt{d\log d}\right)$.

Recall that $r_k^* = \arg\max_{r\in[m]} \Gamma_{r,k}^{(0)}$. Hereafter, we will write $\Gamma_{r^*,k,i}^{(t)} \equiv \Gamma_{r_k^*,k,i}^{(t)}$ for $t \geq 0$ and any $i \in \mathcal{C}_k$. Given $k \in [K]\backslash\mathcal{K}_0$, invoking Induction hypothesis C.1 (b), we know for $s \leq \tau_{k,i}$,

$$\begin{aligned}\ell_{ij}^{(s)} &= \frac{1}{1 + \exp\left\{\sum_{r=1}^m \left[(\alpha^3 - K_0\beta^3)\left(\Gamma_{r,k,i}^{(s)}\right)^3 + \left(\Xi_{r,ij}^{(s)}\right)^3\right]\right\}} \\ &\geq \frac{1}{1 + \exp\left\{1 + \widetilde{O}(d^{3/2}\sigma_\xi^3\sigma_0^3)\right\}} \\ &= \Theta(1).\end{aligned}$$

By the update rule (C.1) and the assumption $K_0\beta^3 \leq \alpha^3/2$, we have

$$\begin{aligned}\Gamma_{r,k,i}^{(t+1)} &= \Gamma_{r,k,i}^{(t)} + 3\eta(\alpha^3 - K_0\beta^3)\nu_i^{(t)} \cdot [\Gamma_{r,k,i}^{(t)}]^2 \\ &= \Gamma_{r,k,i}^{(t)} + \Theta\left(\eta\alpha^3\right) \cdot [\Gamma_{r,k,i}^{(t)}]^2.\end{aligned} \tag{C.5}$$

Now applying the tensor power method in Lemma E.4 to the sequence $\{\Gamma_{r^*,k,i}^{(t)}\}_{t\geq 0}$, we can guarantee

$$\begin{aligned}\tau_{k,i} &\leq \Theta\left(\frac{1}{\eta\alpha^3\Gamma_{r^*,k,i}^{(0)}}\right) + O\left(\log\left(\frac{1}{m[\Gamma_{r^*,k,i}^{(0)}]}\right)\right) \\ &\leq \Theta\left(\frac{1}{\eta\alpha^3\sigma_0}\right) + O\left(\log\left(\frac{1}{m\sigma_0}\right)\right).\end{aligned}$$

By invoking $\sigma_0 = \Theta(d^{-0.52})$, we can verify $\tau_{k,i} \leq \mathrm{T}_v^0 \leq I$.

Given $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, using $\rho^3 = \frac{\alpha^3 - \Theta(K/d)}{K_0}$, we also have

$$\begin{aligned}\Gamma_{r,k,i}^{(t+1)} &= \Gamma_{r,k,i}^{(t)} + 3\eta(\alpha^3 - \rho^3)\nu_i^{(t)} \cdot [\Gamma_{r,k,i}^{(t)}]^2 \\ &= \Gamma_{r,k,i}^{(t)} + \Theta\left(\eta\alpha^3\right) \cdot [\Gamma_{r,k,i}^{(t)}]^2.\end{aligned} \tag{C.6}$$

Applying the tensor power method in Lemma E.4 to the sequence $\{\Gamma_{r^*,k,i}^{(t)}\}_{t\geq 0}$, we can also guarantee $\tau_{k,i} \leq \mathrm{T}_v^0 \leq I$. The rest conclusions of Lemma 6.2 follows from Induction hypothesis C.1. □

C.2.3. PROOF OF THEOREM 2

Lemma C.1. *Given any $i \in \mathcal{C}_k$ with $k \in [K]$. In the training process of GD, for any $\tau_{k,i} < t \leq T$, we maintain:*

$$\eta \sum_{s=\tau_{k,i}}^{t} \nu_i^{(s)} \leq \widetilde{O}(1).$$

Proof. Denote $\tau_{k,i}^{+}$ the first iteration $\max_{r\in[m]} \Gamma_{r,k,i}^{(t)} \geq \Theta(\vartheta^{1/3})$. By the definition of $\tau_{k,i}$, we know $\Gamma_{r,k,i}^{(\tau_{k,i})} \geq \Theta(m^{-\frac{1}{3}}\alpha^{-1})$. Applying Lemma E.8 to (C.1) or (C.2), we can have

$$\eta \sum_{s=\tau_{k,i}}^{\tau_{k,i}^{+}} \nu_i^{(s)} \leq \frac{4}{3(\alpha^3-\beta^3)m^{-\frac{1}{3}}\alpha^{-1}} + 8\left\lceil \frac{\log\left(m\vartheta/\alpha^3\right)}{\log(2)} \right\rceil = \widetilde{O}(1). \tag{C.7}$$

In addition, $\Gamma_{r,k,i}^{(t)} \geq \Theta(\vartheta^{1/3})$ holds for any $t \geq \tau_{k,i}^{+}$. Recalling (A.3), we can guarantee that for any $t > \tau_{k,i}^{+}$,

$$\eta \sum_{s=\tau_{k,i}^{+}}^{t} \nu_i^{(s)} \leq \widetilde{O}(\eta) = \widetilde{O}(1). \tag{C.8}$$

Combining (C.7) and (C.8), we can finish the proof. □

Lemma C.2. *Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{H})$ be a noise vector that is independent of the training data. For any $i \in \mathcal{C}_k$ with $k \in [K]$, we have*

$$\max_{r\in[m]} |\langle \boldsymbol{w}_{r,i}^{(t)}, \boldsymbol{\xi}\rangle| \leq \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) \quad \textit{for any } t \geq 0.$$

Proof. Denote $\Xi_{r,i}^{(t)} = \langle \boldsymbol{w}_{r,i}^{(t)}, \boldsymbol{\xi}\rangle$. In the client $i \in \mathcal{C}_k$, we have

$$\begin{aligned}
\Xi_{r,i}^{(t+1)} &= \Xi_{r,i}^{(t)} + \frac{\eta}{n}\sum_{j\in[n]} y_{ij}\ell_{ij}^{(t)} \sum_{p\in[P]} 3\langle \boldsymbol{w}_r^{(t)}, \mathbf{x}_{ij,p}\rangle^2 \langle \mathbf{x}_{ij,p}, \boldsymbol{\xi}\rangle \\
&= \Xi_{r,i}^{(t)} + \frac{\eta}{n}\sum_{j\in[n]} y_{ij}\ell_{ij}^{(t)} \left(\Xi_{r,ij}^{(t)}\right)^2 \langle \boldsymbol{\xi}_{ij}, \boldsymbol{\xi}\rangle \\
&= \Xi_{r,i}^{(0)} \pm \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{n}\right) \sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2.
\end{aligned} \tag{C.9}$$

If $t \leq \tau_{k,i}$, Lemma C.3 and Induction hypothesis C.1 guarantee that

$$\frac{\eta}{n}\sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 \leq \frac{\eta}{n}\sum_{j\in[n]}\sum_{s=0}^{t} \ell_{ij}^{(s)} \cdot \widetilde{O}(d\sigma_\xi^2\sigma_0^2) \leq \widetilde{O}(\eta \mathrm{T}_v^0 d\sigma_\xi^2\sigma_0^2) = \widetilde{O}(d\sigma_\xi^2\sigma_0). \tag{C.10}$$

If $t > \tau_{k,i}$, Lemma C.5 and Induction hypothesis C.1 guarantee that

$$\frac{\eta}{n}\sum_{j\in[n]}\sum_{s=\tau_{k,i}}^{t} \ell_{ij}^{(s)} \left(\Xi_{r,ij}^{(s)}\right)^2 \leq \eta \sum_{s=\tau_{k,i}}^{t} \nu_i^{(s)} \cdot \widetilde{O}(d\sigma_\xi^2\sigma_0^2) \leq \widetilde{O}(d\sigma_\xi^2\sigma_0^2). \tag{C.11}$$

Plugging (C.10) and (C.11) into (C.9) yields

$$|\Xi_{r,i}^{(t+1)}| \leq \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) + \widetilde{O}(d^{3/2}\sigma_\xi^4\sigma_0) = \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0),$$

which proved the conclusion. □

Proof of Theorem 2. By the update rules (C.1) and (C.2), we know $\{\Gamma_{r,k,i}^{(t)}\}_{t\geq 0}$ is an increasing sequence for any $i \in \mathcal{C}_k$ with $k \in [K]$. Hence we have $\Gamma_{r,k,i}^{(t)} \geq \Gamma_{r,k,i}^{(0)} \geq -\widetilde{O}(\sigma_0)$. Therefore, for any training data $(\mathbf{x}_{ij}, y_{ij})$ with $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$, it holds that

$$\begin{aligned} y_{ij} F(\boldsymbol{W}_i^{(T)}, \mathbf{x}_{ij}) &= y_{ij} \sum_{r\in[m]} \sum_{p\in[P]} \langle \boldsymbol{W}_i^{(T)}, \mathbf{x}_{ij,p} \rangle^3 \\ &= \sum_{r\in[m]} \left[\alpha^3 \left(\Gamma_{r,k,i}^{(T)} \right)^3 - \rho^3 \sum_{k'\in\mathcal{K}_0} \left(\Gamma_{r,k',i}^{(T)} \right)^3 + y \left(\Xi_{r,ij}^{T} \right)^3 \right] \\ &\geq \alpha^3 \cdot \left[\Theta\left(\frac{1}{m\alpha^3} \right) - \widetilde{O}\left(m\sigma_0^3 \right) \right] - \widetilde{O}\left(d^{3/2} \sigma_\xi^3 \sigma_0^3 \right) \\ &= \widetilde{\Omega}(1). \end{aligned} \tag{C.12}$$

Moreover, for any training data $(\mathbf{x}_{ij}, y_{ij})$ with $i \in \mathcal{C}_k$ with $k \in [K] \setminus \mathcal{K}_0$, it holds that

$$\begin{aligned} y_{ij} F(\boldsymbol{W}_i^{(T)}, \mathbf{x}_{ij}) &= \sum_{r\in[m]} \left[(\alpha^3 - K_0 \beta^3) \left(\Gamma_{r,k,i}^{(T)} \right)^3 + y \left(\Xi_{r,ij}^{T} \right)^3 \right] \\ &\geq \alpha^3/2 \cdot \left[\Theta\left(\frac{1}{m\alpha^3} \right) - \widetilde{O}\left(m\sigma_0^3 \right) \right] - \widetilde{O}\left(d^{3/2} \sigma_\xi^3 \sigma_0^3 \right) \\ &= \widetilde{\Omega}(1). \end{aligned} \tag{C.13}$$

Combining (C.12) and (C.13), we can prove the training accuracy in Theorem 2.

Given the new test data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ and $k \in \mathcal{K}_0$, with probability at least $1 - 1/\mathsf{poly}(d)$, we have

$$\begin{aligned} yF(\boldsymbol{W}_i^{(T)}, \mathbf{x}) &= y \sum_{r\in[m]} \sum_{p\in[P]} \langle \boldsymbol{W}_i^{(T)}, \mathbf{x}_p \rangle^3 \\ &= \sum_{r\in[m]} \left[\alpha^3 \left(\Gamma_{r,k,i}^{(T)} \right)^3 - \rho^3 \sum_{k'\in\mathcal{K}_0} \left(\Gamma_{r,k',i}^{(T)} \right)^3 + y \langle \boldsymbol{w}_{r,i}^{(T)}, \boldsymbol{\xi} \rangle^3 \right] \\ &\geq \alpha^3 \cdot \Theta\left(\frac{1}{\alpha^3 m} \right) - K_0 \rho^3 \cdot \widetilde{O}(m\sigma_0^3) - \left| \sum_{r\in[m]} \langle \boldsymbol{w}_{r,i}^{(T)}, \boldsymbol{\xi} \rangle^3 \right| \\ &\geq \widetilde{\Omega}(1) - \widetilde{O}(\sigma_0^3) - \widetilde{O}(d^{3/2} \sigma_\xi^3 \sigma_0^3) \\ &= \widetilde{\Omega}(1), \end{aligned} \tag{C.14}$$

where we used Induction hypothesis C.1 (a) and Lemma C.2. Given the new test data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ and $k \in [K] \setminus \mathcal{K}_0$, with probability at least $1 - 1/\mathsf{poly}(d)$, we have

$$\begin{aligned} yF(\boldsymbol{W}_i^{(T)}, \mathbf{x}) &= y \sum_{r\in[m]} \sum_{p\in[P]} \langle \boldsymbol{W}_i^{(T)}, \mathbf{x}_p \rangle^3 \\ &= \sum_{r\in[m]} \left[(\alpha^3 - K_0 \rho^3) \left(\Gamma_{r,k,i}^{(T)} \right)^3 + y \langle \boldsymbol{w}_{r,i}^{(T)}, \boldsymbol{\xi} \rangle^3 \right] \\ &\geq \widetilde{\Omega}(1) - \widetilde{O}(d^{3/2} \sigma_\xi^3 \sigma_0^3) \\ &= \widetilde{\Omega}(1). \end{aligned} \tag{C.15}$$

The relations (C.14) and (C.15) prove the test accuracy in Theorem 2. □

C.3. Proof of Induction hypothesis C.1

The Induction hypothesis C.1 (a) is proved by Lemma C.4 and Lemma C.6. The Induction hypothesis C.1 (b) is proved by Lemma C.3 and Lemma C.5.

Lemma C.3. *In the training process of Local GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$\max_{i\in[N],j\in[n],r\in[m]} |\Xi_{r,ij}^{(t)}| \leq \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) \quad \textit{for any } t \leq \mathrm{T}_v^0. \tag{C.16}$$

Proof. For the noise $\boldsymbol{\xi}_{ij}$, by the update rule (C.4), we have

$$\begin{aligned}
|\Xi_{r,ij}^{(t+1)}| &= |\Xi_{r,ij}^{(t)}| + \frac{3\eta\|\boldsymbol{\xi}_{ij}\|^2}{n}\left(\Xi_{r,ij}^{(t)}\right)^2 + \frac{3\eta}{n}\sum_{j'\in[n]\setminus\{j\}} \ell_{ij'}^{(t)}\left(\Xi_{r,ij'}^{(t)}\right)^2 |\langle\boldsymbol{\xi}_{ij'},\boldsymbol{\xi}_{ij}\rangle| \\
&\leq |\Xi_{r,ij}^{(t)}| + \Theta\left(\frac{\eta d\sigma_\xi^2}{n}\right)\left(\Xi_{r,ij}^{(t)}\right)^2 + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{n}\right)\sum_{j'\in[n]\setminus\{j\}} \ell_{ij'}^{(t)}\left(\Xi_{r,ij'}^{(t)}\right)^2 \\
&= |\Xi_{r,ij}^{(0)}| + \Theta\left(\frac{\eta d\sigma_\xi^2}{n}\right)\cdot\sum_{s=0}^{t}\left(\Xi_{r,ij}^{(s)}\right)^2 + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{n}\right)\sum_{j'\in[n]\setminus\{j\}}\sum_{s=0}^{t} \ell_{ij'}^{(s)}\left(\Xi_{r,ij'}^{(s)}\right)^2 \\
&\overset{(i)}{\leq} \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) + \left[\Theta\left(\frac{\eta d\sigma_\xi^2}{n}\right) + \widetilde{O}\left(\frac{\eta\sqrt{d}\sigma_\xi^2}{n}\right)\right]\cdot\widetilde{O}\left(\mathrm{T}_v^0 d\sigma_\xi^2\sigma_0^2\right) \\
&= \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)\cdot\left[1 + \widetilde{O}\left(d^{3/2}\sigma_\xi^3\right)\right] \\
&\overset{(ii)}{\leq} \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0),
\end{aligned} \tag{C.17}$$

where (i) follows from the induction hypothesis for iterations $s \leq t$; and (ii) holds due to $\sigma_\xi = \Theta(d^{-0.51})$. □

Lemma C.4. *Given* $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$*, let* $\tau_{k,i}$ *be the first iteration there exists some* $\max_{r\in[m]}\Gamma_{r,k,i}^{(t)} > \Theta\left(\frac{1}{m^{1/3}\alpha}\right)$*. In the training process of Local GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$\max_{r\in[m],k'\in\mathcal{K}_0\setminus\{k\}} |\Gamma_{r,k',i}^{(t)}| \leq \widetilde{O}(\sigma_0) \quad \textit{for any } t \leq \tau_{k,i}. \tag{C.18}$$

Proof. Suppose (C.18) holds for iterations $s \leq t < \tau_{k,i}$. For $i \in \mathcal{C}_k$ with $k \in \mathcal{K}_0$ and $j \in [n]$, we have

$$\begin{aligned}
\ell_{ij}^{(s)} &= \frac{1}{1 + \exp\left\{\sum_{r=1}^m\left[\alpha^3\left(\Gamma_{r,k,i}^{(s)}\right)^3 - \rho^3\sum_{k'\in\mathcal{K}_0}\left(\Gamma_{r,k',i}^{(s)}\right)^3 + y_{ij}\left(\Xi_{r,ij}^{(s)}\right)^3\right]\right\}} \\
&\geq \frac{1}{1 + \exp\left\{\Theta(1) + \widetilde{O}\left(K_0 m\rho^3\sigma_0^3\right) + \widetilde{O}\left(md^{3/2}\sigma_\xi^3\sigma_0^3\right)\right\}} \\
&= \Theta(1).
\end{aligned}$$

Using the update rule (C.2) and $\rho^3 = \frac{\alpha^3 - \Theta(K/d)}{K_0}$, we have

$$\begin{aligned}
\Gamma_{r,k,i}^{(t+1)} &= \Gamma_{r,k,i}^{(t)} + 3\eta(\alpha^3 - \rho^3)\nu_i^{(t)}\cdot[\Gamma_{r,k,i}^{(t)}]^2 \\
&= \Gamma_{r,k,i}^{(t)} + \Theta(\eta\alpha^3)\cdot[\Gamma_{r,k,i}^{(t)}]^2.
\end{aligned} \tag{C.19}$$

Applying Lemma E.4 to the sequence $\{\Gamma_{r^*,k,i}^{(t)}\}_{t\geq 0}$, we can guarantee $\tau_{k,i} \leq \mathrm{T}_v^0$. In addition, for $k' \in \mathcal{K}_0 \setminus \{k\}$, it holds that $\min_{r\in[m]}\Gamma_{r,k',i}^{(0)} \geq -O(\sigma_0\sqrt{\log d})$. Denote $r_{k',i}^- = \arg\min_{r\in[m]}\Gamma_{r,k',i}^{(0)}$. For simplicity, we write $\Gamma_{r^-,k',i}^{(t)} \equiv \Gamma_{r_{k',i}^-,k',i}^{(t)}$. By update rule (C.3) and the fact $\nu_i^{(t)} \leq 1$, we have

$$\begin{aligned}
-\Gamma_{r^-,k',i}^{(t+1)} &= -\Gamma_{r^-,k',i}^{(t)} + 3\eta\rho^3\nu_i^{(t)}\cdot[\Gamma_{r^-,k',i}^{(t)}]^2 \\
&\leq -\Gamma_{r^-,k',i}^{(t)} + 3\eta\rho^3\cdot[\Gamma_{r^-,k',i}^{(t)}]^2.
\end{aligned} \tag{C.20}$$

Notice that $-\Gamma^{(0)}_{r^-,k',i}/\Gamma^{(0)}_{r^*,k,i} \le O\left(\sqrt{\log d}\right)$ and $\alpha^3/\rho^3 = O(K_0) \ge O(\log d)$. Using Lemma E.12 on two sequences in (C.19) and (C.20), we can guarantee $-\Gamma^{(t+1)}_{r^-,k',i} \le -2\Gamma^{(0)}_{r^-,k',i} = \widetilde{O}(\sigma_0)$ since $t+1 \le \tau_{k,i}$. By Lemma E.10, we can guarantee that $\min_{r\in[m]} \Gamma_{r,k',i} \ge \widetilde{O}(\sigma_0)$. For the upper bound, we notice that for any $r \in [m]$,

$$\Gamma^{(t+1)}_{r,k',i} = \Gamma^{(t)}_{r,k',i} - 3\eta\rho^3\nu_i^{(t)} \cdot [\Gamma^{(t)}_{r,k',i}]^2 \le \Gamma^{(t)}_{r,k',i} \le \Gamma^{(0)}_{r,k',i} \le \widetilde{O}(\sigma_0).$$

By induction, we can prove the conclusion. □

Lemma C.5. *In the training process of Local GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$\max_{i\in[N],j\in[n],r\in[m]} |\Xi^{(t)}_{r,ij}| \le \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0) \quad \textit{for any } t > \mathrm{T}_v^0. \tag{C.21}$$

Proof. Suppose (C.21) holds for iterations $\mathrm{T}_v^0 < s \le t$. By the update rule (C.4), we have

$$\begin{aligned}
|\Xi^{(t+1)}_{r,ij}| &\le |\Xi^{(\mathrm{T}_v^0)}_{r,ij}| + \frac{3\eta\|\boldsymbol{\xi}_{ij}\|^2}{n}\sum_{s=\mathrm{T}_v^0}^{t} \ell_{ij}^{(s)}\left(\Xi^{(s)}_{r,ij}\right)^2 + 3\eta\sum_{s=\mathrm{T}_v^0}^{t}\nu_i^{(s)}\left(\Xi^{(s)}_{r,ij'}\right)^2 |\langle\boldsymbol{\xi}_{ij'},\boldsymbol{\xi}_{ij}\rangle| \\
&\le |\Xi^{(\mathrm{T}_v^0)}_{r,ij}| + \Theta\left(\eta d\sigma_\xi^2\right)\sum_{s=\mathrm{T}_v^0}^{t}\nu_i^{(s)}\left(\Xi^{(s)}_{r,ij}\right)^2 + \widetilde{O}\left(\eta\sqrt{d}\sigma_\xi^2\right)\sum_{s=\mathrm{T}_v^0}^{t}\nu_i^{(s)}\left(\Xi^{(s)}_{r,ij'}\right)^2 \\
&\overset{(i)}{\le} |\Xi^{(\mathrm{T}_v^0)}_{r,ij}| + \Theta\left(\eta d^2\sigma_\xi^4\sigma_0^2\right)\sum_{s=\mathrm{T}_v^0}^{t}\nu_i^{(s)} + \widetilde{O}\left(\eta d^{3/2}\sigma_\xi^4\sigma_0^2\right)\sum_{s=\mathrm{T}_v^0}^{t}\nu_i^{(s)} \\
&\overset{(ii)}{\le} \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)\cdot\left[1 + \widetilde{O}((\sqrt{d}\sigma_\xi)^3\sigma_0) + \widetilde{O}\left(d\sigma_\xi^3\sigma_0\right)\right] \\
&\overset{(iii)}{\le} \widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0),
\end{aligned}$$

where (i) follows from the induction hypothesis; (ii) holds due to Lemma (C.1) and the fact $\tau_{k,i} \le \mathrm{T}_v^0$; and (iii) holds due to the setting $\sigma_\xi = \Theta(d^{-0.51})$ and $\sigma_0 = \Theta(d^{-0.52})$. □

Lemma C.6. *Given* $i \in \mathcal{C}_k$ *with* $k \in \mathcal{K}_0$*. In the training process of Local GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$\min_{r\in[m],k'\in\mathcal{K}_0\setminus\{k\}} \Gamma^{(t)}_{r,k',i} \ge -O(\sigma_0\sqrt{\log d}) \quad \textit{for any } t > \tau_{k,i}. \tag{C.22}$$

Proof. Suppose (C.22) holds for iterations $\tau_{k,i} < s \le t$. By update rule (C.3), we have

$$\begin{aligned}
-\Gamma^{(t+1)}_{r^-,k',i} &= -\Gamma^{(t)}_{r^-,k',i} + 3\eta\rho^3\nu_i^{(t)}\cdot[\Gamma^{(t)}_{r^-,k',i}]^2 \\
&\le -\Gamma^{(\tau_{k,i})}_{r^-,k',i} + 3\eta\rho^3\cdot\sum_{s=\tau_{k,i}}^{t}\nu_i^{(s)}[\Gamma^{(s)}_{r^-,k',i}]^2 \\
&\le O(\sigma_0\sqrt{\log d}) + O(\sigma_0^2\log d)\cdot\eta\sum_{s=\tau_{k,i}}^{t}\nu_i^{(s)} \\
&\le O(\sigma_0\sqrt{\log d}) + \widetilde{O}(\sigma_0^2\log d) \\
&= O(\sigma_0\sqrt{\log d}),
\end{aligned}$$

where we used Lemma C.1. By induction, we can prove the conclusion. □

Lemma C.7. *Given any* $k \in [K]$ *and* $i \in \mathcal{C}_k$*, in the training process of Local GD, with probability at least* $1 - 1/\mathsf{poly}(d)$*, we maintain:*

$$\max_{r\in[m]} |\Gamma^{(t)}_{r,k,i}| \le \widetilde{O}(1) \quad \textit{for any } t \le T.$$

Proof. Since $\tau^{+}_{k,i}$ is the first iteration $\max_{r\in[m]}\Gamma^{(t)}_{r,k,i} \geq \Theta(\vartheta^{1/3})$, we only need to prove the conclusion for $t \geq \tau^{+}_{k,i}$. Applying Lemma C.1 to (C.1) and (C.2), for any $t \geq \tau^{+}_{k,i}$ we have

$$\Gamma^{(t)}_{r,k,i} \leq \Gamma^{(\tau^{+}_{k,i})}_{r,k,i} + 3\eta\alpha^3 \sum_{s=\tau^{+}_{k,i}}^{t} \nu_i^{(s)} \leq \widetilde{O}(1).$$

□

D. GD and one-shot Local GD under Parameter 2

D.1. Proof of Theorem 3

Proof. From the signal intensity's update rule in (B.2), we know for any $k \in [K] \setminus \mathcal{K}_0$,

$$\Gamma^{(t+1)}_{r,k} \leq \Gamma^{(t)}_{r,k} + \frac{3\eta\alpha^3}{N}\sum_{i=1}^{N}\nu_i^{(t)}.$$

Let $\tau^{0}_{r,k}$ be the first iteration that $\Gamma^{(t)}_{r,k}$ reaches $2\Gamma^{(0)}_{r,k}$. For $r \in [m]$ such that $\Gamma^{(0)}_{r,k} > 0$, $\{\Gamma^{(t)}_{r,k}\}_{t\geq 0}$ is a positive sequence since it is increasing. Applying Lemma E.5, we have

$$\tau^{0}_{r,k} \geq \frac{N}{12\eta\alpha^3}\frac{1}{\Gamma^{(0)}_{r,k}} \geq \frac{N}{12\eta\alpha^3}\cdot\widetilde{\Omega}(\sigma_0^{-1}) = \widetilde{\Omega}\left(\frac{N}{\eta\alpha^3\sigma_0}\right),$$

where we also used $\max_{r\in[m]}|\Gamma^{(0)}_{r,k}| \leq \widetilde{O}(\sigma_0)$. For $r \in [m]$ such that $\Gamma^{(0)}_{r,k} \leq 0$, we know $\tau^{0}_{r,k} \geq \widetilde{\Omega}\left(\frac{N}{\eta\alpha^3\sigma_0}\right)$ since it needs to exceed zero firstly.

□

D.2. Proof of Theorem 4

Proof. For the conclusion of GD with local updates, the Induction hypothesis C.1 (in Section C.2) still holds since the scale of ρ in Parameter 2 is smaller than that in Parameter 2. Consequently, Lemma 6.2 also holds under Parameter 2. Given the new test data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ and $k \in \mathcal{K}_0$, with probability at least $1 - 1/\mathsf{poly}(d)$, we have

$$\begin{aligned}
yF(\bar{\boldsymbol{W}}, \mathbf{x}) &= y\sum_{r\in[m]}\sum_{p\in[P]}\left\langle \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{W}_i^{(T)}, \mathbf{x}_p\right\rangle^3 \\
&= \sum_{r\in[m]}\left[\alpha^3\left(\frac{1}{N}\sum_{i=1}^{N}\Gamma^{(T)}_{r,k,i}\right)^3 - \rho^3\sum_{k'\in\mathcal{K}_0}\left(\frac{1}{N}\sum_{i=1}^{N}\Gamma^{(T)}_{r,k',i}\right)^3 + y\left\langle\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}^{(T)}_{r,i}, \boldsymbol{\xi}\right\rangle^3\right] \\
&\geq \sum_{r\in[m]}\left[\alpha^3\left(\frac{1}{N}\sum_{i=1}^{N}\Gamma^{(T)}_{r,k,i}\right)^3 - \rho^3\sum_{k'\in\mathcal{K}_0}\left(\frac{1}{N}\sum_{i=1}^{N}\Gamma^{(T)}_{r,k',i}\right)^3\right] - \sum_{r\in[m]}\left(\max_{i\in[N]}\left|\left\langle\boldsymbol{w}^{(T)}_{r,i}, \boldsymbol{\xi}\right\rangle\right|\right)^3 \\
&\geq \widetilde{\Omega}(1) - \widetilde{O}(\rho^3 K_0) - \widetilde{O}(m\sigma_0^3) - \widetilde{O}(d^{3/2}\sigma_\xi^3\sigma_0^3) \\
&= \widetilde{\Omega}(1),
\end{aligned}$$

where we used Lemmas 6.2, C.2 and C.7 in the last inequality; and the last equality holds due to $\rho = 1/\mathsf{poly}(d)$.

Given the new test data $(\mathbf{x}, y) \sim \mathcal{D}_i$ for $i \in \mathcal{C}_k$ and $k \in [K] \setminus \mathcal{K}_0$, with probability at least $1 - 1/\mathsf{poly}(d)$, we have

$$yF(\bar{\boldsymbol{W}}, \mathbf{x}) = y\sum_{r\in[m]}\sum_{p\in[P]}\left\langle \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{W}_i^{(T)}, \mathbf{x}_p\right\rangle^3$$

$$
\begin{aligned}
&= \sum_{r\in[m]} \left[(\alpha^3 - K_0\beta^3)\left(\frac{1}{N}\sum_{i=1}^{N}\Gamma_{r,k,i}^{(T)}\right)^3 + y\left\langle \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_{r,i}^{(T)}, \boldsymbol{\xi}\right\rangle^3 \right] \\
&\geq (\alpha^3 - K_0\beta^3)\cdot\Theta\left(\frac{1}{\alpha^3 m}\right) - \widetilde{O}(m\sigma_0^3) - \sum_{r\in[m]}\left(\max_{i\in[N]}\left|\left\langle \boldsymbol{w}_{r,i}^{(T)}, \boldsymbol{\xi}\right\rangle\right|\right)^3 \\
&\geq \widetilde{\Omega}(1) - \widetilde{O}(\sigma_0^3) - \widetilde{O}(d^{3/2}\sigma_\xi^3\sigma_0^3) \\
&= \widetilde{\Omega}(1),
\end{aligned}
$$

where we used Induction hypothesis C.1 (a) and Lemma C.2. □

E. Auxiliary Lemmas

E.1. Probability inequalities

Lemma E.1. *Let $X \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$ be a Gaussian distributed vector in $\mathbb{R}^d$. With probability at least $1 - 1/\mathsf{poly}(d)$, we have $\|X\| = \Theta(\sigma\sqrt{d})$.*

Proof. Notice that $\|X\|^2/\sigma^2 = \sum_{j\in[d]} X_j^2/\sigma^2 \sim \chi^2(d)$, i.e., the Chi-squared distribution with degree of freedom d. It holds that $\mathbb{E}[\|X\|^2] = 2d\sigma^2$. By the tail probability of $\chi^2(d)$, we know

$$
\mathbb{P}\left(\left|\|X\|^2 - \mathbb{E}[\|X\|^2]\right| > \epsilon\sigma^2\right) \leq 2e^{-\epsilon^2/2}.
$$

Taking $\epsilon = \sqrt{2c\log d}$ for some absolute constant $c \geq 1$, we have

$$
\mathbb{P}\left(\left|\|X\|^2 - 2d\sigma^2\right| > \sqrt{2c\log d}\right) \leq 2d^{-c}.
$$

□

Lemma E.2. *Let $X \sim \mathcal{N}(0, \sigma_x^2\mathbf{I}_d)$ and $Y \sim \mathcal{N}(0, \sigma_y^2\mathbf{I}_d)$ be two independent Gaussian vectors in $\mathbb{R}^d$. With probability at least $1 - 1/\mathsf{poly}(d)$, we have $|\langle X, Y\rangle| \leq O\left(\sigma_x\sigma_y\sqrt{d\log d}\right)$.*

Proof. Let's first fix Y, then we know $\langle X, Y\rangle \sim \mathcal{N}(0, \sigma_x^2\|Y\|^2)$ by independence. By the tail probability of standard Gaussian distribution, we have

$$
\mathbb{P}\left(\frac{|\langle X, Y\rangle|}{\sigma_x\|Y\|} > \epsilon \mid Y\right) \leq 2e^{-\epsilon^2/2}.
$$

Taking $\epsilon = \sqrt{2c\log d}$ for some absolute constant $c \geq 1$, we have

$$
\mathbb{P}\left(|\langle X, Y\rangle| > \sqrt{2c\log d}\cdot\sigma_x\|Y\| \mid Y\right) \leq 2d^{-c}.
$$

Together with the bound on $\|Y\|$ by Lemma E.1, we can prove the conclusion. □

Lemma E.3 (Lemma K.12 in Jelassi & Li (2022))**.** *Let $\{\boldsymbol{w}_r\}_{r=1}^m$ be vectors in $\mathbb{R}^d$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2\mathbf{I}_d)$. If there exists a unit norm vector $\boldsymbol{u}$ such that $|\sum_{r=1}^m \langle \boldsymbol{w}_r, \boldsymbol{u}\rangle^3| \geq 1$, then for any $\epsilon \in (0,1)$, we have*

$$
\mathbb{P}\left(\left|\sum_{r=1}^{m}\langle \boldsymbol{w}_r, \boldsymbol{\xi}\rangle^3\right| \leq \epsilon\sigma_\xi^3\right) \leq O\left(\epsilon^{1/3}\right). \tag{E.1}
$$

E.2. Basic tensor power method

Lemma E.4 (Lemma K.15 in Jelassi & Li (2022))**.** *Let $\{z^{(t)}\}_{t=0}^T$ be a positive sequence defined by the following recursions*

$$
z^{(t+1)} \geq z^{(t)} + h[z^{(t)}]^2,
$$

$$z^{(t+1)} \leq z^{(t)} + H[z^{(t)}]^2,$$

where $z^{(0)} > 0$ is the initialization and $h, H > 0$. Let $v > 0$ such that $z^{(0)} \leq v$ and t_0 be the first iteration $z^{(t)} \geq v$. Then, we have

$$t_0 \leq \frac{3}{hz^{(0)}} + \frac{8H}{h} \left\lceil \frac{\log(v/z^{(0)})}{\log(2)} \right\rceil .$$

Lemma E.5. *Let $\{z^{(t)}\}_{t\geq 0}$ be a positive sequence satisfying the recursive upper bound in Lemma E.4. Let $v > 0$ such that $0 < z^{(0)} \leq v$ and t_0 be the first iteration $z^{(t)} \geq v$. For any $v \geq 2z^{(0)}$, we have the following lower bound*

$$t_0 \geq \frac{1}{4Hz^{(0)}}. \tag{E.2}$$

Proof. Let τ_b be the first iteration that $z^{(t)} \geq 2^b z^{(0)}$ for $b \geq 1$, that is $\tau_b = \inf\{t : z^{(t)} \geq 2^b z^{(0)}\}$. Using the recursive upper bound, we have

$$\begin{aligned} z^{(\tau_1)} &\leq z^{(0)} + H \sum_{s=0}^{\tau_1 - 1} [z^{(s)}]^2 \\ &\leq z^{(0)} + 4H \sum_{s=0}^{\tau_1 - 1} [z^{(0)}]^2 \\ &= z^{(0)} + 4H[z^{(0)}]^2 \cdot \tau_1. \end{aligned}$$

Together with the assumption $C \leq z^{(0)}/8$, it follows that

$$\tau_1 \geq \frac{z^{(\tau_1)} - z^{(0)}}{4H[z^{(0)}]^2} \geq \frac{z^{(0)}}{4A[z^{(0)}]^2} = \frac{1}{4Hz^{(0)}}.$$

Since $v \geq 2z^{(0)}$, the conclusion follows immediately. □

Lemma E.6 (Lemma K.16 in Jelassi & Li (2022))**.** *Let $\{z^{(t)}\}_{t=0}^{T}$ be a positive sequence defined by the following recursions*

$$\begin{aligned} z^{(t)} &\geq z^{(0)} + A \sum_{s=0}^{t-1} [z^{(s)}]^2 - C, \\ z^{(t)} &\leq z^{(0)} + A \sum_{s=0}^{t-1} [z^{(s)}]^2 + C, \end{aligned}$$

where $A, C > 0$ and $z^{(0)} > 0$ is the initialization. Assume that $C \leq z^{(0)}/8$. Let t_0 be the first iteration $z^{(t)} \geq v$. If $v > z^{(0)}$, we have the following upper bound

$$t_0 \leq \frac{21}{Az^{(0)}} + 8 \left\lceil \frac{\log(v/z^{(0)})}{\log(2)} \right\rceil . \tag{E.3}$$

Lemma E.7. *For the same sequence $\{z^{(t)}\}_{t\geq 0}$ be a positive sequence satisfying the recursive upper bound in Lemma E.6. Let $v > 0$ such that $z^{(0)} \leq v$ and t_0 be the first iteration $z^{(t)} \geq v$. For any $v \geq 2z^{(0)}$, we have the following lower bound*

$$t_0 \geq \frac{1}{8Az^{(0)}}. \tag{E.4}$$

Proof. Let τ^b be the first iteration that $z^{(t)} \geq 2^b z^{(0)}$ for $b \geq 1$, that is $\tau_b = \inf\{t : z^{(t)} \geq 2^b z^{(0)}\}$. Using the recursive upper bound, we have

$$z^{(\tau_1)} \leq z^{(0)} + A \sum_{s=0}^{\tau_1 - 1} [z^{(s)}]^2 + C$$

$$\leq z^{(0)} + 4A \sum_{s=0}^{\tau_1 - 1} [z^{(0)}]^2 + C$$
$$= z^{(0)} + 4A[z^{(0)}]^2 \cdot \tau_1 + C.$$

Together with the assumption $C \leq z^{(0)}/8$, it follows that

$$\tau_1 \geq \frac{z^{(\tau_1)} - z^{(0)} - C}{4A[z^{(0)}]^2} \geq \frac{z^{(0)} - C}{4A[z^{(0)}]^2} \geq \frac{1}{8Az^{(0)}}.$$

Since $v \geq 2z^{(0)}$, the conclusion follows immediately. □

E.3. Variants of tensor power method

E.3.1. Bounds for the increment

Lemma E.8. *Let $\{z^{(t)}\}_{t=0}^T$ and $\{a^{(t)}\}_{t=0}^T$ be two positive sequences admitting the following recursions*

$$z^{(t+1)} \geq z^{(t)} + ha^{(t)}[z^{(t)}]^2,$$
$$z^{(t+1)} \leq z^{(t)} + Ha^{(t)}[z^{(t)}]^2.$$

where $0 < h < H$ and $z^{(0)} > 0$. If $\max_{t \leq T} a^{(t)} \leq A$, we have

$$\sum_{s=0}^{T} a^{(s)} \leq \frac{4}{hz^{(0)}} + \frac{8HA}{h} \left\lceil \frac{\log(z^{(T)}/z^{(0)})}{\log(2)} \right\rceil,$$

and

$$\sum_{s=0}^{T} a^{(s)} \geq \frac{z^{(T)} - z^{(0)}}{H[z^{(T)}]^2}.$$

Proof. Let $B = \lceil \frac{\log(z^{(T)}/z^{(0)})}{\log(2)} \rceil$, then we define a sequence of time steps $\{\tau_b\}_{b=0}^B$ where $\tau_b = \inf_t\{t : z^{(t)} \geq 2^b z^{(0)}\}$. Since $z^{(t)}$ is increasing, we know $\tau_1 \leq \ldots \leq \tau_B$. Let's first consider $b = 1$. By the recursive lower bound and the definition of τ_1, we have

$$z^{(\tau_1)} \geq z^{(0)} + h \sum_{s=0}^{\tau_1 - 1} a^{(s)}[z^{(s)}]^2 \geq z^{(0)} + h \sum_{s=0}^{\tau_1 - 1} a^{(s)}[z^{(0)}]^2,$$

which yields

$$\sum_{s=0}^{\tau_1 - 1} a^{(s)} \leq \frac{1}{h} \frac{z^{(\tau_1)} - z^{(0)}}{[z^{(0)}]^2}. \tag{E.5}$$

In addition, invoking the recursive upper bound gives

$$z^{(\tau_1)} \leq z^{(\tau_1 - 1)} + Ha^{(\tau_1 - 1)}[z^{(\tau_1 - 1)}]^2 \leq 2z^{(0)} + HA[2z^{(0)}]^2, \tag{E.6}$$

where we used $z^{(\tau_1 - 1)} \leq 2z^{(0)}$ and $\max_{t \leq T} a^{(t)} \leq A$. Combining (E.5) and (E.6) leads

$$\sum_{s=0}^{\tau_1 - 1} a^{(s)} \leq \frac{1}{h} \frac{2z^{(0)} + 4HA[z^{(0)}]^2 - z^{(0)}}{[z^{(0)}]^2} = \frac{1}{hz^{(0)}} + \frac{4HA}{h}. \tag{E.7}$$

For the case $b > 1$, by the definition of τ_{b-1}, we know $z^{(s)} \geq 2^{b-1} z^{(0)}$ for any $s \geq \tau_{b-1}$. It follows that

$$z^{(\tau_b)} \geq z^{(\tau_{b-1})} + h \sum_{s=\tau_{b-1}}^{\tau_b - 1} a^{(s)}[z^{(s)}]^2 \geq 2^{b-1} z^{(0)} + h \sum_{s=\tau_{b-1}}^{\tau_b - 1} a^{(s)}[2^{b-1} z^{(0)}]^2.$$

It implies that

$$\sum_{s=\tau_{b-1}}^{\tau_b-1} a^{(s)} \leq \frac{z^{(\tau_b)} - 2^{b-1}z^{(0)}}{h[2^{b-1}z^{(0)}]^2}. \tag{E.8}$$

Similar to (E.6), we also have

$$z^{(\tau_b)} \leq z^{(\tau_b-1)} + Ha^{(\tau_b-1)}[z^{(\tau_b-1)}]^2 \leq 2^b z^{(0)} + HA[2^b z^{(0)}]^2. \tag{E.9}$$

Plugging (E.9) into (E.8) gives

$$\sum_{s=\tau_{b-1}}^{\tau_b-1} a^{(s)} \leq \frac{2^b z^{(0)} + HA[2^b z^{(0)}]^2}{h[2^{b-1}z^{(0)}]^2} \leq \frac{1}{2^{b-1}}\frac{2}{hz^{(0)}} + \frac{4HA}{h}. \tag{E.10}$$

Since $\tau_B = T$ and $\tau_0 = 0$, combining (E.7) and (E.10), we conclude that

$$\sum_{s=0}^{T} a^{(s)} = \sum_{b=1}^{B}\sum_{s=\tau_{b-1}}^{\tau_b-1} a^{(s)} \leq \sum_{b=1}^{B} \frac{1}{2^{b-1}}\frac{2}{hz^{(0)}} + B\frac{8HA}{h} \leq \frac{4}{hz^{(0)}} + \frac{8HA}{h}\left\lceil \frac{\log(z^{(T)}/z^{(0)})}{\log(2)} \right\rceil.$$

This proves the upper bound. The lower bound can be obtained by rearranging the following inequality

$$z^{(T)} \leq z^{(0)} + H\sum_{s=0}^{T-1} a^{(s)}[z^{(s)}]^2 \geq z^{(0)} + H\sum_{s=0}^{T-1} a^{(s)}[z^{(T)}]^2.$$

□

Lemma E.9. *Let $\{z^{(t)}\}_{t=0}^{T}$ and $\{a^{(t)}\}_{t=0}^{T}$ be two positive sequences defined by the following recursions*

$$z^{(t)} \geq z^{(0)} + H\sum_{s=0}^{t-1} a^{(s)}[z^{(s)}]^2 - C,$$
$$z^{(t)} \leq z^{(0)} + H\sum_{s=0}^{t-1} a^{(s)}[z^{(s)}]^2 + C,$$

where $H, C > 0$ and $\max_{0\leq s\leq T} a^{(s)} \leq A$. Assume that $C \leq z^{(0)}/8$. Let t_0 be the first iteration $z^{(t)} \geq v$.

- *If $v > z^{(0)}$, we have the following upper bound*

$$\sum_{s=0}^{t_0} a^{(s)} \leq \frac{8}{Hz^{(0)}} + 8A\left\lceil \frac{\log(v/z^{(0)})}{\log(2)} \right\rceil. \tag{E.11}$$

- *If $2z^{(0)} < v < 1/(2HA)$, we have the following lower bound*

$$\sum_{s=0}^{t_0} a^{(s)} \geq \frac{1}{16Hz^{(0)}}. \tag{E.12}$$

Proof. Let τ^b be the first iteration that $z^{(t)} \geq 2^b z^{(0)}$ for $b \geq 0$, that is $\tau_b = \inf\{t : z^{(t)} \geq 2^b z^{(0)}\}$. Denote $B = \left\lceil \frac{\log(v/z^{(0)})}{\log 2} \right\rceil$.

Upper bound (E.11)**.** Since $C > z^{(0)}/8$ and $a^{(t)} > 0$, we know $z^{(t)} \geq z^{(0)}/\sqrt{2}$ holds for any $t \geq 0$. Let's first consider $b = 1$. Using the recursive lower bound, we have

$$\begin{aligned} z^{(\tau_1)} &\geq z^{(0)} + H \sum_{s=0}^{\tau_1 - 1} a^{(s)} [z^{(s)}]^2 - C \\ &\geq z^{(0)} + \frac{H[z^{(0)}]^2}{2} \sum_{s=0}^{\tau_1 - 1} a^{(s)} - C, \end{aligned}$$

which implies

$$\sum_{s=0}^{\tau_1 - 1} a^{(s)} \leq \frac{z^{(\tau_1)} - z^{(0)} + C}{h[z^{(0)}]^2/2}. \tag{E.13}$$

By the recursive upper bound, we also have

$$\begin{aligned} z^{(\tau_1)} &\leq z^{(0)} + H \sum_{s=0}^{\tau_1 - 1} a^{(s)} [z^{(s)}]^2 + C \\ &= z^{(0)} + H \sum_{s=0}^{\tau_1 - 2} a^{(s)} [z^{(s)}]^2 - C + H a^{(\tau_b - 1)} [z^{((\tau_b - 1))}]^2 + 2C \\ &\leq z^{(\tau_1 - 1)} + H a^{(\tau_b - 1)} [z^{(\tau_b - 1)}]^2 + 2C \\ &\leq 2z^{(0)} + 4HA[z^{(0)}]^2 + 2C. \end{aligned} \tag{E.14}$$

Plugging (E.14) into (E.13) leads to

$$\sum_{s=0}^{\tau_1 - 1} a^{(s)} \leq \frac{2z^{(0)} + 4HA[z^{(0)}]^2 + 2C - z^{(0)} + C}{H[z^{(0)}]^2/2} \leq \frac{4}{Hz^{(0)}} + 8A. \tag{E.15}$$

For the case $b \geq 1$, we first notice that when $t \geq \tau_b$,

$$\begin{aligned} z^{(t)} &\geq z^{(\tau_b)} + H \sum_{s=\tau_b}^{t-1} a^{(s)} [z^{(s)}]^2 - 2C \\ &\geq 2^b z^{(0)} - 2C. \end{aligned}$$

It follows that

$$\begin{aligned} z^{(\tau_{b+1})} &\geq z^{(0)} + H \sum_{s=0}^{\tau_{b+1} - 1} a^{(s)} [z^{(s)}]^2 - C \\ &\geq z^{(\tau_b)} + H \sum_{s=\tau_b}^{\tau_{b+1} - 1} a^{(s)} [z^{(s)}]^2 - 2C \\ &\geq z^{(\tau_b)} + H \sum_{s=\tau_b}^{\tau_{b+1} - 1} a^{(s)} [2^b z^{(0)} - 2C]^2 - 2C, \end{aligned}$$

which implies

$$\sum_{s=\tau_b}^{\tau_{b+1} - 1} a^{(s)} \leq \frac{z^{(\tau_{b+1})} - 2^b z^{(0)} + 4C}{H[2^b z^{(0)} - 2C]^2}. \tag{E.16}$$

In addition, we have

$$z^{(\tau_{b+1})} \leq z^{(\tau_{b+1} - 1)} + H a^{(\tau_{b+1} - 1)} [z^{(\tau_{b+1} - 1)}]^2 + 2C$$

$$\leq 2^{b+1}z^{(0)} + HA[2^{b+1}z^{(0)}]^2 + 2C. \tag{E.17}$$

Plugging (E.17) into (E.16) yields

$$\begin{aligned}\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)} &\leq \frac{2^b z^{(0)} + HA[2^{b+1}z^{(0)}]^2 + 6C}{H[2^b z^{(0)} - 2C]^2} \\ &\leq \frac{2^{b+1}z^{(0)} + HA[2^{b+1}z^{(0)}]^2}{2^{2b-1}[z^{(0)}]^2} \\ &\leq \frac{4}{2^b z^{(0)} H} + 8A. \end{aligned} \tag{E.18}$$

Combining (E.15) and (E.18), we have

$$\begin{aligned}\sum_{s=0}^{t_0} a^{(s)} \leq \sum_{b=0}^{B-1}\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)} &\leq \frac{4}{Hz^{(0)}}\sum_{b=0}^{B-1}\frac{1}{2^{b-1}} + 8AB \\ &\leq \frac{8}{Hz^{(0)}} + 8A\left\lceil \frac{\log(v/z^{(0)})}{\log(2)} \right\rceil. \end{aligned}$$

Lower bound (E.12). Let's first consider $b = 1$. Using the recursive upper bound, we have

$$\begin{aligned} z^{(\tau_1)} &\leq z^{(0)} + H\sum_{s=0}^{\tau_1-1} a^{(s)}[z^{(s)}]^2 + C \\ &\leq z^{(0)} + 4H\sum_{s=0}^{\tau_1-1} a^{(s)}[z^{(0)}]^2 + C, \end{aligned} \tag{E.19}$$

which implies that

$$\sum_{s=0}^{\tau_1-1} a^{(s)} \geq \frac{z^{(\tau_1)} - z^{(0)} - C}{4H[z^{(0)}]^2} \geq \frac{z^{(0)} - C}{4H[z^{(0)}]^2} \geq \frac{1}{8Hz^{(0)}}. \tag{E.20}$$

When $b \geq 1$, we have

$$\begin{aligned} z^{(\tau_{b+1})} &\leq z^{(0)} + H\sum_{s=0}^{\tau_{b+1}-1} a^{(s)}[z^{(s)}]^2 + C \\ &= z^{(0)} + H\sum_{s=0}^{\tau_b-1} a^{(s)}[z^{(s)}]^2 - C + H\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)}[z^{(s)}]^2 + 2C \\ &\leq z^{(\tau_b)} + H\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)}[z^{(s)}]^2 + 2C \\ &\leq z^{(\tau_b)} + H\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)}[2^{b+1}z^{(0)}]^2 + 2C, \end{aligned}$$

which implies that

$$\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)} \geq \frac{z^{(\tau_{b+1})} - z^{(\tau_b)} - 2C}{H[2^{b+1}z^{(0)}]^2}. \tag{E.21}$$

In addition, notice that

$$z^{(\tau_b)} \leq z^{(0)} + H\sum_{s=0}^{\tau_b-1} a^{(s)}[z^{(s)}]^2 + C$$

$$
\begin{aligned}
&\le z^{(0)} + H\sum_{s=0}^{\tau_b-2} a^{(s)}[z^{(s)}]^2 - C + Ha^{(\tau_b-1)}[z^{(\tau_b-1)}]^2 + 2C \\
&\le z^{(\tau_b-1)} + Ha^{(\tau_b-1)}[z^{(\tau_b-1)}]^2 + 2C \\
&\le 2^b z^{(0)} + HA[2^b z^{(0)}]^2 + 2C \\
&\le 2^b z^{(0)} + 2^{b-1} z^{(0)} + 2C, \qquad \text{(E.22)}
\end{aligned}
$$

where we used the assumption $HAv \le 1/2$ and the fact $2^b z^{(0)} \le v$. Plugging (E.22) into (E.21) gives

$$
\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)} \ge \frac{2^{b-1}z^{(0)} - 4C}{H[2^{b+1}z^{(0)}]^2} \ge \frac{1}{16Hz^{(0)}} \cdot \frac{1}{2^b}. \qquad \text{(E.23)}
$$

Combining (E.20) and (E.23) leads to

$$
\sum_{s=0}^{t_0} a^{(s)} = \sum_{b=0}^{B-1}\sum_{s=\tau_b}^{\tau_{b+1}-1} a^{(s)} \ge \frac{1}{16Hz^{(0)}} \cdot \sum_{b=0}^{B-1}\frac{1}{2^b} \ge \frac{1}{16Hz^{(0)}}.
$$

□

E.3.2. Competition of the growth between two sequences

Lemma E.10. *Let $\{x^{(t)}\}_{t\ge 0}$ and $\{y^{(t)}\}_{t\ge 0}$ be two positive sequences defined by the following recursions*

$$
\begin{aligned}
x^{(t+1)} &= x^{(t)} + H[x^{(t)}]^2, \\
y^{(t+1)} &= y^{(t)} + H[y^{(t)}]^2,
\end{aligned}
$$

where $H > 0$. If $x^{(0)} \ge y^{(0)}$, we can guarantee that $x^{(t)} \ge y^{(t)}$ holds for any $t \ge 0$.

Proof. The conclusion is trivial. □

Lemma E.11. *Let $\{x^{(t)}\}_{t\ge 0}$ and $\{y^{(t)}\}_{t\ge 0}$ be two positive sequences defined by the following recursions*

$$
\begin{aligned}
x^{(t)} &\ge x^{(0)} + A\sum_{s=0}^{t-1}[x^{(s)}]^2 - C, \\
y^{(t)} &\le y^{(0)} + A\sum_{s=0}^{t-1}[y^{(s)}]^2 + C,
\end{aligned}
$$

where $A, C > 0$. If $x^{(0)} \ge 2y^{(0)}$ and $4C \le x^{(0)}$, we can guarantee that $x^{(t)} \ge y^{(t)}$ holds for any $t \ge 0$.

Proof. We first verify the conclusion at $t = 1$. Using the recursive bounds of $y^{(1)}$ and $x^{(1)}$, we have

$$
\begin{aligned}
y^{(1)} &\le y^{(0)} + A[y^{(0)}]^2 + C \\
&\le \frac{x^{(0)}}{2} + \frac{A}{4}[x^{(0)}]^2 + C \\
&\le x^{(0)} + A[x^{(0)}]^2 - C - \frac{x^{(0)}}{2} + 2C \\
&\le x^{(1)}. \qquad \text{(E.24)}
\end{aligned}
$$

Suppose $x^{(s)} \ge y^{(s)}$ holds for any $s \le t$, then we have

$$
y^{(t+1)} \le y^{(0)} + A\sum_{s=0}^{t}[y^{(s)}]^2 + C
$$

$$
\begin{aligned}
&\leq \frac{x^{(0)}}{2} + A\sum_{s=0}^{t}[x^{(s)}]^2 + C \\
&= x^{(0)} + A\sum_{s=0}^{t}[x^{(s)}]^2 - C - \frac{x^{(0)}}{2} + 2C \\
&\leq x^{(t+1)}. \qquad \text{(E.25)}
\end{aligned}
$$

Combining (E.24) and (E.25), we can finish the proof. □

Lemma E.12. *Let $x^{(0)}, y^{(0)} \leq \frac{1}{\mathsf{poly}(d)}$ and $\{x^{(t)}\}_{t\geq 0}$ and $\{y^{(t)}\}_{t\geq 0}$ be two positive sequences updated as*

$$
\begin{aligned}
x^{(t+1)} &= x^{(t)} + \eta_x \cdot C_t[x^{(t)}]^2, \\
y^{(t+1)} &= y^{(t)} + \eta_y \cdot C_t[y^{(t)}]^2,
\end{aligned}
$$

where $C_t = \Theta(1)$, $\eta_x, \eta_y \leq \widetilde{O}(1)$. For every $v \in (x^{(0)}, O(1)]$, let τ_x be the first iteration such that $x^{(t)} \geq v$. If $y^{(0)}/x^{(0)} \leq O\left(\log^{\varphi}(d)\right)$ and $\eta_x/\eta_y \geq \Omega(\log^{\varrho}(d))$ for $\varrho > \varphi > 0$, then we must have $y_{\tau_x} \leq 2y^{(0)}$.

Proof. Applying Lemma E.4 to $\{x^{(t)}\}$ with $H, h = \Theta(\eta_x)$, we can guarantee

$$
\begin{aligned}
\tau_x &\leq O\left(\frac{1}{\eta_x x^{(0)}}\right) + O\left(\log\left[\frac{v}{x^{(0)}}\right]\right) \\
&\overset{(i)}{\leq} O\left(\frac{\log^{\varphi}(d)}{\eta_x y^{(0)}}\right) + O\left(\log\left[\frac{v}{x^{(0)}}\right]\right) \\
&\overset{(ii)}{\leq} O\left(\frac{\log^{\varphi-\varrho}(d)}{\eta_y y^{(0)}}\right) + O\left(\log\left[\frac{A}{x^{(0)}}\right]\right) \\
&\overset{(iii)}{=} O\left(\frac{\log^{\varphi-\varrho}(d)}{\eta_y y^{(0)}}\right), \qquad \text{(E.26)}
\end{aligned}
$$

where (i) follows from the initial condition $y^{(0)}/x^{(0)} \leq c\log d$; (ii) holds due to $\eta_x/\eta_y \geq \log^{\varrho}(d)$; and (iii) holds due to the assumption $x^{(0)} \leq 1/\mathsf{poly}(d)$. Now denote τ_y be the first iteration such that $y_t \geq 2y_0$. Applying Lemma E.5 to $\{y^{(t)}\}_{t\geq 0}$ with $H, h = \Theta(\eta_y)$, we have

$$
\tau_y \geq O\left(\frac{1}{\eta_y y^{(0)}}\right). \qquad \text{(E.27)}
$$

Comparing (E.26) and (E.27), together with the fact $\varrho > \varphi$, we can guarantee that $\tau_y \geq \tau_x$. Hence it holds that $y^{(\tau_x)} \leq 2y^{(0)}$. □

F. Additional experiments

In this section, we provide an additional experimental result, which is a comparison of Local SGD and Parallel SGD with larger values of I. All details of the experimental setup are identical to the CIFAR-10 experiments in Section 8, other than the setting of I. In Section 8, we evaluated $I \in \{8, 16, 32\}$. Here, we additionally show results for $I \in \{64, 256, 1024\}$, in order to understand whether the optimization/generalization performance of Local SGD degrades for extremely large I.

Note that we run the same number of epochs for all values of I, in order to stay consistent with the main paper. As a result, not all of the training runs reach 100% training accuracy, since extremely large I might require more steps to reach the same training accuracy. All runs for $I \in \{64, 256\}$ reach at least 98.3% training accuracy, and all runs for $I = 1024$ reach at least 95% training accuracy.

The test accuracy and train accuracy of all values of I are shown in Tables 2 and 3, respectively. Across the four settings, test accuracy with $I \in \{64, 256\}$ is nearly as good or better than the smaller values of I that we originally evaluated in the paper, which shows that the generalization benefit of local steps occurs even for large I. For the largest $I = 1024$, the test accuracy is consistently the lowest compared to Local SGD with other I, implying that performance may degrade when

	CIFAR-10 (with augmentation)		CIFAR-10 (no augmentation)	
Algorithm	$h = 0.25$	$h = 0.5$	$h = 0.25$	$h = 0.5$
Parallel SGD	90.17 ± 0.19	90.17 ± 0.19	77.73 ± 0.20	77.73 ± 0.20
Local SGD ($I = 8$)	91.01 ± 0.17	90.71 ± 0.25	80.35 ± 0.14	80.45 ± 0.66
Local SGD ($I = 16$)	$\mathbf{91.21 \pm 0.25}$	90.84 ± 0.07	80.64 ± 0.12	80.77 ± 0.30
Local SGD ($I = 32$)	91.19 ± 0.22	$\mathbf{91.08 \pm 0.25}$	80.86 ± 0.17	81.27 ± 0.36
Local SGD ($I = 64$)	91.14 ± 0.04	90.70 ± 0.07	$\mathbf{81.89 \pm 0.26}$	$\mathbf{81.58 \pm 0.25}$
Local SGD ($I = 256$)	91.19 ± 0.12	90.53 ± 0.18	81.48 ± 0.49	81.04 ± 0.13
Local SGD ($I = 1024$)	90.49 ± 0.14	89.84 ± 0.30	80.10 ± 0.22	79.66 ± 0.07

Table 2: Average test accuracy over three trials, with and without data augmentation and varying $h \in \{0.25, 0.5\}$. The error is the distance from the average to the max/min across three runs. Note that Parallel SGD is unaffected by h, since it does not utilize local steps.

	CIFAR-10 (with augmentation)		CIFAR-10 (no augmentation)	
Algorithm	$h = 0.25$	$h = 0.5$	$h = 0.25$	$h = 0.5$
Parallel SGD	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Local SGD ($I = 8$)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Local SGD ($I = 16$)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Local SGD ($I = 32$)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Local SGD ($I = 64$)	99.78 ± 0.06	99.71 ± 0.05	100.0 ± 0.0	100.0 ± 0.0
Local SGD ($I = 256$)	99.01 ± 0.15	98.56 ± 0.23	100.0 ± 0.0	100.0 ± 0.0
Local SGD ($I = 1024$)	97.47 ± 0.15	96.07 ± 0.18	96.46 ± 0.32	95.28 ± 0.19

Table 3: Average train accuracy over three trials, with and without data augmentation and varying $h \in \{0.25, 0.5\}$. The error is the distance from the average to the max/min across three runs. Note that Parallel SGD is unaffected by h, since it does not utilize local steps.

I becomes extremely large. However, even with $I = 1024$, the test accuracy is still higher than that of Parallel SGD in three of the four settings, showing that the generalization boost from local steps is still present even with extremely large I. Also, we include the training accuracy in Table 3, to consider the fact that Local SGD with large I did not reach 100% training accuracy. Even with this slight decrease in training performance, the testing accuracy of Local SGD with large I is consistently better than Parallel SGD.

To summarize, the generalization benefit of local steps in this setting persists for extremely large I, even up to $I = 1024$. Although the test accuracy of Local SGD with $I = 1024$ is smaller than that of $I \in \{8, 16, 32, 64, 256\}$, it is still larger than that of Parallel SGD. This indicates that the local steps help generalization in real-world datasets even if the number of local steps is large, which is consistent with our theoretical results.