FL-NTK: A Neural Tangent Kernel-based Framework for Federated Learning Convergence Analysis*

Baihe Huang† Xiaoxiao Li‡ Zhao Song§ Xin Yang¶

Abstract

Federated Learning (FL) is an emerging learning scheme that allows different distributed clients to train deep neural networks together without data sharing. Neural networks have become popular due to their unprecedented success. To the best of our knowledge, the theoretical guarantees of FL concerning neural networks with explicit forms and multi-step updates are unexplored. Nevertheless, training analysis of neural networks in FL is non-trivial for two reasons: first, the objective loss function we are optimizing is non-smooth and non-convex, and second, we are even not updating in the gradient direction. Existing convergence results for gradient descent-based methods heavily rely on the fact that the gradient direction is used for updating. This paper presents a new class of convergence analysis for FL, Federated Learning Neural Tangent Kernel (FL-NTK), which corresponds to overparamterized ReLU neural networks trained by gradient descent in FL and is inspired by the analysis in Neural Tangent Kernel (NTK). Theoretically, FL-NTK converges to a global-optimal solution at a linear rate with properly tuned learning parameters. Furthermore, with proper distributional assumptions, FL-NTK can also achieve good generalization.

*A preliminary version of this paper appeared in the Proceedings of the 38th International Conference on Machine Learning (ICML 2021).

†baihehuang@pku.edu.cn. Peking University.

‡xl32@princeton.edu. Princeton University.

§zhaos@princeton.edu. Princeton University.

¶yx1992@cs.washington.edu. The University of Washington.

Contents

1 Introduction

In traditional centralized training, deep learning models learn from the data, and data are collected in a database on the centralized server. In many fields, such as healthcare and natural language processing, models are typically learned from personal data. These personal data are subject to regulations such as California Consumer Privacy Act (CCPA) [Leg18], Health Insurance Portability and Accountability Act (HIPAA) [Act96], and General Data Protection Regulation (GDPR) of European Union. Due to the data regulations, standard centralized learning techniques are not appropriate, and users are much less likely to share data. Thus, the data are only available on the local data owners (i.e. edge devices). Federated learning (FL) is a new type of learning scheme that avoids centralizing data in model training. FL allows local data owners (also known as clients) to locally train the private model and then send the model weights or gradients to the central server. Then central server aggregates the shared model parameters to update new global model, and broadcasts the the parameters of global model to each local client.

Quite different from the centralized training, FL has the following unique properties. First, the training data are distributed on an astonishing number of devices, and the connection between the central server and the device is slow. Thus, the computational cost is a key factor in FL. In communication-efficient FL, local clients are required to update model parameters for a few steps locally then send their parameters to the server [MMR+17]. Second, due to the fact that the data are collected from different clients, the local data points can be sampled from different local distributions. When this happens during training, convergence may not be guaranteed.

The above two unique properties not only bring challenges to algorithm design but also make theoretical analysis much harder. There have been many efforts developing convergence guarantees for FL algorithms based on the assumptions of convexity and smoothness for the objective functions [YYZ19, LHY+20, KMR20]. Although a recent study [LJZ+21] shows theoretical studies of FL on neural networks, its framework fails to generate multiple-local updates analysis, which is a key feature in FL. One of the most conspicuous questions to ask is:

Can we build a unified and generalizable convergence analysis framework for ReLU neural networks in FL?

In this paper, we give an affirmative answer to this question. It was recently proved in a series of papers that gradient descent converges to global optimum if the neural networks are overparameterized (significantly larger than the size of the datasets). Our work is motivated by Neural Tangent Kernel (NTK) that is originally proposed by [JGH18] and has been extensively studied over the last few years.

NTK is defined as the inner product space of pairwise data point gradient (aka Gram matrix). It describes the evolution of deep artificial neural networks during their training by gradient descent. Thus, we propose a novel NTK-based framework for federated learning convergence and generalization analysis on ReLU neural networks, so-called Federated Learning Neural Tangent Kernel(FL-NTK). Unlike the good property of the symmetric Gram matrix in classical NTK, we show the Gram matrix of FL-NTK is asymmetric in Section 3.2. Our techniques address the core question:

How shall we handle the asymmetric Gram matrix in FL-NTK?

Contributions. Our contributions are summarized into the following two folds:

- We proposed a framework to analyze federated learning in neural networks. By appealing to recent advances of over-parameterized neural networks, we prove convergence and generalization

results of federated learning without the assumptions on the convexity of objective functions or distribution of data in the convergence analysis. Thus, we make the first step toward bridging the gap between the empirical success of federated learning and its theoretical understanding in the settings of ReLU neural networks. The results theoretically show that given fixed training instances, the number of communication rounds increases as the number of clients increases, which is also supported by empirical evidence. We show that when the neural networks are sufficiently wide, the training loss across all clients converges to zero at a linear rate. Furthermore, we also prove a data-dependent generalization bound.

- In federated learning, the update in the global model is no longer determined by the gradient directions directly. Indeed, gradients' heterogeneity in multiple local steps hinders the usage of standard neural tangent kernel analysis, which is based on the kernel gradient descent in the function space for a positive semi-definite kernel. We identify the dynamics of training loss by considering all intermediate states of local steps and establishing the tangent kernel space associated with a general non-symmetric Gram matrix to address this issue. We prove that this Gram matrix is close to symmetric at initialization using concentration properties at initialization. Therefore, we guarantee linear convergence results. This technique may further improve our understanding of many different FL optimization and aggregation methods on neural networks.

Organization. In Section 2 we discuss related work. In Section 3 we formulate FL convergence problem. In Section 4 we state our result. In Section 5 we summarize our technique overviews. In Section 6 we give a proof sketch of our result. In Section 7, we conduct experiments that affirmatively support our theoretical results. In Section 8 we conclude this paper and discuss future works.

2 Related Work

Federated Learning Federated learning has emerged as an important paradigm in distributed deep learning. Generally, federated learning can be achieved by two approaches: 1) each party training the model using private data and where only model parameters being transferred and 2) using encryption techniques to allow safe communications between different parties [YLCT19]. In this way, the details of the data are not disclosed in between each party. In this paper, we focus on the first approach, which has been studied in [DCM+12, SS15, MMRAyA16, MMR+17]. Federated average (FedAvg) [MMR+17] firstly addressed the communication efficiency problem by introducing a global model to aggregate local stochastic gradient descent updates. Later, different variations and adaptations have arisen. This encompasses a myriad of possible approaches, including developing better optimization algorithms [LSZ+20, WYS+20] and generalizing model to heterogeneous clients under special assumptions [ZLL+18, KMA+21, LJZ+21].

Federated learning has been widely used in different fields. Healthcare applications have started to utilize FL for multi-center data learning to solve small data, and privacy in data sharing issues [LGD+20, RHL+20, LMX+19, AdTBT20]. We have also seen new FL algorithms popping up [WTS+19, LLH+20, CYS+20] in mobile edge computing. FL also has promising applications in autonomous driving [LLC+19], financial filed [YZY+19], and so on.

Convergence of Federated Learning Despite its promising benefits, FL comes with new challenges to tackle, especially for its convergence analysis under communication-efficiency algorithms and data heterogeneity. The convergence of the general FL framework on neural networks is underexplored. A recent work [LJZ+21] studies FL convergence on one-layer neural networks. Nevertheless,

it is limited by the assumption that each client performs a single local update epoch. Another line of approaches does not directly work on neural network setting [LHY+20, KMR20, YYZ19]. Instead, they make assumptions on the convexity and smoothness of the objective functions, which are not realistic for non-linear neural networks.

3 Problem Formulation as Neural Tangent Kernel

To capture the training dynamic of FL on ReLU neural networks, we formulate the problem in Neural Tangent Kernel regime.

Notations We use N to denote the number of clients and use c to denote its index. We use T to denote the number of communication rounds, and use t to denote its index. We use K to denote the number of local update steps, and we use k to denote its index. We use $u(t)$ to denote the aggregated server model after round t. We use $w_{k,c}(t)$ to denote c-th client's model in round t and step k.

Let $S_1 \cup S_2 \cup \cdots \cup S_N = [n]$ and $S_i \cap S_j = \emptyset$. Given n input data points and labels $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ in $\mathbb{R}^d \times \mathbb{R}$, the data of each client c is $\{(x_i, y_i) : i \in S_c\}$. Let $\phi(z) = \max\{z, 0\}$ denote the ReLU activation.

For each client $c \in [N]$, we use $y_c \in \mathbb{R}^{|S_c|}$ to denote the ground truth with regard to its data, and denote $y_c^{(k)}(t) \in \mathbb{R}^{|S_c|}$ to be the (local) model's output of its data in the t-th global round and k-th local step. For simplicity, we also use $y^{(k)}(t) \in \mathbb{R}^n$ to denote aggregating all (local) model's outputs in the t-th global round and k-th local step.

3.1 Preliminaries

In this subsection we introduce Algorithm 1, the brief algorithm of our federated learning (under NTK setting):

- In the t-th global round, server broadcasts $u(t) \in \mathbb{R}^{d\times m}$ to every client.
- Each client c then starts from $w_{0,c}(t) = u(t)$ and takes K (local) steps gradient descent to find $w_{K,c}(t)$.
- Each client sends $\Delta u_c(t) = w_{K,c}(t) - w_{0,c}(t)$ to server.
- Server computes a new $u(t+1)$ based on the average of all $\Delta u_c(t)$. Specifically, the server updates $u(t)$ by the average of all local updates $\Delta u_c(t)$ and arrives at $u(t+1) = u(t) + \eta_{\text{global}} \cdot \sum_{c\in[N]} \Delta u_c(t)/N$.
- We repeat the above four steps by T times.

Setup We define one-hidden layer neural network function $f : \mathbb{R}^d \to \mathbb{R}$ similar as [DZPS19, SY19, BPSW21, SZ21]

$$f(u, x) := \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \phi(u_r^\top x),$$

where $u \in \mathbb{R}^{d\times m}$ and $u_r \in \mathbb{R}^d$ is the r-th column of matrix u.

Definition 3.1 (Initialization)**.** *We initialize $u \in \mathbb{R}^{d\times m}$ and $a \in \mathbb{R}^m$ as follows:*

Algorithm 1 Training Neural Network with FedAvg under NTK setting.

$u_r(0) \sim \mathcal{N}(0, I_d)$ for $r \in [m]$. ▷ $u \in \mathbb{R}^{d \times m}$
for $t = 1, \ldots, T$ **do**
  for $c = 1, \ldots, N$ **do**
    $w_{0,c}(t) \leftarrow u(t)$ ▷ $w_{0,c}(t), u(t) \in \mathbb{R}^{d \times m}$
    for $k = 1, \ldots, K$ **do**
      for $i \in S_c$ **do**
        $y_c^{(k)}(t)_i \leftarrow \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \phi(w_{k,c,r}(t)^\top x_i)$ ▷ $y_c^{(k)}(t) \in \mathbb{R}^{|S_c|}$
      end for
      for $r = 1 \to m$ **do**
        for $i \in S_c$ **do**
          $J_{i,:} \leftarrow \frac{1}{\sqrt{m}} a_r \phi'(w_{k,c,r}(t)^\top x_i) x_i^\top$ ▷ $J_{i,:} = \frac{\partial f(w_{k,c}(t), x_i, a)}{\partial w_{k,c,r}(t)} \in \mathbb{R}^{1 \times d}$
        end for
        $\mathrm{grad}_r \leftarrow -J^\top (y_c - y_c^{(k)}(t))$ ▷ $J \in \mathbb{R}^{|S_c| \times d}$
        $w_{k,c,r}(t) \leftarrow w_{k-1,c,r}(t) - \eta_{\mathrm{local}} \cdot \mathrm{grad}_r$
      end for
    end for
    $\Delta u_c \leftarrow w_{k,c}(t) - u(t)$
  end for
  $\Delta u \leftarrow \frac{1}{N} \sum_{c \in [N]} \Delta u_c$ ▷ $\Delta u \in \mathbb{R}^{d \times m}$
  $u(t+1) \leftarrow u(t) + \eta_{\mathrm{global}} \Delta u$ ▷ $u(t+1) \in \mathbb{R}^{d \times m}$
end for

- *For each $r \in [m]$, u_r is sampled from $\mathcal{N}(0, \sigma^2 I)$.*
- *For each $r \in [m]$, a_r is sampled from $\{-1, +1\}$ uniformly at random (we don't need to train).*

We define the loss function for $j \in [N]$,

$$L_j(u, x) := \frac{1}{2} \sum_{i \in S_j} (f(u, x_i) - y_i)^2,$$

$$L(u, x) := \frac{1}{N} \sum_{j=1}^N L_j(u, x).$$

We can compute the gradient $\frac{\partial f(u,x)}{\partial u_r} \in \mathbb{R}^d$ (of function f),

$$\frac{\partial f(u, x)}{\partial u_r} = \frac{1}{\sqrt{m}} a_r x \mathbf{1}_{u_r^\top x \geq 0}.$$

We can compute the gradient $\frac{\partial L_c(u)}{\partial u_r} \in \mathbb{R}^d$ (of function L),

$$\frac{\partial L_c(u)}{\partial u_r} = \frac{1}{\sqrt{m}} \sum_{i=1}^n (f(u, x_i) - y_i) a_r x_i \mathbf{1}_{u_{r,c}^\top x_i \geq 0}.$$

We formalize the problem as minimizing the sum of loss functions over all clients:

$$\min_{u \in \mathbb{R}^{d \times M}} L(u).$$

Local update In each local step, we update $w_{k,c}$ by gradient descent.

$$w_{k+1,c} \leftarrow w_{k,c} - \eta_{\text{local}} \cdot \frac{\partial L_c(w_{k,c})}{\partial w_{k,c}}.$$

Note that

$$\frac{\partial L_c(w_{k,c})}{\partial w_{k,c,r}} = \frac{1}{\sqrt{m}} \sum_{i \in S_c} (f(w_{k,c}, x_i) - y_i) a_r x_i \mathbf{1}_{w_{k,c,r}^\top x_i \geq 0}.$$

Global aggregation In each global communication round we aggregate all local updates of clients by taking a simple average

$$\Delta u(t) = \sum_{c \in [N]} \Delta u_c(t)/N,$$

where $\Delta u_c(t) = w_{K,c} - w_{0,c}$ for all $c \in [N]$.

Global steps in total Then global model simply add $\Delta u(t)$ to its parameters.

$$u(t+1) \leftarrow u(t) + \eta_{\text{global}} \cdot \Delta u(t).$$

3.2 NTK Analysis

The neural tangent kernel $H^\infty \in \mathbb{R}^{n \times n}$, introduced in [JGH18], is given by

$$H_{i,j}^\infty := \mathop{\mathbb{E}}_{w \sim \mathcal{N}(0,I)} \left[x_i^\top x_j \mathbf{1}_{w^\top x_i \geq 0, w^\top x_j \geq 0} \right]$$

At round t, let $y(t) = (y_1(t), y_2(t), \cdots, y_n(t)) \in \mathbb{R}^n$ be the prediction vector where $y_i(t) \in \mathbb{R}$ is defined as

$$y_i(t) = f(u(t), x_i).$$

Recall that we denote labels $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$, $y_c = \{y_i\}_{i \in S_c}$, predictions $y(t) = (y(t)_1, \cdots, y(t)_n)$ and $y_c(t) = \{y(t)_i\}_{i \in S_c}$, we can then rewrite $\|y - y(t)\|_2^2$ as follows:

$$\begin{aligned}
&\|y - y(t+1)\|_2^2 \\
=& \|y - y(t) - (y(t+1) - y(t))\|_2^2 \\
=& \|y - y(t)\|_2^2 - 2(y - y(t))^\top (y(t+1) - y(t)) \ + \|y(t+1) - y(t)\|_2^2. \qquad (1)
\end{aligned}$$

Now we focus on $y(t+1) - y(t)$, notice for each $i \in [n]$

$$\begin{aligned}
&y_i(t+1) - y_i(t) \\
=& \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r (\phi(u_r(t+1)^\top x_i) - \phi(u_r^\top(t) x)) \\
=& \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \left(\phi\big((u_r(t+1) + \eta_{\text{global}} \Delta u_r(t))^\top x_i\big) - \phi(u_r^\top(t) x) \right) \qquad (2)
\end{aligned}$$

where

$$\Delta u_r(t) := \frac{a_r}{N} \sum_{c\in[N]} \sum_{k\in[K]} \frac{\eta_{\text{local}}}{\sqrt{m}} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j) x_j \mathbf{1}_{w_{k,c,r}(t)^\top x_j \geq 0}.$$

In order to further analyze Eq (2), we separate neurons into two sets. One set contains neurons with the activation pattern changing over time and another set contains neurons with activation pattern holding the same. Specifically for each $i \in [n]$, we define the set $Q_i \subset [m]$ of neurons whose activation pattern is certified to hold the same throughout the algorithm

$$Q_i := \Big\{ r \in [m] : \forall w \in \mathbb{R}^d \text{ s.t. } \|w - w_r(0)\|_2 \leq R, \ \ \mathbf{1}_{w_r(0)^\top x_i \geq 0} = \mathbf{1}_{w^\top x_i \geq 0} \Big\},$$

and use $\overline{Q}_i$ to denote its complement. Then $y_i(t+1) - y_i(t) = v_{1,i} + v_{2,i}$ where

$$\begin{aligned} v_{1,i} &= \frac{1}{\sqrt{m}} \sum_{r\in Q_i} a_r \Big(\phi\big((u_r(t) + \eta_{\text{global}} \Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i) \Big), \\ v_{2,i} &= \frac{1}{\sqrt{m}} \sum_{r\in \overline{Q}_i} a_r \Big(\phi\big((u_r(t) + \eta_{\text{global}} \Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i) \Big). \end{aligned} \tag{3}$$

The benefit of this procedure is that v_1 can be written in closed form

$$v_{1,i} = \frac{\eta_{\text{global}} \eta_{\text{local}}}{Nm} \sum_{k\in[K], c\in[N]} \sum_{j\in S_c} \sum_{r\in Q_i} q_{k,c,j,r},$$

where

$$q_{k,c,j,r} := -(y_c^{(k)}(t)_j - y_j) x_i^\top x_j \mathbf{1}_{w_{k,c,r}(t)^\top x_j, u_r(t)^\top x_i \geq 0},$$

and v_2 is sufficiently small which we will show later.

Now, we extend the NTK analysis to FL. We start with defining the Gram matrix for f as follows.

Definition 3.2. *For any $t \in [0,T], k \in [K], c \in [N]$, we define matrix $H(t,k,c) \in \mathbb{R}^{n\times n}$ as follows*

$$\begin{aligned} H(t,k,c)_{i,j} &= \frac{1}{m} \sum_{r=1}^{m} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}, \\ H(t,k,c)^{\perp}_{i,j} &= \frac{1}{m} \sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}. \end{aligned}$$

This Gram matrix is crucial for the analysis of error dynamics. When $t = 0$ and the width m approaches infinity, the H matrix becomes the NTK, and with infinite width, neural networks just behave like kernel methods with respect to the NTK [ADH+19b, LSS+20]. It turns out that in the finite width case [LL18, AZLS19a, AZLS19b, DZPS19, DLL+19, SY19, OS20, HY20, CX20, ZPD+20, BPSW21, SZ21], the spectral property of the gram matrix also governs convergence guarantees for neural networks.

We can then decompose $-2(y - y(t))^\top (y(t+1) - y(t))$ into

$$-2(y - y(t))^\top (y(t+1) - y(t)) = \ -2(y - y(t))^\top (v_1 + v_2)$$

$$= -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c} p_{i,k,c,j} - 2\sum_{i\in[n]}(y_i - y_i(t))v_{2,i} \tag{4}$$

where

$$p_{i,k,c,j} := (y_i - y_i(t)) \cdot (y_j - y_c^{(k)}(t)_j) \cdot (H(t,k,c)_{i,j} - H(t,k,c)^{\perp}_{i,j}).$$

Now it remains to analyze Eq (1) and Eq (4). Our analysis leverages several key observations in the classical Neural Tangent Kernel theory throughout the learning process:

- Weights change lazily, namely

$$\|u(t+1) - u(t)\|_2 \le O(1/n).$$

- Activation patterns remain roughly the same, namely

$$\|H(t,k,c)^{\perp}\|_F \le O(1),$$

  and

$$\|v_2\|_2 \le O(\|y - y(t)\|_2).$$

- Error controls model updates, namely

$$\|y(t+1) - y(t)\|_2 \le O(\|y - y(t)\|_2).$$

Based on the above observations, we show that the dynamics of federated learning is dominated in the following way

$$\|y - y(t+1)\|_2^2 \approx \|y - y(t)\|_2^2 - 2\sum_{k\in[K]}(y - y(t))^\top H(t,k)(y - y^{(k)}(t)),$$

where the Gram matrix $H(t,k) \in \mathbb{R}^{n\times n}$ comes from combining the S_c columns of $H(t,k,c)$ for all $c \in [N]$. Since $S_1 \cup S_2 \cup \cdots \cup S_N = [n]$ and $S_i \cap S_j = \emptyset$, every $j \in [n]$ belongs to one unique S_c for some c and $H(t,k)_{i,j} = H(t,k,c)_{i,j}, j \in S_c$.

There are two difficulties lies in the analysis of these dynamics. First, unlike the symmetric Gram matrix in the standard NTK theory for centralized training, our FL framework's Gram matrix is asymmetric. Secondly, model update in each global round is influenced by all intermediate model states of all the clients.

To address these difficulties, we bring in two new techniques to facilitate our understanding of the learning dynamics.

- First, we generalize Theorem 4.2 in [DZPS19] to non-symmetric Gram matrices. We show in Lemma B.11 that with good initialization $H(t,k)$ is close to the original Gram matrix $H(0)$, so that model could benefit from a linear learning rate determined by the smallest eigenvalue of $H(0)$.

- Secondly, we leverage concentration properties at initialization to bound the difference between the errors in local steps and the errors in the global step. Specifically, we can show that $y - y^{(k)}(t) \approx y - y(t)$ for all $k \in [K]$.

4 Our Results

We first present the main result on the convergence of federated learning in neural networks by the following theorem.

Theorem 4.1 (Informal version of Theorem B.3)**.** *Let $m = \Omega(\lambda^{-4} n^4 \log(n/\delta))$, we iid initialize $u_r(0)$, a_r as Definition 3.1 where $\sigma = 1$. Let λ denote $\lambda_{\min}(H(0))$. Let κ denote the condition number of $H(0)$. For N clients, for any ϵ, let*

$$T = O\Big(\frac{N}{\lambda \eta_{\text{local}} \eta_{\text{global}} K} \cdot \log(1/\epsilon)\Big),$$

there is an algorithm (FL-NTK) runs in T global steps and each client runs K local steps with choosing

$$\eta_{\text{local}} = O\Big(\frac{\lambda}{\kappa K n^2}\Big) \quad \text{and} \quad \eta_{\text{global}} = O(1)$$

and outputs weight U with probability at least $1 - \delta$ such that the training loss $L(u, x)$ satisfies

$$L(u, x) = \frac{1}{2N} \sum_{i=1}^{N} (f(u, x_i) - y_i)^2 \leq \epsilon.$$

We note that our theoretical framework is very powerful. With additional assumptions on the training data distribution and test data distribution, we can also show an upper bound for the generalization error of federated learning in neural networks. We first introduce a distributional assumption, which is standard in the literature (e.g, see [ADH+19a, SY19]).

Definition 4.2 (Non-degenerate Data Distribution, Definition 5.1 in [ADH+19a])**.** *A distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$ is (λ, δ, n)-*non-degenerate*, if with probability at least $1 - \delta$, for n i.i.d. samples $(x_i, y_i)_{i=1}^n$ chosen from $\mathcal{D}$, $\lambda_{\min}(H(0)) \geq \lambda > 0$.*

Our result on generalization bound is stated in the following theorem.

Theorem 4.3 (Informal, see Appendix C for details)**.** *Fix failure probability $\delta \in (0, 1)$. Suppose the training data $S = \{(x_i, y_i)\}_{i=1}^n$ are i.i.d samples from a $(\lambda, \delta/3, n)$-non-degenerate distribution $\mathcal{D}$, and*

- $\sigma \leq O(\lambda \cdot \text{poly}(\log n, \log(1/\delta))/n)$,
- $m \geq \Omega(\sigma^{-2} \cdot \text{poly}(n, \log m, \log(1/\delta), \lambda^{-1}))$,
- $T \geq \Omega((\eta_{\text{local}} \eta_{\text{global}} K \lambda)^{-1} N \log(n/\delta))$.

We initialize $u \in \mathbb{R}^{d \times m}$ and $a \in \mathbb{R}^m$ as Definition 3.1. Consider any loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, 1]$ that is 1-Lipschitz in its first argument. Then with probability at least $1 - \delta$ the following event happens: after T global steps, the generalization loss

$$L_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\ell(f(u, x), y)]$$

is upper bounded as

$$L_{\mathcal{D}}(f) \leq (2 y^\top (H^\infty)^{-1} y / n)^{1/2} + O(\sqrt{\log(n/(\lambda\delta))/n}).$$

5 Techniques Overview

In our algorithm, when $K = 1$, i.e., we only perform one local step per global step, essentially we are performing gradient descent on all the n data points with step size $\eta_{\text{global}}\eta_{\text{local}}/N$. As the norm of the gradient is proportional to $1/\sqrt{m}$, when the neural network is sufficiently wide, we can control the norm of the gradient. Then by the update rule of gradient descent, we hence upper bound the movement of the first layer weight u. By anti-concentration of the normal distribution, this implies that for each input x, the activation status of most of the ReLU gates remains the same as initialized, which enables us to apply the standard convergence analysis of gradient descent on convex and smooth functions. Finally, we can handle the effect of ReLU gates whose activation status have changed by carefully choosing the step size.

However, the analysis becomes much more complicated when $K \geq 2$, where the movement of u is no longer determined by the gradient directly. Nevertheless, on each client, we are still performing gradient descent for K local steps. So we can handle the movement of the local weight w. The major technical bulk of this work is then proving that the training error shrinks when we set the global weight movement as the average of the local weight. Our argument is inspired by that of [DZPS19] but is much more involved.

6 Proof Sketch

In this section we sketch our proof of Theorem 4.1 and Theorem 4.3. The detailed proof is deferred to Appendix B.

In order to prove the linear convergence rate in Theorem 4.1, it is sufficient to show that the training loss shrinks in each round, or formally for each $\tau = 0, 1, \cdots$,

$$\|y(\tau+1) - y\|_2^2 \leq \Big(1 - \frac{\lambda\eta_{\text{global}}\eta_{\text{local}}K}{2N}\Big) \cdot \|y(\tau) - y\|_2^2. \tag{5}$$

We prove Eq. (5) by induction. Assume that we have proved for $\tau \leq t-1$ and we want to prove Eq. (5) for $\tau = t$. We first show that the movement of the weight u is bounded under the induction hypothesis.

Lemma 6.1 (Movement of global weight, informal version of Lemma B.12)**.** *For any $r \in [m]$,*

$$\|u_r(t) - u_r(0)\|_2 = O\Big(\frac{\sqrt{n}\|y - y(0)\|_2}{\sqrt{m}\lambda}\Big).$$

The detailed proof can be found in Appendix B.5.

We then turn to the proof of Eq. (5) by decomposing the loss in $t+1$-th global round

$$\begin{aligned} &\|y - y(t+1)\|_2^2 \\ =\ &\|y - y(t)\|_2^2 - 2(y - y(t))^\top (y(t+1) - y(t)) + \|y(t+1) - y(t)\|_2^2 \end{aligned} \tag{6}$$

To this end, we need to investigate the change of prediction in consecutive rounds, which is described in Eq. (2). For the sake of simplicity, we introduce the notation

$$z_{i,r} := \phi\big((u_r(t) + \eta_{\text{global}}\Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i),$$

then we have

$$y(t+1)_i - y(t)_i = \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r z_{i,r} = \frac{1}{\sqrt{m}}\sum_{r\in Q_i} a_r z_{i,r} + v_{2,i},$$

where $v_{2,i}$ is introduced in Eq. (3).

For client $c \in [N]$ let $y_c^{(k)}(t)_j$ $(j \in S_c)$ be defined by $y_c^{(k)}(t)_j = f(w_{k,c}(t), x_j)$. By the gradient-averaging scheme described in Algorithm 1, $\Delta u_r(t)$, the change in the global weights is

$$\frac{a_r}{N} \sum_{c\in[N]} \sum_{k\in[K]} \frac{\eta_{\text{local}}}{\sqrt{m}} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j) x_j \mathbf{1}_{w_{k,c,r}(t)^\top x_j \geq 0}.$$

Therefore, we can calculate $\frac{1}{\sqrt{m}} \sum_{r\in Q_i} a_r z_{i,r}$

$$\begin{aligned}
&\frac{1}{\sqrt{m}} \sum_{r\in Q_i} a_r z_{i,r} \\
=& \frac{1}{\sqrt{m}} \sum_{r\in Q_i} a_r \Big(\phi\big((u_r(t) + \eta_{\text{global}} \Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i) \Big) \\
=& \frac{\eta_{\text{global}}\eta_{\text{local}}}{mN} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j) x_i^\top x_j \cdot \sum_{r\in Q_i} \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0} \\
=& \frac{\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(t,k,c)^\perp_{i,j}).
\end{aligned}$$

Further, we write $-2(y - y(t))^\top (y(t+1) - y(t))$ as follows:

$$\begin{aligned}
&-2(y - y(t))^\top (y(t+1) - y(t)) \\
=& -2(y - y(t))^\top (v_1 + v_2) \\
=& -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(t,k,c)^\perp_{i,j}) \\
&-2 \sum_{i\in[n]} (y_i - y_i(t)) v_{2,i}.
\end{aligned}$$

To summarize, we can decompose the loss as

$$\|y - y(t+1)\|_2^2 = \|y - y(t)\|_2^2 + C_1 + C_2 + C_3 + C_4.$$

where

$$\begin{aligned}
C_1 =& -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j) H(t,k,c)_{i,j}, \\
C_2 =& +\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j) H(t,k,c)^\perp_{i,j}, \\
C_3 =& -2 \sum_{i\in[n]} (y_i - y_i(t)) v_{2,i}, \\
C_4 =& +\|y(t+1) - y(t)\|_2^2.
\end{aligned}$$

Let $R = \max_{r\in[m]} \|u_r(t) - u_r(0)\|_2$ be the maximal movement of global weights. Note that by Lemma 6.1, R can be made arbitrarily small as long as the width m is sufficiently large. Next, we bound C_1, C_2, C_3, C_4, by arguing that they are bounded from above if R is small and the learning rate is properly chosen. The detailed proof is deferred to Appendix B.2.

Lemma 6.2 (Bounding of C_1, C_2, C_3, C_4, informal version of Claim B.4, Claim B.5, Claim B.6 and Claim B.7)**.** *Assume that*

- $R = O(\lambda/n)$,
- $\eta_{\text{local}} = O(1/(\kappa K n^2))$,
- $\eta_{\text{global}} = O(1)$.

Then with high probability we have

$$\begin{aligned} C_1 \leq & -\eta_{\text{global}}\eta_{\text{local}}\lambda K\|y - y(t)\|_2^2/N, \\ C_2 \leq & +\eta_{\text{global}}\eta_{\text{local}}\lambda K\|y - y(t)\|_2^2/(40N), \\ C_3 \leq & +\eta_{\text{global}}\eta_{\text{local}}\lambda K\|y - y(t)\|_2^2/(40N), \\ C_4 \leq & +\eta_{\text{global}}\eta_{\text{local}}\lambda K\|y - y(t)\|_2^2/(40N). \end{aligned}$$

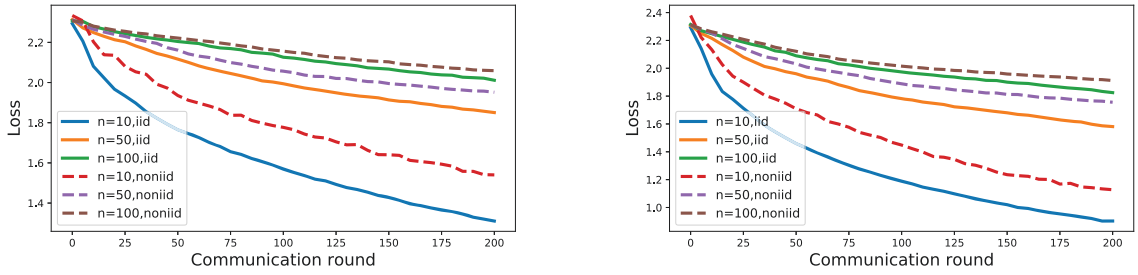
Combining the above lemma, we arrive to

$$\|y - y(t+1)\|_2^2 \leq \left(1 - \frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{2N}\right) \cdot \|y - y(t)\|_2^2,$$

which completes the proof of Eq. (5). Finally Theorem 4.1 follows from

$$\|y - y(T)\|_2^2 \leq \left(1 - \frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{2N}\right)^T \leq \epsilon.$$

7 Experiment



(a) Local update epoch K =2 (b) Local update epoch K =5

Figure 1: Training loss vs. communication rounds when number of clients $N = 10, 50, 100$ with iid and non-iid setting using mini-batch SGD optimizer.

Models and datasets We examine our theoretical results on a benchmark dataset - Cifar10 in FL study. We perform 10 class classification tasks using ResNet56 [HZRS16]. For fair convergence comparison, *we fixed the total number of samples* n. Based on our main result Theorem 4.1, we show the convergence with respect to the number of client N. To clearly evaluate the effects on N, we set all the clients to be activated (sampling all the clients) at each aggregation. We examine the settings of both non-iid and iid clients:

iid Data distribution is homogeneous in all the clients. Specifically, the label distribution over 10 classes is a uniform distribution.

non-iid Data distribution is heterogeneous in all the clients. For non-IID splits, we utilize the Dirichlet distribution as in [YAG+19, WYS+20, HAA20]. First, each of these datasets is heterogeneously divided into J batches. The heterogeneous partition allocates $p_z \sim \text{Dir}_J(\alpha)$ proportion of the instances of label z to batch j. Then one label is sampled based on these vectors for each device, and an image is sampled without replacement based on the label. For all the classes, we repeat this process until all data points are assigned to devices.

Setup The experiment is conducted on one Ti2080 Nvidia GPU. We use SGD with a learning rate of 0.03 to train the neural network for 160 communication rounds[1]. We set batch size as 128. We set the local update epoch $K = 2$ and $K = 5$[2], and Dirichlet distribution parameter $\alpha = 0.5$. There are 50,000 training instances in Cifar10. We vary $N = 10, 50, 100$ and record the training loss over communication rounds. The implementation is based on FedML [HLS+20].

Impact of N Our theory suggests that given fixed training instances (fixed n) a smaller N requires less communication rounds to converge to a given ϵ. In other words, a smaller N may accelerate the convergence of the global model. Intuitively, a large N on a fixed amount of data means a heavier communication burden. Figure 1 shows the training loss curve of different choices of N. We empirically observe that a smaller N converges faster given a fix number of total training samples, which is affirmative to our theoretical results.

8 Discussion

Overall, we provide the first comprehensive proof of convergence of gradient descent and generalization bound for over-parameterized ReLU neural network in federated learning. We consider the training dynamic with respect to local update steps K and number of clients N.

Different from most of existing theoretical analysis work on FL, FL-NTK prevails over the ability to exploit neural network parameters. There is a great potential for FL-NTK to understand the behavior of different neural network architectures, i.e., how Batch Normalization affects convergence in FL, like FedBN [LJZ+21]. Different from FedBN, whose analysis is limited to $K = 1$, we provide the first general framework for FL convergence analysis by considering $K \geq 2$. The extension is non-trivial, as parameter update does not follow the gradient direction due to the heterogeneity of local data. To tackle this issue, we establish the training trajectory by considering all intermediate states and establishing an asymmetric Gram matrix related to local gradient aggregations. We show that with quartic network width, federated learning can converge to a global-optimal at a linear rate. We also provide a data-dependent generalization bound of over-parameterized neural networks trained with federated learning.

It will be interesting to extend the current FL-NTK framework for multi-layer cases. Existing NTK results of two-layer neural networks (NNs) have shown to be generalized to multi-layer NNs with various structures such as RNN, CNN, ResNet, etc. [AZLS19a, AZLS19b, DLL+19]. The key techniques for analyzing FL on wide neural networks are addressed in our work. Our result can be generalized to multi-layer NNs by controlling the perturbation propagation through layers using

[1] It is difficult to run real NTK experiment or full-batch gradient descent (GD) due to memory contrains.

[2] Local update *step* equals to *epoch* in GD. But the number of *steps* of **one** *epoch* in SGD is equal to the number of mini-batches, J.

well-developed tools (rank-one perturbation analysis, randomness decomposition, and extended McDiarmid's Inequality). We hope our results and techniques will provide insights for further study of distributed learning and other optimization algorithms.

Acknowledgements

Baihe Huang is supported by the Elite Undergraduate Training Program of School of Mathematical Sciences at Peking University, and work done while interning at Princeton University and Institute for Advanced Study (advised by Zhao Song). Xin Yang is supported by NSF grant CCF-2006359. This project is partially done while Xin was a Ph.D. student at University of Washington. The authors would like to thank the ICML reviewers for their valuable comments.

Appendix

Roadmap: In Appendix A, we list several probability results. In Appendix B we prove our convergence result of FL-NTK. In Appendix C, we prove our generalization result of FL-NTK.

A Probability Tools

Lemma A.1 (Bernstein inequality [Ber24]). *Let $X_1, \dots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all $i \in [n]$. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^n X_i \leq t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + Mt/3}\right).$$

Lemma A.2 (Anti-concentration inequality of Gaussian). *Let $X \sim N(0, \sigma^2)$, then for any $0 < t \leq \sigma$*

$$\Pr\left[|X| \leq t\right] \in (\frac{2t}{3\sigma}, \sqrt{\frac{2}{\pi}} \cdot \frac{t}{\sigma}).$$

Proof. For completeness, we provide a short proof. Since $X \sim N(0, \delta^2)$, the CDF of X^2 is $\Pr[X^2 \leq t^2] = \frac{\gamma(1/2, t^2/2\sigma^2)}{\Gamma(1/2)}$ where $\gamma(\cdot,\cdot)$ is the incomplete lower gamma function. This can be further simplified to $\Pr[X^2 \leq t^2] = \mathrm{erf}(\sqrt{t^2/2\sigma^2})$ where erf is the error function. For $z \leq 1$, we can sandwich the erf function by $2z/3 \leq \mathrm{erf}(z/\sqrt{2}) \leq \sqrt{2/\pi}z$, thus letting $z = t/\sigma$ complete the proof. □

B Convergence of Neural Networks in Federated Learning

Definition B.1. *We let κ to denote the condition number of Gram matrix $H(0)$.*

Assumption B.2. *We assume $\|x_i\|_2 = 1$ and $\lambda = \lambda_{\min}(H(0)) \in (0, 1]$.*

Notation	Dimension	Meaning
N	$\mathbb{N}$	#clients
c	$[N]$	its index
T	$\mathbb{N}$	#communication rounds
t	$[T]$	its index
K	$\mathbb{N}$	#local update steps
k	$[K]$	its index
$y(t)$	$\mathbb{R}^n$	aggregated server model after global round t
y_c	$\mathbb{R}^{\lvert S_c\rvert}$	ground truth of c-th client
$y_c^{(k)}(t)$	$\mathbb{R}^{\lvert S_c\rvert}$	c-th client's model in global round t and local step k
$y^{(k)}(t)$	$\mathbb{R}^n$	all client's model in global round t and local step k
$w_{k,c}(t)$	$\mathbb{R}^{d\times m}$	c-th client's model parameter in global round t and local step k
$u(t)$	$\mathbb{R}^{d\times m}$	aggregated server model parameter in global round t and local step k

Table 1: Summary of several notations

B.1 Convergence Result

Theorem B.3. *Recall that $\lambda = \lambda_{\min}(H(0)) > 0$. Let $m = \Omega(\lambda^{-4} n^4 \log(n/\delta))$, we iid initialize $u_r(0) \sim \mathcal{N}(0, I)$, a_r sampled from $\{-1, +1\}$ uniformly at random for $r \in [m]$, and we set the step size $\eta_{\text{local}} = O(\lambda/(\kappa K n^2))$ and $\eta_{\text{global}} = O(1)$, then with probability at least $1 - \delta$ over the random initialization we have for $t = 0, 1, 2, \cdots$*

$$\|y(t) - y\|_2^2 \leq (1 - \frac{\eta_{\text{global}} \eta_{\text{local}} \lambda K}{2N})^t \cdot \|y(0) - y\|_2^2. \tag{7}$$

Proof. We prove by induction. The base case is $t = 0$ and it is trivially true. Assume for $\tau = 0, \cdots, t$ we have proved Eq. (7) to be true. We show Eq. (7) holds for $\tau = t + 1$.

Recall that the set $Q_i \subset [m]$ is defined as follow

$$Q_i := \Big\{ r \in [m] : \forall w \in \mathbb{R}^d \text{ s.t. } \|w - w_r(0)\|_2 \leq R, \;\; \mathbf{1}_{w_r(0)^\top x_i \geq 0} = \mathbf{1}_{w^\top x_i \geq 0} \Big\},$$

and $\overline{Q}_i$ denotes its complement.

Let $v_{1,i}, v_{2,i}$ be defined as follows

$$v_{1,i} = \frac{1}{\sqrt{m}} \sum_{r \in Q_i} a_r \Big(\phi\big((u_r(t) + \eta_{\text{global}} \Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i) \Big),$$
$$v_{2,i} = \frac{1}{\sqrt{m}} \sum_{r \in \overline{Q}_i} a_r \Big(\phi\big((u_r(t) + \eta_{\text{global}} \Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i) \Big).$$

Let $H(t, k, c)_{i,j}, H(t, k, c)^\perp_{i,j}$ be defined as follows

$$H(t, k, c)_{i,j} = \frac{1}{m} \sum_{r=1}^m x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0},$$
$$H(t, k, c)^\perp_{i,j} = \frac{1}{m} \sum_{r \in \overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}.$$

Define $H(t)$ and $H(t)^\perp \in \mathbb{R}^{n \times n}$ as

$$H(t)_{i,j} = \frac{1}{m} \sum_{r=1}^m x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \geq 0, u_r(t)^\top x_j \geq 0},$$
$$H(t)^\perp_{i,j} = \frac{1}{m} \sum_{r \in \bar{Q}_i} x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \geq 0, u_r(t)^\top x_j \geq 0}.$$

Let $y_c^{(k)}(t)_j$ $(j \in S_c)$ be defined by

$$y_c^{(k)}(t)_j = f(w_{k,c}(t), x_j).$$

We can write $\Delta u_r(t)$ as follow

$$\Delta u_r(t) = \frac{a_r}{N} \sum_{c \in [N]} \sum_{k \in [K]} \frac{\eta_{\text{local}}}{\sqrt{m}} \sum_{j \in S_c} (y_j - y_c^{(k)}(t)_j) x_j \mathbf{1}_{w_{k,c,r}(t)^\top x_j \geq 0}.$$

Thus we have

$$\begin{aligned} v_{1,i} &= \frac{\eta_{\text{global}}\eta_{\text{local}}}{mN}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_j - y_c^{(k)}(t)_j)x_i^\top x_j \sum_{r\in Q_i}\mathbf{1}_{u_r^\top x_i\geq 0, w_{k,c,r}(t)^\top x_j\geq 0} \\ &= \frac{\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(t,k,c)^{\perp}_{i,j}). \end{aligned}$$

We can therefore write $-2(y-y(t))^\top(y(t+1)-y(t))$ as follow

$$\begin{aligned} & -2(y-y(t))^\top(y(t+1)-y(t)) \\ =\ & -2(y-y(t))^\top(v_1+v_2) \\ =\ & -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(t,k,c)^{\perp}_{i,j}) \\ & -2\sum_{i\in[n]}(y_i - y_i(t))v_{2,i}. \end{aligned}$$

Let

$$\begin{aligned} C_1 =\ & -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)H(t,k,c)_{i,j} \\ C_2 =\ & \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)H(t,k,c)^{\perp}_{i,j} \\ C_3 =\ & -2\sum_{i\in[n]}(y_i - y_i(t))v_{2,i} \\ C_4 =\ & \|y(t+1)-y(t)\|_2^2. \end{aligned}$$

Then

$$\begin{aligned} & \|y-y(t+1)\|_2^2 \\ =\ & \|y-y(t)\|_2^2 - 2(y-y(t))^\top(y(t+1)-y(t)) + \|y(t+1)-y(t)\|_2^2 \\ =\ & \|y-y(t)\|_2^2 + C_1 + C_2 + C_3 + C_4. \end{aligned}$$

By Claim B.4, Claim B.5, Claim B.6 and Claim B.7 we have

$$\begin{aligned} \|y-y(t+1)\|_2^2 \leq\ & \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}(-K\lambda + 4nRK(1+2\eta_{\text{local}}Kn) + 2\eta_{\text{local}}\kappa\lambda K^2 n)\|y-y(y)\|_2^2 \\ & + \frac{16\eta_{\text{global}}\eta_{\text{local}}}{N}K(1+2\eta_{\text{local}}Kn)nR\|y-y(y)\|_2^2 \\ & + \frac{16\eta_{\text{global}}\eta_{\text{local}}}{N}K(1+2\eta_{\text{local}}Kn)nR\|y-y(y)\|_2^2 \\ & + \frac{4\eta_{\text{global}}^2\eta_{\text{local}}^2 n^2K^2(1+2\eta_{\text{local}}Kn)^2}{N^2}\|y-y(y)\|_2^2. \end{aligned}$$

By the choice of $\eta_{\text{local}} \leq \frac{\lambda}{1000\kappa n^2 K}$ and $\eta_{\text{local}}\eta_{\text{global}} \leq \frac{\lambda}{1000\kappa n^2 K}$ and $R \leq \lambda/(1000n)$ we come to

$$\|y-y(t+1)\|_2^2 \leq \|y-y(t)\|_2^2$$

$$
\begin{aligned}
&- \frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{N}\|y - y(t)\|_2^2 \qquad (8)\\
&+ 40\frac{\eta_{\text{global}}\eta_{\text{local}}KnR}{N}\|y - y(t)\|_2^2\\
&+ 40\frac{\eta_{\text{global}}\eta_{\text{local}}KnR}{N}\|y - y(t)\|_2^2\\
&+ \frac{\eta_{\text{local}}^2\eta_{\text{global}}^2 n^2K^2}{N^2}\|y - y(t)\|_2^2\\
\le\ & \|y - y(t)\|_2^2\\
&- (1 - 1/10)\frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{N}\|y - y(t)\|_2^2\\
&+ 80\frac{\eta_{\text{global}}\eta_{\text{local}}KnR}{N}\|y - y(t)\|_2^2\\
\le & \|y - y(t)\|_2^2 - \frac{1}{2}\frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{N}\|y - y(t)\|_2^2 \qquad (9)
\end{aligned}
$$

where the second step follows from $\eta_{\text{local}} \le \frac{\lambda}{1000\kappa n^2 K}$, the third step follows from $R \le \lambda/(1000n)$. □

B.2 Bounding C_1, C_2, C_3, C_4

Claim B.4. *We have with probability at least $1 - n^2 \cdot \exp(-mR/10)$ over random initialization*

$$
C_1 \le \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N}\|y - y(t)\|_2^2(-K\lambda + 4nRK(1 + 2\eta_{\text{local}}Kn) + 2\eta_{\text{local}}\kappa\lambda K^2 n).
$$

Proof. We first calculate

$$
\begin{aligned}
&\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)H(t,k,c)_{i,j}\\
=&\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(0)_{i,j})\\
&+\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j(t) - y_c^{(k)}(t)_j)H(0)_{i,j}\\
&+K\sum_{i\in[n]}\sum_{j\in[n]}(y_i - y_i(t))(y_j - y_j(t))H(0)_{i,j}.
\end{aligned}
$$

From Lemma B.12 and Lemma B.9 we have $\|u_r(t) - u(0)\|_2 \le R$ and $\|w_{k,c,r}(t) - u(0)\|_2 \le R$. Let $H(t,k)$ be defined by

$$
H(t,k)_{i,j} = H(t,k,c)_{i,j}
$$

for $j \in S_c$. Then from Lemma B.11 we obtain

$$
\|H(t,k) - H(0)\|_F \le 2nR
$$

with probability at least $1 - n^2 \cdot \exp(-mR/10)$ over random initialization.

Therefore from direct calculations we have

$$
\left|\sum_{i\in[n]}\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y_i - y_i(t))(y_j - y_c^{(k)}(t)_j)(H(t,k,c)_{i,j} - H(0)_{i,j})\right|
$$

$$
\begin{aligned}
&= \sum_{k\in[K]} (y - y(t))^\top (H(t,k) - H(0))(y - y^{(k)}(t)) \\
&\le \sum_{k\in[K]} \|y - y(t)\|_2 \|y - y^{(k)}(t)\|_2 \|H(t,k) - H(0)\|_F \\
&\le 4nRK(1 + 2\eta_{\text{local}} Kn)\|y - y(t)\|_2^2.
\end{aligned}
$$

where the last step comes from Eq (16).

By Lemma B.10 we have

$$
\begin{aligned}
\left| \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j(t) - y_c^{(k)}(t)_j) H(0)_{i,j} \right| &\le \sum_{k\in[K]} \|y - y(t)\|_2 \|H(0)\| \|y(t) - y^{(k)}(t)\|_2 \\
&\le 2\eta_{\text{local}} \kappa\lambda K^2 n \|y - y(t)\|_2^2.
\end{aligned}
$$

Finally we have

$$
K \sum_{i\in[n]} \sum_{j\in[n]} (y_i - y_i(t))(y_j - y_j(t)) H(0)_{i,j} \ge K\lambda \|y - y(t)\|_2^2.
$$

Combining the above we conclude the proof with

$$
\begin{aligned}
&C_1 \\
=& -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j) H(t,k,c)_{i,j} \\
\le& -\frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} (-4nRK(1 + 2\eta_{\text{local}} K^2 n)\|y - y(t)\|_2^2 + K\lambda\|y - y(t)\|_2^2 - 2\eta_{\text{local}}\kappa\lambda K^2 n \|y - y(t)\|_2^2) \\
\le& \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \|y - y(t)\|_2^2 (-K\lambda + 4nRK(1 + 2\eta_{\text{local}} K^2 n) + 2\eta_{\text{local}}\kappa\lambda K^2 n).
\end{aligned}
$$

□

Claim B.5. *The following holds with probability at least* $1 - n\exp(-mR)$ *over random initialization*

$$
C_2 \le \frac{16\eta_{\text{global}}\eta_{\text{local}}}{N} K(1 + 2\eta_{\text{local}} nK) nR \|y - y(t)\|_2^2.
$$

Proof. We define matrix $H(t,k)^\perp \in \mathbb{R}^{n\times n}$ such that $H(t,k)^\perp_{i,j} = H(t,k,c)^\perp_{i,j}, j \in S_c$. Notice that

$$
\begin{aligned}
C_2 &= \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{i\in[n]} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} (y_i - y_i(t))(y_j - y_c^{(k)}(t)_j) H(t,k,c)^\perp_{i,j} \\
&= \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{k\in[K]} (y - y(t))^\top H(t,k)^\perp (y - y^{(k)}(t)) \\
&\le \frac{2\eta_{\text{global}}\eta_{\text{local}}}{N} \sum_{k\in[K]} \|y - y(t)\|_2 \|H(t,k)^\perp\|_F \|y - y^{(k)}(t)\|_2 \\
&\le \frac{4\eta_{\text{global}}\eta_{\text{local}}}{N} K(1 + 2\eta_{\text{local}} nK) \|y - y(t)\|_2^2 \|H(t,k)^\perp\|_F
\end{aligned}
$$

where the last step comes from Eq (16).

It thus suffices to upper bound $\|H(t,k)^{\perp}\|_F$.

For each $i \in [n]$, we define ζ_i as follows

$$\zeta_i = \sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i}.$$

It then follows from direct calculations that

$$\begin{aligned}
\|H(t,k)^{\perp}\|_F^2 =& \sum_{i=1}^{n}\sum_{j=1}^{n}(H(t,k)^{\perp}_{i,j})^2 \\
=& \sum_{i=1}^{n}\sum_{c\in[N]}\sum_{j\in S_c}\Big(\frac{1}{m}\sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}\Big)^2 \\
=& \sum_{i=1}^{n}\sum_{c\in[N]}\sum_{j\in S_c}\Big(\frac{1}{m}\sum_{r=1}^{m} x_i^\top x_j \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}\cdot \mathbf{1}_{r\in\overline{Q}_i}\Big)^2 \\
=& \sum_{i=1}^{n}\sum_{c\in[N]}\sum_{j\in S_c}(\frac{x_i^\top x_j}{m})^2\Big(\sum_{r=1}^{m} \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}\cdot \mathbf{1}_{r\in\overline{Q}_i}\Big)^2 \\
\leq& \frac{1}{m^2}\sum_{i=1}^{n}\sum_{c\in[N]}\sum_{j\in S_c}\Big(\sum_{r=1}^{m} \mathbf{1}_{u_r(t)^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}\cdot \mathbf{1}_{r\in\overline{Q}_i}\Big)^2 \\
=& \frac{n}{m^2}\sum_{i=1}^{n}\Big(\sum_{r=1}^{m}\mathbf{1}_{r\in\overline{Q}_i}\Big)^2 \\
=& \frac{n}{m^2}\sum_{i=1}^{n}\zeta_i^2.
\end{aligned}$$

Fix $i \in [n]$. The plan is to use Bernstein inequality to upper bound ζ_i with high probability.

First by Eq. (11) we have $\mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}] \leq R$. We also have

$$\begin{aligned}
\mathbb{E}\Big[(\mathbf{1}_{r\in\overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}])^2\Big] =& \mathbb{E}[\mathbf{1}^2_{r\in\overline{Q}_i}] - \mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}]^2 \\
\leq& \mathbb{E}[\mathbf{1}^2_{r\in\overline{Q}_i}] \\
\leq& R.
\end{aligned}$$

Finally we have $|\mathbf{1}_{r\in\overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r\in\overline{Q}_i}]| \leq 1$.

Notice that $\{\mathbf{1}_{r\in\overline{Q}_i}\}_{r=1}^m$ are mutually independent, since $\mathbf{1}_{r\in\overline{Q}_i}$ only depends on $w_r(0)$. Hence from Bernstein inequality (Lemma A.1) we have for all $t > 0$,

$$\Pr\left[\zeta_i > m\cdot R + t\right] \leq \exp\left(-\frac{t^2/2}{m\cdot R + t/3}\right).$$

By setting $t = 3mR$, we have

$$\Pr\left[\zeta_i > 4mR\right] \leq \exp(-mR). \tag{10}$$

Hence by union bound, with probability at least $1 - n\exp(-mR)$,

$$\|H(t,k)^{\perp}\|_F^2 \leq \frac{n}{m^2}\cdot n\cdot (4mR)^2 = 16n^2R^2.$$

Putting all together we have

$$\|H(t,k)^{\perp}\|_F \le 4nR$$

with probability at least $1 - n\exp(-mR)$ over random initialization. □

Claim B.6. *With probability at least $1 - n\exp(-mR)$ over random initialization the following holds*

$$C_3 \le \frac{16\eta_{\text{global}}\eta_{\text{local}}K}{N}(1 + 2\eta_{\text{local}}nK)nR\|y - y(t)\|_2^2$$

Proof. We can upper bound $\|v_2\|_2$ in the following sense

$$\begin{aligned}
\|v_2\|_2^2 &\le \sum_{i=1}^n \left(\frac{\eta_{\text{global}}}{\sqrt{m}} \sum_{r\in\overline{Q}_i} \left| \Delta u_r(t)^\top x_i \right| \right)^2 \\
&= \frac{\eta_{\text{global}}^2}{m} \sum_{i=1}^n \Big(\sum_{r=1}^m \mathbf{1}_{r\in\overline{Q}_i} |\Delta u_r(t)^\top x_i| \Big)^2 \\
&\le \frac{\eta_{\text{global}}^2\eta_{\text{local}}^2}{m} \cdot \Big(\frac{2K(1+2\eta_{\text{local}}nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2 \Big)^2 \cdot \sum_{i=1}^n \left(\sum_{r=1}^m \mathbf{1}_{r\in\overline{Q}_i} \right)^2
\end{aligned}$$

where the last step comes from Lemma B.10.

It is previously shown that $\sum_{r=1}^m \mathbf{1}_{r\in\overline{Q}_i} \le 4mR$ holds with probability at least $1 - n\exp(-mR)$ over random initialization, thus with probability at least $1 - n\exp(-mR)$ over random initialization

$$\begin{aligned}
\|v_2\|_2^2 &\le \frac{\eta_{\text{global}}^2\eta_{\text{local}}^2}{m} \cdot \frac{4K^2(1+2\eta_{\text{local}}nK)^2 n}{N^2 m} \|y - y(t)\|_2^2 \cdot n(4mR)^2 \\
&\le \Big(\frac{8\eta_{\text{global}}\eta_{\text{local}}K}{N} (1 + 2\eta_{\text{local}}nK)nR\|y - y(t)\| \Big)^2.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we complete the proof with

$$\begin{aligned}
C_3 = & -2\sum_{i\in[n]} (y_i - y_i(t))v_{2,i} \\
&\le 2\|y - y(t)\|_2 \cdot \|v_2\|_2 \\
&\le \frac{16\eta_{\text{global}}\eta_{\text{local}}K}{N} (1 + 2\eta_{\text{local}}nK)nR\|y - y(t)\|_2^2.
\end{aligned}$$

□

Claim B.7. *We have*

$$C_4 \le \frac{4\eta_{\text{local}}^2\eta_{\text{global}}^2 n^2K^2(1+2\eta_{\text{local}}nK)^2}{N^2} \|y - y(t)\|_2^2.$$

Proof. Recall that $y(t+1) - y(t) = v_1 + v_2$, we have

$$\|y(t+1) - y(t)\|_2^2 \le \sum_{i=1}^n \left(\frac{\eta_{\text{global}}}{\sqrt{m}} \sum_{r=1}^m \left| \Delta u_r(t)^\top x_i \right| \right)^2$$

$$
\begin{aligned}
&= \frac{\eta_{\text{global}}^2}{m} \sum_{i=1}^{n} \Big(\sum_{r=1}^{m} |\Delta u_r(t)^\top x_i| \Big)^2 \\
&\le \frac{\eta_{\text{global}}^2 \eta_{\text{local}}^2}{m} \cdot \Big(\frac{2K(1+2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2 \Big)^2 \cdot nm^2 \\
&\le \frac{4\eta_{\text{local}}^2 \eta_{\text{global}}^2 n^2 K^2 (1+2\eta_{\text{local}} nK)^2}{N^2} \|y - y(t)\|_2^2
\end{aligned}
$$

where the penultimate step comes from Lemma B.10. □

B.3 Random Initialization

Lemma B.8. *Let events E_1, E_2, E_3 be defined as follows*

$$
\begin{aligned}
E_1 &= \Big\{ \phi(w_r(0)^\top x_i) \le \sqrt{2\log(6mn/\delta)}, \forall r \in [m], \forall i \in [n] \Big\} \\
E_2 &= \Big\{ \Big| \sum_{r=1}^{m} \frac{1}{\sqrt{m}} a_r \phi(w_r(0)^\top x_i) \mathbf{1}_{w_r(0)^\top x_i \le \sqrt{2\log(6mn/\delta)}} \Big| \le \sqrt{2\log(2mn/\delta)} \cdot \log(8n/\delta), \forall i \in [n] \Big\} \\
E_3 &= \Big\{ \sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i} \le 4mR, \forall i \in [n] \Big\}.
\end{aligned}
$$

Then $E_1 \cap E_2 \cap E_3$ is true with probability at least $1-\delta$ over the random initialization. Furthermore given $E_1 \cap E_2 \cap E_3$ the following holds

$$
\|y - y(0)\|_2^2 = O(n \log(m/\delta) \log^2(n/\delta)).
$$

Proof. First we bound $\Pr[\neg E_3]$. For each $i \in [n]$, we define ζ_i as follows

$$
\zeta_i = \sum_{r=1}^{m} \mathbf{1}_{r \in \overline{Q}_i}.
$$

We use w as shorthand for $w(0)$. Define the event

$$
A_{i,r} = \Big\{ \exists u : \|u - w_r\|_2 \le R, \mathbf{1}_{x_i^\top w_r \ge 0} \ne \mathbf{1}_{x_i^\top u \ge 0} \Big\}.
$$

Note this event happens if and only if $|w_r^\top x_i| < R$. Recall that $w_r \sim \mathcal{N}(0, I)$. By anti-concentration inequality of Gaussian (Lemma A.2), we have

$$
\Pr[A_{i,r}] = \Pr_{z \sim \mathcal{N}(0,1)}[|z| < R] \le \frac{2R}{\sqrt{2\pi}}. \tag{11}
$$

It thus follows from Eq. (11) that $\mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}] \le R$. We also have

$$
\begin{aligned}
\mathbb{E}\Big[(\mathbf{1}_{r \in \overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}])^2 \Big] &= \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}^2] - \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}]^2 \\
&\le \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}^2] \\
&\le R.
\end{aligned}
$$

Therefore $|\mathbf{1}_{r \in \overline{Q}_i} - \mathbb{E}[\mathbf{1}_{r \in \overline{Q}_i}]| \le 1$.

Notice that $\{\mathbf{1}_{r\in\overline{Q}_i}\}_{r=1}^m$ are mutually independent, since $\mathbf{1}_{r\in\overline{Q}_i}$ only depends on w_r. Hence from Bernstein inequality (Lemma A.1) we have for all $t > 0$,

$$\Pr\left[\zeta_i > m\cdot R + t\right] \le \exp\left(-\frac{t^2/2}{m\cdot R + t/3}\right).$$

By setting $t = 3mR$, we have

$$\Pr\left[\zeta_i > 4mR\right] \le \exp(-mR). \tag{12}$$

Taking union bound and note the choice of R and m we have

$$\Pr\left[\neg E_3\right] \le n\exp(-mR) \le \delta/3.$$

Next we bound $\Pr[\neg E_1]$. Fix $r \in [m]$ and $i \in [n]$. Since $w_r \sim \mathcal{N}(0, I)$ and $\|x_i\|_2 = 1$, $w_r^\top x_i$ follows distribution $\mathcal{N}(0, 1)$. From concentration of Gaussian distribution, we have

$$\Pr_{w_r}[w_r^\top x_i \ge \sqrt{2\log(6mn/\delta)}] \le \frac{\delta}{6mn}.$$

Let E_1 be the event that for all $r \in [m]$ and $i \in [n]$ we have $\phi(w_r^\top x_i) \le \sqrt{2\log(6mn/\delta)}$. Then by union bound, $\Pr[\neg E_1] \le \frac{\delta}{3}$,

Finally we bound $\Pr[\neg E_2]$. Fix $i \in [n]$. For every $r \in [m]$, we define random variable $z_{i,r}$ as

$$z_{i,r} := \frac{1}{\sqrt{m}}\cdot a_r \cdot \phi(w_r^\top x_i)\cdot \mathbf{1}_{w_r^\top x_i \le \sqrt{2\log(6mn/\delta)}}.$$

Then $z_{i,r}$ only depends on $a_r \in \{-1, 1\}$ and $w_r \sim \mathcal{N}(0, I)$. Notice that $\mathbb{E}_{a_r,w_r}[z_{i,r}] = 0$, and $|z_{i,r}| \le \sqrt{2\log(6mn/\delta)}$. Moreover,

$$\begin{aligned}
\mathop{\mathbb{E}}_{a_r,w_r}[z_{i,r}^2] &= \mathop{\mathbb{E}}_{a_r,w_r}\left[\frac{1}{m}a_r^2\phi^2(w_r^\top x_i)\mathbf{1}^2_{w_r^\top x_i\le\sqrt{2\log(6mn/\delta)}}\right] \\
&= \frac{1}{m}\mathop{\mathbb{E}}_{a_r}[a_r^2]\cdot\mathop{\mathbb{E}}_{w_r}\left[\phi^2(w_r^\top x_i)\mathbf{1}^2_{w_r^\top x_i\le\sqrt{2\log(6mn/\delta)}}\right] \\
&\le \frac{1}{m}\cdot 1\cdot\mathop{\mathbb{E}}_{w_r}[(w_r^\top x_i)^2] \\
&= \frac{1}{m},
\end{aligned}$$

where the second step uses independence between a_r and w_r, the third step uses $a_r \in \{-1, 1\}$ and $\phi(t) = \max\{t, 0\}$, and the last step follows from $w_r^\top x_i \sim \mathcal{N}(0, 1)$.

Now we are ready to apply Bernstein inequality (Lemma A.1) to get for all $t > 0$,

$$\Pr\left[\sum_{r=1}^m z_{i,r} > t\right] \le \exp\left(-\frac{t^2/2}{m\cdot\frac{1}{m} + \sqrt{2\log(6mn/\delta)}\cdot t/3}\right).$$

Setting $t = \sqrt{2\log(6mn/\delta)}\cdot\log(8n/\delta)$, we have with probability at least $1 - \frac{\delta}{8n}$,

$$\sum_{r=1}^m z_{i,r} \le \sqrt{2\log(6mn/\delta)}\cdot\log(8n/\delta).$$

Notice that we can also apply Bernstein inequality (Lemma A.1) on $-z_{i,r}$ to get

$$\Pr\left[\sum_{r=1}^m z_{i,r} < -t\right] \leq \exp\left(-\frac{t^2/2}{m\cdot\frac{1}{m} + \sqrt{2\log(6mn/\delta)}\cdot t/3}\right).$$

Let E_2 be the event that for all $i \in [n]$,

$$\left|\sum_{r=1}^m z_{i,r}\right| \leq \sqrt{2\log(2mn/\delta)}\cdot\log(8n/\delta).$$

By applying union bound on all $i \in [n]$, we have $\Pr[\neg E_2] \leq \delta/3$.

By union bound, $E_1 \cap E_2 \cap E_3$ will happen with probability at least $1-\delta$.

If both E_1 and E_2 happen, we have

$$\begin{aligned}
\|y - u(0)\|_2^2 =& \sum_{i=1}^n (y_i - f(W(0), a, x_i))^2 \\
=& \sum_{i=1}^n \Big(y_i - \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\phi(w_r^\top x_i)\Big)^2 \\
=& \sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n \frac{y_i}{\sqrt{m}}\sum_{r=1}^m a_r\phi(w_r^\top x_i) + \sum_{i=1}^n \frac{1}{m}\Big(\sum_{r=1}^m a_r\phi(w_r^\top x_i)\Big)^2 \\
=& \sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n y_i\sum_{r=1}^m z_{i,r} + \sum_{i=1}^n \Big(\sum_{r=1}^m z_{i,r}\Big)^2 \\
\leq& \sum_{i=1}^n y_i^2 + 2\sum_{i=1}^n |y_i|\sqrt{2\log(2mn/\delta)}\cdot\log(4n/\delta) + \sum_{i=1}^n \Big(\sqrt{2\log(2mn/\delta)}\cdot\log(4n/\delta)\Big)^2 \\
=& O(n\log(m/\delta)\log^2(n/\delta)),
\end{aligned}$$

where the first step uses E_1, the second step uses E_2, and the last step follows from $|y_i| = O(1), \forall i \in [n]$.

□

B.4 Local Steps

The following theorem is standard in neural tangent kernel theory (see e.g. [SY19]).

Lemma B.9. *With probability at least $1-\delta$ over the random initialization, the following holds for all $k \in [K]$ and $c \in [N]$ and $r \in [m]$ in step t*

$$\|y_c^{(k)}(t) - y_c\|_2^2 \leq (1 - \eta_{\text{local}}\lambda/2)^k \cdot \|y_c^{(0)}(t) - y_c\|_2^2, \tag{13}$$

$$\|w_{k,c,r}(t+1) - w_{0,c,r}(t)\|_2 \leq \frac{4\sqrt{n}\|y_c^{(0)}(t) - y_c\|_2}{\sqrt{m}\lambda}, \tag{14}$$

$$\|y_c^{(k+1)}(t) - y_c^{(k)}(t)\|_2^2 \leq \eta_{\text{local}}^2 n^2 \cdot \|y_c^{(k)}(t) - y_c\|_2^2. \tag{15}$$

We then prove a Lemma that controls the updates in local steps.

Lemma B.10. *Given Eq (15) for all $k \in [K], c \in [N]$ in step t the following holds for all $k \in [K], c \in [N]$*

$$\|y_c(t) - y_c^{(k)}(t)\|_2 \leq 2\eta_{\text{local}} nK \|y_c(t) - y_c\|,$$
$$\|\Delta u_r(t)\|_2 \leq \frac{2\eta_{\text{local}} K(1 + 2\eta_{\text{local}} nK)\sqrt{n}}{N\sqrt{m}} \|y - y(t)\|_2.$$

Proof. For the first inequality, from Eq (15) we have

$$\begin{aligned}
\|y_c - y_c^{(k)}(t)\|_2 &\leq \|y_c^{(k)}(t) - y_c^{(k-1)}(t)\|_2 + \|y_c^{(k-1)}(t) - y_c\|_2 \\
&\leq (\eta_{\text{local}} n + 1)\|y_c - y_c^{(k-1)}(t)\|_2 \\
&\leq (\eta_{\text{local}} n + 1)^k \|y_c - y_c(t)\|_2.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|y_c(t) - y_c^{(k)}(t)\|_2 &\leq \sum_{i=1}^{k} \|y_c^{(j)}(t) - y_c^{(j-1)}(t)\|_2 \\
&\leq \sum_{i=1}^{k} \eta_{\text{local}} n \|y_c - y_c^{(j-1)}(t)\|_2 \\
&\leq \sum_{i=1}^{k} \eta_{\text{local}} n (\eta_{\text{local}} n + 1)^{j-1} \|y_c - y_c(t)\|_2 \\
&\leq 2\eta_{\text{local}} nK \|y_c(t) - y_c\|_2
\end{aligned}$$

where the last step comes from the choice of η_{local}.

For the second inequality, notice that

$$\begin{aligned}
\|\Delta u_r(t)\|_2 &= \eta_{\text{local}} \Big\| \frac{a_r}{N} \sum_{c\in[N]} \sum_{k\in[K]} \frac{1}{\sqrt{m}} \sum_{j\in S_c} (y_j - y^{(k)}(t)_j) x_j \mathbf{1}_{w_{r,k,c}(t)^\top x_j \geq 0} \Big\|_2 \\
&\leq \frac{\eta_{\text{local}}}{N\sqrt{m}} \sum_{k\in[K]} \sum_{c\in[N]} \sum_{j\in S_c} |y_j - y^{(k)}(t)_j| \\
&\leq \frac{\eta_{\text{local}}\sqrt{n}}{N\sqrt{m}} \sum_{k\in[K]} \|y - y^{(k)}(t)\|_2
\end{aligned}$$

where the second step comes form triangle inequality and $\|x_i\|_2 = 1$ and the last step comes from Cauchy-Schwartz inequality. From the $\|y_c(t) - y_c^{(k)}(t)\|_2 \leq 2\eta_{\text{local}} nK \|y_c(t) - y_c\|_2$ we have

$$\begin{aligned}
\|y - y^{(k)}(t)\|_2^2 &= \sum_{c\in[N]} \|y_c - y_c^{(k)}(t)\|_2^2 = \sum_{c\in[N]} 2(\|y_c - y_c(t)\|_2^2 + \|y_c(t) - y_c^{(k)}(t)\|_2^2) \\
&\leq \sum_{c\in[N]} 2(\|y_c - y_c(t)\|_2^2 + (2\eta_{\text{local}} nK)^2 \|y_c(t) - y_c\|_2^2) \\
&\leq 2(1 + 2\eta_{\text{local}} nK)^2 \|y - y(t)\|_2^2. \qquad (16)
\end{aligned}$$

It thus follows that

$$\|\Delta u_r(t)\|_2 \leq \frac{\eta_{\text{local}}\sqrt{n}}{N\sqrt{m}} \sum_{k\in[K]} \|y - y^{(k)}(t)\|_2$$

$$\leq \frac{2\eta_{\text{local}}K(1+2\eta_{\text{local}}nK)\sqrt{n}}{N\sqrt{m}}\|y-y(t)\|_2.$$

□

B.5 Technical Lemma

Lemma B.11. *For any set of weight vectors $\widetilde{w}_1,\cdots,\widetilde{w}_m \in \mathbb{R}^d$ and $\widehat{w}_1,\cdots,\widehat{w}_m \in \mathbb{R}^d$ define $H(\widetilde{w},\widehat{w}) \in \mathbb{R}^{n\times n}$ as*

$$H(\widetilde{w},\widehat{w})_{i,j} = \frac{1}{m}x_i^\top x_j \sum_{r=1}^m \mathbf{1}_{\widetilde{w}_r^\top x_i\geq 0,\widehat{w}_r^\top x_j\geq 0}.$$

Let $R\in(0,1)$ and $w_1,\cdots,w_m$ be iid generated from $\mathcal{N}(0,I)$. Then we have with probability at least $1-n^2\cdot\exp(-mR/10)$ the following holds

$$\|H(w,w)-H(\widetilde{w},\widehat{w})\|_F < 2nR$$

for any $\widetilde{w}_1,\cdots,\widetilde{w}_m\in\mathbb{R}^d$ and $\widehat{w}_1,\cdots,\widehat{w}_m\in\mathbb{R}^d$ such that $\|\widehat{w}_r-w_r\|_2\leq R$ and $\|\widetilde{w}_r-w_r\|_2\leq R$ for any $r\in[m]$.

Proof. For each $r\in[m]$ and $i,j\in[n]$, we define

$$s_{r,i,j} := \mathbf{1}_{\widetilde{w}_r^\top x_i\geq 0,\widehat{w}_r^\top x_j\geq 0} - \mathbf{1}_{w_r^\top x_i\geq 0,w_r^\top x_j\geq 0}.$$

The random variable we consider can be rewritten as follows

$$\begin{aligned}
&\sum_{i=1}^n\sum_{j=1}^n |H(\widetilde{w},\widehat{w})_{i,j}-H(w,w)_{i,j}|^2 \\
\leq\ & \frac{1}{m^2}\sum_{i=1}^n\sum_{j=1}^n\left(\sum_{r=1}^m \mathbf{1}_{\widetilde{w}_r^\top x_i\geq 0,\widehat{w}_r^\top x_j\geq 0} - \mathbf{1}_{w_r^\top x_i\geq 0,w_r^\top x_j\geq 0}\right)^2 \\
=\ & \frac{1}{m^2}\sum_{i=1}^n\sum_{j=1}^n\Big(\sum_{r=1}^m s_{r,i,j}\Big)^2.
\end{aligned}$$

It thus suffices to bound $\frac{1}{m^2}(\sum_{r=1}^m s_{r,i,j})^2$.

Fix i,j and we simplify $s_{r,i,j}$ to s_r. Then $\{s_r\}_{r=1}^m$ are mutually independent random variables. We define the event

$$A_{i,r} = \left\{\exists u : \|u-w_r\|_2\leq R, \mathbf{1}_{x_i^\top w_r\geq 0}\neq \mathbf{1}_{x_i^\top u\geq 0}\right\}.$$

If $\neg A_{i,r}$ and $\neg A_{j,r}$ happen, then

$$\left|\mathbf{1}_{\widetilde{w}_r^\top x_i\geq 0,\widehat{w}_r^\top x_j\geq 0} - \mathbf{1}_{w_r^\top x_i\geq 0,w_r^\top x_j\geq 0}\right| = 0.$$

If $A_{i,r}$ or $A_{j,r}$ happen, then

$$\left|\mathbf{1}_{\widetilde{w}_r^\top x_i\geq 0,\widehat{w}_r^\top x_j\geq 0} - \mathbf{1}_{w_r^\top x_i\geq 0,w_r^\top x_j\geq 0}\right| \leq 1.$$

So we have

$$\begin{aligned}\mathop{\mathbb{E}}_{w_r}[s_r] \le \mathop{\mathbb{E}}_{w_r}\left[\mathbf{1}_{A_{i,r}\vee A_{j,r}}\right] &\le \Pr[A_{i,r}] + \Pr[A_{j,r}] \\ &\le \frac{4R}{\sqrt{2\pi}} \\ &\le 2R,\end{aligned}$$

and

$$\begin{aligned}\mathop{\mathbb{E}}_{w_r}\left[\left(s_r - \mathop{\mathbb{E}}_{w_r}[s_r]\right)^2\right] &= \mathop{\mathbb{E}}_{w_r}[s_r^2] - \mathop{\mathbb{E}}_{w_r}[s_r]^2 \\ &\le \mathop{\mathbb{E}}_{w_r}[s_r^2] \\ &\le \mathop{\mathbb{E}}_{w_r}\left[\left(\mathbf{1}_{A_{i,r}\vee A_{j,r}}\right)^2\right] \\ &\le \frac{4R}{\sqrt{2\pi}} \\ &\le 2R.\end{aligned}$$

We also have $|s_r| \le 1$. So we can apply Bernstein inequality (Lemma A.1) to get for all $t > 0$,

$$\begin{aligned}\Pr\left[\sum_{r=1}^m s_r \ge 2mR + mt\right] &\le \Pr\left[\sum_{r=1}^m (s_r - \mathbb{E}[s_r]) \ge mt\right] \\ &\le \exp\left(-\frac{m^2t^2/2}{2mR + mt/3}\right).\end{aligned}$$

Choosing $t = R$, we get

$$\begin{aligned}\Pr\left[\sum_{r=1}^m s_r \ge 3mR\right] &\le \exp\left(-\frac{m^2R^2/2}{2mR + mR/3}\right) \\ &\le \exp\left(-mR/10\right).\end{aligned}$$

It follows that

$$\Pr\left[\frac{1}{m}\sum_{r=1}^m s_r \ge 3R\right] \le \exp(-mR/10).$$

Similarly

$$\Pr\left[\frac{1}{m}\sum_{r=1}^m s_r \le -3R\right] \le \exp(-mR/10).$$

Therefore we complete the proof. □

Lemma B.12. *If Eq. (7) holds for $i = 0, \cdots, k$, then we have for all $r \in [m]$*

$$\|u_r(t) - u_r(0)\|_2 \le \frac{8\sqrt{n}\|y - y(0)\|_2}{\sqrt{m}\lambda} := D.$$

Proof. We have

$$\begin{aligned}\|u_r(t)-u_r(0)\|_2 &\le \eta_{\text{global}}\sum_{\tau=0}^{t}\|\Delta u_r(\tau)\|_2\\ &\le \eta_{\text{global}}\sum_{\tau=0}^{t}\frac{2\eta_{\text{local}}K(1+2\eta_{\text{local}}nK)\sqrt{n}}{N\sqrt{m}}\|y-y(\tau)\|_2\\ &\le \eta_{\text{global}}\frac{2\eta_{\text{local}}K(1+2\eta_{\text{local}}nK)\sqrt{n}}{N\sqrt{m}}\sum_{\tau=0}^{t}(1-\frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{2N})^{\tau}\|y-y(0)\|_2\\ &\le \frac{8\sqrt{n}\|y-y(0)\|_2}{\sqrt{m}\lambda}.\end{aligned}$$

where the second step comes from Lemma B.10 and the last step comes from the choice of η_{local}. □

C Generalization

In this section, we generalize our initialization scheme to each $w_r(0) \sim \mathcal{N}(0,\sigma^2 I)$. Notice that this just introduces an extra σ^{-2} term to every occurrence of m. In addition, we use $U(t) = [u_1(t),\cdots,u_m(t)]^\top \in \mathbb{R}^{d\times m}$ to denote parameters in a matrix form. For convenience, we first list several definitions and results which will be used in the proof our generalization theorem. Our setting mainly follows [ADH+19a].

C.1 Definitions

Definition C.1 (Non-degenerate Data Distribution, Definition 5.1 in [ADH+19a]). *A distribution $\mathcal{D}$ over $\mathbb{R}^d\times\mathbb{R}$ is (λ,δ,n)-non-degenerate, if with probability at least $1-\delta$, for n iid samples $\{(x_i,y_i)\}_{i=1}^n$ chosen from $\mathcal{D}$, $\lambda_{\min}(H^\infty)\ge\lambda>0$.*

Definition C.2 (Loss Functions). *Let $\ell:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be the loss function. For function $f:\mathbb{R}^d\to\mathbb{R}$, for distribution $\mathcal{D}$ over $\mathbb{R}^d\times\mathbb{R}$, the* population loss *is defined as*

$$L_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(f(x),y)].$$

Let $S=\{(x_i,y_i)\}_{i=1}^n$ be n samples. The empirical loss over S is defined as

$$L_S(f) := \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i),y_i).$$

Definition C.3 (Rademacher Complexity). *Let $\mathcal{F}$ be a class of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$. Given n samples $S=\{x_1,\cdots,x_n\}$ where $x_i\in\mathbb{R}^d$ for $i\in[n]$, the* empirical Rademacher complexity *of $\mathcal{F}$ is defined as*

$$\mathcal{R}_S(\mathcal{F}) := \frac{1}{n}\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f(x_i)\right].$$

where $\epsilon\in\mathbb{R}^d$ and each entry of ϵ are drawn from independently uniform at random from $\{\pm1\}$.

C.2 Tools from Previous Work

Theorem C.4 (Theorem B.1 in [ADH+19a]). *Suppose the loss function $\ell(\cdot,\cdot)$ is bounded in $[0,c]$ for some $c>0$ and is ρ-Lipschitz in its first argument. Then with probability at least $1-\delta$ over samples S of size n,*

$$\sup_{f\in\mathcal{F}}\{L_{\mathcal{D}}(f)-L_S(f)\}\le 2\rho\mathcal{R}_S(\mathcal{F})+3c\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Lemma C.5 (Lemma 5.4 in [ADH+19a]). *Given $R>0$, with probability at least $1-\delta$ over the random initialization on $U(0)\in\mathbb{R}^{m\times d}$ and $a\in\mathbb{R}^m$, for all $B>0$, the function class*

$$\mathcal{F}_{R,B}^{U(0),a}=\{f(U,\cdot,a)=\frac{1}{\sqrt{m}}\sum_{r=1}^{m}a_r\phi(u_r^\top x):\|u_r-u_r(0)\|_2\le R,\forall r\in[m];\|U-U(0)\|_F\le B\}$$

has bounded empirical Rademacher complexity

$$\mathcal{R}_S(\mathcal{F}_{R,B}^{U(0),a})\le\frac{B}{\sqrt{2n}}\left(1+\left(\frac{2\log(2/\delta)}{m}\right)^{1/4}\right)+\frac{2R^2\sqrt{m}}{\sigma}+R\sqrt{2\log(2/\delta)}.$$

Lemma C.6 (Lemma C.3 in [ADH+19a]). *With probability at least $1-\delta$ we have*

$$\|H(0)-H^\infty\|_F\le O(n\sqrt{\log(n/\delta)/\sqrt{m}}).$$

C.3 Complexity Bound

To simplify the proof in the following sections, we define $\rho:=\eta_{\text{local}}\eta_{\text{global}}K/N$.

Now we prove a key technical lemma which will be used to prove the main result.

Lemma C.7. *Let $\lambda=\lambda_{\min}(H^\infty)>0$. Fix $\sigma>0$, let $m=\Omega(\lambda^{-4}\sigma^{-2}n^4\log(n/\delta))$, we iid initialize $w_r\sim\mathcal{N}(0,\sigma^2I)$, a_r sampled from $\{-1,+1\}$ uniformly at random for $r\in[m]$ and set $\eta_{\text{local}}=O(\frac{\lambda}{n^2K\kappa}),\eta_{\text{global}}=O(1)$. For weights $w_1,\cdots,w_m\in\mathbb{R}^d$, let $\text{vec}(W)=[w_1^\top\ w_2^\top\ \cdots w_m^\top]^\top\in\mathbb{R}^{md}$ be the concatenation of $w_1,\cdots,w_m$. Then with probability at least $1-6\delta$ over the random initialization, we have for all $t\ge0$,*

- $\|u_r(t)-u_r(0)\|_2\le\frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$,
- $\|U(t)-U(0)\|_F\le(y^\top(H^\infty)^{-1}y)^{1/2}+O\left((\frac{n\sigma}{\lambda}+\frac{n^{7/2}}{\sigma^{1/2}m^{1/4}})\cdot\text{poly}(\log(m/\delta))\right)$.

Proof. Similarly to Appendix B, $\|u_r(t)-u_r(0)\|_2\le\frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$ and $\|w_{k,c,r}(t)-u(0)\|_2\le\frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$. For integer $k\ge0$, define $J(k,t)\in\mathbb{R}^{md\times n}$ as the matrix

$$J(k,t)=\frac{1}{\sqrt{m}}\begin{pmatrix}a_1x_1\mathbf{1}_{w_{k,c_1,1}(t)^\top x_1\ge0} & \cdots & a_1x_n\mathbf{1}_{w_{k,c_n,1}(t)^\top x_n\ge0}\\ \vdots & \ddots & \vdots\\ a_mx_1\mathbf{1}_{w_{k,c_1,m}(t)^\top x_1\ge0} & \cdots & a_mx_n\mathbf{1}_{w_{k,c_n,m}(t)^\top x_n\ge0}\end{pmatrix}$$

where $c_i\in[N]$ denotes the unique client such that $i\in c_i$. We claim that

$$\|J(k,t)-J(0,0)\|_F\le O\Big(n\cdot\big(\delta+\frac{n\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\sigma\lambda\sqrt{m}}\big)\Big)^{1/2}.$$

In fact, we can calculate $\|J(k,t)-J(0,0)\|_F^2$ in the following

$$\begin{aligned}\|J(k,t)-J(0,0)\|_F^2 &= \frac{1}{m}\sum_{r=1}^{m}\sum_{c\in[N]}\sum_{i\in S_c}\Big(\|x_i\|_2\cdot a_i(\mathbf{1}_{w_{k,c,r}(t)^\top x_i\geq 0}-\mathbf{1}_{u_r(0)^\top x_i\geq 0})\Big)^2\\ &= \frac{1}{m}\sum_{r=1}^{m}\sum_{c\in[N]}\sum_{i\in S_c}\Big(\mathbf{1}_{w_{k,c,r}(t)^\top x_i\geq 0}-\mathbf{1}_{u_r(0)^\top x_i\geq 0}\Big)^2\\ &= \frac{1}{m}\sum_{r=1}^{m}\sum_{c\in[N]}\sum_{i\in S_c}\mathbf{1}_{\mathbf{1}_{w_{k,c,r}(t)^\top x_i\geq 0}\neq\mathbf{1}_{u_r(0)^\top x_i\geq 0}}.\end{aligned}$$

Fix $c\in[N], i\in S_c$ and for $r\in[m]$ define t_r as follows

$$t_r=\mathbf{1}_{\mathbf{1}_{w_{k,c,r}(t)^\top x_i\geq 0}\neq\mathbf{1}_{u_r(0)^\top x_i\geq 0}}.$$

Consider the event

$$A_{i,r}=\{\exists w:\ \|u_r(0)-w\|_2\leq R, \mathbf{1}_{w^\top x_i\geq 0}\neq\mathbf{1}_{u_r(0)^\top x_i\geq 0}\}$$

where $R=\frac{Cn\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\lambda\sqrt{m}}$ for sufficiently small constant $C>0$. If $t_r=1$ then either $A_{i,r}$ happens or $\|w_{k,c,r}(t)-u_r(0)\|_2\leq R$, otherwise $t_r=0$. Therefore

$$\mathbb{E}[t_r]\leq\Pr[A_{i,r}]+\Pr[\|u_r(0)-w_{k,c,r}(t)\|_2<R]\leq R\sigma^{-1}+\delta.$$

And similarly $\mathbb{E}[(t_r-\mathbb{E}[t_r])^2]\leq\mathbb{E}[t_r^2]=R\sigma^{-1}+\delta$. Applying Bernstein inequality, we have for all $t>0$,

$$\Pr\Big[\sum_{r=1}^{m}t_r\geq mR\sigma^{-1}+m\delta+mt\Big]\leq\exp\left(-\frac{m^2t^2}{2(mR\sigma^{-1}+m\delta+mt/3)}\right).$$

Choosing $t=R\sigma^{-1}+\delta$,

$$\Pr\Big[\sum_{r=1}^{m}t_r\geq 2m(R\sigma^{-1}+\delta)\Big]\leq\exp(-m(R\sigma^{-1}+\delta)/10).$$

By applying union bound over $i\in[n]$, we have with probability at least $1-n\exp(-m(R\sigma^{-1}+\delta)/10)$, $\|J(k,t)-J(0,0)\|_F\leq 2n(R\sigma^{-1}+\delta)$. This is exactly what we need.

Notice that we can rewrite the update rule in federated learning as

$$\begin{aligned}\mathrm{vec}(U(t+1)) &= \mathrm{vec}(U(t))-\frac{\eta_{\mathrm{global}}}{N}\sum_{k\in[K]}\eta_{\mathrm{local}}J(k,c)(y^{(k)}(t)-y)\\ &= \mathrm{vec}(U(t))-\rho\frac{1}{K}\sum_{k\in[K]}J(k,c)(y^{(k)}(t)-y)\end{aligned}\tag{17}$$

where the last step follows from definition of $\rho=\eta_{\mathrm{global}}\eta_{\mathrm{local}}K/N$.

Recall from Appendix B that

$$\Delta u_r(t)=\frac{a_r}{N}\sum_{c\in[N]}\sum_{k\in[K]}\frac{\eta_{\mathrm{local}}}{\sqrt{m}}\sum_{j\in S_c}(y_j-y_c^{(k)}(t)_j)x_j\mathbf{1}_{w_{k,c,r}(t)^\top x_j\geq 0},$$

and

$$H(t,k,c)_{i,j} = \frac{1}{m}\sum_{r=1}^{m} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0},$$
$$H(t,k,c)^\perp_{i,j} = \frac{1}{m}\sum_{r\in\overline{Q}_i} x_i^\top x_j \mathbf{1}_{u_r^\top x_i \geq 0, w_{k,c,r}(t)^\top x_j \geq 0}.$$

We have

$$\begin{aligned}
v_{1,i} &= \frac{1}{\sqrt{m}}\sum_{r\in Q_i} a_r\Big(\phi\big((u_r(t)+\eta_{\text{global}}\Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i)\Big) \\
&= -\frac{\eta_{\text{local}}\eta_{\text{global}}K}{N}\sum_{j\in S_c}(y(t)_j - y_j)H^\infty_{i,j} + (-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y(t)_j)H^\infty_{i,j} \\
&\quad + (-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y_j)(H(t,k,c)_{i,j} - H^\infty_{i,j}) \\
&\quad + (-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y_j)H(t,k,c)^\perp_{i,j}, \\
v_{2,i} &= \frac{1}{\sqrt{m}}\sum_{r\in\overline{Q}_i} a_r\Big(\phi\big((u_r(t)+\eta_{\text{global}}\Delta u_r(t))^\top x_i\big) - \phi(u_r(t)^\top x_i)\Big).
\end{aligned}$$

Following the proof of Appendix B, let

$$\begin{aligned}
\xi_i(t) = v_{2,i}(t) &+ (-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y(t)_j)H^\infty_{i,j} \\
&+ (-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y_j)(H(t,k,c)_{i,j} - H^\infty_{i,j}) \\
&+ (-\frac{\eta_{\text{local}}\eta_{\text{global}}}{N})\sum_{k\in[K]}\sum_{c\in[N]}\sum_{j\in S_c}(y^{(k)}(t)_j - y_j)H(t,k,c)^\perp_{i,j}.
\end{aligned}$$

Notice that

$$\sum_{i=1}^{n}|\overline{Q}_i| = \sum_{r=1}^{m}\sum_{i=1}^{n}\mathbf{1}_{r\in\overline{Q}_i} = \sum_{i=1}^{n}(\sum_{r=1}^{m}\mathbf{1}_{r\in\overline{Q}_i}).$$

Hence by Eq. (10), with probability at least $1 - n\exp(-mR\sigma^{-1})$ we have

$$\sum_{i=1}^{n}|\overline{Q}_i| \leq 4mnR\sigma^{-1}.$$

Similar to Appendix B, by the choice of $R = \frac{8\sqrt{n}\|y-u(0)\|_2}{\sqrt{m}\lambda}$ and

$$\|y - y(0)\|_2 = O\Big(\sqrt{n\log(m/\delta)\log^2(n/\delta)}\Big),$$

we can bound $\xi(t) = [\xi_1(t), \cdots, \xi_n]^\top \in \mathbb{R}^n$ as

$$\|\xi(t)\|_2 \leq O\Big(\frac{\eta_{\text{global}}\eta_{\text{local}}n^{5/2}K\kappa\sqrt{\log(m/\delta)\log^2(n/\delta)}}{N\sigma\lambda\sqrt{m}}\|y - y(t)\|_2\Big)$$

$$= O\bigg(\frac{\rho n^{5/2}\kappa\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\sigma\lambda\sqrt{m}}\|y-y(t)\|_2\bigg) \tag{18}$$

where the last step follows from definition of $\rho = \eta_{\text{global}}\eta_{\text{local}}K/N$.

Notice that with probability at least $1-\delta$, for all $i \in [n]$,

$$|y_i(0)| \le \sigma \cdot \sqrt{2\log(2mn/\delta)} \cdot \log(4n/\delta),$$

which implies

$$\|y(0)\|_2^2 \le n\sigma^2 \cdot 2\log(2mn/\delta) \cdot \log^2(4n/\delta). \tag{19}$$

Therefore we can explicitly write the dynamics of the global model as

$$\begin{aligned} y(t) - y &= (I - \frac{\eta_{\text{global}}\eta_{\text{local}}K}{N}H^\infty)(y(t-1)-y) + \xi(t-1) \\ &= (I-\rho H^\infty)(y(t-1)-y) + \xi(t-1) \\ &= (I-\rho H^\infty)^t(y(0)-y) + \sum_{\tau=0}^{t-1}(I-\rho H^\infty)^\tau \xi(t-1-\tau) \\ &= -(I-\rho H^\infty)^t y + e(t). \end{aligned}$$

where the second step follows from definition of $\rho = \eta_{\text{global}}\eta_{\text{local}}K/N$, the third step comes from recursively applying the former step.y.

By Eq (18) and Eq (19) we have

$$\begin{aligned} e(t) &= (I-\rho H^\infty)^t y(0) + \sum_{\tau=0}^{t-1}(I-\rho H^\infty)^\tau \xi(t-1-\tau) \\ &= O\bigg((1-\rho)^t \cdot \sqrt{n\sigma^2} \cdot \sqrt{2\log(2mn/\delta)} \cdot \log(8n/\delta) + t(1-\rho)^t \cdot \frac{\rho n^3 \log(m/\delta)\log^2(n/\delta)}{\lambda\sigma\sqrt{m}}\bigg) \end{aligned}$$

where we used $\|y(t)-y\|_2^2 \le (1-\frac{\eta_{\text{global}}\eta_{\text{local}}\lambda K}{2N})^t \cdot \|y(0)-y\|_2^2$ from Theorem B.3.

By Eq (17),

$$\begin{aligned} \text{vec}(U(T)) - \text{vec}(U(0)) &= \sum_{t=0}^{T-1}(\text{vec}(U(t+1)) - \text{vec}(U(t))) \\ &= \sum_{t=0}^{T-1}\bigg(-\rho \cdot \frac{1}{K}\sum_{k\in[K]} J(k,t)(y^{(k)}(t)-y)\bigg) \\ &= \sum_{t=0}^{T-1}\rho \cdot J(0,0)(I-\rho H^\infty)^t y \\ &\quad + \sum_{t=0}^{T-1}\rho \cdot \frac{1}{K}\sum_{k\in[K]}(J(k,t)-J(0,0))(I-\rho H^\infty)^t y \\ &\quad - \sum_{t=0}^{T-1}\rho \cdot \frac{1}{K}\sum_{k\in[K]} J(k,t)(y^{(k)}(t) - y(t) + e(k)) \end{aligned}$$

$$= B_1 + B_2 + B_3$$

where

$$\begin{aligned}
B_1 :=& + \rho \cdot \sum_{t=0}^{T-1} J(0,0)(I - \rho H^\infty)^t y, \\
B_2 :=& + \rho \cdot \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k \in [K]} (J(k,t) - J(0,0))(I - \rho H^\infty)^t y, \\
B_3 :=& - \rho \cdot \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k \in [K]} J(k,t)(y^{(k)}(t) - y(t) + e(k)).
\end{aligned}$$

We bound these terms separately.

Putting Claim C.8, C.9 and C.10 together we have

$$\begin{aligned}
& \|U(T) - U(0)\|_F \\
=& \|\operatorname{vec}(U(T)) - \operatorname{vec}(U(0))\|_2 \\
=& B_1 + B_2 + B_3 \\
\leq& (y^\top (H^\infty)^{-1} y)^{1/2} + O\left((\frac{n^2 \sqrt{\log(n/\delta)}}{\lambda^2 \sqrt{m}})^{1/2} + (\frac{n^{3/2}}{m^{1/4} \sigma^{1/2} \lambda^{3/2}} + \frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\lambda^3 \sigma \sqrt{m}}) \cdot \operatorname{poly}(\log(m/\delta)) \right) \\
\leq& (y^\top (H^\infty)^{-1} y)^{1/2} + O\left(\frac{n\sigma}{\lambda} \cdot \operatorname{poly}(\log(m/\delta)) + \frac{n^{7/2}}{\sigma^{1/2} m^{1/4}} \cdot \operatorname{poly}(\log(m/\delta)) \right)
\end{aligned}$$

which completes the proof of Lemma C.7. □

C.4 Technical Claims

Claim C.8 (Bounding B_1)**.** *With probability at least $1 - \delta$ over the random initialization, we have*

$$\|B_1\|_2^2 \leq y^\top D^\top H^\infty D y + O(\frac{n^2 \sqrt{\log(n/\delta)}}{\lambda^2 \sqrt{m}})$$

where $D = \sum_{t=0}^{T-1} \rho (I - \rho H^\infty)^t \in \mathbb{R}^{n \times n}$.

Proof. Recall $D = \sum_{t=0}^{T-1} \rho (I - \rho H^\infty)^t \in \mathbb{R}^{n \times n}$, then we have

$$\begin{aligned}
\|B_1\|_2^2 =& \Big\| \sum_{t=0}^{T-1} \rho \cdot J(0,0) \cdot (I - \rho \cdot H^\infty)^t y \Big\|_2^2 \\
=& y^\top D^\top J(0,0)^\top J(0,0) D y \\
=& y^\top D^\top H^\infty D y + y^\top D^\top (H(0) - H^\infty) D y \\
\leq& y^\top D^\top H^\infty D y + \|H(0) - H^\infty\|_F \cdot \|D\|_2^2 \|y\|^2 \\
\leq& y^\top D^\top H^\infty D y + O(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}) \left(\sum_{t=0}^{T-1} \rho (1 - \rho\lambda)^t \right)^2 n \\
\leq& y^\top D^\top H^\infty D y + O(\frac{n^2 \sqrt{\log(n/\delta)}}{\lambda^2 \sqrt{m}})
\end{aligned}$$

where the penultimate step comes from Lemma C.6 and $y_i = O(1)$. □

Claim C.9 (Bounding B_2)**.** *With probability at least $1-\delta$ over the random initialization, we have*

$$\|B_2\|_2 \leq \frac{n^{3/2}\operatorname{poly}(\log(m/\delta))}{m^{1/4}\sigma^{1/2}\lambda^{3/2}}.$$

Proof. For B_2, we have

$$\begin{aligned}
\|B_2\|_2 &= \Big\| \sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} (J(k,t) - J(0,0))(I-\rho H^\infty)^t y \Big\|_2 \\
&\leq \sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} \|J(k,t) - J(0,0)\|_F \cdot \|I-\rho H^\infty\|_2^t \cdot \|y\|_2 \\
&\leq O\left(\frac{n\operatorname{poly}(\log(m/\delta))}{m^{1/4}\sigma^{1/2}\lambda^{1/2}} \cdot \rho \cdot \sum_{k=0}^{K-1} (1-\rho\lambda)^k \cdot \sqrt{n} \right) \\
&= O\left(\frac{n^{3/2}\operatorname{poly}(\log(m/\delta))}{m^{1/4}\sigma^{1/2}\lambda^{3/2}} \right).
\end{aligned}$$

where in the third step we use

$$\|J(k,t) - J(0,0)\|_F \leq O\Big(n \cdot \big(\delta + \frac{n\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\sigma\lambda\sqrt{m}} \big) \Big)^{1/2}$$

and without loss of generality, we can set δ sufficiently small. □

Claim C.10 (Bounding B_3)**.** *With probability at least $1-\delta$ over the random initialization, we have*

$$\|B_3\|_2 \leq (\frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\lambda^3\sigma\sqrt{m}}) \cdot \operatorname{poly}(\log(m/\delta)).$$

Proof. Notice that for $k, t \geq 0$, $\|J(k,t)\|_F^2 \leq \frac{mn}{m} = n$. By Eq (18) and Eq (19) we have

$$\begin{aligned}
\|B_3\|_2 &= \left\| -\sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} J(k,t)(y^{(k)}(t) - y(t) + e(k)) \right\|_2 \\
&\leq \rho \frac{1}{K} \cdot \sqrt{n} \cdot \sum_{t=0}^{T-1} O\Big((1-\rho)^t \cdot \sqrt{n\sigma^2} \cdot \sqrt{2\log(2mn/\delta)} \cdot \log(8n/\delta) \\
&\quad + t(1-\rho)^t \cdot \frac{\rho n^3 \log(m/\delta)\log^2(n/\delta)}{\lambda\sigma\sqrt{m}} \Big) \\
&\leq (\frac{n\sigma}{\lambda} + \frac{n^{7/2}}{\lambda^3\sigma\sqrt{m}}) \cdot \operatorname{poly}(\log(m/\delta)),
\end{aligned}$$

here in the first step $-\sum_{t=0}^{T-1} \rho \cdot \frac{1}{K} \sum_{k\in[K]} J(k,t)e(k)$ is the dominant term.

□

C.5 Main Results

Now we can present our main result in this section.

Theorem C.11. *Fix failure probability $\delta \in (0,1)$. Set $\sigma = O(\lambda \operatorname{poly}(\log n, \log(1/\delta))/n)$, $m = \Omega\left(\sigma^{-2}(n^{14}\operatorname{poly}(\log m, \log(1/\delta), \lambda^{-1}))\right)$, let the two layer neural network be initialized with w_r i.i.d sampled from $\mathcal{N}(0, \sigma^2 I)$ and a_r sampled from $\{-1,+1\}$ uniformly at random for $r \in [m]$. Suppose the training data $S = \{(x_i, y_i)\}_{i=1}^n$ are i.i.d samples from a $(\lambda, \delta/3, n)$-non-degenerate distribution $\mathcal{D}$. Let $\rho = \eta_{\text{local}}\eta_{\text{global}}K/N$ and train the two layer neural network $f(U(t), \cdot, a)$ by federated learning for*

$$T \geq \Omega\left(\rho^{-1}\lambda^{-1}\operatorname{poly}(\log(n/\delta))\right)$$

iterations. Consider loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,1]$ that is 1-Lipschitz in its first argument. Then with probability at least $1-\delta$ over the random initialization on $U(0) \in \mathbb{R}^{d\times m}$ and $a \in \mathbb{R}^m$ and the training samples,the population loss $L_{\mathcal{D}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(U(T), x, a), y)]$ is upper bounded by

$$L_{\mathcal{D}}(f) \leq \sqrt{2y^\top (H^\infty)^{-1} y/n} + O(\sqrt{\log(n/(\lambda\delta))/(2n)}).$$

Proof. We will define a sequence of failing events and bound these failure probability individually, then we can apply the union bound to obtain the desired result.

Let E_1 be the event that $\lambda_{\min}(H^\infty) < \lambda$. Because $\mathcal{D}$ is $(\lambda, \delta/3, n)$-non-degenerate, $\Pr[E_1] \leq \epsilon/3$. In the remaining of the proof we assume E_1 does not happen.

Let E_2 be the event that $L_S(f(U(T), \cdot, a)) = \frac{1}{n}\sum_{i=1}^n \ell(f(U(T), x_i, a), y_i) > \frac{1}{\sqrt{n}}$. By Theorem B.3 with scaling δ properly, with probability $1-\delta/9$ we have $L_S(f(U(T), \cdot, a)) \leq \frac{1}{\sqrt{n}}$. So we have $\Pr[E_2] \leq \delta/9$.

Set $R, B > 0$ as

$$R = O(\frac{n\sqrt{\log(m/\delta)\log^2(n/\delta)}}{\lambda\sqrt{m}}),$$
$$B = (y^\top (H^\infty)^{-1} y)^{1/2} + O\left(\frac{n\sigma}{\lambda} \cdot \operatorname{poly}(\log(m/\delta)) + \frac{n^{7/2}}{\sigma^{1/2}m^{1/4}} \cdot \operatorname{poly}(\log(m/\delta))\right).$$

Notice that $\|y\|_2 = O(\sqrt{n})$ and $\|(H^\infty)^{-1}\|_2 = 1/\lambda$. By our setting of $\sigma = O(\frac{\lambda \operatorname{poly}(\log n, \log(1/\delta))}{n})$ and $m\sigma^2 \geq n^{14} > n^{12}$, $B = O(\sqrt{n/\lambda})$. Let E_3 be the event that there exists $r \in [m]$ so that $\|u_r - u_r(0)\|_2 > R$, or $\|U - U(0)\|_F > B$. By Lemma C.7, $\Pr[E_3] \leq \delta/9$.

For $i = 1, 2, \cdots,$, let $B_i = i$. Let E_4 be the event that there exists $i > 0$ so that

$$\mathcal{R}_S(\mathcal{F}_{R,B_i}^{U(0),a}) > \frac{B_i}{\sqrt{2n}}\left(1 + (\frac{2\log(18/\delta)}{m})^{1/4}\right) + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{2\log(18/\delta)}.$$

By Lemma C.5, $\Pr[E_4] \leq 1 - \delta/9$.

Assume neither of E_3, E_4 happens. Let i^* be the smallest integer so that $B_{i^*} = i^* \geq B$, then we have $B_{i^*} \leq B + 1$ and $i^* = O(\sqrt{n/\lambda})$. Since E_3 does not happen, we have $f(U(T), \cdot, a) \in \mathcal{F}_{R,B_{i^*}}^{U(0),a}$. Moreover,

$$\begin{aligned}\mathcal{R}_S(\mathcal{F}_{R,B_{i^*}}^{U(0),a}) &\leq \frac{B+1}{\sqrt{2n}}\left(1 + (\frac{2\log(18/\delta)}{m})^{1/4}\right) + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{\log(18/\delta)} \\ &= \sqrt{\frac{y^\top (H^\infty)^{-1} y}{2n}} + \frac{1}{\sqrt{n}} + O(\frac{\sqrt{n}\sigma \cdot \operatorname{poly}(\log(m/\delta))}{\lambda})\end{aligned}$$

$$
\begin{aligned}
&+ \frac{n^3 \operatorname{poly}(\log m, \log(1/\delta), \lambda^{-1})}{m^{1/4}\sigma^{1/2}} + \frac{2R^2\sqrt{m}}{\sigma} + R\sqrt{\log(18/\delta)} \\
&= \sqrt{\frac{y^\top (H^\infty)^{-1} y}{2n}} + \frac{1}{\sqrt{n}} + O(\frac{\sqrt{n}\sigma \cdot \operatorname{poly}(\log(m/\delta))}{\lambda}) + \frac{n^3 \operatorname{poly}(\log m, \log(1/\delta), \lambda^{-1})}{m^{1/4}\sigma^{1/2}} \\
&= \sqrt{\frac{y^\top (H^\infty)^{-1} y}{2n}} + \frac{2}{\sqrt{n}}
\end{aligned}
$$

where the first step follows from E_4 does not happen and the choice of B, the second step follows from the choice of R, and the last step follows from the choice of m and σ.

Finally, let E_5 be the event so that there exists $i \in \{1, 2, \cdots, O(\sqrt{n/\lambda})\}$ so that

$$
\sup_{f \in \mathcal{F}_{R,B_i}^{U(0),a}} \{L_{\mathcal{D}}(f) - L_S(f)\} > 2\mathcal{R}_S(\mathcal{F}_{R,B_i}^{U(0),a}) + \Omega\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right).
$$

By Theorem C.4 and applying union bound on i, we have $\Pr[E_5] \leq \delta/3$.

In the case that all of the bad events E_1, E_2, E_3, E_4, E_5 do not happen,

$$
\begin{aligned}
L_{\mathcal{D}}(f(U(T), \cdot, a)) &\leq L_S(f(U(T), \cdot, a)) + 2\mathcal{R}_S(\mathcal{F}_{R,B_{i^*}}^{U(0),a}) + O\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right) \\
&\leq \sqrt{\frac{2y^\top (H^\infty)^{-1} y}{n}} + \frac{5}{\sqrt{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right) \\
&= \sqrt{\frac{2y^\top (H^\infty)^{-1} y}{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda\delta})}{2n}}\right).
\end{aligned}
$$

which is exactly what we need. □

References

[Act96] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

[ADH+19a] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, pages 322–332, 2019.

[ADH+19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8141–8150, 2019.

[AdTBT20] Mathieu Andreux, Jean Ogier du Terrail, Constance Beguier, and Eric W Tramel. Siloed federated learning for multi-centric histopathology datasets. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 129–139. Springer, 2020.

[AZLS19a] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, pages 242–252, 2019.

[AZLS19b] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[Ber24] Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

[BPSW21] Jan van den Brand, Binghui Peng, Zhao Song, and Omri Weinstein. Training (overparametrized) neural networks in near-linear time. In *Innovations in Theoretical Computer Science (ITCS)*, 2021.

[CX20] Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020.

[CYS+20] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications*, 2020.

[DCM+12] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. In *Advances in neural information processing systems (NeurIPS)*, pages 1223–1231, 2012.

[DLL+19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 1675–1685. PMLR, 2019.

[DZPS19] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/pdf/1810.02054, 2019.

[HAA20] Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[HLS+20] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, et al. Fedml: A research library and benchmark for federated machine learning. *NeurIPS Federated Learning workshop*, 2020.

[HY20] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning (ICML)*, pages 4542–4551, 2020.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems (NeurIPS)*, pages 8571–8580, 2018.

[KMA+21] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021.

[KMR20] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on indentical and heterogeneous data. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2020.

[Leg18] California State Legislature. California consumer privacy act (ccpa). https://oag.ca.gov/privacy/ccpa, 2018.

[LGD+20] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 2020.

[LHY+20] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *International Conference on Learning Representations (ICLR)*, 2020.

[LJZ+21] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations (ICLR)*, 2021.

[LL18] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8157–8166, 2018.

[LLC+19] Xinle Liang, Yang Liu, Tianjian Chen, Ming Liu, and Qiang Yang. Federated transfer reinforcement learning for autonomous driving. *arXiv preprint arXiv:1910.06001*, 2019.

[LLH+20] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

[LMX+19] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 133–141. Springer, 2019.

[LSS+20] Jason D Lee, Ruoqi Shen, Zhao Song, Mengdi Wang, and Zheng Yu. Generalized leverage score sampling for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[LSZ+20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems (MLSys)*, 2020.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282. PMLR, 2017.

[MMRAyA16] H. McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. 2016.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. In *arXiv preprint*. https://arxiv.org/pdf/1902.04674.pdf, 2020.

[RHL+20] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

[SS15] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (CCS)*, pages 1310–1321. ACM, 2015.

[SY19] Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

[SZ21] Zhao Song and Ruizhe Zhang. Hybrid quantum-classical implementation of training over-parameterized neural networks. In *manuscript*, 2021.

[WTS+19] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

[WYS+20] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

[YAG+19] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning (ICML)*, pages 7252–7261, 2019.

[YLCT19] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):12, 2019.

[YYZ19] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5693–5700, 2019.

[YZY+19] Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. Ffd: a federated learning based method for credit card fraud detection. In *International Conference on Big Data*, pages 18–32. Springer, 2019.

[ZLL+18] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[ZPD+20] Yi Zhang, Orestis Plevrakis, Simon S Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.