Widening the Network Mitigates the Impact of Data Heterogeneity on FedAvg

Like Jian [1] **Dong Liu** [1]

Abstract

Federated learning (FL) enables decentralized clients to train a model collaboratively without sharing local data. A key distinction between FL and centralized learning is that clients' data are non-independent and identically distributed, which poses significant challenges in training a global model that generalizes well across heterogeneous local data distributions. In this paper, we analyze the convergence of overparameterized FedAvg with gradient descent (GD). We prove that the impact of data heterogeneity diminishes as the width of neural networks increases, ultimately vanishing when the width approaches infinity. In the infinite-width regime, we further prove that both the global and local models in FedAvg behave as linear models, and that FedAvg achieves the same generalization performance as centralized learning with the same number of GD iterations. Extensive experiments validate our theoretical findings across various network architectures, loss functions, and optimization methods.

1. Introduction

Federated Learning (FL) is a distributed machine learning paradigm that enables collaborative model training across distributed clients while preserving data locality (McMahan et al., 2017), a critical feature for privacy-sensitive domains such as healthcare, finance, and mobile computing, where regulatory or infrastructural constraints prohibit data centralization. However, a fundamental challenge in FL arises from the intrinsic non-independent and identically distributed (non-IID) nature of client data (Li et al., 2021b), where local datasets exhibit significant distributional shifts due to user-specific behaviors, geographic variations, or device-specific usage patterns. Such statistical heterogeneity leads to divergent local optimizations, degrading model convergence and generalization (Li et al., 2019; Zhao et al., 2018).

To address the challenges posed by non-IID data in FL, numerous research efforts have emerged, including client regularization (Li et al., 2020), adaptive optimization frameworks (Karimireddy et al., 2020; Reddi et al., 2021), personalized model architectures (Jeong & Hwang, 2022; T Dinh et al., 2020), and etc. While these approaches have demonstrated empirical success, they often require intricate hyperparameter tuning or restrictive assumptions about convexity, data similarity, or gradient boundedness, limiting their applicability in practical highly heterogeneous environments.

In parallel, overparameterized neural networks have garnered prominence in centralized learning for their remarkable ability to achieve strong generalization despite nonconvex optimization landscapes, underpinned by theoretical frameworks such as the neural tangent kernel (NTK) (Jacot et al., 2018). These networks exhibit implicit regularization properties, enabling interpolation of complex data distributions while maintaining robust generalization (Neyshabur et al., 2015; 2019a;b; Lee et al., 2018), motivating a pivotal question: Can increasing the network width inherently mitigate the effects of data heterogeneity in FL?

In this work, we analyze the convergence of FedAvg with gradient descent (GD) for multi-layer overparameterized neural networks and establish that the impact of data heterogeneity can indeed be reduced by widening the network. Further, we prove that as the network width approaches infinity, both global and local models behave as linear models. Strikingly, in this regime, FedAvg and centralized GD yield identical model parameters and outputs under matched iterations, achieving equivalent generalization performance. To the best of our knowledge, this is the first work to provide a quantitative analysis explicitly linking the width of neural networks to the impact of data heterogeneity on both FL training and generalization. Our key contributions are:

- **Theoretical guarantees for heterogeneity reduction:** We prove that the model divergence is bounded and decreases inversely proportional to the square root of the network width asymptotically, without relying on restrictive assumptions on convexity or gradient similarity/boundedness. This bound is vital in the convergence analysis of FedAvg, revealing that the impact

[1]School of Cyber Science and Technology, Beihang University, Beijing, China. Correspondence to: Dong Liu <dliu@buaa.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

of data heterogeneity slows the convergence rate, but vanishes when the width approaches infinity, allowing the convergence rate to recover linearity.

- **Bridging federated and centralized learning:** We extend the NTK theory from centralized learning to FL with multi-layer networks, showing that infinite network width induces constant global and local NTKs, further linearizes both global and local models. Notably, we prove the equivalence between infinite-width FedAvg and centralized GD, thereby achieving the same generalization performance, bridging decentralized and centralized learning paradigms.

- **Empirical validation:** We conducted numerous experiments on MINST and CIFAR-10 datasets, spanning diverse network structures, loss functions, and optimizers to validate our theoretical analysis.

2. Related Work

Data Heterogeneity. While prior works have provided convergence analyses of federated learning with non-IID data, many rely on restrictive assumptions. For example, some studies assume convex loss functions (Cho et al., 2020; Khaled et al., 2019; Li et al., 2019), others require bounded gradient dissimilarity (Li et al., 2020; Zhang et al., 2023; Wang et al., 2020b), and some assume bounded gradients (Li et al., 2019; Cho et al., 2020). These conditions are often difficult to satisfy in practice. In this paper, we analyze the convergence of FedAvg via NTK without imposing convexity, bounded gradients, or bounded gradient dissimilarity.

To address data heterogeneity, various techniques have been proposed. Regularization-based methods (Li et al., 2020; Durmus et al., 2021) introduce a regularization term during local updates to mitigate client drift. Client selection approaches (Cho et al., 2020; Goetz et al., 2019; Zhang et al., 2023) choose a subset of clients whose aggregated updates approximate those of all clients. Personalized federated learning (Jeong & Hwang, 2022; Jiang et al., 2024) allows clients to leverage aggregated knowledge while fine-tuning on their local data. Other methods, such as SCAFFOLD (Karimireddy et al., 2020) and FedNova (Wang et al., 2020b), correct optimization bias between clients and the global model to improve convergence. FedMA (Wang et al., 2020a) constructs the global model layer by layer to diminish the impact of heterogeneous data, and MOON (Li et al., 2021a) compares local and global models to correct client drift.

Overparameterized FL. Recent works have made substantial progress in overparameterized federated learning. For instance, Li et al. (2021b) proposed FedBN, which applies local batch normalization to mitigate heterogeneity, and analyzed its convergence using NTK. However, their analysis is limited to two-layer networks, limiting the model capacity. Under the same two-layer assumption, Jiang et al. (2024) proposed a local-global update mixing method and analyzed its convergence via NTK, while Huang et al. (2021) showed that overparameterized FedAvg converges to the global optimal solution with linear convergence rate.

Moving beyond the depth constraint on networks, Deng et al. (2022) proved that overparameterized FedAvg with ReLU activation converges in polynomial time with stochastic gradient descent (SGD). Fed-ensemble (Shi et al., 2024) employs model ensembling to enhance the generalization of FL, supported by an NTK-based convergence analysis. Yet, data heterogeneity is not considered. To address data heterogeneity, Yue et al. (2022) proposed to let clients transmit Jacobian matrices rather than weights or gradients, providing a more expressive data representation. Yu et al. (2022) observed that the learning of final layers in FL is strongly influenced by non-convexity and propose the train-convexify-train (TCT) method to alleviate these issues.

Among the most relevant works, Song et al. (2023) established that overparameterized FedAvg achieves linear convergence to zero training loss, and empirically observed that wide neural networks achieve better and more stable performance in FL. By contrast, we theoretically prove that the model divergence in FL caused by data heterogeneity is bounded by $\mathcal{O}(n^{-\frac{1}{2}})$, where n is the network width, thereby establishing the first quantitative relationship between network width and the mitigation of heterogeneity. Beyond this, we unveil the generalization performance of overparameterized FL by proving that infinite-width FedAvg and centralized learning yield identical model outputs under matched training iterations, while the theoretical analysis in (Song et al., 2023) solely focused on training loss.

3. Notations and Problem Formulation

In this section, we establish the basic notations, formulate the FL problem, and introduce the metric quantifying the impact of data heterogeneity.

We consider a standard FL setup consisting of a server and M clients. The local training dataset of client i is denoted by $\mathcal{D}_i$ with $\mathcal{X}_i = \{x|(x,y) \in \mathcal{D}_i\}$ and $\mathcal{Y}_i = \{y|(x,y) \in \mathcal{D}_i\}$ representing the set of inputs and labels of client i, respectively, where $x \in \mathbb{R}^{n_0}$ and $y \in \mathbb{R}^k$. The global dataset is defined as the union of all clients dataset, i.e., $\mathcal{D} = \cup_{i=1}^M \mathcal{D}_i$ with $\mathcal{X} = \cup_{i=1}^M \mathcal{X}_i$ and $\mathcal{Y} = \cup_{i=1}^M \mathcal{Y}_i$.

Suppose that each client trains a L-layer fully-connected neural network (FNN), where the width of the l-th layer is denoted by n_l. Then, the output of the l-th layer of client i's

model can be expressed as

$$f_{i,l}(x) = \begin{cases} x, & l = 0 \\ \sigma\left(W_{i,l} f_{i,l-1}(x) + b_{i,l}\right), & 0 < l < L \\ W_{i,L} f_{i,L-1}(x) + b_{i,L}, & l = L \end{cases} \quad (1)$$

where $W_{i,l} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $b_{i,l} \in \mathbb{R}^{n_l}$ are the weight and bias of the l-th layer of client i, respectively, and $\sigma(\cdot)$ is the activation function.

We define $\theta_i = \operatorname{vec}\left(\{W_{i,l}, b_{i,l}\}_{l=1}^{L}\right) \in \mathbb{R}^w$ as the vector of all trainable parameters of client i's model. Then, the model output of a single input sample x can be expressed as $f_i(x, \theta_i)$ and the concatenated output of all samples in $\mathcal{X}_i$ is denoted by $f_i(\mathcal{X}_i, \theta_i) = \operatorname{vec}\big(\{f(x, \theta_i)\}_{x \in \mathcal{X}_i}\big) \in \mathbb{R}^{k|\mathcal{D}_i|}$. Similarly, we further define the global model f having the same structure as f_i but with parameter $\theta \in \mathbb{R}^w$. Analogously, its concatenated output of all samples in $\mathcal{X}$ is denoted by $f(\mathcal{X}, \theta) = \operatorname{vec}\big(\{f(x, \theta)\}_{x \in \mathcal{X}}\big) \in \mathbb{R}^{k|\mathcal{D}|}$. To simplify the notations, we use the short hand $f_i(\theta_i) \triangleq f_i(\mathcal{X}_i, \theta_i)$ and $f(\theta) \triangleq f(\mathcal{X}, \theta)$ in the following.

We consider the mean square error (MSE) loss function, and hence the loss function of client i is expressed as

$$\Phi_i = \frac{1}{2|\mathcal{D}_i|} \sum_{(x,y) \in \mathcal{D}_i} \left\| f_i(x, \theta_i) - y \right\|_2^2, \quad (2)$$

The goal is to minimize the global loss function defined by $\Phi = \sum_{i=1}^{M} p_i \Phi_i$, where $p_i = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}$.

At initialization, each client's model parameters are sampled from the Gaussian distribution as follows

$$W_{1,l}^0 = \cdots = W_{M,l}^0 \sim \mathcal{N}\big(0, \tfrac{\sigma_{W_l}^2}{n_l}\big), \quad (3)$$

$$b_{1,l}^0 = \cdots = b_{M,l}^0 \sim \mathcal{N}\left(0, \sigma_b^2\right). \quad (4)$$

Upon initialization, each client updates its local model minimizing the loss function by GD for τ iterations. For every τ local iterations, each client uploads its local model to the server for model aggregation, and then the server broadcasts the aggregated model to each client for the next round. Let t denote the number of global rounds. Then, the model parameters of client i after the r-th ($1 \le r \le \tau$) local iteration in the t-th global round can be denoted by $\theta_i^{t\tau+r}$.

Specifically, during the t-th and $(t+1)$-th global round, say in the $(t\tau + r + 1)$-th total iteration, the model parameters are updated by GD as

$$\theta_i^{t\tau+r+1} \leftarrow \theta_i^{t\tau+r} - \frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^{t\tau+r}\right) g_i\left(\theta_i^{t\tau+r}\right) \quad (5)$$

where η is the learning rate and

$$J_i(\theta_i) = \nabla_{\theta_i} f_i(\theta_i) \in \mathbb{R}^{w \times k|\mathcal{D}_i|}, \quad (6)$$

$$g_i(\theta_i) = f_i(\theta_i) - \operatorname{vec}(\mathcal{Y}_i) \quad (7)$$

are the local Jacobian matrix and error vector, respectively. Similarly, we define the global Jacobian matrix and error vector respectively, as

$$J(\theta) \triangleq \nabla_\theta f(\theta) \in \mathbb{R}^{w \times k|\mathcal{D}|}, \quad (8)$$

$$g(\theta) \triangleq f(\theta) - \operatorname{vec}(\mathcal{Y}). \quad (9)$$

In the $(t+1)$-th global round, the model is aggregated by FedAvg, i.e.,

$$\theta^{(t+1)\tau} = \sum_{i=1}^{M} p_i \theta_i^{t\tau+\tau}, \quad (10)$$

where we let $t\tau + \tau$ and $(t+1)\tau$ to denote the time instants before and after the $(t+1)$-th global aggregation, respectively. Then, the aggregated parameters $\theta^{(t+1)\tau}$ are broadcast to all clients, which yields $\theta_i^{(t+1)\tau} = \theta^{(t+1)\tau}, \forall i$. Consequently, the relation between the global and local Jacobian and error at the t-th global round can be described by

$$g_i\left(\theta_i^{t\tau}\right) = P_i g\left(\theta^{t\tau}\right), \quad (11)$$

$$J\left(\theta^{t\tau}\right) = \sum_{i=1}^{M} J_i\left(\theta_i^{t\tau}\right) P_i, \quad (12)$$

where $P_i \in \mathbb{R}^{k|\mathcal{D}_i| \times k|\mathcal{D}|}$ is a projection matrix defined as

$$P_i = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \underbrace{0 \ \cdots \ 0}_{k|\mathcal{D}_i| \times \sum_{j=1}^{i-1} k|\mathcal{D}_j|} & & & \underbrace{0 \ 0 \ \cdots \ 1}_{k|\mathcal{D}_i| \times k|\mathcal{D}_i|} & & & \underbrace{0 \ \cdots \ 0}_{k|\mathcal{D}_i| \times \sum_{j=i+1}^{M} k|\mathcal{D}_j|} & & \end{pmatrix}, \quad (13)$$

whose operator norm $\|P_i\|_{\mathrm{op}} = 1$.

To facilitate convergence analysis, the following notations are also introduced:

$$f\left(\theta_i^{t\tau+r}\right) = f\left(\mathcal{X}, \theta_i^{t\tau+r}\right), \quad (14)$$

$$J\left(\theta_i^{t\tau+r}\right) = \nabla_{\theta_i^{t\tau+r}} f\left(\theta_i^{t\tau+r}\right), \quad (15)$$

$$g\left(\theta_i^{t\tau+r}\right) = f\left(\theta_i^{t\tau+r}\right) - \operatorname{vec}(\mathcal{Y}), \quad (16)$$

where $f\left(\theta_i^{t\tau+r}\right)$ denotes the output of the global model when its parameters are replaced with client i's parameters in round $t\tau + r$.

In the $(t+1)$-th global round, the degree to which client i's model deviates from the global model is characterized by

$$\left\| \Delta\theta_i^{(t+1)\tau} \right\|_2 = \left\| \theta_i^{t\tau+\tau} - \theta^{(t+1)\tau} \right\|_2 \quad (17)$$

Therefore, we use $\sum_{i=1}^{M} p_i \left\| \Delta\theta_i^{(t+1)\tau} \right\|_2$ to quantify the degree of data heterogeneity and term it as *model divergence*. Apparently, when the data is IID, $\sum_{i=1}^{M} p_i \left\| \Delta\theta_i^{(t+1)\tau} \right\|_2$ approaches zero as the number of local data increases. By contrast, when the data is non-IID, $\sum_{i=1}^{M} p_i \left\| \Delta\theta_i^{(t+1)\tau} \right\|_2$ remains non-zero and increases with the degree of data heterogeneity.

4. Convergence Analysis

In this section, we analyze the convergence of overparameterized FedAvg. We derive the bound on the model divergence explicitly and analyze how it influences the convergence rate and error.

We first introduce several notations regarding overparameterized neural networks. Let $n = \min\{n_1, n_2, \cdots, n_L\}$ and define the global NTK matrix (Lee et al., 2019) in the t-th global round as

$$\Theta^{t\tau} = \frac{1}{n} J\left(\theta^{t\tau}\right)^T J\left(\theta^{t\tau}\right). \tag{18}$$

Meanwhile, the analytic NTK matrix is defined as

$$\Theta = \lim_{n\to\infty} \Theta^0. \tag{19}$$

Analogously, the local NTK matrix of the standard parameterization in the $(t\tau + r)$-th iteration is defined as

$$\Theta_i^{t\tau+r} = \frac{1}{n} J\left(\theta_i^{t\tau+r}\right)^T J_i\left(\theta_i^{t\tau+r}\right) P_i. \tag{20}$$

The following assumptions are made to facilitate the convergence analysis.

Assumption 1. The minimum width among all hidden layers n is sufficiently large such that the terms of order $\mathcal{O}(n^{-1})$ and higher are omitted.

Assumption 2. The analytic NTK Θ is full rank, i.e., the minimum eigenvalue λ_m of Θ satisfies $\lambda_m > 0$.

Assumption 3. The norm of every input data is bounded, i.e., $\|x\|_2 \leq 1$.

Assumption 4. The activation function σ satisfies

$$|\sigma(0)|, \|\sigma'\|_\infty, \sup_{x\neq x'} \frac{|\sigma(x) - \sigma(x')|}{|x - x'|} < \infty$$

Assumptions $1 \sim 4$ are common in analyzing the overparameterized neural network (Lee et al., 2019; Shi et al., 2024).

The learning rate is set to $\eta = \frac{\eta_0}{n}$ and η_0 is a constant independent of n, which results in infinitesimally updates during each gradient descent step when n is sufficiently large. Consequently, we adopt gradient flow as an approximation of gradient descent, which can be expressed as

$$\frac{d\theta_i^{t\tau+r}}{dr} = -\frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^{t\tau+r}\right) g_i\left(\theta_i^{t\tau+r}\right). \tag{21}$$

Next, we present the main theorem regarding the bound of model divergence and the convergence of overparameterized FedAvg.

Theorem 1. *Under Assumptions 1 to 4, for any small $\delta_0 > 0$, there exist $R_0 > 0$, $N > 0$, $\eta_0 > 0$, $C > 0$ and $C_1 > 0$, such that for any $n \geq N$, the following holds with probability at least $(1 - \delta_0)$ over random initialization:*

$$\sum_{i=1}^M p_i \left\|\Delta\theta_i^{(t+1)\tau}\right\|_2 \leq \zeta \triangleq \frac{2\eta_0\tau C R_0}{\sqrt{n}\,(1-q)}, \tag{22}$$

$$\left\|g(\theta^{t\tau})\right\|_2 \leq q^t R_0 + \frac{2\eta_0\tau C C_1 R_0 \zeta\,(1-q^t)}{(1-q)^2}, \tag{23}$$

$$\left\|\theta^{t\tau} - \theta^0\right\|_2 \leq \frac{\eta_0\tau C R_0\,(1-q^t)}{\sqrt{n}\,(1-q)}, \tag{24}$$

$$\left\|\Theta^{t\tau} - \Theta^0\right\|_F \leq \frac{2\eta_0\tau C^3 R_0\,(1-q^t)}{\sqrt{n}\,(1-q)}, \tag{25}$$

$$\begin{aligned}&\left\|\Theta_i^{t\tau+r} - \Theta_i^0\right\|_F \\ &\leq \frac{2\eta_0 r q^t C^3 R_0\sqrt{k}}{\sqrt{n|\mathcal{D}_i|}} + \frac{2\eta_0\tau C^3 R_0\,(1-q^t)\sqrt{k|\mathcal{D}_i|}}{(1-q)\sqrt{n}}\end{aligned} \tag{26}$$

where $q = 1 - \frac{\eta_0\tau\lambda_m}{3|\mathcal{D}|} + \frac{\eta_0^2\tau^2 C^4}{2} e^{\eta_0\tau C^2}$.

The detailed proof is provided in appendix B and we present the proof sketch in the following.

Proof Sketch. We use mathematical induction to prove Theorem 1 and the induction hypotheses are (23) and (24). It is trivial that (23) and (24) hold when $t = 0$, and our aim is to prove (23) and (24) for $t + 1$.

[Step 1] We first present several essential lemmas in Appendix A, including proving the Lipschitz continuity of the global and local Jacobians and some properties regarding the Taylor series expansion.

[Step 2] Prove induction hypothesis (24) holds for $t + 1$. Due to the small learning rate $\eta = \frac{\eta_0}{n}$ for large n, we treat time as continuous and use gradient flow to approximate GD, which yields

$$\frac{dg\left(\theta_i^{t\tau+r}\right)}{dr} = -\frac{\eta_0\Theta_i^{t\tau+r}}{|\mathcal{D}_i|} g\left(\theta_i^{t\tau+r}\right). \tag{27}$$

Applying the mean value theorem of integral, we can obtain

$$g\left(\theta_i^{t\tau+\tau}\right) = e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}} g\left(\theta_i^{t\tau}\right), \bar{r}_i \in (0,\tau). \tag{28}$$

Based on (28) and the induction hypotheses, the local model parameters variation can be bounded by:

$$\left\|\theta_i^{t\tau+\tau} - \theta_i^0\right\|_2 \leq \frac{\eta_0\tau C\left\|g\left(\theta^{t\tau}\right)\right\|_2}{\sqrt{n}|\mathcal{D}_i|} + \left\|\theta^{t\tau} - \theta^0\right\|_2. \tag{29}$$

Notably, (29) captures the relationship between the dynamics of local model and global model, based on which we can obtain the recursive relationship between $\|\theta^{t\tau} - \theta^0\|_2$ and $\|\theta^{(t+1)\tau} - \theta^0\|_2$ from

$$\left\|\theta^{(t+1)\tau}-\theta^{0}\right\|_2 \leq \sum_{i=1}^{M} p_i \left\|\theta_i^{t\tau+\tau}-\theta^{0}\right\|_2 \leq \sum_{i=1}^{M} \frac{\eta_0 \tau C \left\|g\left(\theta^{t\tau}\right)\right\|_2}{\sqrt{n}} + \left\|\theta^{t\tau}-\theta^{0}\right\|_2 \quad (30)$$

Using (30), we can prove (24) holds for $t+1$.

[**Step 3**] Based on the inductions hypotheses and a variation of (29), we are able to obtain (22).

[**Step 4**] Prove induction hypothesis (23) holds for $t+1$. By taking the Taylor series expansion of $\sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right)$ at $\theta^{(t+1)\tau}$, we are able to derive

$$\left\|g\left(\theta^{(t+1)\tau}\right)\right\|_2 \leq \left\|\sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right)\right\|_2 + \left\|\sum_{i=1}^{M} p_i \Omega_i\right\|_2, \quad (31)$$

where Ω_i represents the remainder terms. Based on the results of [**Step 1**] $\sim$ [**Step 3**], we can further bound $\|\sum_{i=1}^{M} p_i g(\theta_i^{t\tau+\tau})\|_2$ and $\|\sum_{i=1}^{M} p_i \Omega_i\|$, respectively, as

$$\left\|\sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right)\right\|_2 \leq q \left\|g\left(\theta_i^{t\tau}\right)\right\|_2 \quad (32)$$

$$\left\|\sum_{i=1}^{M} p_i \Omega_i\right\|_2 \leq \frac{2\eta_0 \tau C C_1 R_0 \zeta}{(1-q)} \quad (33)$$

Plugging (32) and (33) into (31), (23) holds for $t+1$.

[**Step 5**] Based on the Lipschitzness of the global and local jacobians as well as the results in [**Step 2**]$\sim$[**Step 4**], we prove (25) and (26).

Remark 1 (**Bound on the model divergence**). *Inequation (22) establishes an upper bound ζ on the model divergence caused by data heterogeneity. Since $\zeta = \mathcal{O}(n^{-\frac{1}{2}})$, increasing the network width can reduce the effect of data heterogeneity. Note that we do not impose any strict assumptions on the convexity of the loss function (Cho et al., 2020; Khaled et al., 2019; Li et al., 2019), the bound of local gradients (Li et al., 2019; Cho et al., 2020) or the divergence between local and global gradient (Li et al., 2020; Zhang et al., 2023; Wang et al., 2020b). Instead, we prove that model the divergence is indeed bounded as long as the network is sufficiently wide.*

Remark 2 (**Impact of data heterogeneity on the convergence rate**). *Inequation (23) characterizes the evolution of training error across the global aggregation rounds. Different from Song et al. (2023); Huang et al. (2021), the presence of $\zeta > 0$ here slows down the convergence, making the convergence rate no longer linear and the convergence error no longer zero. Recalling that $\zeta = \mathcal{O}(n^{-\frac{1}{2}})$, widening the network enhances the convergence rate by mitigating the model divergence. When $n \to \infty$, we have $\zeta \to 0$ and the impact of data heterogeneity vanishes, resulting in a linear convergence rate and zero training error as shown in (23).*

Remark 3 (**Lazy training**). *Inequality (24) shows that as the network width increases, each global update in overparameterized FedAvg remains confined within an increasingly smaller neighborhood of size $\mathcal{O}(n^{-\frac{1}{2}})$ around its initialization, thereby extending the lazy-training phenomenon observed in centralized settings (Chizat et al., 2019) to FL settings.*

Remark 4 (**Constant global and local NTKs**). *Inequation (25) shows that as the network width increases, the global and local NTK experiences less variation during training. When the width approaches infinity, the both the global and local NTKs are constant, which extends the findings in centralized learning (Jacot et al., 2018) to FL settings.*

Next, we will investigate the training dynamics of FedAvg in the infinite-width regime and compare it with centralized learning to further investigate the generalization performance of overparameterized FedAvg.

5. Generalization Performance

In this section, we analyze the training dynamics and generalization performance of overparameterized FedAvg as the network width $n \to \infty$. First, we prove that both the global and local models behave as linear models during the training process. Then, we derive the closed-form expression of those linear models and establish the equivalence between infinite-width FedAvg and centralized GD.

We define the linear models $f^{\text{lin}}\left(\theta^{t\tau}\right)$ and $f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)$ as the first-order Taylor expansion of the global model $f\left(\theta^{t\tau}\right)$ and local model $f_i\left(\theta_i^{t\tau+r}\right)$, respectively:

$$f^{\text{lin}}\left(\theta^{t\tau}\right) = f\left(\theta^{0}\right) + J\left(\theta^{0}\right)^{T}\left(\theta^{t\tau}-\theta^{0}\right) \quad (34)$$

$$f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right) = f_i\left(\theta^{0}\right) + J_i\left(\theta^{0}\right)^{T}\left(\theta_i^{t\tau+r}-\theta^{0}\right) \quad (35)$$

Our main results are as follows.

Theorem 2. *Under Assumptions 2 to 4, when $n \to \infty$, we have*

$$\sup_{t\geq 0}\left\|f^{\text{lin}}(\theta^{t\tau}) - f(\theta^{t\tau})\right\|_2 = \mathcal{O}\left(n^{-\frac{1}{2}}\right), \quad (36)$$

$$\sup_{\substack{t\geq 0,\\ 1\leq r\leq \tau}}\left\|f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right) - f_i\left(\theta_i^{t\tau+r}\right)\right\|_2 = \mathcal{O}\left(n^{-\frac{1}{2}}\right), \forall i \quad (37)$$

The detailed proof is provided in appendix C.

Remark 5 (**Infinite-width FedAvg induces linearized global/local models**). *Theorem 2 suggests that as the network width approaches infinity, the global and local models become linear models. This extends the findings of Shi et al. (2024), which demonstrated that the global model can be*

approximated by a linear model, to show that both local and global models can be well approximated by linear models.

Therefore, we can analyze the training dynamic of those linear models instead. The main theorem describing their training dynamics is presented as follows.

Theorem 3. *Under Assumptions 2 to 4, when $n \to \infty$ and $\eta_0 \tau$ is sufficiently small such that the terms of $\mathcal{O}\left(\eta_0^2 \tau^2\right)$ and higher are neglected, the linear model has closed-form expressions for the global parameters and outputs throughout the training process:*

$$\theta^{t\tau} = -\frac{1}{n} J(\theta^0)(\Theta^0)^{-1}\Big(I - e^{-\frac{\eta_0 t\tau}{|\mathcal{D}|}\Theta^0}\Big)\, g(\theta^0) + \theta^0, \quad (38)$$

$$f^{\text{lin}}(x, \theta^{t\tau}) = f(x, \theta^0) - \Theta^0(x)(\Theta^0)^{-1}\left(I - e^{-\frac{\eta_0 t\tau}{|\mathcal{D}|}\Theta^0}\right) g(\theta^0), \quad (39)$$

where $\Theta^0(x) \triangleq \frac{1}{n} J(x, \theta^0)^T J(\theta^0)$.

The detailed proof is provided in appendix D.

Suppose there is a model having the same structure that trains on the global dataset $\mathcal{D}$ via centralized GD, whose model parameters at the t'-th GD iteration is denoted by $\theta_{\text{cen}}^{t'}$ and the model output is $f(x, \theta_{\text{cen}}^{t'})$. When the initialization of θ_{cen} and θ are the same, i.e., $\theta_{\text{cen}}^0 = \theta^0$, the following can be obtained according to Lee et al. (2019, Equations (8), (10), (11)):

$$\theta_{\text{cen}}^{t'} = -\frac{1}{n} J(\theta^0)(\Theta^0)^{-1}\left(I - e^{-\frac{\eta_0 \Theta^0 t'}{|\mathcal{D}|}}\right) g(\theta^0) + \theta^0, \quad (40)$$

$$f_{\text{cen}}\big(x, \theta_{\text{cen}}^{t'}\big) = f\left(x, \theta^0\right) - \Theta^0(x)\big(\Theta^0\big)^{-1}\left(I - e^{-\frac{\eta_0 \Theta^0 t'}{|\mathcal{D}|}}\right) g(\theta^0). \quad (41)$$

When $t' = t\tau$, by comparing (38), (39) with (40), (41) and employing Theorem 2, we can obtain

$$\theta_{\text{cen}}^{t'} = \theta^{t\tau}, \quad f(x, \theta_{\text{cen}}^{t'}) = f(x, \theta^{t\tau}) \quad (42)$$

Remark 6 (Infinite-width FedAvg generalizes the same as centralized GD). *Equations (42) suggest that when the total number iterations of centralized GD and FedAvg are the same, both models share the same model parameters in the infinite-width regime, thereby producing the same output for an arbitrary test input and achieving the same generalization performance. This means that the impacts of data heterogeneity on the generalization performance vanishes.*

6. Numerical Experiments

In this section, we verify our theoretical findings by numerical experiments spanning various network architectures, loss functions, and optimization methods. Specifically, we evaluate the impact of data heterogeneity under different network widths, verify that both the local and global models of overparameterized FedAvg can be well approximated by linear models, and demonstrate that overparameterized FedAvg generalizes the same as centralized learning. The number of clients in our experiments are set to $M = 10$, and the dataset as well as the model settings are provided as follows.[1]

Non-IID Data Generation. 1) Standard dataset: We conduct experiments on two widely used image classification datasets: MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al.). To partition the datasets among different clients and generate non-IID data, we follow the approach proposed by Hsu et al. (2019), which employs the Dirichlet distribution with a concentration parameter α controlling data heterogeneity. Specifically, a smaller value of α indicates a higher degree of data heterogeneity and we set $\alpha = 0.1$ throughout our experiments.

2) Small dataset: To facilitate MSE loss minimization using gradient descent (GD) as required by our theoretical derivations, we also use the mini-MNIST and mini-CIFAR-10 datasets for binary image classification tasks. The mini-MNIST dataset is created by randomly selecting two classes from the MNIST dataset, followed by randomly sampling 50 images from each class for the training set and 10 images from each class for the test set. A similar approach is used to generate the mini-CIFAR-10 dataset, which contains 500 training images and 100 test images. To generate non-IID data, inspired by McMahan et al. (2017); Zhang et al. (2021), we assign each class exclusively to specific clients: half of the clients receive all images from one class, while the remaining clients receive all images from the others.

Experimental Models. We employ three types of models: FNN, convolutional neural networks (CNNs), and residual networks (ResNets).

1) FNN: The structure of FNNs is detailed in Table 1 of Appendix E, where a width factor k is introduced to adjust the network width. By setting $k = 1, 2, 4, 16$, we construct networks of varying widths, named FNN1, FNN2, FNN4, and FNN16.

2) CNN: We adopt the approach described by Park et al. (2019) to obtain CNNs with varying widths having the base architecture in (LeCun et al., 1998, Figure 2). The details of the architecture are presented in Table 2 of Appendix E, where the width factor k is used to scale the channel size. By setting $k = 1, 2, 8, 32$, we generate CNN1, CNN2, CNN8, and CNN32.

[1] Codes to reproduce the main results are available at https://github.com/kkhuge/ICML2025.

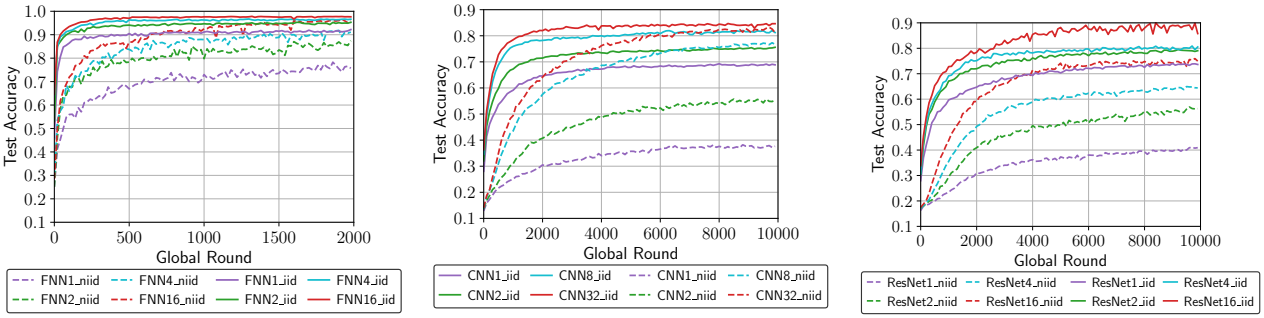

Figure 1. Test accuracy of different network families. Each global round consists of $\tau = 5$ local SGD iterations. The left figure shows the test accuracy of FNNs on both IID and non-IID MNIST datasets. The middle and the right figures show the test accuracy of CNNs and ResNets, respectively, on both IID and non-IID CIFAR-10 datasets.

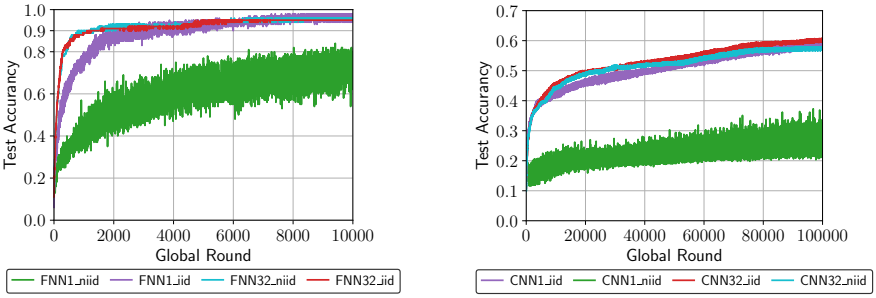

Figure 2. Test accuracy of large networks. Each global round consists of $\tau = 5$ local SGD iterations, $\sigma_W = 1$, $\sigma_b = 0.1$. The left figure shows the test accuracy of the FNN32 and FNN1 on both IID and non-IID MNIST datasets. The right figure shows the test accuracy of the CNN32 and CNN1 on both IID and non-IID CIFAR-10 datasets.

3) ResNet: The network architecture is based on the work of Zagoruyko (2016), as shown in Table 3 of Appendix E. The parameter ψ represents the number of blocks in each group, which is set to $\psi = 1$ in our experiments. The channel size is fixed at 16, and we vary the width factor $k = 1, 2, 4, 16$ to obtain ResNet1, ResNet2, ResNet4, and ResNet16.

6.1. Impact of Non-IID Versus Network Width

Although our theoretical analysis is based on GD with learning rate $\eta = \mathcal{O}(n^{-1})$, to show our conclusions can be extended to more practical settings, we use SGD with batch size 64 and set a common learning rate $\eta = 0.1$ with a weight decay of 0.0005. Moreover, for MNIST and CIFAR-10, we use the practical cross-entropy loss instead of the MSE loss required in the theoretical analysis. As shown in Figure 1, in the non-IID cases, the test accuracy of FNN1, FNN2, FNN4, and FNN16 decreases by 17.4%, 9.5%, 6.3%, and 2.0%, respectively, compared to the IID cases. Similarly, the test accuracy of CNN1, CNN2, CNN8, and CNN32 drops by 44.9%, 26.7%, 5.1%, and 2.4%, while the test accuracy of ResNet1, ResNet2, ResNet4, and ResNet16 decreases by 44.6%, 29.1%, 18.7%, and 14.8%, respectively. These results verify that the impact of data heterogeneity diminishes as the network width increases.

To further verify that the impact of data heterogeneity vanishes as the network width approaches infinity, we set the learning rate with $\eta = \frac{\eta_0}{n}$ in line with our theoretical analysis. As shown in Figure 2, the convergence rate and final accuracy of FNN32 are nearly identical for both IID and non-IID data, and a similar trend is observed for CNN32. In contrast, a noticeable gap exists in FNN1 between IID and non-IID data, which is also evident in CNN1.

Additionally, to monitor the evolution of model parameters and the global/local NTKs during training, we train FNNs on the non-IID mini-MNIST dataset using GD for binary classification with MSE loss. As shown in Figure 3, increasing the network width diminishes the variations in both the global/local NTKs and the model parameters. For sufficiently wide networks, the global and local NTKs as well as the model parameters remain nearly constant, exhibiting a lazy training behavior.

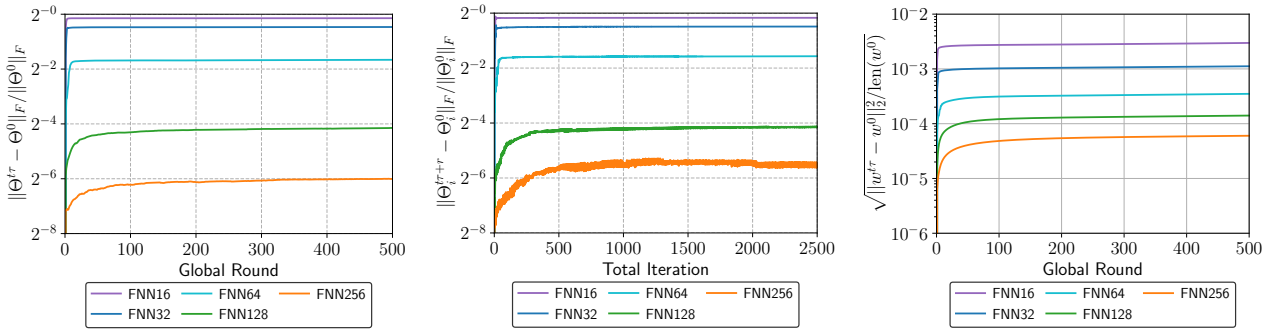

Figure 3. Training dynamic of NTK and model parameters. Each global round consists of $\tau = 5$ local GD iterations, $\eta_0 = 1$, $\sigma_W = 1.5$, $\sigma_b = 0.1$. The left figure shows the variation in the global NTK, the middle figure show the variation in a randomly chosen local NTK, while the right figure shows the model parameters' update during the training process.

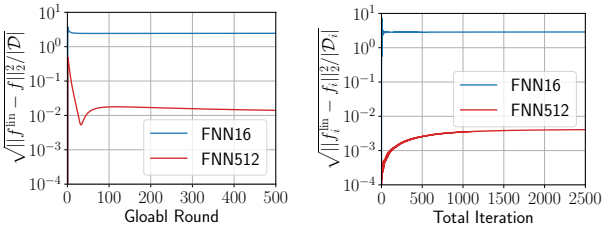

Figure 4. Output difference between FedAvg and linear model. Each global round consists of $\tau = 5$ local GD iterations, $\eta_0 = 1$, $\sigma_W = 1.5$, $\sigma_b = 0.1$.

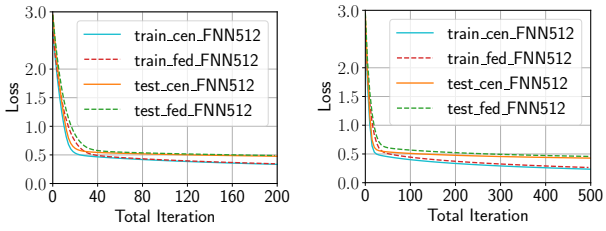

Figure 5. Training and testing loss of FedAvg and centralized learning. $\eta_0 = 0.1$, $\sigma_W = 1.5$, $\sigma_b = 0.1$. The global round of FedAvg consists of $\tau = 2$ and $\tau = 5$ local GD iterations in the left and right figures, respectively.

6.2. Linear Approximation of FedAvg

To show that overparameterized FedAvg can be well approximated by linear models, we train FNN16 and FNN512 on the non-IID mini-MNIST dataset with MSE loss for binary image classification using GD. In Figure 4, we analyze the difference in the outputs between the global model f in FedAvg and the global linear model f^{lin} throughout the training process. Additionally, a randomly selected local model is examined by comparing its outputs with those of the corresponding linear model f_i^{lin}.

As shown in Figure 4, we can observe that, for FNN512, the outputs of the global model f and the linear model f^{lin} remain nearly identical throughout training. By contrast, the narrower FNN16 exhibits noticeable difference between f and f^{lin}. A similar trend is observed for the local models f_i and f_i^{lin}. These findings confirm Theorem 2 that wider networks enable linear approximations to align more closely with the dynamics of FedAvg.

6.3. Comparison of FedAvg with Centralized Learning

To compare overparameterized FedAvg with centralized learning, in Figure 5 we evaluate the loss of $f_{\text{cen}}(x, \theta^{t'})$ and $f(x, \theta^{t\tau})$ on both the training and testing dataset of mini-CIFAR-10, by ensuring $t' = t\tau$ for a fair comparison. It can be observed that the outputs of FedAvg and centralized learning are almost identical under the same number of GD iterations, empirically confirming Theorem 3 that overparameterized FedAvg generalized the same as centralized learning.

7. Conclusion and Future Directions

In this work, we established a quantitative relationship between neural network width and the impact of data heterogeneity in FedAvg. We proved that the impact of data heterogeneity on the convergence of FedAvg diminishes at a rate of $\mathcal{O}(n^{-\frac{1}{2}})$ with increasing network width n and vanishes entirely in the infinite-width limit. In that regime, we extended NTK theory from centralized learning to FL, showing that both the global and local models in FedAvg are linear and have constant NTKs. Furthermore, we derived closed-form expressions for the model outputs of FedAvg, revealing the equivalence between infinite-width FedAvg and centralized GD in both training dynamics and generalization performance, under matched training iterations.

Extensive experiments on MINST and CIFAR-10 datasets validated our conclusions across different network architectures, loss functions, and optimization methods.

These theoretical findings provide valuable insights for practical federated learning. Notably, the linear dependence between model outputs and parameters suggests a potential communication-efficient FL strategy: clients may transmit only the model outputs instead of the complete model parameters for aggregation. This approach may significantly reduce the communication overhead of FL and deserves further investigation. Another essential direction for future research involves extending these analyses to more realistic FL settings by relaxing the idealized assumptions, such as infinite network width and continuous-time gradient flow.

Acknowledgments

We sincerely thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by National Natural Science Foundation of China under Grant 62301015.

Impact Statement

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

References

Cho, Y. J., Wang, J., and Joshi, G. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

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

Durmus, A. E., Yue, Z., Ramon, M., Matthew, M., Paul, W., and Venkatesh, S. Federated learning based on dynamic regularization. In *International conference on learning representations*, 2021.

Goetz, J., Malik, K., Bui, D., Moon, S., Liu, H., and Kumar, A. Active federated learning. *arXiv preprint arXiv:1909.12641*, 2019.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

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

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jacot, A., Gabriel, F., and Hongler, C. The asymptotic spectrum of the hessian of DNN throughout training. In *International Conference on Learning Representations*, 2020.

Jeong, W. and Hwang, S. J. Factorized-FL: Personalized federated learning with parameter factorization & similarity matching. *Advances in Neural Information Processing Systems*, 35:35684–35695, 2022.

Jiang, M., Le, A., Li, X., and Dou, Q. Heterogeneous personalized federated learning by local-global updates mixing via convergence rate. In *International Conference on Learning Representations*, 2024.

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

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local GD on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10 (canadian institute for advanced research). http://www.cs.toronto.edu/~kriz/cifar.html.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous

networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

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

Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021b.

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

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019a.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019b.

Park, D., Sohl-Dickstein, J., Le, Q., and Smith, S. The effect of network width on stochastic gradient descent and generalization: an empirical study. In *International Conference on Machine Learning*, pp. 5042–5051. PMLR, 2019.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Shi, N., Lai, F., Kontar, R. A., and Chowdhury, M. Fed-ensemble: Ensemble models in federated learning for improved generalization and uncertainty quantification. *IEEE Transactions on Automation Science and Engineering*, 21(3):2792–2803, 2024. doi: 10.1109/TASE.2023.3269639.

Song, B., Khanduri, P., Zhang, X., Yi, J., and Hong, M. FedAvg converges to zero training loss linearly for over-parameterized multi-layer neural networks. In *International Conference on Machine Learning*, pp. 32304–32330. PMLR, 2023.

T Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020b.

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

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

Zagoruyko, S. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, J., Li, A., Tang, M., Sun, J., Chen, X., Zhang, F., Chen, C., Chen, Y., and Li, H. Fed-CBS: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction. In *International Conference on Machine Learning*, pp. 41354–41381. PMLR, 2023.

Zhang, X., Hong, M., Dhople, S., Yin, W., and Liu, Y. FedPD: A federated learning framework with adaptivity to non-IID data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*, 2018.

A. Essential Lemmas

In this section, we introduce the necessary lemmas used in deriving Theorems 1 $\sim$ 3.

Lemma 1 (Local Lipschitzness of the Jacobian). *There exists a constant $C > 0$, such that for any $C' > 0$, with high probability over random initialization the following holds:*

$$\begin{cases} \frac{1}{\sqrt{n}} \left\| J(\theta) - J(\theta') \right\|_F \leq C \left\| \theta - \theta' \right\|_2, \\ \frac{1}{\sqrt{n}} \left\| J(\theta) \right\|_F \leq C, \end{cases} \quad \forall \theta, \theta' \in B(\theta_0, C' n^{-\frac{1}{2}}) \tag{43}$$

where $B(\theta_0, R) \triangleq \{\theta : \left\| \theta - \theta_0 \right\|_2 < R\}$.

Lemma 1 has been proved by Lee et al. (2019, Lemma 1) and we will apply it directly.

Lemma 2 (Local Lipschitzness of the Local Jacobian). *There exists a constant $C > 0$, such that for any $C' > 0$, with high probability over random initialization the following holds:*

$$\begin{cases} \frac{1}{\sqrt{n}} \left\| J_i(\theta) - J_i(\theta') \right\|_F \leq C \left\| \theta - \theta' \right\|_2, \\ \frac{1}{\sqrt{n}} \left\| J_i(\theta) \right\|_F \leq C, \end{cases} \quad \forall \theta, \theta' \in B(\theta_0, C' n^{-\frac{1}{2}}) \tag{44}$$

where $B(\theta_0, R) \triangleq \{\theta : \left\| \theta - \theta_0 \right\|_2 < R\}$.

Proof. Since $J(\theta) - J(\theta')$ is the concatenation of all $J_i(\theta) - J_i(\theta')$ for $i = 1, \cdots, M$, we have

$$\frac{1}{\sqrt{n}} ||J_i(\theta) - J_i(\theta')||_F \leq \frac{1}{\sqrt{n}} ||J(\theta) - J(\theta')||_F \leq C ||\theta - \theta'||_2, \tag{45}$$

where the last step applies Lemma 1. Similarly, since $J(\theta)$ is the concatenation of all $J_i(\theta)$, we have

$$\frac{1}{\sqrt{n}} ||J_i(\theta)||_F \leq \frac{1}{\sqrt{n}} ||J(\theta)||_F \leq C. \tag{46}$$

□

Lemma 3. *For a square matrix A whose norm satisfies $||A|| \leq \rho_A$, the remainder term $\Omega(e^{-A}) \triangleq \sum_{k=2}^{\infty} (-1)^k \frac{A^k}{k!}$, i.e., the sum of second-order and higher terms in the Taylor series expansion of e^{-A}, satisfies*

$$\Omega\left(e^{-A}\right) \leq \frac{\rho_A^2}{2} e^{\rho_A}. \tag{47}$$

Proof. By taking the norm on both sides of $\Omega(e^{-A}) = \sum_{k=2}^{\infty} (-1)^k \frac{A^k}{k!}$, we can readily obtain

$$\left\| \Omega\left(e^{-A}\right) \right\| \leq \sum_{k=2}^{\infty} \frac{\|A\|^k}{k!} \leq \frac{\|A\|^2}{2} \sum_{k=2}^{\infty} \frac{\|A\|^{k-2}}{(k-2)!} \leq \frac{\rho_A^2}{2} e^{\|A\|} \leq \frac{\rho_A^2}{2} e^{\rho_A}. \tag{48}$$

□

Lemma 4. *For any non-negative aggregation weights combination $\{p_i\}_{i=1,\cdots,M}$, if $\eta_0 \tau$ is sufficiently small such that the terms of $\mathcal{O}(\eta_0^2 \tau^2)$ and higher are neglected, the following equation holds:*

$$\sum_{i=1}^{M} p_i e^{-\eta_0 \tau \Theta_i} = e^{-\eta_0 \tau \sum_{i=1}^{M} p_i \Theta_i}. \tag{49}$$

Proof. Employing the Taylor series expansion on both sides of (49) yields

$$\sum_{i=1}^{M} p_i e^{-\eta_0 \tau \Theta_i} = \sum_{i=1}^{M} p_i \left[I - \eta_0 \tau \Theta_i + \mathcal{O}\left(\eta_0^2 \tau^2\right) \right] = I - \eta_0 \tau \sum_{i=1}^{M} p_i \Theta_i, \tag{50}$$

and

$$e^{-\eta_0 \tau \sum_{i=1}^{M} p_i \Theta_i} = I - \eta_0 \tau \sum_{i=1}^{M} p_i \Theta_i + \mathcal{O}\left(\eta_0^2 \tau^2\right) = I - \eta_0 \tau \sum_{i=1}^{M} p_i \Theta_i, \tag{51}$$

respectively. Comparing (50) with (51), we can obtain

$$\sum_{i=1}^{M} p_i e^{-\eta_0 \tau \Theta_i} = e^{-\eta_0 \tau \sum_{i=1}^{M} p_i \Theta_i}. \tag{52}$$

□

B. Proof of Theorem 1

Since the width of network is large, with the initialization described in Section 3, the output $f(\theta^0)$ converges to Gaussian process $\mathcal{N}(0, \mathcal{K}(\mathcal{X}, \mathcal{X}))$ according to the central limit theorem, where $\mathcal{K}(\mathcal{X}, \mathcal{X}) = \lim_{n \to \infty} \mathbb{E}\left[f(\theta^0) f(\theta^0)^T \right]$. Therefore, for arbitrarily small $\delta_0 > 0$, there exist constants $R_0 > 0$ and n' such that for any $n \geq n'$, with probability at least $1 - \delta_0$ over random initialization, we have

$$\left\| g\left(\theta^0\right) \right\|_2 \leq R_0. \tag{53}$$

In the following, we prove Theorem 1 by mathematical induction, where the induction hypotheses are

$$\left\| g(\theta^{t\tau}) \right\|_2 \leq q^t R_0 + \frac{2\eta_0 \tau C C_1 R_0 \zeta \left(1 - q^t\right)}{\left(1 - q\right)^2}, \tag{54}$$

$$\left\| \theta^{t\tau} - \theta^0 \right\|_2 \leq \frac{\eta_0 \tau C R_0 \left(1 - q^t\right)}{\sqrt{n}\left(1 - q\right)}. \tag{55}$$

When $t = 0$, (54) and (55) trivially hold. Then, we aim to prove the hypotheses hold in the $(t+1)$-th global round, i.e.,

$$\left\| g(\theta^{(t+1)\tau}) \right\|_2 \leq q^{t+1} R_0 + \frac{2\eta_0 \tau C C_1 R_0 \zeta \left(1 - q^{t+1}\right)}{\left(1 - q\right)^2}, \tag{56}$$

$$\left\| \theta^{(t+1)\tau} - \theta^0 \right\|_2 \leq \frac{\eta_0 \tau C R_0 \left(1 - q^{t+1}\right)}{\sqrt{n}\left(1 - q\right)}. \tag{57}$$

Referring to (5), the local update step of client i in the $(t\tau + r + 1)$-th iteration is

$$\theta_i^{t\tau + r + 1} = \theta_i^{t\tau + r} - \frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^{t\tau + r}\right) g_i\left(\theta_i^{t\tau + r}\right). \tag{58}$$

Since $\eta = \frac{\eta_0}{n}$ is small, we can approximate the local update with gradient flow by making time continuous, yielding

$$\frac{d\theta_i^{t\tau + r}}{dr} = -\frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^{t\tau + r}\right) g_i\left(\theta_i^{t\tau + r}\right). \tag{59}$$

Further applying the chain rule, we can obtain

$$\begin{aligned} \frac{dg\left(\theta_i^{t\tau + r}\right)}{dr} &= J\left(\theta_i^{t\tau + r}\right)^T \frac{d\theta_i^{t\tau + r}}{dr} \\ &= -\frac{\eta}{|\mathcal{D}_i|} J\left(\theta_i^{t\tau + r}\right)^T J_i\left(\theta_i^{t\tau + r}\right) g_i\left(\theta_i^{t\tau + r}\right) \\ &= -\frac{\eta_0}{n|\mathcal{D}_i|} J\left(\theta_i^{t\tau + r}\right)^T J_i\left(\theta_i^{t\tau + r}\right) P_i g\left(\theta_i^{t\tau + r}\right) \end{aligned}$$

$$= -\frac{\eta_0 \Theta_i^{t\tau+r}}{|\mathcal{D}_i|} g\left(\theta_i^{t\tau+r}\right), \tag{60}$$

where the last step follows from the definition of the local NTK in (20). By integrating both sides of (60) from 0 to r and applying the mean value theorem for integrals, we have

$$g\left(\theta_i^{t\tau+r}\right) = e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\hat{r}_i}} g\left(\theta_i^{t\tau}\right), \ \hat{r}_i \in (0, r) \tag{61}$$

Replacing r with τ yields

$$g\left(\theta_i^{t\tau+\tau}\right) = e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}} g\left(\theta_i^{t\tau}\right), \ \bar{r}_i \in (0, \tau) \tag{62}$$

B.1. Proof of (57)

We first prove (57). To bound $\left\|\theta^{(t+1)\tau} - \theta^0\right\|_2$, we proceed the following derivations.

$$\begin{aligned}
\frac{d\left\|\theta_i^{t\tau+r} - \theta_i^0\right\|_2}{dr} &\leq \left\|\frac{d\theta_i^{t\tau+r}}{dr}\right\|_2 \\
&= \frac{\eta}{|\mathcal{D}_i|}\left\|J_i\left(\theta_i^{t\tau+r}\right) g_i\left(\theta_i^{t\tau+r}\right)\right\|_2 \\
&= \frac{\eta}{|\mathcal{D}_i|}\left\|J_i\left(\theta_i^{t\tau+r}\right) P_i g\left(\theta_i^{t\tau+r}\right)\right\|_2 \\
&\leq \frac{\eta}{|\mathcal{D}_i|}\left\|J_i\left(\theta_i^{t\tau+r}\right)\right\|_F \left\|P_i\right\|_{\text{op}} \left\|g\left(\theta_i^{t\tau+r}\right)\right\|_2 \\
&\overset{(a)}{\leq} \frac{\eta C\sqrt{n}}{|\mathcal{D}_i|}\left\|g\left(\theta_i^{t\tau+r}\right)\right\|_2 \\
&= \frac{\eta_0 C}{\sqrt{n}|\mathcal{D}_i|}\left\|e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\hat{r}_i}} g\left(\theta^{t\tau}\right)\right\|_2 \\
&\leq \frac{\eta_0 C}{\sqrt{n}|\mathcal{D}_i|}\left\|e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\hat{r}_i}}\right\|_{\text{op}} \left\|g\left(\theta^{t\tau}\right)\right\|_2 \\
&\leq \frac{\eta_0 C}{\sqrt{n}|\mathcal{D}_i|}\left\|g\left(\theta^{t\tau}\right)\right\|_2,
\end{aligned} \tag{63}$$

where the first step is obtained by applying the chain rule and Cauchy-Schwarz inequality, step (a) holds because of Lemma 2 and $\|P_i\|_{\text{op}} = 1$, and the last step holds because $\Theta_i^{t\tau+\hat{r}_i}$ is not full rank from its definition (20), yielding $\left\|e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\hat{r}_i}}\right\|_{\text{op}} \leq e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\lambda_{\min}(\Theta_i^{t\tau+\hat{r}_i})} = 1$. Integrating from 0 to r on both sides of (63) yields

$$\left\|\theta_i^{t\tau+r} - \theta_i^0\right\|_2 \leq \frac{\eta_0 r C}{\sqrt{n}|\mathcal{D}_i|}\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \left\|\theta^{t\tau} - \theta^0\right\|_2. \tag{64}$$

Further replacing r with τ yields

$$\left\|\theta_i^{t\tau+\tau} - \theta_i^0\right\|_2 \leq \frac{\eta_0 \tau C}{\sqrt{n}|\mathcal{D}_i|}\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \left\|\theta^{t\tau} - \theta^0\right\|_2. \tag{65}$$

According to (10), we can obtain

$$\begin{aligned}
\left\|\theta^{(t+1)\tau} - \theta^0\right\|_2 &= \left\|\sum_{i=1}^M p_i\left(\theta_i^{t\tau+\tau} - \theta^0\right)\right\|_2 \\
&\leq \sum_{i=1}^M p_i\left\|\theta_i^{t\tau+\tau} - \theta^0\right\|_2 \\
&\overset{(a)}{\leq} \sum_{i=1}^M p_i\left[\frac{\eta_0 \tau C}{\sqrt{n}|\mathcal{D}_i|}\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \left\|\theta^{t\tau} - \theta^0\right\|_2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{M\eta_0\tau C}{\sqrt{n}|\mathcal{D}|}\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \left\|\theta^{t\tau}-\theta^0\right\|_2 \\
&\overset{\text{(b)}}{\leq} \frac{\eta_0\tau C}{\sqrt{n}}\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \left\|\theta^{t\tau}-\theta^0\right\|_2 \\
&\overset{\text{(c)}}{\leq} \frac{\eta_0\tau C}{\sqrt{n}}\left[q^t\left(R_0 - \frac{2\eta_0\tau C C_1 R_0\zeta}{(1-q)^2}\right) + \frac{2\eta_0\tau C C_1 R_0\zeta}{(1-q)^2}\right] + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} \\
&= \frac{\eta_0\tau C R_0\left(1-q^{t+1}\right)}{\sqrt{n}\left(1-q\right)},
\end{aligned} \tag{66}
$$

where step (a) comes from (65), step (b) holds because $|\mathcal{D}| \geq M$, step (c) applies the induction hypothesis (54) and (55), and the last step omits $\frac{\zeta}{\sqrt{n}} = \mathcal{O}(n^{-1})$ according to Assumption 1. Therefore, (57) is proved.

B.2. Bounding the Model Divergence

To prove (56), we first bound $\left\|\Delta\theta_i^{(t+1)\tau}\right\|_2$. According to (64), we have:

$$
\begin{aligned}
\left\|\theta_i^{t\tau+r}-\theta_i^0\right\|_2 &\leq \frac{\eta_0 r C}{\sqrt{n}|\mathcal{D}_i|}\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \left\|\theta^{t\tau}-\theta^0\right\|_2 \\
&\leq \frac{\eta_0 r C}{\sqrt{n}|\mathcal{D}_i|}\left[q^t\left(R_0 - \frac{2\eta_0\tau C C_1 R_0\zeta}{(1-q)^2}\right) + \frac{2\eta_0\tau C C_1 R_0\zeta}{(1-q)^2}\right] + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} \\
&= \frac{\eta_0 r C q^t R_0}{\sqrt{n}|\mathcal{D}_i|} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)},
\end{aligned} \tag{67}
$$

where the second step employs the induction hypotheses (54) and (55), and the last step omits $\frac{\zeta}{\sqrt{n}} = \mathcal{O}(n^{-1})$. Referring to (17), we have

$$
\begin{aligned}
\left\|\Delta\theta_i^{(t+1)\tau}\right\|_2 &= \left\|\theta_i^{t\tau+\tau}-\theta^{(t+1)\tau}\right\|_2 \\
&= \left\|\left(\theta_i^{t\tau+\tau}-\theta^0\right) - \left(\theta^{(t+1)\tau}-\theta^0\right)\right\|_2 \\
&= \left\|\left(\theta_i^{t\tau+\tau}-\theta^0\right) - \sum_{i=1}^M p_i\left(\theta_i^{t\tau+\tau}-\theta^0\right)\right\|_2 \\
&\leq \left\|\left(\theta_i^{t\tau+\tau}-\theta^0\right)\right\|_2 + \sum_{i=1}^M p_i\left\|\left(\theta_i^{t\tau+\tau}-\theta^0\right)\right\|_2 \\
&\overset{\text{(a)}}{\leq} \frac{\eta_0\tau C q^t R_0}{\sqrt{n}|\mathcal{D}_i|} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} + \sum_{i=1}^M p_i\left[\frac{\eta_0\tau C q^t R_0}{\sqrt{n}|\mathcal{D}_i|} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)}\right] \\
&= \frac{\eta_0\tau C q^t R_0}{\sqrt{n}|\mathcal{D}_i|} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} + \frac{M\eta_0\tau C q^t R_0}{\sqrt{n}|\mathcal{D}|} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} \\
&\overset{\text{(b)}}{\leq} \frac{\eta_0\tau C q^t R_0}{\sqrt{n}} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} + \frac{\eta_0\tau C q^t R_0}{\sqrt{n}} + \frac{\eta_0\tau C R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)} \\
&= \frac{2\eta_0\tau C R_0}{\sqrt{n}}\left(q^t + \frac{1-q^t}{1-q}\right) \\
&= \frac{2\eta_0\tau C R_0\left(1-q^{t+1}\right)}{\sqrt{n}\left(1-q\right)} \\
&\leq \frac{2\eta_0\tau C R_0}{\sqrt{n}\left(1-q\right)},
\end{aligned} \tag{68}
$$

where the step (a) is from (67), step (b) holds because $|\mathcal{D}_i| \geq 1$ and $|\mathcal{D}| \geq M$, and the last step hold because we require $0 \leq q < 1$.[2] Therefore, the data heterogeneity term can be bounded by

$$\sum_{i=1}^{M} p_i \left\| \Delta\theta_i^{(t+1)\tau} \right\|_2 \leq \sum_{i=1}^{M} p_i \frac{2\eta_0 \tau C R_0}{\sqrt{n}\,(1-q)} = \frac{2\eta_0 \tau C R_0}{\sqrt{n}\,(1-q)} = \zeta. \tag{69}$$

Notably, we have proven (22).

B.3. Proof of (56)

Finally, we prove (56) to finish the induction. Taking the Taylor series expansion of $g\left(\theta_i^{t\tau+\tau}\right)$ at $\theta^{(t+1)\tau}$ yields

$$g\left(\theta_i^{t\tau+\tau}\right) = g\left(\theta^{(t+1)\tau}\right) + J\left(\theta^{(t+1)\tau}\right)^T \Delta\theta_i^{(t+1)\tau} + \Omega_i, \tag{70}$$

where Ω_i represents the remainder terms of order two and above. Taking the sum of both sides yields

$$\begin{aligned}
\sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right) &= \sum_{i=1}^{M} p_i g\left(\theta^{(t+1)\tau}\right) + \sum_{i=1}^{M} p_i J\left(\theta^{(t+1)\tau}\right)^T \Delta\theta_i^{(t+1)\tau} + \sum_{i=1}^{M} p_i \Omega_i \\
&= \sum_{i=1}^{M} p_i g\left(\theta^{(t+1)\tau}\right) + \sum_{i=1}^{M} p_i \Omega_i \\
&= g\left(\theta^{(t+1)\tau}\right) + \sum_{i=1}^{M} p_i \Omega_i,
\end{aligned} \tag{71}$$

where the second step holds because $\sum_{i=1}^{M} p_i J\left(\theta^{(t+1)\tau}\right)^T \Delta\theta_i^{(t+1)\tau} = 0$ according to the model aggregation (10). Rewriting (71), we have

$$g\left(\theta^{(t+1)\tau}\right) = \sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right) - \sum_{i=1}^{M} p_i \Omega_i. \tag{72}$$

Taking the norm of both sides yields

$$\left\| g\left(\theta^{(t+1)\tau}\right) \right\|_2 \leq \left\| \sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right) \right\|_2 + \left\| \sum_{i=1}^{M} p_i \Omega_i \right\|_2. \tag{73}$$

In the following, we bound $\left\| \sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right) \right\|_2$ and $\left\| \sum_{i=1}^{M} p_i \Omega_i \right\|_2$, respectively.

1) Bounding $\left\| \sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right) \right\|_2$: According to (62), we have

$$\begin{aligned}
\left\| \sum_{i=1}^{M} p_i g\left(\theta_i^{t\tau+\tau}\right) \right\|_2 &= \left\| \sum_{i=1}^{M} p_i e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}} g\left(\theta_i^{t\tau}\right) \right\|_2 \\
&= \left\| \sum_{i=1}^{M} p_i \left[I - \frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i} + \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}}\right) \right] g\left(\theta_i^{t\tau}\right) \right\|_2 \\
&\leq \left\| \sum_{i=1}^{M} p_i \left[I - \frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i} + \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}}\right) \right] \right\|_{\text{op}} \left\| g\left(\theta_i^{t\tau}\right) \right\|_2 \\
&\leq \left\| \sum_{i=1}^{M} p_i \left(I - \frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i} \right) \right\|_{\text{op}} \left\| g\left(\theta_i^{t\tau}\right) \right\|_2 + \sum_{i=1}^{M} p_i \left\| \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}}\right) \right\|_{\text{op}} \left\| g\left(\theta_i^{t\tau}\right) \right\|_2
\end{aligned}$$

[2] We prove that there exists $\eta_0 > 0$ such that $0 \leq q < 1$ at the end of Section B.3.

$$= \left\| I - \frac{\eta_0 \tau}{|\mathcal{D}|} \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}} \left\| g\left(\theta_i^{t\tau}\right) \right\|_2 + \sum_{i=1}^{M} p_i \left\| \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|} \Theta_i^{t\tau+\bar{r}_i}} \right) \right\|_{\text{op}} \left\| g\left(\theta_i^{t\tau}\right) \right\|_2, \tag{74}$$

where the second step is obtained by employing the Taylor series expansion and $\Omega(e^{-A}) = \sum_{k=2}^{\infty} \frac{(-1)^k}{k!} A^k$. Then, we bound $\left\| I - \frac{\eta_0 \tau}{|\mathcal{D}|} \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}}$ and $\sum_{i=1}^{M} p_i \left\| \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|} \Theta_i^{t\tau+\bar{r}_i}} \right) \right\|_{\text{op}}$, respectively.

$\left\| I - \frac{\eta_0 \tau}{|\mathcal{D}|} \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}}$ can be rewritten as

$$\begin{aligned}
\left\| I - \frac{\eta_0 \tau}{|\mathcal{D}|} \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}} &= \left\| I - \frac{\eta_0 \tau \Theta}{|\mathcal{D}|} + \frac{\eta_0 \tau \Theta}{|\mathcal{D}|} - \frac{\eta_0 \tau \Theta^0}{|\mathcal{D}|} + \frac{\eta_0 \tau \Theta^0}{|\mathcal{D}|} - \frac{\eta_0 \tau}{|\mathcal{D}|} \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}} \\
&\leq \left\| I - \frac{\eta_0 \tau \Theta}{|\mathcal{D}|} \right\|_{\text{op}} + \left\| \frac{\eta_0 \tau \Theta}{|\mathcal{D}|} - \frac{\eta_0 \tau \Theta^0}{|\mathcal{D}|} \right\|_{\text{op}} + \left\| \frac{\eta_0 \tau \Theta^0}{|\mathcal{D}|} - \frac{\eta_0 \tau}{|\mathcal{D}|} \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}} \\
&= \left(1 - \frac{\eta_0 \tau \lambda_m}{|\mathcal{D}|} \right) + \frac{\eta_0 \tau}{|\mathcal{D}|} \left\| \Theta - \Theta^0 \right\|_{\text{op}} + \frac{\eta_0 \tau}{|\mathcal{D}|} \left\| \Theta^0 - \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}} \\
&\leq \left(1 - \frac{\eta_0 \tau \lambda_m}{|\mathcal{D}|} \right) + \frac{\eta_0 \tau}{|\mathcal{D}|} \left\| \Theta - \Theta^0 \right\|_F + \frac{\eta_0 \tau}{|\mathcal{D}|} \left\| \Theta^0 - \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} \right\|_{\text{op}}.
\end{aligned} \tag{75}$$

According to Lee et al. (2019, Section G.1), there exists n'', such that the following event

$$\left\| \Theta^0 - \Theta \right\|_F \leq \frac{\lambda_m}{3} \tag{76}$$

with probability at least $1 - \frac{\delta_0}{5}$ hold for any $n \geq n''$. As for $\left\| \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} - \Theta^0 \right\|_{\text{op}}$, according to the definition of local NTK (20), we can obtain

$$\begin{aligned}
\left\| \sum_{i=1}^{M} \Theta_i^{t\tau+\bar{r}_i} - \Theta^0 \right\|_{\text{op}} &= \frac{1}{n} \left\| \sum_{i=1}^{M} J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) P_i - \sum_{i=1}^{M} J\left(\theta^0\right)^T J_i\left(\theta_i^0\right) P_i \right\|_{\text{op}} \\
&= \frac{1}{n} \left\| \sum_{i=1}^{M} \left[J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) P_i - J\left(\theta^0\right)^T J_i\left(\theta_i^0\right) P_i \right] \right\|_{\text{op}} \\
&\leq \frac{1}{n} \sum_{i=1}^{M} \left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) - J\left(\theta^0\right)^T J_i\left(\theta_i^0\right) \right\|_{\text{op}} \left\| P_i \right\|_{\text{op}} \\
&= \frac{1}{n} \sum_{i=1}^{M} \left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) - J\left(\theta_i^0\right)^T J_i\left(\theta_i^0\right) \right\|_{\text{op}} \\
&= \frac{1}{n} \sum_{i=1}^{M} \left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T \left[J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) - J_i\left(\theta_i^0\right) \right] + \left[J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T - J\left(\theta_i^0\right)^T \right] J_i\left(\theta_i^0\right) \right\|_{\text{op}} \\
&\leq \frac{1}{n} \sum_{i=1}^{M} \left[\left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right) \right\|_{\text{op}} \left\| J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) - J_i\left(\theta_i^0\right) \right\|_{\text{op}} + \left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T - J\left(\theta_i^0\right)^T \right\|_{\text{op}} \left\| J_i\left(\theta_i^0\right) \right\|_{\text{op}} \right] \\
&\leq \frac{1}{n} \sum_{i=1}^{M} \left[\left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right) \right\|_F \left\| J_i\left(\theta_i^{t\tau+\bar{r}_i}\right) - J_i\left(\theta_i^0\right) \right\|_F + \left\| J\left(\theta_i^{t\tau+\bar{r}_i}\right)^T - J\left(\theta_i^0\right)^T \right\|_F \left\| J_i\left(\theta_i^0\right) \right\|_F \right] \\
&\leq 2C^2 \sum_{i=1}^{M} \left\| \theta_i^{t\tau+\bar{r}_i} - \theta_i^0 \right\|_2,
\end{aligned} \tag{77}$$

where the last step holds because of Lemma 1 and Lemma 2. Plugging (67) into (77) yields

$$\begin{aligned}\left\|\sum_{i=1}^{M}\Theta_i^{t\tau+\bar{r}_i}-\Theta^0\right\|_{\text{op}} &\leq \frac{2\eta_0\bar{r}_iC^3q^tR_0}{\sqrt{n}|\mathcal{D}_i|}+\frac{2\eta_0\tau C^3R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)}\\ &\leq \frac{2\eta_0\tau C^3R_0}{\sqrt{n}}+\frac{2\eta_0\tau C^3R_0}{\sqrt{n}\left(1-q\right)}\\ &= \frac{4\eta_0\tau C^3R_0\left(2-q\right)}{\sqrt{n}\left(1-q\right)}\\ &\leq \frac{\lambda_m}{3},\end{aligned} \tag{78}$$

where the second step holds because $\bar{r}_i \leq \tau$, $|\mathcal{D}_i| \geq 1$ and $0 \leq q < 1$, and the last step holds when $n \geq n'''$ with $n''' = \frac{144\eta_0^2\tau^2C^6R_0^2(2-q)^2}{\lambda_m^2(1-q)^2}$. Then, plugging (76) and (78) into (75), we can obtain

$$\left\|I-\frac{\eta_0\tau}{|\mathcal{D}|}\sum_{i=1}^{M}\Theta_i^{t\tau+\bar{r}_i}\right\|_{\text{op}} \leq 1-\frac{\eta_0\tau\lambda_m}{3|\mathcal{D}|}. \tag{79}$$

To bound $\sum_{i=1}^{M}p_i\left\|\Omega\left(e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}}\right)\right\|_{\text{op}}$, we first bound $\left\|\Theta_i^{t\tau+\bar{r}_i}\right\|_{\text{op}}$. According to (20), we have

$$\begin{aligned}\left\|\Theta_i^{t\tau+\bar{r}_i}\right\|_{\text{op}} &= \frac{1}{n}\left\|J\left(\theta_i^{t\tau+\bar{r}_i}\right)^TJ_i\left(\theta_i^{t\tau+\bar{r}_i}\right)P_i\right\|_{\text{op}}\\ &\leq \frac{1}{n}\left\|J\left(\theta_i^{t\tau+\bar{r}_i}\right)\right\|_F\left\|J_i\left(\theta_i^{t\tau+\bar{r}_i}\right)\right\|_F\left\|P_i\right\|_{\text{op}}\\ &= \frac{1}{n}\left\|J\left(\theta_i^{t\tau+\bar{r}_i}\right)\right\|_F\left\|J_i\left(\theta_i^{t\tau+\bar{r}_i}\right)\right\|_F\\ &\leq C^2,\end{aligned} \tag{80}$$

where the last step holds beacuse of Lemma 1 and Lemma 2. Then, according to Lemma 3, we can obtain

$$\begin{aligned}\sum_{i=1}^{M}p_i\left\|\Omega\left(e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+\bar{r}_i}}\right)\right\|_{\text{op}} &\leq \sum_{i=1}^{M}\frac{p_i}{2}\left(\frac{\eta_0\tau C^2}{|\mathcal{D}_i|}\right)^2e^{\frac{\eta_0\tau C^2}{|\mathcal{D}_i|}}\\ &= \sum_{i=1}^{M}\frac{\eta_0^2\tau^2C^4}{2|\mathcal{D}_i||\mathcal{D}|}e^{\frac{\eta_0\tau C^2}{|\mathcal{D}_i|}}\\ &\leq \sum_{i=1}^{M}\frac{\eta_0^2\tau^2C^4}{2|\mathcal{D}|}e^{\eta_0\tau C^2}\\ &= \frac{M\eta_0^2\tau^2C^4}{2|\mathcal{D}|}e^{\eta_0\tau C^2}\\ &\leq \frac{\eta_0^2\tau^2C^4}{2}e^{\eta_0\tau C^2}.\end{aligned} \tag{81}$$

Further plugging (79) and (81) to (74), we have

$$\left\|\sum_{i=1}^{M}p_ig\left(\theta_i^{t\tau+\tau}\right)\right\|_2 \leq \left(1-\frac{\eta_0\tau\lambda_m}{3|\mathcal{D}|}+\frac{\eta_0^2\tau^2C^4}{2}e^{\eta_0\tau C^2}\right)\left\|g\left(\theta_i^{t\tau}\right)\right\|_2 \triangleq q\left\|g\left(\theta_i^{t\tau}\right)\right\|_2 \tag{82}$$

2) Bounding $\left\|\sum_{i=1}^{M}p_i\Omega_i\right\|_2$: By defining $\Gamma\left(\beta\right)=g\left(\theta^{(t+1)\tau}+\beta\Delta\theta_i^{(t+1)\tau}\right), \beta\in[0,1]$, we can obtain

$$\Gamma\left(0\right)=g\left(\theta^{(t+1)\tau}\right),\quad \Gamma'\left(0\right)=\nabla g\left(\theta^{(t+1)\tau}\right)\Delta\theta_i^{(t+1)\tau},\quad \Gamma\left(1\right)=g\left(\theta_i^{t\tau+\tau}\right). \tag{83}$$

We further define

$$u(\beta) = \Gamma(\beta) + (1-\beta)\Gamma'(\beta). \tag{84}$$

Then, we can obtain

$$\begin{aligned}\|u(1) - u(0)\|_2 &= \|\Gamma(1) - \Gamma(0) - \Gamma'(0)\|_2 \\ &= \left\| g\left(\theta_i^{t\tau+\tau}\right) - g\left(\theta^{(t+1)\tau}\right) - \nabla g\left(\theta^{(t+1)\tau}\right)\Delta\theta_i^{(t+1)\tau}\right\|_2 \\ &= \|\Omega_i\|_2 .\end{aligned} \tag{85}$$

According to mean value inequality, we have

$$\|u(1) - u(0)\|_2 \leq \sup_{0\leq\beta\leq 1} \|u'(\beta)\|_2 (1-0). \tag{86}$$

Combining (85) and (86) yields

$$\|\Omega_i\|_2 \leq \sup_{0\leq\beta\leq 1} \|u'(\beta)\|_2 . \tag{87}$$

Taking the derivative on both side of (84) yields

$$\begin{aligned}u'(\beta) &= \Gamma'(\beta) - \Gamma'(\beta) + (1-\beta)\Gamma''(\beta) \\ &= (1-\beta)\Gamma''(\beta) \\ &= (1-\beta)\left(\Delta\theta_i^{(t+1)\tau}\right)^T \nabla^2 g\left(\theta^{(t+1)\tau} + \beta\Delta\theta_i^{(t+1)\tau}\right)\Delta\theta_i^{(t+1)\tau}.\end{aligned} \tag{88}$$

By plugging (88) into (87), and then (87) into $\left\|\sum_{i=1}^M p_i\Omega_i\right\|_2$, we can obtain

$$\begin{aligned}\left\|\sum_{i=1}^M p_i\Omega_i\right\|_2 &\leq \sum_{i=1}^M p_i \|\Omega_i\|_2 \\ &\leq \sum_{i=1}^M p_i \sup_{0\leq\beta\leq 1} \|u'(\beta)\|_2 \\ &= \sum_{i=1}^M p_i (1-\beta) \sup_{0\leq\beta\leq 1} \left\|\left(\Delta\theta_i^{(t+1)\tau}\right)^T \nabla^2 g\left(\theta^{(t+1)\tau} + \beta\Delta\theta_i^{(t+1)\tau}\right)\Delta\theta_i^{(t+1)\tau}\right\|_2 \\ &\leq \sum_{i=1}^M p_i (1-\beta) \left\|\Delta\theta_i^{(t+1)\tau}\right\|_2 \sup_{0\leq\beta\leq 1} \left\|\nabla^2 g\left(\theta^{(t+1)\tau} + \beta\Delta\theta_i^{(t+1)\tau}\right)\right\|_{\text{op}} \left\|\Delta\theta_i^{(t+1)\tau}\right\|_2 \\ &\overset{\text{(a)}}{\leq} \sum_{i=1}^M p_i (1-\beta) \left\|\Delta\theta_i^{(t+1)\tau}\right\|_2 C_1\sqrt{n} \left\|\Delta\theta_i^{(t+1)\tau}\right\|_2 \\ &\overset{\text{(b)}}{\leq} \sum_{i=1}^M p_i (1-\beta) \frac{2\eta_0\tau C R_0}{\sqrt{n}(1-q)} \cdot C_1\sqrt{n} \cdot \frac{2\eta_0\tau C R_0}{\sqrt{n}(1-q)} \\ &\overset{\text{(c)}}{\leq} \frac{4\eta_0^2\tau^2 C^2 R_0^2 C_1}{\sqrt{n}(1-q)^2} \\ &= \frac{2\eta_0\tau C C_1 R_0 \zeta}{(1-q)}.\end{aligned} \tag{89}$$

where step (a) applies $\frac{1}{\sqrt{n}}\left\|\nabla^2 g(\theta)\right\|_{\text{op}} \leq C_1$, which is proved by Jacot et al. (2020, Lemma 1) under Assumption 4, step (b) comes from (68), step (c) holds because $\beta \in [0,1]$, and the last step comes from (69). Plugging (82) and (89) into (73):

$$\left\| g\left(\theta^{(t+1)\tau}\right)\right\|_2 \leq q\left\|g\left(\theta^{t\tau}\right)\right\|_2 + \frac{2\eta_0\tau C C_1 R_0 \zeta}{(1-q)}$$

$$\overset{(\mathsf{a})}{=} q\left[q^t\left(R_0-\frac{2\eta_0\tau CC_1R_0\zeta}{(1-q)^2}\right)+\frac{2\eta_0\tau CC_1R_0\zeta}{(1-q)^2}\right]+(1-q)\frac{2\eta_0\tau CC_1R_0\zeta}{(1-q)^2}$$
$$=q^{t+1}\left(R_0-\frac{2\eta_0\tau CC_1R_0\zeta}{(1-q)^2}\right)+\frac{2\eta_0\tau CC_1R_0\zeta}{(1-q)^2}, \tag{90}$$

where step (a) applies induction hypothesis (54). At this point, we have proven (56). Recall that (57) have also been proven and we require $n\geq n',n'',n'''$ during the derivations, and hence the induction hypotheses (54) and (55) hold for $n\geq N\triangleq\max\{n',n'',n'''\}$.

Noting that we also require $0\leq q<1$ to complete the proof, we prove there exists $\eta_0>0$ such that $0\leq q<1$ in the following. Referring to (82), we rewrite q as a function of η_0 as

$$q\left(\eta_0\right)=1-\frac{\eta_0\tau\lambda_m}{3|\mathcal{D}|}+\frac{\eta_0^2\tau^2C^4}{2}e^{\eta_0\tau C^2}. \tag{91}$$

Taking the derivative of $q\left(\eta_0\right)$ with respect to η_0, we obtain

$$q'\left(\eta_0\right)=-\frac{\tau\lambda_m}{3|\mathcal{D}|}+\eta_0\tau^2C^4e^{\eta_0\tau C^2}+\frac{\eta_0^2\tau^3C^6}{2}e^{\eta_0\tau C^2}. \tag{92}$$

Further taking the second derivative, we have

$$q''\left(\eta_0\right)=\tau^2C^4e^{\eta_0\tau C^2}+\eta_0\tau^3C^6e^{\eta_0\tau C^2}+\frac{\eta_0^2\tau^4C^8}{2}e^{\eta_0\tau C^2}>0. \tag{93}$$

Therefore, $q'\left(\eta_0\right)$ is monotonically increasing. According to Assumption 2, we have $\lambda_m>0$ and hence it obvious that $q'\left(0\right)<0$ and $\lim_{\eta_0\to\infty}q'\left(\eta_0\right)>0$, and hence there exists $\eta_0'>0$ such that $q'\left(\eta_0'\right)=0$ holds. Consequently, $q\left(\eta_0\right)$ is monotonically decreasing on $(0,\eta_0']$, and monotonically increasing on (η_0',∞). Additionally, from (91), we have $\lim_{\eta_0\to0}q\left(\eta_0\right)=1$. Consequently, if $q\left(\eta_0'\right)\geq0$, $0\leq q\left(\eta_0\right)<1$ holds for $0<\eta_0\leq\eta_0'$. Otherwise, if $q\left(\eta_0'\right)<0$, there exists $\eta_0''\in(0,\eta_0')$ such that $q\left(\eta_0''\right)=0$ holds. Then, $0\leq q\left(\eta_0\right)<1$ holds for $0<\eta_0\leq\eta_0''$. To sum up, as long as $0<\eta_0\leq\min\{\eta_0',\eta_0''\}$, $0\leq q\left(\eta_0\right)<1$ holds.

B.4. Bounding the Variation on Global and Local NTKs

We continue to prove (25) and (26) in Theorem 1. Referring to (18), we have

$$\begin{aligned}\left\|\Theta^{t\tau}-\Theta^0\right\|_F&=\frac{1}{n}\left\|J\left(\theta^{t\tau}\right)^TJ\left(\theta^{t\tau}\right)-J\left(\theta^0\right)^TJ\left(\theta^0\right)\right\|_F\\&=\frac{1}{n}\left\|\left[J\left(\theta^{t\tau}\right)^T-J\left(\theta^0\right)^T\right]J\left(\theta^{t\tau}\right)-J\left(\theta^0\right)^T\left[J\left(\theta^{t\tau}\right)-J\left(\theta^0\right)\right]\right\|_F\\&\leq\frac{1}{n}\left\|J\left(\theta^{t\tau}\right)-J\left(\theta^0\right)\right\|_F\left\|J\left(\theta^{t\tau}\right)\right\|_F+\frac{1}{n}\left\|J\left(\theta^0\right)\right\|_F\left\|J\left(\theta^{t\tau}\right)-J\left(\theta^0\right)\right\|_F\\&\leq2C^2\left\|\theta^{t\tau}-\theta^0\right\|_2\\&\leq\frac{2\eta_0\tau C^3R_0\left(1-q^t\right)}{\sqrt{n}\left(1-q\right)}.\end{aligned} \tag{94}$$

where the fourth step holds because of Lemma 1 and the last step holds because of (55).

Referring to (20), we have

$$\begin{aligned}\left\|\Theta_i^{t\tau+r}-\Theta_i^0\right\|_F&=\frac{1}{n}\left\|J\left(\theta_i^{t\tau+r}\right)^TJ_i\left(\theta_i^{t\tau+r}\right)P_i-J\left(\theta_i^0\right)^TJ_i\left(\theta_i^0\right)P_i\right\|_F\\&\leq\frac{1}{n}\left\|J\left(\theta_i^{t\tau+r}\right)^TJ_i\left(\theta_i^{t\tau+r}\right)-J\left(\theta_i^0\right)^TJ_i\left(\theta_i^0\right)\right\|_F\left\|P_i\right\|_F\\&\leq\frac{\sqrt{k|\mathcal{D}_i|}}{n}\left\|J\left(\theta_i^{t\tau+r}\right)^TJ_i\left(\theta_i^{t\tau+r}\right)-J\left(\theta_i^0\right)^TJ_i\left(\theta_i^0\right)\right\|_F\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{k|\mathcal{D}_i|}}{n} \left\| \left[J\left(\theta_i^{t\tau+r}\right)^T - J\left(\theta_i^0\right)^T \right] J_i\left(\theta_i^{t\tau+r}\right) + J\left(\theta_i^0\right)^T \left[J_i\left(\theta_i^{t\tau+r}\right) - J_i\left(\theta_i^0\right) \right] \right\|_F \\
&\leq \frac{\sqrt{k|\mathcal{D}_i|}}{n} \left\| J\left(\theta_i^{t\tau+r}\right) - J\left(\theta_i^0\right) \right\|_F \left\| J_i\left(\theta_i^{t\tau+r}\right) \right\|_F + \frac{\sqrt{k|\mathcal{D}_i|}}{n} \left\| J\left(\theta_i^0\right) \right\|_F \left\| J_i\left(\theta_i^{t\tau+r}\right) - J_i\left(\theta_i^0\right) \right\|_F \\
&\overset{(a)}{\leq} 2C^2 \sqrt{k|\mathcal{D}_i|} \left\| \theta_i^{t\tau+r} - \theta_i^0 \right\|_2 \\
&\leq \frac{2\eta_0 r q^t C^3 R_0 \sqrt{k}}{\sqrt{n|\mathcal{D}_i|}} + \frac{2\eta_0 \tau C^3 R_0 \left(1-q^t\right) \sqrt{k|\mathcal{D}_i|}}{(1-q)\sqrt{n}},
\end{aligned} \tag{95}
$$

where step (a) employs Lemma 1 and Lemma 2, and the last step holds because of (67). We have now completed the proof of Theorem 1.

C. Proof of Theorem 2

To prove Theorem 2, we first bound $\left\| g^{\text{lin}}\left(\theta^{(t+1)\tau}\right) - g\left(\theta^{(t+1)\tau}\right) \right\|_2$. According to (34), we have

$$
g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) = g\left(\theta^0\right) + J\left(\theta^0\right)^T \left(\theta_i^{t\tau+r} - \theta^0\right). \tag{96}
$$

Taking the derivative with respect to $\theta_i^{t\tau+r}$ on both sides yields

$$
J^{\text{lin}}\left(\theta_i^{t\tau+r}\right) = J\left(\theta^0\right). \tag{97}
$$

Then, we consider

$$
\begin{aligned}
&\frac{d}{dr}\left(e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left[g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) - g\left(\theta_i^{t\tau+r}\right) \right] \right) \\
&= \frac{\eta_0 \Theta_i^0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left[g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) - g\left(\theta_i^{t\tau+r}\right) \right] \\
&\quad + \frac{\eta_0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left[-\frac{1}{n} J(\theta^0)^T J_i\left(\theta_i^0\right) g_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right) + \frac{1}{n} J\left(\theta_i^{t\tau+r}\right)^T J_i\left(\theta_i^{t\tau+r}\right) g_i\left(\theta_i^{t\tau+r}\right) \right] \\
&= \frac{\eta_0 \Theta_i^0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) - g\left(\theta_i^{t\tau+r}\right) \right) \\
&\quad + \frac{\eta_0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left[-\frac{1}{n} J(\theta^0)^T J_i\left(\theta_i^0\right) P_i g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) + \frac{1}{n} J\left(\theta_i^{t\tau+r}\right)^T J_i\left(\theta_i^{t\tau+r}\right) P_i g\left(\theta_i^{t\tau+r}\right) \right] \\
&= \frac{\eta_0 \Theta_i^0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left[g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) - g\left(\theta_i^{t\tau+r}\right) \right] + \frac{\eta_0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left[-\Theta_i^0 g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) + \Theta_i^{t\tau+r} g\left(\theta_i^{t\tau+r}\right) \right] \\
&= \frac{\eta_0}{|\mathcal{D}_i|} e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(\Theta_i^{t\tau+r} - \Theta_i^0 \right) g\left(\theta_i^{t\tau+r}\right).
\end{aligned} \tag{98}
$$

By integrating both sides from 0 to τ, we obtain

$$
g^{\text{lin}}\left(\theta_i^{t\tau+\tau}\right) - g\left(\theta_i^{t\tau+\tau}\right) = e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \left[g^{\text{lin}}\left(\theta^{t\tau}\right) - g\left(\theta^{t\tau}\right) \right] + \frac{\eta_0}{|\mathcal{D}_i|} e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(\Theta_i^{t\tau+r} - \Theta_i^0 \right) g\left(\theta_i^{t\tau+r}\right) dr. \tag{99}
$$

The aggregation of linear local models yields

$$
g^{\text{lin}}\left(\theta^{(t+1)\tau}\right) = \sum_{i=1}^M p_i g^{\text{lin}}\left(\theta_i^{t\tau+\tau}\right). \tag{100}
$$

Further considering equation (72), we have

$$
g^{\text{lin}}\left(\theta^{(t+1)\tau}\right) - g\left(\theta^{(t+1)\tau}\right) = \sum_{i=1}^M p_i \left[g^{\text{lin}}\left(\theta_i^{t\tau+\tau}\right) - g\left(\theta_i^{t\tau+\tau}\right) \right] + \sum_{i=1}^M p_i \Omega_i. \tag{101}
$$

By taking the norm on both sides, we obtain

$$\begin{aligned}
&\left\|g^{\text{lin}}\left(\theta^{(t+1)\tau}\right)-g\left(\theta^{(t+1)\tau}\right)\right\|_2 \\
&=\left\|\sum_{i=1}^M p_i\left[g^{\text{lin}}\left(\theta_i^{t\tau+\tau}\right)-g\left(\theta_i^{t\tau+\tau}\right)\right]+\sum_{i=1}^M p_i\Omega_i\right\|_2 \\
&\leq\left\|\sum_{i=1}^M p_i\left[g^{\text{lin}}\left(\theta_i^{t\tau+\tau}\right)-g\left(\theta_i^{t\tau+\tau}\right)\right]\right\|_2+\left\|\sum_{i=1}^M p_i\Omega_i\right\|_2 \\
&\leq\left\|\sum_{i=1}^M p_i e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\left[g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right]\right\|_2+\left\|\sum_{i=1}^M p_i\frac{\eta_0}{|\mathcal{D}_i|}e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0}\left(\Theta_i^{t\tau+r}-\Theta_i^0\right)g\left(\theta_i^{t\tau+r}\right)dr\right\|_{\text{op}} \\
&\quad+\left\|\sum_{i=1}^M p_i\Omega_i\right\|_2, \qquad (102)
\end{aligned}$$

where the last step is obtained by substituting (99), and $\left\|\sum_{i=1}^M p_i\Omega_i\right\|_2$ is bounded in (89). In the following, we bound $\left\|\sum_{i=1}^M p_i e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\left(g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right)\right\|_2$ and $\left\|\sum_{i=1}^M p_i\frac{\eta_0}{|\mathcal{D}_i|}e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0}\left(\Theta_i^{t\tau+r}-\Theta_i^0\right)g\left(\theta_i^{t\tau+r}\right)dr\right\|_{\text{op}}$, respectively, and finally bound $\left\|f^{\text{lin}}\left(\theta^{t\tau+r}\right)-f\left(\theta^{t\tau+r}\right)\right\|_2$ as well as $\left\|f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-f_i\left(\theta_i^{t\tau+r}\right)\right\|_2$.

C.1. Bounding $\left\|\sum_{i=1}^M p_i e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\left(g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right)\right\|_2$

By taking the Taylor series expansion of $e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}$, we obtain

$$\begin{aligned}
&\left\|\sum_{i=1}^M p_i e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\left(g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right)\right\|_2 \\
&\leq\left\|\sum_{i=1}^M p_i e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\right\|_{\text{op}}\left\|g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right\|_2 \\
&=\left\|\sum_{i=1}^M p_i\left[I-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0+\Omega\left(e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\right)\right]\right\|_{\text{op}}\left\|g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right\|_2 \\
&=\left\|I-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_{\text{op}}\left\|g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right\|_2+\sum_{i=1}^M p_i\left\|\Omega\left(e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\right)\right\|_{\text{op}}\left\|g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right\|_2. \qquad (103)
\end{aligned}$$

Next, we bound $\left\|I-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_{\text{op}}$ and $\sum_{i=1}^M p_i\left\|\Omega\left(e^{-\frac{\eta_0\tau}{|\mathcal{D}_i|}\Theta_i^0}\right)\right\|_{\text{op}}$, respectively.

$$\begin{aligned}
\left\|I-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_{\text{op}}&=\left\|I-\frac{\eta_0\Theta\tau}{|\mathcal{D}|}+\frac{\eta_0\Theta\tau}{|\mathcal{D}|}-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_{\text{op}} \\
&\leq\left\|I-\frac{\eta_0\Theta\tau}{|\mathcal{D}|}\right\|_{\text{op}}+\left\|\frac{\eta_0\Theta\tau}{|\mathcal{D}|}-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_{\text{op}} \\
&=\left(1-\frac{\eta_0\tau\lambda_m}{|\mathcal{D}|}\right)+\left\|\frac{\eta_0\Theta\tau}{|\mathcal{D}|}-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_{\text{op}} \\
&\leq\left(1-\frac{\eta_0\tau\lambda_m}{|\mathcal{D}|}\right)+\left\|\frac{\eta_0\Theta\tau}{|\mathcal{D}|}-\frac{\eta_0\Theta^0\tau}{|\mathcal{D}|}\right\|_F \\
&\overset{\text{(a)}}{\leq}\left(1-\frac{\eta_0\tau\lambda_m}{|\mathcal{D}|}\right)+\frac{\eta_0\tau\lambda_m}{3|\mathcal{D}|} \\
&\leq 1-\frac{\eta_0\tau\lambda_m}{3|\mathcal{D}|} \qquad (104)
\end{aligned}$$

where step (a) applies (76). To bound $\sum_{i=1}^{M} p_i \left\| \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) \right\|_{\text{op}}$, we first bound $\left\|\Theta_i^0\right\|_{\text{op}}$ as

$$\begin{aligned}
\left\|\Theta_i^0\right\|_{\text{op}} &= \left\| \frac{1}{n} J(\theta^0)^T J_i(\theta_i^0) P_i \right\|_{\text{op}} \\
&\leq \frac{1}{n} \left\|J(\theta^0)\right\|_F \left\|J_i(\theta_i^0)\right\|_F \left\|P_i\right\|_{\text{op}} \\
&= \frac{1}{n} \left\|J(\theta^0)\right\|_F \left\|J_i(\theta_i^0)\right\|_F \\
&\leq C^2, \qquad (105)
\end{aligned}$$

where the last step applies Lemmas 1 and 2. Then, we have

$$\begin{aligned}
\sum_{i=1}^{M} p_i \left\| \Omega\left(e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) \right\|_{\text{op}} &\leq \sum_{i=1}^{M} \frac{p_i}{2} \left(\frac{\eta_0 \tau C^2}{|\mathcal{D}_i|}\right)^2 e^{\frac{\eta_0 \tau C^2}{|\mathcal{D}_i|}} \\
&= \sum_{i=1}^{M} \frac{\eta_0^2 \tau^2 C^4}{2|\mathcal{D}_i||\mathcal{D}|} e^{\frac{\eta_0 \tau C^2}{|\mathcal{D}_i|}} \\
&\leq \sum_{i=1}^{M} \frac{\eta_0^2 \tau^2 C^4}{2|\mathcal{D}|} e^{\eta_0 \tau C^2} \\
&= \frac{M \eta_0^2 \tau^2 C^4}{2|\mathcal{D}|} e^{\eta_0 \tau C^2} \\
&\leq \frac{\eta_0^2 \tau^2 C^4}{2} e^{\eta_0 \tau C^2}. \qquad (106)
\end{aligned}$$

where the first step employs Lemma 3. Further, plugging (104) and (106) into (103) yields

$$\left\| \sum_{i=1}^{M} p_i e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \left(g^{\text{lin}}\left(\theta^{t\tau}\right) - g\left(\theta^{t\tau}\right)\right) \right\|_2 \leq q \left\|g^{\text{lin}}\left(\theta^{t\tau}\right) - g\left(\theta^{t\tau}\right)\right\|_2. \qquad (107)$$

C.2. Bounding $\left\| \sum_{i=1}^{M} p_i \frac{\eta_0}{|\mathcal{D}_i|} e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(\Theta_i^{t\tau+r} - \Theta_i^0\right) g\left(\theta_i^{t\tau+r}\right) dr \right\|_{\text{op}}$

Considering $p_i = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}$, we can obtain

$$\begin{aligned}
&\left\| \sum_{i=1}^{M} p_i \frac{\eta_0}{|\mathcal{D}_i|} e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(\Theta_i^{t\tau+r} - \Theta_i^0\right) g\left(\theta_i^{t\tau+r}\right) dr \right\|_{\text{op}} \\
&= \left\| \sum_{i=1}^{M} \frac{\eta_0}{|\mathcal{D}|} e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(\Theta_i^{t\tau+r} - \Theta_i^0\right) g\left(\theta_i^{t\tau+r}\right) dr \right\|_{\text{op}} \\
&\overset{\text{(a)}}{=} \left\| \sum_{i=1}^{M} \frac{\eta_0 \tau}{|\mathcal{D}|} \left(\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right) e^{\frac{\eta_0 (\tilde{r}_i - \tau)}{|\mathcal{D}_i|}\Theta_i^0} g\left(\theta_i^{t\tau+\tilde{r}_i}\right) \right\|_2 \\
&\leq \sum_{i=1}^{M} \frac{\eta_0 \tau}{|\mathcal{D}|} \left\|\left(\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right)\right\|_{\text{op}} \left\| e^{\frac{\eta_0 (\tilde{r}_i - \tau)}{|\mathcal{D}_i|}\Theta_i^0} \right\|_{\text{op}} \left\|g\left(\theta_i^{t\tau+\tilde{r}_i}\right)\right\|_2 \\
&\overset{\text{(b)}}{\leq} \sum_{i=1}^{M} \frac{\eta_0 \tau}{|\mathcal{D}|} \left\|\left(\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right)\right\|_{\text{op}} \left\|g\left(\theta_i^{t\tau+\tilde{r}_i}\right)\right\|_2 \\
&\overset{\text{(c)}}{\leq} \sum_{i=1}^{M} \frac{\eta_0 \tau}{|\mathcal{D}|} \left\|\left(\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right)\right\|_{\text{op}} \left\| e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^{t\tau+r_i'}} \right\|_{\text{op}} \left\|g\left(\theta_i^{t\tau}\right)\right\|_2
\end{aligned}$$

$$\leq \sum_{i=1}^{M} \frac{\eta_0 \tau}{|\mathcal{D}|} \left\| \left(\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right) \right\|_{\text{op}} \left\| g\left(\theta_i^{t\tau}\right) \right\|_2, \tag{108}$$

where step (a) holds according to the mean value theorem of integrals and $\tilde{r}_i \in (0, \tau)$, step (b) holds because Θ_i^0 is not full rank according to the definition of local NTK (20) and hence $\left\| e^{-\frac{\eta_0(\tilde{r}_i - \tau)}{|\mathcal{D}_i|}\Theta_i^0} \right\|_{\text{op}} \leq e^{-\frac{\eta_0(\tilde{r}_i - \tau)}{|\mathcal{D}_i|}\lambda_{\min}\left(\Theta_i^0\right)} = 1$, step (c) comes from (61) and $r_i' \in (0, \tilde{r}_i)$, the last step holds because $\Theta_i^{t\tau+r_i'}$ is not full rank. Then, we proceed to bound $\left\|\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right\|_{\text{op}}$, which can be derived as

$$\begin{aligned}
&\left\|\Theta_i^{t\tau+\tilde{r}_i} - \Theta_i^0\right\|_{\text{op}} \\
&= \frac{1}{n} \left\| J\left(\theta_i^{t\tau+\tilde{r}_i}\right)^T J_i\left(\theta_i^{t\tau+\tilde{r}_i}\right) - J\left(\theta_i^0\right)^T J_i\left(\theta_i^0\right) \right\|_{\text{op}} \\
&= \frac{1}{n} \left\| [J\left(\theta_i^{t\tau+\tilde{r}_i}\right)^T - J\left(\theta_i^0\right)^T] J_i\left(\theta_i^{t\tau+\tilde{r}_i}\right) + J\left(\theta_i^0\right)^T \left[J\left(\theta_i^{t\tau+\tilde{r}_i}\right) - J_i\left(\theta_i^0\right)\right] \right\|_{\text{op}} \\
&\leq \frac{1}{n} \left\| J\left(\theta_i^{t\tau+\tilde{r}_i}\right)^T - J\left(\theta_i^0\right)^T \right\|_F \left\| J_i\left(\theta_i^{t\tau+\tilde{r}_i}\right) \right\|_F + \left\| J\left(\theta_i^0\right)^T \right\|_F \left\| J\left(\theta_i^{t\tau+\tilde{r}_i}\right) - J_i\left(\theta_i^0\right) \right\|_F \\
&\leq 2C^2 \left\| \theta_i^{t\tau+\tilde{r}_i} - \theta^0 \right\|_2 \\
&\overset{\text{(a)}}{\leq} \frac{2\eta_0 \tilde{r}_i C^3 R_0 q^t}{\sqrt{n}|\mathcal{D}_i|} + \frac{2\eta_0 \tau C^3 R_0 \left(1 - q^t\right)}{\sqrt{n}\left(1 - q\right)} \\
&\leq \frac{2\eta_0 \tau C^3 q^t R_0}{\sqrt{n}} + \frac{2\eta_0 \tau C^3 R_0 \left(1 - q^t\right)}{\sqrt{n}\left(1 - q\right)},
\end{aligned} \tag{109}$$

where step (a) comes from (67) and the last step holds because $\tilde{r}_i \leq \tau$ and $|\mathcal{D}_i| \geq 1$. Plugging (109) into (108) yields

$$\begin{aligned}
&\left\| \sum_{i=1}^{M} p_i \frac{\eta_0}{|\mathcal{D}_i|} e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0} \int_0^\tau e^{\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0} \left(\Theta_i^{t\tau+r} - \Theta_i^0\right) g\left(\theta_i^{t\tau+r}\right) dr \right\|_{\text{op}} \\
&\leq \frac{M\eta_0\tau}{|\mathcal{D}|} \left[\frac{2\eta_0 \tau C^3 q^t R_0}{\sqrt{n}} + \frac{2\eta_0 \tau C^3 R_0 \left(1 - q^t\right)}{\sqrt{n}\left(1 - q\right)} \right] \left\| g\left(\theta_i^{t\tau}\right) \right\|_2 \\
&\overset{\text{(a)}}{\leq} \frac{M\eta_0\tau}{|\mathcal{D}|} \left[\frac{2\eta_0 \tau C^3 q^t R_0}{\sqrt{n}} + \frac{2\eta_0 \tau C^3 R_0 \left(1 - q^t\right)}{\sqrt{n}\left(1 - q\right)} \right] \left[q^t \left(R_0 - \frac{2\eta_0 \tau C C_1 R_0 \zeta}{\left(1 - q\right)^2} \right) + \frac{2\eta_0 \tau C C_1 R_0 \zeta}{\left(1 - q\right)^2} \right] \\
&\overset{\text{(b)}}{=} \frac{M\eta_0\tau}{|\mathcal{D}|} \left[\frac{2\eta_0 \tau C^3 q^t R_0}{\sqrt{n}} + \frac{2\eta_0 \tau C^3 R_0 \left(1 - q^t\right)}{\sqrt{n}\left(1 - q\right)} \right] q^t R_0 \\
&= \frac{2M\eta_0^2\tau^2 C^3 R_0^2 q^t \left(1 - q^{t+1}\right)}{\sqrt{n}|\mathcal{D}|\left(1 - q\right)} \\
&\leq \frac{2\eta_0^2\tau^2 C^3 R_0^2 q^t \left(1 - q^{t+1}\right)}{\sqrt{n}\left(1 - q\right)},
\end{aligned} \tag{110}$$

where step (a) comes from (23) in Theorem 1 and step (b) omits $\frac{\zeta}{\sqrt{n}} = \mathcal{O}(n^{-1})$.

C.3. Bounding $\left\| f^{\text{lin}}\left(\theta^{t\tau+r}\right) - f\left(\theta^{t\tau+r}\right) \right\|_2$ and $\left\| f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right) - f_i\left(\theta_i^{t\tau+r}\right) \right\|_2$

Plugging (110) and (107) into (102) yields

$$\begin{aligned}
\left\| g^{\text{lin}}\left(\theta^{(t+1)\tau}\right) - g\left(\theta^{(t+1)\tau}\right) \right\|_2 &\leq q \left\| g^{\text{lin}}\left(\theta^{t\tau}\right) - g\left(\theta^{t\tau}\right) \right\|_2 + \frac{2\eta_0^2\tau^2 C^3 R_0^2 q^t \left(1 - q^{t+1}\right)}{\sqrt{n}\left(1 - q\right)} + \left\| \sum_{i=1}^{M} p_i \Omega_i \right\|_2 \\
&\leq q \left\| g^{\text{lin}}\left(\theta^{t\tau}\right) - g\left(\theta^{t\tau}\right) \right\|_2 + \frac{2\eta_0^2\tau^2 C^3 R_0^2 q^t \left(1 - q^{t+1}\right)}{\sqrt{n}\left(1 - q\right)} + \frac{2\eta_0 \tau C C_1 R_0 \zeta}{\left(1 - q\right)},
\end{aligned} \tag{111}$$

where the last step comes from (89). By recursively employing (111) and considering the fact $\left\|g^{\text{lin}}\left(\theta^0\right)-g\left(\theta^0\right)\right\|_2=0$, we obtain

$$\begin{aligned}\left\|g^{\text{lin}}\left(\theta^{(t+1)\tau}\right)-g\left(\theta^{(t+1)\tau}\right)\right\|_2 &\leq \frac{1-q^{t+1}}{1-q}\left[\frac{2\eta_0^2\tau^2C^3R_0^2q^t\left(1-q^{t+1}\right)}{\sqrt{n}\left(1-q\right)}+\frac{2\eta_0\tau CC_1R_0\zeta}{(1-q)}\right]\\ &=\frac{2\eta_0^2\tau^2C^3R_0^2q^t\left(1-q^{t+1}\right)^2}{\sqrt{n}\left(1-q\right)^2}+\frac{2\eta_0\tau CC_1R_0\zeta\left(1-q^{t+1}\right)}{\left(1-q\right)^2}.\end{aligned} \tag{112}$$

Considering $\zeta=\frac{2\eta_0\tau CR_0}{\sqrt{n}(1-q)}$ and replacing $t+1$ with t, we have

$$\sup_{t\geq 0}\left\|g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right\|_2=\sup_{t\geq 0}\left\|f^{\text{lin}}\left(\theta^{t\tau}\right)-f\left(\theta^{t\tau}\right)\right\|_2=\mathcal{O}\big(n^{-\frac{1}{2}}\big). \tag{113}$$

Finally, we bound $\left\|f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-f_i\left(\theta_i^{t\tau+r}\right)\right\|_2$ as

$$\begin{aligned}\left\|f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-f_i\left(\theta_i^{t\tau+r}\right)\right\|_2 &=\left\|g_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-g_i\left(\theta_i^{t\tau+r}\right)\right\|_2\\ &=\left\|P_ig^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-P_ig\left(\theta_i^{t\tau+r}\right)\right\|_2\\ &\leq\left\|P_i\right\|_{\text{op}}\left\|g^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-g\left(\theta_i^{t\tau+r}\right)\right\|_2\\ &=\left\|g^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-g\left(\theta_i^{t\tau+r}\right)\right\|_2\\ &\overset{(\mathsf{a})}{\leq} q\left\|g^{\text{lin}}\left(\theta^{t\tau}\right)-g\left(\theta^{t\tau}\right)\right\|_2+\frac{2\eta_0^2\tau^2C^3R_0^2q^t\left(1-q^{t+1}\right)}{\sqrt{n}\left(1-q\right)}\\ &\leq q\left[\frac{2\eta_0^2\tau^2C^3R_0^2q^{t-1}\left(1-q^t\right)^2}{\sqrt{n}\left(1-q\right)^2}+\frac{2\eta_0\tau CC_1R_0\zeta\left(1-q^t\right)}{\left(1-q\right)^2}\right]+\frac{2\eta_0^2\tau^2C^3R_0^2q^t\left(1-q^{t+1}\right)}{\sqrt{n}\left(1-q\right)}\\ &=\frac{2\eta_0^2\tau^2C^3R_0^2q^t\left(1-q^t\right)^2}{\sqrt{n}\left(1-q\right)^2}+\frac{2\eta_0\tau CC_1R_0\zeta q\left(1-q^t\right)}{\left(1-q\right)^2}+\frac{2\eta_0^2\tau^2C^3R_0^2q^t\left(1-q^{t+1}\right)}{\sqrt{n}\left(1-q\right)},\end{aligned} \tag{114}$$

where step (a) holds because of (99), (107) and (110). The third inequation holds because of (112). Further considering $\zeta=\frac{2\eta_0\tau CR_0}{\sqrt{n}(1-q)}$, we have

$$\sup_{t\geq 0,\ 1\leq r\leq\tau}\left\|f_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)-f_i\left(\theta_i^{t\tau+r}\right)\right\|_2=\mathcal{O}\big(n^{-\frac{1}{2}}\big),\ \forall i \tag{115}$$

Thus, Theorem 2 is proved.

D. Proof of Theorem 3

The Jacobians of the linear global and local models are

$$J^{\text{lin}}\left(\theta^{t\tau}\right)=J\left(\theta^0\right),\quad J_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)=J_i\left(\theta_i^0\right). \tag{116}$$

The local update process of the linear model can be expressed as

$$\theta_i^{t\tau+r+1}=\theta_i^{t\tau+r}-\eta J_i\left(\theta_i^0\right)g_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right). \tag{117}$$

Via continuous time gradient flow, we obtain

$$\frac{d\theta_i^{t\tau+r}}{dr}=-\frac{\eta}{|\mathcal{D}_i|}J_i\left(\theta_i^0\right)g_i^{\text{lin}}\left(\theta_i^{t\tau+r}\right)=-\frac{\eta}{|\mathcal{D}_i|}J_i\left(\theta_i^0\right)P_ig^{\text{lin}}\left(\theta_i^{t\tau+r}\right). \tag{118}$$

Employing the chain rule yields

$$\frac{dg^{\text{lin}}\left(\theta_i^{t\tau+r}\right)}{dr}=J\left(\theta_i^0\right)^T\frac{d\theta_i^{t\tau+r}}{dr}$$

$$= -\frac{\eta}{|\mathcal{D}_i|} J\left(\theta_i^0\right)^T J_i\left(\theta_i^0\right) P_i g^{\text{lin}}\left(\theta_i^{t\tau+r}\right)$$
$$= -\frac{\eta_0 \Theta_i^0}{|\mathcal{D}_i|} g^{\text{lin}}\left(\theta_i^{t\tau+r}\right). \tag{119}$$

Integrating from 0 to r on both sides yields

$$g^{\text{lin}}\left(\theta_i^{t\tau+r}\right) = e^{\frac{-\eta_0 r}{\mathcal{D}_i}\Theta_i^0} g^{\text{lin}}\left(\theta^{t\tau}\right). \tag{120}$$

Plugging (120) into (118) yields

$$\frac{d\theta_i^{t\tau+r}}{dr} = -\frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^0\right) P_i e^{\frac{-\eta_0 r}{\mathcal{D}_i}\Theta_i^0} g^{\text{lin}}\left(\theta^{t\tau}\right). \tag{121}$$

Integrating from 0 to τ on both sides yields

$$\theta_i^{t\tau+\tau} - \theta^{t\tau} = -\frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^0\right) P_i \left(\int_0^\tau e^{\frac{-\eta_0 r}{\mathcal{D}_i}\Theta_i^0} dr\right) g^{\text{lin}}\left(\theta^{t\tau}\right), \tag{122}$$

where $\int_0^\tau e^{\frac{-\eta_0 r}{\mathcal{D}_i}\Theta_i^0} dr$ can be further derived as

$$\begin{aligned}
\int_0^\tau e^{\frac{-\eta_0 \Theta_i^0 r}{\mathcal{D}_i}} dr &= \int_0^\tau \sum_{k=0}^\infty \frac{1}{k!}\left(-\frac{\eta_0 r}{|\mathcal{D}_i|}\Theta_i^0\right)^k dr \\
&= \sum_{k=0}^\infty \frac{1}{k!}\left(-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0\right)^k \frac{\tau}{k+1} \\
&= \frac{-|\mathcal{D}_i|}{\eta_0}\left(\Theta_i^0\right)^{-1} \sum_{k=0}^\infty \frac{1}{(k+1)!}\left(-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0\right)^{k+1} \\
&= \frac{-|\mathcal{D}_i|}{\eta_0}\left(\Theta_i^0\right)^{-1} \sum_{k=1}^\infty \frac{1}{k!}\left(-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0\right)^{k} \\
&= \frac{|\mathcal{D}_i|}{\eta_0}\left(\Theta_i^0\right)^{-1}\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right).
\end{aligned} \tag{123}$$

Plugging (123) into (122) yields

$$\begin{aligned}
\theta_i^{t\tau+\tau} - \theta^{t\tau} &= -\frac{\eta}{|\mathcal{D}_i|} J_i\left(\theta_i^0\right) P_i \frac{|\mathcal{D}_i|}{\eta_0}\left(\Theta_i^0\right)^{-1}\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) g^{\text{lin}}\left(\theta^{t\tau}\right) \\
&= -\frac{1}{n}\left(J\left(\theta^0\right)^T\right)^{-1} J\left(\theta^0\right)^T J_i\left(\theta_i^0\right) P_i \left(\Theta_i^0\right)^{-1}\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) g^{\text{lin}}\left(\theta^{t\tau}\right) \\
&= -\left(J\left(\theta^0\right)^T\right)^{-1}\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) g^{\text{lin}}\left(\theta^{t\tau}\right).
\end{aligned} \tag{124}$$

Considering the model aggregation process, we have

$$\begin{aligned}
\theta^{(t+1)\tau} - \theta^{t\tau} &= \sum_{i=1}^M p_i\left(\theta_i^{t\tau+\tau} - \theta^{t\tau}\right) \\
&= -\left(J\left(\theta^0\right)^T\right)^{-1} \sum_{i=1}^M p_i\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) g^{\text{lin}}\left(\theta^{t\tau}\right) \\
&\overset{\text{(a)}}{=} -\left(J\left(\theta^0\right)^T\right)^{-1}\left(I - e^{-\sum_{i=1}^M p_i \frac{\eta_0 \tau}{|\mathcal{D}_i|}\Theta_i^0}\right) g^{\text{lin}}\left(\theta^{t\tau}\right) \\
&= -\left(J\left(\theta^0\right)^T\right)^{-1}\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}|}\Theta^0}\right) g^{\text{lin}}\left(\theta^{t\tau}\right),
\end{aligned} \tag{125}$$

where step (a) employs Lemma 4. Considering the aggregation process of the linear model, we have

$$\begin{aligned} g^{\text{lin}}\left(\theta^{(t+1)\tau}\right) &= \sum_{i=1}^{M} p_i g^{\text{lin}}\left(\theta_i^{t\tau+\tau}\right) \\ &= \sum_{i=1}^{M} p_i e^{\frac{-\eta_0 \tau}{\mathcal{D}_i}\Theta_i^0} g^{\text{lin}}\left(\theta_i^{t\tau}\right) \\ &\overset{(\mathsf{a})}{=} e^{-\eta_0 \tau \sum_{i=1}^{M} \frac{p_i}{\mathcal{D}_i}\Theta_i^0} g^{\text{lin}}\left(\theta_i^{t\tau}\right) \\ &= e^{-\frac{\eta_0 \tau}{\mathcal{D}}\Theta^0} g^{\text{lin}}\left(\theta^{t\tau}\right), \end{aligned} \tag{126}$$

where step (a) employs Lemma 4. By cursively employing (126), we can obtain

$$g^{\text{lin}}\left(\theta^{(t+1)\tau}\right) = e^{-\frac{\eta_0 (t+1)\tau}{\mathcal{D}}\Theta^0} g^{\text{lin}}\left(\theta^0\right). \tag{127}$$

Replace $t+1$ with t yields $g^{\text{lin}}\left(\theta^{t\tau}\right) = e^{-\frac{\eta_0 \Theta^0 t\tau}{\mathcal{D}}} g^{\text{lin}}\left(\theta^0\right)$, plugging which into (125) further yields

$$\begin{aligned} \theta^{(t+1)\tau} - \theta^{t\tau} &= -\left(J\left(\theta^0\right)^T\right)^{-1}\left(I - e^{-\frac{\eta_0 \tau}{|\mathcal{D}|}\Theta^0}\right) e^{\frac{-\eta_0 t\tau}{\mathcal{D}}\Theta^0} g^{\text{lin}}\left(\theta^0\right) \\ &= -\left(J\left(\theta^0\right)^T\right)^{-1}\left(e^{\frac{-\eta_0 t\tau}{\mathcal{D}}\Theta^0} - e^{-\frac{\eta_0 (t+1)\tau}{|\mathcal{D}|}\Theta^0}\right) g^{\text{lin}}\left(\theta^0\right). \end{aligned} \tag{128}$$

Employing (128) iteratively yields

$$\theta^{(t+1)\tau} - \theta^0 = -\left(J\left(\theta^0\right)^T\right)^{-1}\left(I - e^{-\frac{\eta_0 (t+1)\tau\Theta^0}{|\mathcal{D}|}}\right) g^{\text{lin}}\left(\theta^0\right). \tag{129}$$

Replacing $t+1$ with t yields

$$\theta^{t\tau} - \theta^0 = -\frac{1}{n} J\left(\theta^0\right)\left(\Theta^0\right)^{-1}\left(I - e^{-\frac{\eta_0 \Theta^0 t\tau}{|\mathcal{D}|}}\right) g^{\text{lin}}\left(\theta^0\right). \tag{130}$$

Therefore, for an arbitrary input x, we can obtain the closed form of $g^{\text{lin}}\left(x, \theta^{t\tau}\right)$ and $f^{\text{lin}}\left(x, \theta^{t\tau}\right)$ as

$$\begin{aligned} g^{\text{lin}}\left(x, \theta^{t\tau}\right) &= g\left(x, \theta^0\right) + J\left(x, \theta^0\right)^T\left(\theta^{t\tau} - \theta^0\right) \\ &= g\left(x, \theta^0\right) - \Theta^0(x)\left(\Theta^0\right)^{-1}\left(I - e^{-\frac{\eta_0 \Theta^0 t\tau}{|\mathcal{D}|}}\right) g^{\text{lin}}\left(\theta^0\right) \end{aligned} \tag{131}$$

and

$$f^{\text{lin}}\left(x, \theta^{t\tau}\right) = f\left(x, \theta^0\right) - \Theta^0(x)\left(\Theta^0\right)^{-1}\left(I - e^{-\frac{\eta_0 \Theta^0 t\tau}{|\mathcal{D}|}}\right)\left(f\left(\mathcal{X}, \theta^0\right) - \mathcal{Y}\right), \tag{132}$$

respectively.

E. Network Details

Table 1. Fully-Connected Network

KAYER	INPUT	IUTPUT
FC1	INPUT SIZE	$8 \times k$
FC2	$8 \times k$	$8 \times k$
FC3	$8 \times k$	OUTPUT SIZE

Table 2. Convolution Network

LAYER	CHANNEL	KERNEL	OUTPUT	STRIDE
CONV1	$6 \times k$	5×5	28×28	1
POOLING	$6 \times k$	2×2	14×14	2
CONV2	$16 \times k$	5×5	10×10	1
POOLING	$16 \times k$	2×2	5×5	2
FC1	–	–	$120 \times k$	–
FC2	–	–	$84 \times k$	–
FC3	–	–	10	–

Table 3. Residual Network

LAYER	OUTPUT	BLOCK TYPE
CONV1	32×32	$[3 \times 3, \text{CHANNEL SIZE} \times k]$
CONV2	32×32	$\begin{bmatrix} 3 \times 3, \text{CHANNEL SIZE} \times k \\ 3 \times 3, \text{CHANNEL SIZE} \times k \end{bmatrix} \times \psi$
CONV3	16×16	$\begin{bmatrix} 3 \times 3, \text{CHANNEL SIZE} \times k \\ 3 \times 3, \text{CHANNEL SIZE} \times k \end{bmatrix} \times \psi$
CONV4	8×8	$\begin{bmatrix} 3 \times 3, \text{CHANNEL SIZE} \times k \\ 3 \times 3, \text{CHANNEL SIZE} \times k \end{bmatrix} \times \psi$
AVG-POOL	1×1	$[8 \times 8]$