LINEAR CONVERGENCE OF DECENTRALIZED FEDAVG FOR NON-CONVEX OBJECTIVES: THE INTERPOLATION REGIME

Anonymous authors
Paper under double-blind review

ABSTRACT

In the age of Bigdata, Federated Learning (FL) provides machine learning (ML) practitioners with an indispensable tool for solving large-scale learning problems. FL is a distributed optimization paradigm where multiple nodes each having access to a local dataset collaborate (with or without a server) to solve a joint problem. Federated Averaging (FedAvg) although the algorithm of choice for many FL applications is not very well understood especially in the interpolation regime, a common phenomenon observed in modern overparameterized neural networks. In this work, we address this challenge and perform a thorough theoretical performance analysis of FedAvg in the interpolation regime. Specifically, we analyze the performance of FedAvg in two settings: (i) *Server*: When the network has access to a server that coordinates the information sharing among nodes, and (ii) *Decentralized:* The server-less setting, where the local nodes communicate over an undirected graph. We consider a class of non-convex functions satisfying the Polyak-Lojasiewicz (PL) condition, a condition that is satisfied by overparameterized neural networks. For the first time, we establish that FedAvg under both *Server* and *Decentralized* settings achieves linear convergence rates of $\mathcal{O}(T^{3/2}\log(1/\epsilon))$ and $\mathcal{O}(T^2\log(1/\epsilon))$, respectively, where ϵ is the desired solution accuracy, and T is the number of local updates at each node. In contrast to the standard FedAvg analysis, our work does not require bounded heterogeneity, variance, and gradient assumptions. Instead, we show that sample-wise (and local) smoothness of the local loss functions suffices to capture the effect of heterogeneity in FL training. Finally, we conduct experiments on multiple real datasets to corroborate our theoretical findings.

1 INTRODUCTION

Federated learning (FL) is a distributed machine learning scenario which allows the edge devices to learn a shared model while maintaining the training data decentralized at the edge devices Konečnỳ et al. (2016); McMahan et al. (2017). This avoids the need to share the data with a central server and hence preserves privacy of the individual clients (edge devices). Assuming a supervised learning setting, where each of the N clients having access to some local data $(\boldsymbol{x}, y) \sim \mathcal{D}_k$ from distribution $\mathcal{D}_k$ with $k \in [N]$ aim to solve the `FL Problem:` $\min_{\boldsymbol{w}\in\mathbb{R}^d} \Phi(\boldsymbol{w}) := \frac{1}{N}\sum_{k=1}^{N} \Phi_k(\boldsymbol{w})$, where $\Phi_k(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_k} l_k(f_{\boldsymbol{w}}(\boldsymbol{x}), y)$ is the average loss at the client $k \in [N]$ for the input feature vector $\boldsymbol{x} \in \mathcal{X}$, and the corresponding output label $y \in \mathcal{Y}$. Here, $f_{\boldsymbol{w}}(\boldsymbol{x})$ is the output of the model parameterized by $\boldsymbol{w} \in \mathbb{R}^d$.

The de-facto standard for solving the above `FL Problem` is the simple Federated Averaging (FedAvg) algorithm McMahan et al. (2017). In recent years, many works have attempted to characterize the convergence of FedAvg under different settings (Stich, 2018; Li et al., 2019; Woodworth et al., 2020a; Ma et al., 2018; Yu et al., 2019b). For example, the authors in Stich (2018) show a convergence rate of $\mathcal{O}\left(1/N\epsilon\right)$ for minimizing strongly convex functions while Haddadpour & Mahdavi (2019) establishes similar rates for minimizing functions satisfying Polyak-Lojasiewicz (PL) condition. Here, ϵ refers to the desired solution accuracy. For minimizing non-convex smooth functions, FedAvg achieves a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ (Karimireddy et al., 2020b; Yang et al., 2021). However, in practice, it has been observed that FedAvg converges at a much faster rate compared

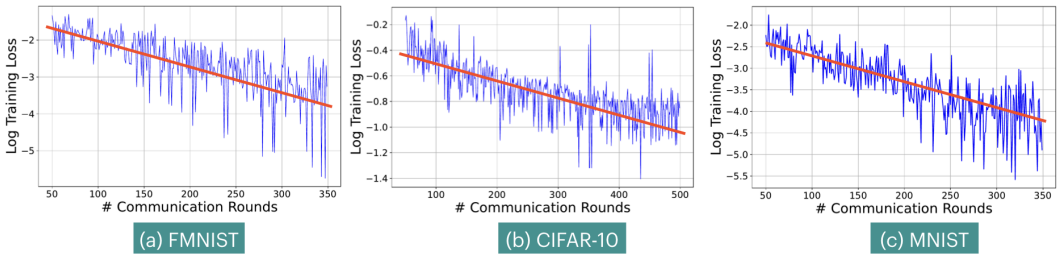

Figure 1: log-training loss versus communication rounds for overparameterized Deep Neural Networks (DNNs) and a simple regression model.

to the rates demonstrated in these works. To illustrate this fact, in Figure 1 we plot the behavior of the training loss (on a log scale) as a function of communication rounds for FedAvg to solve classification tasks on MNIST, FMNIST and CIFAR-10 data sets (for experimental details please see Section 4). It is clear from the plots that the losses decrease linearly as a function of communication rounds. This implies that the standard analyses of FedAvg lacks theoretical explanation of this linear convergence as shown in Fig. 1. In this work, we attempt to fill this gap and perform a thorough theoretical analysis of FedAvg in the interpolation regime under two settings; the *Server* setting, when the network has access to a server that coordinates the information sharing among clients, and the *Decentralized* setting also referred to as the server-less setting, where the local nodes communicate over an undirected graph. Under both these settings we establish the linear convergence of FedAvg for minimizing a class of non-convex functions satisfying the PL inequality. We note that PL inequality plays a key role in the training of overparameterized systems. Specifically, many works have shown that the loss function of an overparamterized neural network satisfies the PL inequality (Bassily et al., 2018; Liu et al., 2020). Furthermore, our analysis reveals that the standard but restrictive assumptions of bounded gradients (Yu et al., 2019b; Stich, 2018; Li et al., 2019; Koloskova et al., 2020), heterogeneity (Yu et al., 2019a; Woodworth et al., 2020b; Yu et al., 2019a), and variance (Woodworth et al., 2020b; Qu et al., 2020) can be avoided while guaranteeing this linear convergence of FedAvg. Next, we list the major contributions of our work.

• We analyze the convergence of FedAvg in the *Server* setting which consists of a single server communicating with several clients (McMahan et al., 2017). In this setting, we show that to achieve an ϵ-accurate solution FedAvg in the interpolation regime requires $R \sim \mathcal{O}\left(T^{3/2}\log\left(1/\epsilon\right)\right)$ rounds of communication, where T is the number of local SGD updates.

• We consider the *Decentralized* setting where N distributed clients communicate over an undirected graph $\mathcal{G}$. Similar to the *Server* setting, we show that to achieve an ϵ-accurate solution *Decentralized* FedAvg requires $R \sim \mathcal{O}\left(T^{2}\log\left(1/\epsilon\right)\right)$ rounds of communication. We also characterize the effect of the network topology on the performance of *Decentralized* FedAvg. Moreover, our proof technique utilizes a novel induction based analysis for establishing this convergence.

• Our theoretical results under both the *Server* and the *Decentralized* settings do not require assumptions such as bounded heterogeneity, gradient and variance. We show that sample-wise smoothness of the stochastic loss functions suffices to capture the effect of data heterogeneity among different clients (Bassily et al., 2018) while avoiding the need to impose these restrictive assumptions.

• Finally, to corroborate our theoretical findings we present experimental results using various datasets such as MNIST, FMNIST, CIFAR-10 and Shakespeare, under different settings.

Related Work: FedAvg first proposed in (McMahan et al., 2017), has been extensively studied in the *server* setting and with homogeneous data (see Stich (2018); Wang & Joshi (2018); Khaled et al. (2019); Yu et al. (2019b); Wang et al. (2019); Yang et al. (2021)). Recently, many works have adapted the analyses of FedAvg for minimizing the non-convex losses in the heterogeneous data settings. We note that most of the works (see Yu et al. (2019a); Yang et al. (2021); Karimireddy et al. (2020b)) establish the convergence rate of $\mathcal{O}(1/N\epsilon^2)$ to an ϵ-stationary point under the bounded heterogeneity setting. It is also worth noting that numerous works have proposed variants of FedAvg with different local update rules (e.g., variance reduction, momentum SGD, adaptive updates, etc.) with the goal of improving the performance of FedAvg (Karimireddy et al., 2020b; Sharma et al., 2019; Liang et al., 2019; Khanduri et al., 2021; Karimireddy et al., 2020a). However, in practice FedAvg remains the algorithm of choice for training large FL systems.

Table 1: Comparisons with the existing work. Here, (s), (d), SC, C and NC represent *Server*, *Decentralized*, Strongly convex, Convex and Non-convex settings, respectively.

ALGORITHM	CONVERGENCE	EXTRA ASSUMPTIONS	SETTING
Local SGD (Stich, 2018) (s)	$\mathcal{O}\left(1/N\epsilon\right)$	Bounded gradient	SC
Local SGD (Yu et al., 2019b) (s)	$\mathcal{O}\left(1/N\epsilon^2\right)$	Bounded variance, smoothness	NC
Local SGD (Haddadpour et al., 2019) (s)	$\mathcal{O}\left(1/N\epsilon\right)$	Bounded variance, smoothness	NC
FedAvg (Qu et al., 2020) (s)	$\mathcal{O}\left(T\log(1/\epsilon)\right)$	Bounded Gradient, Bounded Variance	Overparameterized SC
Local SGD (Woodworth et al., 2020b) (s)	$\mathcal{O}\left(1/N\epsilon^2\right)$	Bounded Variance	C
Local SGD (Woodworth et al., 2020a) (s)	$\mathcal{O}\left(1/N\epsilon^2\right)$	Bounded Variance	C
PR-SGD (Yu et al., 2019a) (s)	$\mathcal{O}\left(1/N\epsilon^2\right)$	Bounded Variance	NC
FedAvg (Karimireddy et al., 2020b) (s)	$\mathcal{O}\left(1/N\epsilon^2\right)$	Bounded gradient dissimilarity Bounded heterogeneity	NC
OUR WORK (s)	$\mathcal{O}(T^{3/2}\log(1/\epsilon))$	Smoothness	Overparameterized NC
NFSGD (Haddadpour & Mahdavi, 2019) (d)	$\mathcal{O}\left(\frac{1}{N\epsilon^2}\right)$	Bounded local variance	NC
DECENTRALIZED SGD (Koloskova et al., 2020) (d)	$\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$	Bounded Variance, Bounded heterogeneity	Overparameterized SC
OUR WORK (d)	$\mathcal{O}(T^2\log(1/\epsilon))$	Smoothness	Overparameterized NC

A few works that have also analyzed the performance of Fedvg in the *decentralized* setting as well. In (Haddadpour & Mahdavi, 2019; Yu et al., 2019a), the authors analyze the convergence of FedAvg under both server and decentralized setting with bounded gradient dissimilarity assumption and establish a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ for minimizing smooth non-convex objectives. We note that all the above works provide a sublinear rate of convergence for FedAvg, however, as illustrated in Fig. 1, FedAvg converges at a much faster rate in practice. To understand this behavior of FedAvg, in this work we analyze the performance of FedAvg under both server and decentralized settings for minimizing a special class of non-convex functions satisfying PL inequality under the interpolation regime. We note that overparameterized neural networks/systems usually operate in the interpolation regime while their loss functions have been shown to satisfy the PL inequality.

The linear convergence of centralized SGD in the interpolation regime for minimizing PL objectives was first established in Bassily et al. (2018). Recently, (Qu et al., 2020) showed linear convergence rate of FedAvg in the server setting for minimizing strongly-convex objectives in the overparameterized regime. Similarly, the authors in (Koloskova et al., 2020) have also established the linear convergence of FedAvg in the decentralized setting for minimizing strongly-convex losses in an overparameterized setting. The above works only focus on analysis of FedAvg for the strongly-convex objectives in the overparameterized regime while we focus on the more general class of non-convex functions satisfying the PL inequality. Moreover, compared to other works that assume restrictive bounded gradient, heterogeneity, and variance assumptions, we show that such assumptions can be avoided by using a sample-wise smoothness assumption. Table 1 presents a summary of the above discussion. Please refer to Appendix A.2 for a detailed literature review.

2 *Server* FEDAVG

In this section, we present the classical FedAvg algorithm in the server setting and prove its convergence. The FL setup consists of a set of N clients and a central server. The clients collaborate with the help of the server to solve the `FL Problem` stated in Section 1. As noted earlier, a standard algorithm to solve the `FL Problem` is FedAvg (McMahan et al., 2017). FedAvg executes in two major steps, namely the local SGD step and the aggregation and broadcast step by the server. The main steps of FedAvg are discussed below, and are outlined in Algorithm 1.

1. `Initial Broadcast`: The central server broadcasts the initial model parameters denoted $\underline{\boldsymbol{w}}^0$ to all the clients at the beginning, i.e., in the round $r = 0$. See steps 1 and 3 of Algorithm 1.
2. `Local updates`: Each client performs T local SGD updates starting from the initial model received from the server. To compute the stochastic gradient, each client $k \in [N]$ uniformly samples a batch of size b denoted by $\mathcal{B}_k^{r,t}$. The resulting model after T local updates in the r-th communication round, $\boldsymbol{w}_k^{r,T}$, is shared with the server. See steps 6 and 7 of Algorithm 1.
3. `Aggregation and Broadcast`: In the r-th global communication round, the central server computes an average, $\underline{\boldsymbol{w}}^r$, of the received models from the individual clients. The server then broadcasts the aggregated model $\underline{\boldsymbol{w}}^r$ to all the clients. The steps (2) and (3) are repeated for R communication rounds. See steps 11, 12 and 3 of Algorithm 1.

Algorithm 1: FedAvg McMahan et al. (2017)

INITIALIZE $\{\boldsymbol{w}_k^{0,0} = \underline{\boldsymbol{w}}^0\}$, $\boldsymbol{w}_k \in \mathbb{R}^d$ for $k \in [N]$
for $r = 0, 1, \ldots, R-1$ **do**
  BROADCAST $\underline{\boldsymbol{w}}^r$ to the nodes $k \in [N]$
  for $t = 0, 1, \ldots, T-1$ **do**
    for *devices* $k \in [N]$ **do**
      SAMPLE A BATCH $\mathcal{B}_k^{r,t}$ and $|\mathcal{B}_k^{r,t}| = b$
      SGD step on $\boldsymbol{w}_k^{r,t}$ for $k \in [N]$: $\boldsymbol{w}_k^{r,t+1} = \boldsymbol{w}_k^{r,t} - \frac{\eta}{b} \sum_{j \in \mathcal{B}^{r,t}} \nabla \Phi_{k,j} \left(\boldsymbol{w}_k^{r,t} \right)$
    end
  end
  RECEIVE $\boldsymbol{w}_k^{r,T}$ from nodes $k \in [N]$
  AGGREGATION step : $\underline{\boldsymbol{w}}^{r+1} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{w}_k^{r,T}$
end

In this work, we for the first time establish linear convergence of FedAvg in the interpolation setting for non-convex loss functions (satisfying PL condition) (Bassily et al., 2018). It is important to note that the proof of convergence provided in Bassily et al. (2018) for centralized SGD cannot be directly extended to the FL setting of Algorithm 1. The major challenge in the analysis is presented by the phenomenon referred to as *client drift* where the local models drift apart with multiple local SGD updates because of data heterogeneity. Naturally, the analysis involves bounding the client drift in terms of the loss function, which is challenging, and non-trivial. Moreover, without making bounded gradient and bounded heterogeneity assumptions, it remains an open challenge on how to control this client drift. In this work, we show that this drift can in fact be controlled with sample-wise smoothness assumption and a careful analysis of the drift term. To analyze the convergence of *Server* FedAvg, we make some standard assumptions as discussed next.

2.1 ASSUMPTIONS

In this section, we present the assumptions and definitions used in the paper.

Definition 1. *(L-Smoothness): The function $\Phi(\boldsymbol{u})$ is said to be L smooth if there exist a constant $L > 0$ such that $\|\nabla \Phi(\boldsymbol{u}_1) - \nabla \Phi(\boldsymbol{u}_2)\|_2 \leq L \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2$ for any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$. Note that this implies $\Phi(\boldsymbol{u}_1) \leq \Phi(\boldsymbol{u}_2) + \langle \nabla \Phi(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2 \rangle + \frac{L}{2} \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2$ for any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$.*

Definition 2. *(ϵ-accurate solution): A point $\boldsymbol{w} \in \mathbb{R}^d$ is called an ϵ-accurate solution if $\Phi(\boldsymbol{w}) - \Phi(\boldsymbol{w}^*) \leq \epsilon$, where $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \Phi(\boldsymbol{w})$. A stochastic algorithm is said to achieve an ϵ-accurate solution in r rounds if $\mathbb{E}[\Phi(\boldsymbol{w}^r) - \Phi(\boldsymbol{w}^*)] \leq \epsilon$, where expectation is taken over the stochasticity of the algorithm.*

Assumption 1. *(Interpolation (Bassily et al., 2018)): We assume that there exists a $\boldsymbol{w}^* \in \mathbb{R}^d$ such that the per sample loss $\Phi_{k,j}(\boldsymbol{w}^*) = 0$ for all samples $j \in [b]$.*

Assumption 2. *(PL inequality): The loss function $\Phi(\boldsymbol{v})$ satisfies the PL inequality, i.e., $\|\nabla \Phi(\boldsymbol{v})\|^2 \geq \mu \Phi(\boldsymbol{v})$ for some $\mu > 0$ and for all $\boldsymbol{v} \in \mathbb{R}^d$. Further, the local loss functions $\Phi_k(\boldsymbol{v})$ for all $k \in [N]$ are also assumed to satisfy the PL inequality, henceforth referred to as local PL inequality, i.e., $\|\nabla \Phi_k(\boldsymbol{v})\|^2 \geq \mu_k \Phi_k(\boldsymbol{v})$ for some $\mu_k > 0$ and for all $\boldsymbol{v} \in \mathbb{R}^d$.*

Assumption 3. *(Sample-wise, Local and Global smoothness): The functions $\Phi_{k,j}(\cdot)$ for all $j \in [b], k \in [N]$ are assumed to be $l_{k,j}$-smooth. The local functions $\Phi_k(\cdot)$ for all $k \in [N]$ are assumed to be L_k-smooth. The above assumptions imply $\|\nabla \Phi_{k,j}(\boldsymbol{v})\|^2 \leq 2 l_{k,j} \Phi_{k,j}(\boldsymbol{v})$ and $\|\nabla \Phi_k(\boldsymbol{v})\|^2 \leq 2 L_k \Phi_k(\boldsymbol{v})$ for all $k \in [N]$ and $j \in [b]$. We also assume the global loss $\Phi(\cdot)$ to be L-smooth.*

Assumption 4. *(Unbiased Gradient and Loss function): We assume that the estimates of the gradient and the loss function at each client $k \in [N]$ is unbiased, i.e., $\mathbb{E}[\nabla \Phi_{k,j}(\boldsymbol{w})] = \nabla \Phi_k(\boldsymbol{w})$ and $\mathbb{E}[\Phi_{k,j}(\boldsymbol{w})] = \Phi_k(\boldsymbol{w})$ for any $j \in [b]$ and $\boldsymbol{w} \in \mathbb{R}^d$.*

Note that the above assumptions, including interpolation, PL inequality and smoothness are standard assumptions made by several authors in the past (Bassily et al., 2018; Karimi et al., 2016; Ma

et al., 2018; Nguyen & Mondelli, 2020). For example, the authors in Bassily et al. (2018); Liu et al. (2022); Karimi et al. (2016); Haddadpour et al. (2019) assume PL inequality along with sample-wise smoothness to prove linear convergence of SGD in a centralized setting for the interpolation regime. Importantly, it must be noted that the PL inequality plays an important role in the analysis of overparameterized neural networks. Specifically, many works have shown that the overparameteried neural networks satisfy the PL inequality (Liu et al., 2020; Nguyen & Mondelli, 2020; Nguyen et al., 2021; Allen-Zhu et al., 2019). Moreover, we note that the assumption on sample-wise smoothness is not very strict since any neural network with smooth activation will satisfy this assumption. Although the above does not cover neural networks with non-smooth activations (like ReLU), it is known that such activations can be approximated to arbitrary accuracy by a kernel based smooth activation function (Nguyen & Mondelli, 2020, Eq. 2). Next, we present the first main result of the paper. The detailed proof is presented in Appendix A.4.

2.2 Convergence of *Server* FedAvg

Theorem 1. *Under Assumptions 1–4, choosing* $\eta \leq \min\left\{\frac{4}{\mu}, \frac{2}{\mu_{min}}, \frac{\mu}{\zeta_1}, \frac{L^2}{\zeta_2}, \frac{\mu_{min}}{\zeta_3}, \left(\frac{\mu}{T^3\zeta_4}\right)^{\frac{1}{2}}\right\}$, *the iterates generated by Algorithm 1 satisfy*

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right] \leq \left(1 - \frac{\eta\mu}{8}\right)^r \Phi\left(\underline{\boldsymbol{w}}^0\right) . \tag{1}$$

where, $\zeta_1 := 4\left(\frac{2Ll_{max}}{bN} + \frac{2LL_{max}}{N} + 2LL_{max}\right)$, $\zeta_2 := 2\left(\frac{Ll_{max}^2}{bN} + \frac{LL_{max}^2}{N} + LL_{max}^2\right)$, $\zeta_3 := 2\left(\frac{l_{max}L_{max}}{b} + \frac{L_{max}^2 b(b-1)}{b^2}\right)$ *and* $\zeta_4 := 8L^2\left(\frac{2l_{max}}{b} + \frac{2b(b-1)L_{max}}{b^2}\right)$, $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$, $l_{max} := \max_{k,j} l_{k,j}$ *and* $L_{max} := \max_k L_k$.

Next, we characterize the sample complexity of FedAvg.

Corollary 1. *Under the setting of Theorem 1 to achieve an* ϵ*-accurate solution, Algorithm 1 requires* $R = \mathcal{O}\left(\left(T^{3/2}/\mu\right)\log\left(\Phi\left(\underline{\boldsymbol{w}}^0\right)/\epsilon\right)\right)$ *rounds of communication.*

Corollary 1 establishes the linear convergence of FedAvg in terms of the number of rounds required to achieve an ϵ-accurate solution. From Corollary 1, we also observe that as the number of local updates, T, increase the global communication rounds also increase. Note that this is expected since more local updates result in client drift, thereby, requiring more communication rounds to reach the ϵ-accurate solution. Note that we can control the effect of client drift by choosing the learning rate appropriately, i.e., $\eta \sim \mathcal{O}(1/T^{3/2})$, and as a consequence, we need to choose the global rounds to be of the order of $\mathcal{O}(T^{3/2})$. However, note that Corollary 1 captures the worst case behavior of FedAvg. Ideally, one would expect the convergence performance (in terms of communication rounds) to first improve with T and then worsen as T increases beyond a threshold as also observed in Yu et al. (2019b;b). A more fine grained study reveals that this is in fact the case for our analysis as well. Below, we discuss the effect of the local updates on the performance of FedAvg.

Effect of Local Updates: Our analysis shows that the convergence performance is not always a monotonic function of T. To see this, we refer the readers to equation 21 in the Appendix, which reveals that for a fixed $\eta \leq 4/\mu$, the first term decreases with T while the second term increases. This implies that if $T \leq T_{\text{th}}$ the first term dominates (for η small enough) and the convergence performance improves as T increases, however, as $T \geq T_{\text{th}}$ the second term dominates and the convergence performance degrades as T increases. This implies that an optimal choice of $T = T_{\text{th}}$ exists, and is a solution to $(1-\eta\mu/4)^T = \eta^3 L^2 (T-1)^3 \left(2l_{max}/b + 2b(b-1)L_{max}/b^2\right)$. Note that for a fixed η, the optimal T is independent of the communication rounds r. Later we present experimental results to corroborate this behavior. For the clarity of presentation and better interpretability of the result, Corollary 1 only focuses on the general result establishing linear convergence of FedAvg.

We also note that the variance in our theoretical results is implicitly controlled by the smoothness assumption of the loss function. In the theoretical analysis, the dependence on the batch size, b is captured by the bound on the learning rate η in Theorem 1. It is clear from above that a smaller batch size implies smaller η and, thereby, a slower convergence rate. In contrast, a larger batch-size allows us to choose a larger step-size, thereby, leading to a faster convergence rate.

The best known rate for FedAvg for minimizing the strongly-convex objectives in the overaprameterized regime is $\mathcal{O}(\exp(-NR/T))$ with bounded variance assumption Qu et al. (2020). These assumptions are seldom satisfied, and hence are of limited applicability. In contrast, our analysis without the need of any variance or heterogeneity assumptions achieves $\mathcal{O}(\exp(-\mu R/8T^{3/2})$ rate for minimizing a much harder class of non-convex problems satisfying PL inequality. Further, note that the theorem also reveals the effect of data heterogeneity on the performance of FedAvg, for example, the choice of step-size will be limited by the worst case smoothness constants as well as the PL inequality parameter μ_{min}'s among all clients. Finally, we present the proof sketch of Theorem 1 and discuss the major challenges in the proof.

Proof Sketch of Theorem 1 and Challenges: In FedAvg setting, the individual nodes execute multiple local updates. This leads to a phenomenon termed as "client drift", wherein, the local models drift apart from each other because of heterogeneity in the local datasets. The first step is to establish descent in $\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right]$ in terms of the client drift, see equation 18. Note that this expression is independent of the variance, and proving this requires a careful application of the smoothness and the PL inequality, as in equation 13 and equation 21. Next we bound the drift term in Lemma 1. In standard FedAvg analyses, this drift is controlled using the bounded heterogeneity (or gradient) assumptions. However, guaranteeing convergence of FedAvg without making bounded heterogeneity (or gradients) assumption is unclear. Therefore, the challenge lies in bounding this drift in terms of the local loss (see equation 7), and then relate it to the global loss (see Lemma 2). These steps require us to again carefully use the smoothness and PL inequality. Note that our analysis is the first to establish fast (linear) convergence of FedAvg for minimizing PL functions in the interpolation regime without imposing any assumptions on the local data distributions. In the next section, we extend our analysis and results to a decentralized learning scenario.

3 *Decentralized* FEDAVG

In this section, we present the *Decentralized* FedAvg algorithm, and prove its convergence. The *Decentralized* setting consists of N distributed edge devices which are represented using a connectivity graph $\mathcal{G} \in (\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V} \in [N]$ is the vertex set or clients, and $\mathcal{E} \subseteq \{\mathcal{V} \times \mathcal{V}\}$ represents the edges of the graph. Any edge $(i, j) \in \mathcal{E}$ represents a connection between node i and j. Further, the connections are represented using mixing matrix $P = [p_{i,k}] \in \mathbb{R}^{N \times N}$, where $p_{i,k} = 0$ if there is no edge between node i and k i.e., $(i, k) \notin \mathcal{E}$, else $p_{k,i} > 0$.[1] The algorithm for *Decentralized* FedAvg is presented in **Algorithm** 2 while in the following, we provide an outline:

1. `Initialization`: Each client $k \in [N]$ initializes the model parameters denoted by $\underline{\boldsymbol{w}}_k^0$ at the beginning, i.e., in the round $r = 0$. See Step 1 of Algorithm 2.
2. `Local updates`: Each client performs T local SGD updates starting from the model parameters obtained by the aggregation of the updates from neighbouring clients. To compute the stochastic gradient, each client $k \in [N]$ uniformly samples a batch size b denoted by $\mathcal{B}_k^{r,t}$. The resulting model parameters after T local rounds in the r-th global round are denoted by $\boldsymbol{w}_k^{r,T}$ which is sent to all the neighbouring clients of k. See Steps 6 and 7 of Algorithm 2. Note that this procedure is similar to the local update step in the FedAvg case.
3. `Aggregation`: In the r-th global communication round, each client $k \in [N]$ computes a local average of the model parameters received by its neighbors. The aggregate model is denoted by $\underline{\boldsymbol{w}}_k^r$. The steps (2) and (3) above are repeated for R rounds. See Steps 11, 12 of Algorithm 2.

Below, we present the convergence result of *Decentralized* FedAvg algorithm.

3.1 CONVERGENCE OF *Decentralized* FEDAVG

In this section, we prove that the *Decentralized* FedAvg algorithm converges linearly to the global optimum for any smooth non-convex function satisfying PL inequality in the interpolated regime. Compared to the FedAvg this case poses several challenges. In particular, for Decentralized FedAvg we need to handle two drift terms, namely *local drift* and the *global drift*. Local drift refers to the update at each client drifting away from the average obtained from the neighboring clients while

[1] In our analysis, we have assumed $p_{i,k} = \frac{1}{d_i}$ whenever $p_{i,k} > 0$. Here, d_i is the degree of the i-th node.

Algorithm 2: *Decentralized* Federated Averaging Algorithm (*Decentralized* FedAvg)

INITIALIZE $\{\boldsymbol{w}_k^{0,0} = \underline{\boldsymbol{w}}_k^0\}$, $\boldsymbol{w}_k \in \mathbb{R}^d$ for $k \in [N]$
for $r = 0, 1, \ldots, R-1$ **do**
  INITIALIZE $\underline{\boldsymbol{w}}_k^r$ to the device $k \in [N]$
  for $t = 0, 1, \ldots, T-1$ **do**
    for *devices* $k \in [N]$ **do**
      SAMPLE A BATCH $\mathcal{B}_k^{r,t}$ and $|\mathcal{B}_k^{r,t}| = b$
      SGD step on $\boldsymbol{w}_k^{r,t}$ for $k \in [N]$: $\boldsymbol{w}_k^{r,t+1} = \boldsymbol{w}_k^{r,t} - \frac{\eta}{b} \sum_{j \in \mathcal{B}_k^{r,t}} \nabla \Phi_{k,j} \left(\boldsymbol{w}_k^{r,t} \right)$
    end
  end
  RECEIVE $\boldsymbol{w}_k^{r,T}$ from clients $k \in [N]$
  AGGREGATION step : $\underline{\boldsymbol{w}}_k^{r+1} = \sum_{i \in \mathcal{N}_k} p_{k,i} \boldsymbol{w}_i^{r,T}$
end

the global drift refers to the average obtained from the neighboring clients drifting away from the global average. To derive the convergence of FedAvg, we employ a novel induction based proof technique and perform a carefully calibrated analysis which can be of independent interest to the research community. In addition to the Assumptions 1-4, our analysis also relies on the following assumption on the mixing matrix (Koloskova et al., 2020).

Assumption 5. *The mixing matrix P is assumed to be symmetric, i.e., $P = P^T$, and doubly stochastic, i.e., $P\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T P = \mathbf{1}^T$.*

The above assumption covers all networks that are symmetric, for example, fully connected, ring topology etc. In the following, we provide our main result for the decentralized FedAvg. The detailed proof is presented in Appendix A.5.

Theorem 2. *Under Assumptions 1-5, choosing*

$$\eta \leq \min \left\{ \frac{4}{\mu}, \frac{1}{L}, \frac{2}{\mu_{min}}, \frac{\mu}{4\zeta_1}, \frac{\mu}{2\zeta_2}, \frac{1}{8} \left(\frac{\mu}{\zeta_3 T^3} \right)^{\frac{1}{3}}, \left(\frac{1}{\zeta_4 T^2} \right)^{\frac{1}{2}}, \frac{1}{\zeta_5}, \frac{N\mu_{min}}{\zeta_6 T^2}, \frac{\mu_{min}}{\zeta_7}, \left(\frac{1}{T^2 \zeta_8} \right)^{\frac{1}{2}}, \left(\frac{1}{T^3 \zeta_9} \right)^{\frac{1}{4}} \right\}, \tag{2}$$

the average total loss of the Decentralized *FedAvg algorithm satisfies*

$$\mathbb{E}[\Phi \left(\underline{\boldsymbol{w}}^{r+1} \right)] \leq \Lambda^r \left(\Lambda + 4\eta^4 L L_m \beta T^3 \lambda^2 (r+1)^2 \right) \Phi \left(\underline{\boldsymbol{w}}^0 \right), \tag{3}$$

where $\lambda \triangleq \left(1 + \frac{1}{\psi}\right) \lambda_2^2$, $\psi > \frac{1}{\frac{1}{\lambda_2^2} - 1}$, $L_m := \max \left\{ L_{max}^2, 2L_{max} N \right\}$, $\beta := \frac{4 l_{max} \psi^2 N}{(1+\psi)\mu_{min}}$, $\mu_{min} := \min_{k \in [N]} \{\mu_k\}$, $L_{max} := \max_k L_k$ *and* $\Lambda := \max \left(\left(1 - \frac{\eta\mu}{8}\right), \lambda \right)$. *In the above,* $\zeta_1 := 4 \left(\frac{2Ll_{max}}{bN} + \frac{2LL_{max}}{N} + 2LL_{max} \right)$, $\zeta_2 := 2 \left(\frac{Ll_{max}^2}{bN} + \frac{LL_{max}^2}{N} + LL_{max}^2 \right)$, $\zeta_3 := \frac{64 l_{max} L L_{max}}{\mu_{min}} + 16\gamma L \lambda_2^2 L_{max}$, $\zeta_4 := \frac{16 l_{max} L_{max}^2}{\mu_{min}} + 4\lambda_2^2 \gamma L^2$, $\zeta_5 := 2 \left(\frac{8 l_{max} T^2 L_{max}}{\mu_{min}} + L \right)$, $\zeta_6 := 4 l_{max} L_{max}^2$, $\zeta_7 := 2 \left[\frac{l_{max} L_{max}}{b} + \frac{L_{max}^2 b(b-1)}{b^2} \right]$, $\zeta_8 := \beta L_m N R \lambda$ *and* $\zeta_9 := 4 L L_m \beta (r+1)^3 \lambda$.

Now, in the following, we characterize the sample complexity of Decentralized FedAvg.

Corollary 2. *Under Assumptions 1-5, to achieve an ϵ accurate solution, Algorithm 2 requires $R = \mathcal{O}\left(T^2/\mu_{min} \log \left(\Phi \left(\underline{\boldsymbol{w}}^0 \right)/\epsilon \right) \right)$ number of communication rounds.*

Corollary 2 shows that even in the *Decentralized* setting, FedAvg is capable of achieving linear convergence. Observe from Theorem 2 and the corollary above that an ϵ-accurate solution can be achieved if the number of global communications rounds R scales as $\mathcal{O}(T^2)$. We know that for the *Server* FedAvg algorithm (complete graph), the second maximum eigenvalue λ_2 is 0. In that case,

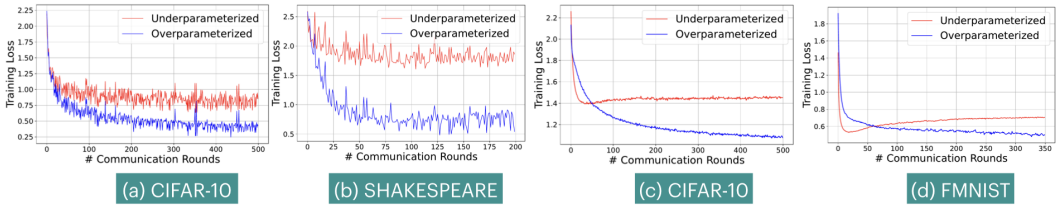

Figure 2: Training loss on different datasets versus the communication rounds for FedAvg in the *Server* (see (a) and (b)) and *Decentralized* (see (c) and (d)) setting.

the second term on the right hand side of equation 3 goes to zero and we are able to recover the result in Theorem 1 for the *Server* FedAvg algorithm. However, R is of the order T^2 as apposed to $T^{3/2}$ in the *Server* case. This is an artifact of our analysis, and we believe that an order of $T^{3/2}$ can be recovered, which is relegated to the future work. By doing a fine grained analysis, one can show that the effect of local updates T on the convergence is quite similar to the *server* setting. We do not present the result that we obtain after the fine grained analysis in Theorem 2 and Corollary 2.

Effect of Network Topology : The effect of *Decentralized* clients is captured through the term involving λ_2. In particular, if $\lambda_2 \neq 0$, then the convergence is relatively slower due to the second term, i.e., $4\eta^4 L L_m \beta T^2 \lambda^2 (r+1)^2$ and Γ_p. Although the above result holds good only for networks with symmetric and doubly stochastic mixing matrix, we believe that similar results hold good even in the general setting as well, and can be proved using the technique developed in this paper.

Again, this is the first result establishing linear convergence of FedAvg in the decentralized setting when minimizing non-convex functions satisfying PL-inequality in the interpolation regime. Next, we present a sketch of the proof and its challenges.

Proof Sketch of Theorem 2 and Challenges: In addition to the challenges mentioned in section 2.2, the *Decentralized* setting poses several new challenges. Unlike server setting, as a consequence of the execution of local updates within each communication round, the nodes do not have consensus. This implies that in the decentralized setting, we need to control the consensus error in addition to the client drift (see Sec. 2.2). We handle this challenge by bounding the loss in terms of the drift term that captures both local and global drifts as in Lemma 3. Further, we bound the drift term which depends on the average loss, leading to two coupled equations; this leads to a new challenge in proving the result. We use a novel application of induction principle to prove linear convergence of both the average loss and the drift. Many distributed machine learning problems lead to such coupled equation, which can be handled in the way described in this paper. Hence, the technique can be of independent interest. In the next section, we present the experimental results.

4 Experimental Results

In this section, we experimentally validate our theoretical findings for the server and the decentralized versions of FedAvg. First, we present the experimental setup for both the settings.

***Server* FedAvg:** Here we consider the FedAvg algorithm with a central server that communicates with 60 or 25 edge devices that run several rounds of local SGD before sharing the model with the server. We consider the following models to validate our theory: (i) a simple regression and Deep Neural Network (DNN) models under overparameterized setting for image classification tasks on CIFAR-10, MNIST, and FMNIST datasets (no. of devices = 60), and (ii) DNN under underparameterized and overparameterized settings for next-character prediction task using Shakespeare dataset (no. of devices = 25). Refer to Appendix A.7 for more details of the experimental setup.

***Decentralized* FedAvg:** In this setting, we run Algorithm 2 for the following networks (i) ring, (ii) random doubly stochastic, and (iii) torus topologies. The neural network architecture and the datasets considered are same as the *server* setting discussed above.

In the above settings, we compare (a) the performance of server/decentralized FedAvg with both underparametrized and overparameterized neural network models, (b) effect of topology on the convergence (*only* decentralized case), and (c) effect of local updates on the convergence. Figure 2 plots the training loss for FedAvg in the *Server* and *Decentralized* setting for underparameterized

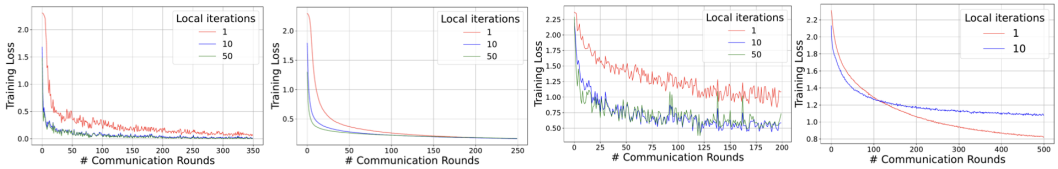

Figure 3: Effect of T on the convergence of FedAvg in the *Server* (see (a) and (b)) and *Decentralized* (see (c) and (d)) setting for simple regression and CNN model.

and overparameterized models on different datasets. As established in Theorems 1 and 2, the loss of FedAvg in the *Server* and *Decentralized* settings diminishes rapidly for the overparameterized models compared to the underparameterized models. This is due to the fact that the PL inequality is satisfied for overparameterized systems which helps to reach the global optimum at a linear rate as demonstrated by Theorems 1 and 2. Additional experiments depicting the testing performance of FedAvg on different models and under different settings are included in Appendix A.7.

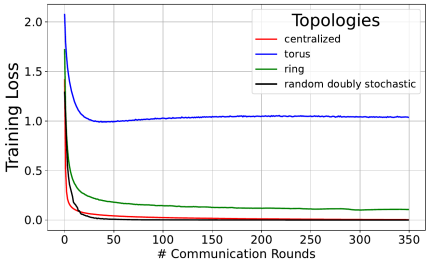

Figure 4: Training loss versus the communication rounds for FedAvg in the *decentralized* setting. Here, random doubly stochastic case has 5 clients while for others we have used 60 clients.

Effect of local updates T: Figure 3 shows plots of the training losses on MNIST dataset for the FedAvg under the *Server* and *Decentralized* settings on the overparameterized regression model and the CNN. From equation 21, we see that as T increases, the convergence speed either decreases or increases depending on the coefficient of T^3 in the second term. We capture this phenomenon in Fig. 3. In particular, as T increases, the rate of convergence increases for simple regression model while it decreases/saturates for the CNN based DNN model. One plausible explanation is that the smoothness constants of simple regression is small, and hence results in smaller second term in equation 21. However, for CNN based DNN, the second term dominates, and hence results in slower convergence with T.

Comparison with different topologies in the *Decentralized* case: Figure 4 shows the training loss versus the communication rounds R for overparameterized CNN model using MNIST dataset with $T = 10$ for four different topologies. Since centralized topology has $\lambda_2 = 0$, it outperforms the network with ring topology and a random (doubly) stochastic matrix. However, the torus topology does not satisfy the conditions required, i.e., symmetric and doubly stochastic matrix, and hence cannot be used for corroborating our theoretical findings. Nevertheless, we have conducted experiments with torus topology, and Figure 4 shows that the torus has the worst convergence performance. One reason for this could be that the ring topology has more structure, i.e., it has a symmetric and doubly stochastic mixing matrix P as opposed to the torus topology. The theoretical analysis of networks with general topology is relegated to our future work.

Conclusion: In this work, we performed a theoretical analysis of the well known FedAvg algorithm for the class of smooth non-convex overparameterized systems in the interpolation regime. We considered two settings, namely (i) *Server* setting where the central server coordinates the exchange of information, and (ii) *Decentralized* setting where nodes communicate over an undirected graph. In this regime, it is well know that neural networks with non-convex loss functions typically satisfy an inequality called Polyak-Lojasiewicz (PL) condition. Assuming PL condition, we showed that in both the settings, the FedAvg algorithm achieves linear convergence rates of $\mathcal{O}(T^{3/2}\log(1/\epsilon))$ and $\mathcal{O}(T^2\log(1/\epsilon))$, respectively, where ϵ is the desired solution accuracy, and T is the number of local SGD updates at each node. As opposed to standard analysis of FedAvg algorithm, we showed that our approach does not require bounded heterogeneity, variance, and gradient assumptions. We captured the heterogeneity in FL training through sample-wise and local smoothness of loss functions. Finally, we carried out experiments on multiple real datasets to confirm our theoretical observations.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pp. 496–506. PMLR, 2022.

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

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

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

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning convergence analysis. *arXiv preprint arXiv:2105.05001*, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020b.

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

Prashant Khanduri, Pranay Sharma, Swatantra Kafle, Saikiran Bulusu, Ketan Rajawat, and Pramod K Varshney. Distributed stochastic non-convex optimization: Momentum-based variance reduction. *arXiv preprint arXiv:2005.00224*, 2020.

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34:6050–6061, 2021.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

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

Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pp. 8119–8129. PMLR, 2021.

Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

Zhaonan Qu, Kaixiang Lin, Jayant Kalagnanam, Zhaojian Li, Jiayu Zhou, and Zhengyuan Zhou. Federated learning's blessing: Fedavg has linear speedup. *arXiv preprint arXiv:2007.05690*, 2020.

Pranay Sharma, Swatantra Kafle, Prashant Khanduri, Saikiran Bulusu, Ketan Rajawat, and Pramod K Varshney. Parallel restarted spider–communication efficient distributed nonconvex optimization with optimal computation complexity. *arXiv preprint arXiv:1912.06036*, 2019.

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

Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *arXiv preprint arXiv:2104.11375*, 2021.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

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

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020b.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pp. 7184–7193. PMLR, 2019a.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019b.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

A APPENDIX

A.1 NOTATION

We use bold small letters to denote vectors, and capital bold letters for matrices. We denote the expected value of a random variable X by $\mathbb{E}[X]$. We denote l_2-norm by $\|.\|_2$ and the Frobenius norm by $\|.\|_F$. Also $\langle ., . \rangle$ denotes the inner product space. The cardinality of a any set $\mathcal{B}$ is represented by $|\mathcal{B}|$. We use the standard notation $\mathcal{O}(n)$ to denote the order of n. For a vector valued function $\Phi(\boldsymbol{w})$, the gradient is denoted by $\nabla\Phi(\boldsymbol{w})$, and the Hessian is denoted by $\nabla^2\Phi(\boldsymbol{w})$. We use $\mathbf{1}$ to represent a column vector with all ones. We use $[N]$ to denote the set $\{1, 2, \ldots, N\}$.

A.2 RELATED WORK

After the introduction of the FedAvg (McMahan et al., 2017), multiple works have analyzed the convergence of FedAvg in the server setting and with homogeneous data, i.e., when the data is i.i.d across clients (see Stich (2018); Wang & Joshi (2018); Khaled et al. (2019); Yu et al. (2019b); Wang et al. (2019); Yang et al. (2021)). The authors in (Stich, 2018) were the first to obtain a rate of $\mathcal{O}(1/N\epsilon)$ for strongly convex and smooth problems. Later (Haddadpour et al., 2019; Haddadpour & Mahdavi, 2019) proved a similar result but for non-convex functions satisfying PL inequality. In (Woodworth et al., 2020a), the authors analyzed the trade-off between Minibatch and Local SGD in the homogeneous settings and established $\mathcal{O}(1/N\epsilon^2)$ convergence rates for minimizing smooth convex objectives. The analysis of FedAvg for the general non-convex settings was first performed in Yu et al. (2019b) where the authors establish a rate of $\mathcal{O}(1/N\epsilon^2)$. Recently, many works have adapted the analyses of FedAvg for minimizing the non-convex losses in the heterogeneous data settings. For example, Yu et al. (2019a) extended the results of Yu et al. (2019b) for the heterogeneous data setting. Specifically, the authors in (Yu et al., 2019a) utilized a Momentum SGD updates and established the convergence rate of $\mathcal{O}(1/N\epsilon^2)$ under bounded heterogeneity setting. The work (Karimireddy et al., 2020b) also provided a tight analysis for FedAvg and established linear speed-up with the number of clients. Recently, (Yang et al., 2021) analyzed the linear speed-up effect of FedAvg while (Khanduri et al., 2021) analyzed the trade-off between the batch sizes and the local updates. We note that all these works establish a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ for minimizing non-convex smooth losses in the bounded heterogeneity setting. It is also worth noting that numerous works have proposed variants of FedAvg with different local update rules (e.g., variance reduction, momentum SGD, adaptive updates, etc.) with the goal of improving the performance of FedAvg (Karimireddy et al., 2020b; Sharma et al., 2019; Liang et al., 2019; Khanduri et al., 2021; 2020; Karimireddy et al., 2020a; Das et al., 2022). However, in practice FedAvg remains the algorithm of choice for training large FL systems.

There are a few works that have analyzed the performance of Fedvg in the decentralized settings. One of the initial works, (Lian et al., 2017) considered a decentralized parallel SGD (D-PSGD) and provided convergence rate of $\mathcal{O}(1/N\epsilon^2)$ for minimizing smooth non-convex functions. Later, (Haddadpour & Mahdavi, 2019) analyzed the convergence of FedAvg under both server and decentralized setting with bounded gradient dissimilarity assumption. The authors showed a convergence rate of $\mathcal{O}(1/N\epsilon^2)$ for minimizing non-convex functions in both the server and decentralized settings. The authors in Yu et al. (2019a) also extended the analysis of Momentum SGD to decentralized networks and established a convergence of $\mathcal{O}(1/N\epsilon^2)$ for minimizing non-convex functions. All the above works provide a sublinear rate of convergence for FedAvg, however, as illustrated in Fig. 1, FedAvg converges at a much faster rate in practice. To understand this behavior of FedAvg, in this work we analyze the performance of FedAvg under both server and decentralized settings for minimizing a special class of non-convex functions satisfying PL inequality under the interpolation regime. We note that overparameterized neural networks/systems usually operate in the interpolation regime while their loss functions have been shown to satisfy the PL inequality (Liu et al., 2022).

The linear convergence of centralized SGD in the interpolation regime for minimizing PL objectives was first established in Bassily et al. (2018). Recently, (Qu et al., 2020) showed linear convergence rate of FedAvg in the server setting for minimizing strongly-convex objectives in the overparameterized regime. Similarly, the authors in (Koloskova et al., 2020) have also established the linear convergence of FedAvg in the decentralized setting for minimizing strongly-convex losses in an overparameterized setting. The above works only focus on analysis of FedAvg for the strongly-convex objectives in the overparameterized regime while we focus on the more general class of

non-convex functions satisfying the PL inequality. Moreover, compared to other works that assume restrictive bounded gradient, heterogeneity, and variance assumptions, we show that such assumptions can be avoided by using a sample-wise smoothness assumption. Table 1 presents a summary of the above discussion.

In a separate line of work, the linear convergence of SGD (and GD) for optimizing overparameterized neural networks/systems with specific activation functions, network widths, and assumptions on data and loss functions has been established (Zou et al., 2020; Li & Liang, 2018; Allen-Zhu et al., 2019; Jacot et al., 2018; Du et al., 2018; Chizat et al., 2019; Nguyen & Mondelli, 2020). Recently, the works in (Huang et al., 2021; Deng et al., 2022) have extended some of these specific neural network architectures to FL settings. However, we note that these works are orthogonal to our setting as we consider a general setting without assuming a specific model to be learned.

A.3 Useful Lemmas

In this section, we state two Lemmas that will be used in proving our main results.

Lemma 1. *For any matrices $A \in \mathbb{C}^{N\times N}$ and $B \in \mathbb{C}^{N\times d}$, we have $\|AB\|_F^2 \leq N \|A\|_{op}^2 \|B\|_F^2$.*

Lemma 2. *(See Lemma 1 in Sun et al. (2021)) For any $m \in \mathbb{N}$, the mixing matrix P satisfies $\|P^m - Q\|_{op} \leq \lambda_2^m$, where λ_2 is the second largest eigenvalue of the mixing matrix P, and $Q := \frac{1}{N}\mathbf{1}\mathbf{1}^T$.*

A.4 Proof of theorem 1

In this section, we present the proofs for the convergence of Algorithm 1.

A.4.1 Useful Lemmas to Prove Theorem 1

To start with, we briefly discuss some Lemmas to prove the main result. Using the following Lemmas, theorem 1 will be proved in Sec. A.4.2. The local model drift is bounded in terms of local loss. The local model drifts away from the global averaged model during the local updates which is the essence of the following lemma.

Lemma 1. *The local drift $\frac{1}{N}\sum_{k=1}^N \mathbb{E}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2$ is bounded in terms of local weight i.e., $\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)$ as follows*

$$\frac{1}{N}\sum_{k=1}^N \mathbb{E}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 \leq \frac{\eta^2 t}{N}\left[\frac{2l_{max}}{b} + \frac{2b(b-1)L_{max}}{b^2}\right]\sum_{k=1}^N\sum_{\tau=0}^{t-1}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right].$$

where $l_{max} := \max_{k,j} l_{k,j}$ and $L_{max} := \max_k L_k$

Proof: Using the step 7 of Algorithm 1, we have

$$\boldsymbol{w}_k^{r,t} = \boldsymbol{w}_k^{r,t-1} - \frac{\eta}{b}\sum_{j\in\mathcal{B}_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t-1}\right).$$

Performing the telescopic sum over $\boldsymbol{w}$, we get

$$\boldsymbol{w}_k^{r,t} = \boldsymbol{w}_k^{r,0} - \eta\sum_{\tau=0}^{t-1}\frac{1}{b}\sum_{j\in\mathcal{B}_k^{r,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right). \tag{4}$$

Averaging over $k \in [N]$ results in

$$\underline{\boldsymbol{w}}^{r,t} = \underline{\boldsymbol{w}}^{r,0} - \frac{\eta}{N}\sum_{k=1}^N\sum_{\tau=0}^{t-1}\frac{1}{b}\sum_{j\in\mathcal{B}_k^{r,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right). \tag{5}$$

Using equation 4 and equation 5 in $\frac{1}{N}\sum_{k-1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2$ and noting the fact that $\underline{\boldsymbol{w}}^{r,0}=\boldsymbol{w}_k^{r,0}$, we get

$$\frac{1}{N}\sum_{k-1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2\leq\frac{1}{N}\sum_{k=1}^{N}\|\eta\sum_{\tau=0}^{t-1}\frac{1}{b}\sum_{j\in\mathcal{B}_k^{r,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)-\frac{\eta}{N}\sum_{k'=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j'\in\mathcal{B}_{k'}^{r,\tau}}\nabla\Phi_{k',j'}\left(\boldsymbol{w}_{k'}^{r,\tau}\right)\|^2.$$

For a sequence X_k for $k\in[N]$, we have $\sum_{k=1}^{N}\|X_k-\underline{X}\|^2\leq\sum_{k=1}^{N}\|X_k\|^2$. Applying this in the above results in

$$\begin{aligned}\frac{1}{N}\sum_{k-1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2\leq&\frac{1}{N}\sum_{k=1}^{N}\|\eta\sum_{\tau=0}^{t-1}\frac{1}{b}\sum_{j\in\mathcal{B}_k^{r,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\|^2 \qquad (6)\\ \leq&\frac{\eta^2 t}{N}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\left[\frac{1}{b^2}\sum_{j\in B_k^{r,\tau}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right\|^2+\frac{1}{b^2}\sum_{j\neq j'}\mathcal{F}_k^{r,\tau}\right].\end{aligned}$$

where $\mathcal{F}_k^{r,\tau}:=\left\langle\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right),\nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,\tau}\right)\right\rangle$. Taking expectation, we get

$$\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2\leq\frac{\eta^2 t}{N}\sum_{k=1}^{N}\sum_{t=0}^{t-1}\left[\frac{1}{b}\mathbb{E}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right\|^2+\frac{b(b-1)}{b^2}\mathbb{E}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right\|^2\right].$$

Further, using smoothness assumption (see assumption 3), we have

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2&\leq\frac{\eta^2 t}{N}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\left[\frac{2l_{k,j}}{b}\mathbb{E}\left[\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right]+\frac{2L_k b(b-1)}{b^2}\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]\\&\leq\frac{\eta^2 t}{N}\sum_{k,\tau=1,0}^{N,t-1}\left[\frac{2\max_{k,j}l_{k,j}}{b}\mathbb{E}\left[\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right]+\frac{2\max_k L_k b(b-1)}{b^2}\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]\\&\overset{(a)}{\leq}\frac{\eta^2 t}{N}\left[\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right]\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]. \qquad (7)\end{aligned}$$

where (a) follows from the fact that $l_{max}:=\max_{k,j}l_{k,j}$ and $L_{max}:=\max_k L_k$. □

Next, we show that the local loss is bounded in terms of global average weight. This is necessary to obtain linear convergence of Algorithm 1.

Lemma 2. *The local average loss $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]$ is bounded in terms of global average weight i.e., $\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$ as follows*

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]\leq\left(1-\frac{\eta\mu_k}{2}\right)^{\tau}\Phi_k\left(\underline{\boldsymbol{w}}^r\right). \qquad (8)$$

Proof: Applying the smoothness assumption (see 3) for $\Phi_k\left(\boldsymbol{u}\right)$, we have

$$\begin{aligned}\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)&\leq\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)+\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right),\boldsymbol{w}_k^{r,\tau}-\boldsymbol{w}_k^{r,\tau-1}\right\rangle+\frac{L_k}{2}\left\|\boldsymbol{w}_k^{r,\tau}-\boldsymbol{w}_k^{r,\tau-1}\right\|^2\\&=\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)+\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right),\frac{1}{b}\sum_{j\in\mathcal{B}_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle+\frac{\eta^2 L_k}{2b^2}\left\|G_k^{r,\tau-1}\right\|^2.\end{aligned}$$

where $G_k^{r,\tau-1}:=\sum_{j\in\mathcal{B}_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)$. The last equality follows from step 7 of **Algorithm 1**, i.e., $\boldsymbol{w}_k^{r,\tau}-\boldsymbol{w}_k^{r,\tau-1}=-\frac{\eta}{b}\sum_{j\in\mathcal{B}_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)$. Taking expectation on both sides in the above, we get

$$\begin{aligned}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]&\leq\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)-\eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|^2+\frac{L_k\eta^2}{2b^2}\left\|G_k^{r,\tau-1}\right\|^2\right]\\&\leq\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)-\eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|^2+\frac{L_{\max}\eta^2}{2b^2}\left\|G_k^{r,\tau-1}\right\|^2\right], \qquad (9)\end{aligned}$$

where $L_{max} := \max_k L_k$. The last term on the right side in equation 9 can be bounded as

$$\frac{1}{b^2}\mathbb{E}\left\|\sum_{j\in\mathcal{B}_k^{r,\tau-1}}\nabla\Phi_{kj}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|^2 \leq \mathbb{E}\left[\frac{1}{b^2}\sum_{j\in B_k^{r,\tau-1}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|^2+\frac{1}{b^2}\sum_{j\neq j'}\mathcal{F}_k^{r,\tau-1}\right]$$
$$\overset{(a)}{\leq} \left[\frac{2l_{max}}{b}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right]+\frac{2L_{\max}b(b-1)}{b^2}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right]\right],$$

where $\mathcal{F}_k^{r,\tau-1} := \left\langle\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right),\nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle$, and (a) follows from smoothness assumption and the fact that $l_{max} := \max_{k,j} l_{k,j}$ and $L_{max} := \max_k L_k$. Now, plugging the above in equation 9, we get

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)-\eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|^2+\eta^2\left(\frac{l_{max}L_{max}}{b}+\frac{L_{max}^2b(b-1)}{b^2}\right)L_k^{r,\tau-1}\right].$$

where $L_k^{r,\tau-1} := \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)$. Using the local PL inequality i.e., $\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|^2 \geq \mu_{min}\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)$, where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$. (see definition 2), the above can be further bounded as

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left[1-\eta\mu_{min}+\eta^2\left(\frac{l_{max}L_{max}}{b}+\frac{L_{max}^2b(b-1)}{b^2}\right)\right]\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].$$

Choosing $\eta \leq \frac{\mu_{min}}{2\left(\frac{l_{max}L_{max}}{b}+\frac{L_{max}^2b(b-1)}{b^2}\right)}$ results in the following

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].$$

It is easy to see that the above implies

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\Phi_k\left(\underline{\boldsymbol{w}}^r\right).$$

□

In the next subsection, we provide the proof of 1 using Lemmas proved above.

A.4.2 Completing the Proof of Theorem 1

From the Assumption 1, $\Phi\left(\boldsymbol{w}\right)$ can be written as

$$\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right) \leq \Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)+\langle\nabla\Phi(\underline{\boldsymbol{w}}^{r,t}),\underline{\boldsymbol{w}}^{r,t+1}-\underline{\boldsymbol{w}}^{r,t}\rangle+\frac{L}{2}\left\|\underline{\boldsymbol{w}}^{r,t+1}-\underline{\boldsymbol{w}}^{r,t}\right\|^2.$$

Now, using the stochastic gradient descent update $\underline{\boldsymbol{w}}^{r,t+1}-\underline{\boldsymbol{w}}^{r,t} = -\frac{\eta}{bN}\left(\sum_{k=1}^N\sum_{j\in\mathcal{B}^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right)$ in the above and using Assumption 4, we get

$$\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right) \leq \Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)-\eta\left\langle\nabla\Phi(\underline{\boldsymbol{w}}^{r,t}),\frac{1}{bN}\sum_{k=1}^N\sum_{j\in\mathcal{B}^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle+\frac{\eta^2L}{2}\left\|\mathcal{G}^{r,t}\right\|^2.$$

where $\mathcal{G}^{r,t} := \frac{1}{bN}\sum_{k=1}^N\sum_{j\in\mathcal{B}^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)$. Taking the expectation conditioning on $\boldsymbol{w}_k^{r,t}$, we get[2]

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq \mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]-\eta\underbrace{\left\langle\nabla\Phi(\underline{\boldsymbol{w}}^{r,t}),\frac{1}{N}\sum_{k=1}^N\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle}_{\mathcal{A}_1}+$$
$$\frac{\eta^2L}{2}\left(\underbrace{\frac{1}{b^2N^2}\sum_{k=1}^N\left\|\sum_{j\in\mathcal{B}^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{\mathcal{A}_2}+\underbrace{\frac{1}{b^2N^2}\sum_{k\neq k'}\left\langle\sum_{j\in\mathcal{B}^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right),\sum_{i\in\mathcal{B}^{r,t}}\nabla\Phi_{k',i}\left(\boldsymbol{w}_{k'}^{r,t}\right)\right\rangle}_{\mathcal{A}_3}\right). \tag{10}$$

[2]The conditional term is not explicitly written. However, it be clear from the context.

The term $\mathcal{A}_2$ in equation 10 can be bounded as

$$\mathcal{A}_2 = \frac{1}{b^2N^2}\sum_{k=1}^{N}\sum_{j\in\mathcal{B}^{r,t}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 + \frac{1}{b^2N^2}\sum_{k=1}^{N}\sum_{j\neq j'}\left\langle\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right),\nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle.$$

Now taking the expectation conditioning on $\boldsymbol{w}_k^{r,t}$, we get

$$\mathbb{E}[\mathcal{A}_2] = \frac{1}{bN^2}\sum_{k=1}^{N}\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \frac{b(b-1)}{b^2N^2}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2. \tag{11}$$

Taking the expectation of $\mathcal{A}_3$ in equation 10, we get

$$\begin{aligned}
\mathbb{E}[\mathcal{A}_3] &= \frac{1}{N^2}\sum_{k\neq k'}\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right),\nabla\Phi_{k'}\left(\boldsymbol{w}_{k'}^{r,t}\right)\right\rangle \\
&\overset{(a)}{\leq} \frac{1}{2N^2}\sum_{k\neq k'}\left[\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \|\nabla\Phi_{k'}\left(\boldsymbol{w}_{k'}^{r,t}\right)\|^2\right] \\
&= \frac{2(N-1)}{2N^2}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 \\
&\leq \frac{1}{N}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2,
\end{aligned} \tag{12}$$

where (a) follows from the fact that $\langle a,b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. Next, the inner product term $\mathcal{A}_1$ in equation 10 can be written as

$$\begin{aligned}
\mathcal{A}_1 &= \frac{1}{2}\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 + \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \frac{1}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\
&\overset{(a)}{\geq} \frac{1}{2}\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 + \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 - \frac{L^2}{2N}\sum_{k=1}^{N}\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\|^2,
\end{aligned} \tag{13}$$

where (a) follows from smoothness assumption (see 1). Substituting equation 11, equation 12 and equation 13 in equation 10, we get the following

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq\ & \mathbb{E}\Bigg[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 - \frac{\eta}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 \\
&+ \frac{\eta L^2}{2N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + \frac{\eta^2 L}{2bN^2}\underbrace{\sum_{k=1}^{N}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{:=\mathcal{A}_4} \\
&+ \left(\frac{\eta^2 Lb(b-1)}{2b^2N^2} + \frac{\eta^2 L}{2N}\right)\underbrace{\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{:=\mathcal{A}_5}\Bigg].
\end{aligned} \tag{14}$$

The term $\mathcal{A}_4$ in equation 14 can be upper bounded as follows

$$\begin{aligned}
\mathcal{A}_4 &\overset{(a)}{\leq} 2\sum_{k=1}^{N}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 + 2\sum_{k=1}^{N}\left\|\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\
&\overset{(b)}{\leq} 2\sum_{k=1}^{N}l_{k,j}^2\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4\sum_{k=1}^{N}l_{k,j}\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right) \\
&\overset{(c)}{\leq} 2l_{max}^2\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4l_{max}\sum_{k=1}^{N}\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right),
\end{aligned}$$

where (a) follows by adding and subtracting $\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)$ and using the fact that, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (b) follows from Assumption 3, and (c) follows from the fact that $l_{max} := \max_{k,j} l_{k,j}$. Taking the expectation of $\mathcal{A}_4$, we get the following bound

$$\mathbb{E}\left[\mathcal{A}_4\right] \leq 2l_{max}^2 \sum_{k=1}^{N} \mathbb{E}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4l_{max} \sum_{k=1}^{N} \Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right). \tag{15}$$

Now, let us upper bound the term $\mathcal{A}_5$ in equation 14 as

$$\begin{aligned}
\mathcal{A}_5 &\overset{(a)}{\leq} 2\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 + 2\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\
&\overset{(b)}{\leq} 2\sum_{k=1}^{N} L_k^2\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4\sum_{k=1}^{N} L_k\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right) \\
&\overset{(c)}{\leq} 2L_{max}^2\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4L_{max}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right).
\end{aligned} \tag{16}$$

In the above, (a) follows by adding and subtracting $\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)$ and using the fact that, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and (b) follows from Assumption 3 and (c) follows from the fact that $L_{max} := \max_k L_k$. Substituting upper bounds from equation 15 and equation 16 in equation 14, we get

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq\ & \mathbb{E}\Bigg[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 - \frac{\eta}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 \\
&+ \left(\frac{\eta L^2}{2N} + \frac{\eta^2 L l_{max}^2}{bN^2} + \frac{\eta^2 L L_{max}^2}{N^2} + \frac{\eta^2 L L_{max}^2}{N}\right)\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 \\
&+ \left(\frac{2\eta^2 L l_{max}}{bN} + \frac{2\eta^2 L L_{max}}{N} + 2\eta^2 L L_{max}\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\Bigg].
\end{aligned} \tag{17}$$

Now, using PL inequality, i.e., $\|\nabla\Phi(\boldsymbol{w})\|^2 \geq \mu\Phi(\boldsymbol{w}),\ \forall \boldsymbol{w} \in \mathbb{R}^d$ and rearranging the terms, we get

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq\ & \mathbb{E}\Bigg[\left(1 - \frac{\eta\mu}{2} + \left(\frac{2\eta^2 L l_{max}}{bN} + \frac{2\eta^2 L L_{max}}{N} + 2\eta^2 L L_{max}\right)\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) \\
&+ \left(\frac{\eta L^2}{2} + \frac{\eta^2 L l_{max}^2}{bN} + \frac{\eta^2 L L_{max}^2}{N} + \eta^2 L L_{max}^2\right)\frac{1}{N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2\Bigg].
\end{aligned}$$

Choosing $\eta \leq \min\left\{\frac{\mu}{4\left(\frac{2Ll_{max}}{bN} + \frac{2LL_{max}}{N} + 2LL_{max}\right)}, \frac{L^2}{2\left(\frac{Ll_{max}^2}{bN} + \frac{LL_{max}^2}{N} + LL_{max}^2\right)}\right\}$, the above can be further bounded as

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq \mathbb{E}\left(1 - \frac{\eta\mu}{4}\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) + \frac{\eta L^2}{N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2. \tag{18}$$

In order to prove linear convergence, it suffices to show that the second term above, i.e., $\frac{1}{N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2$ is exponential in $\Phi(\boldsymbol{w})$. From Lemma 1, it follows that the second term on the right hand side in equation 18, becomes

$$\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 \leq \frac{\eta^2 t}{N}\left[\frac{2l_{max}}{b} + \frac{2b(b-1)L_{max}}{b^2}\right]\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]. \tag{19}$$

Substituting equation 8 of Lemma 2, i.e. $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]$, in the above results in

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2 &\leq \frac{\eta^2 t}{N}\left[\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right]\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\mathbb{E}\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\\ &\overset{(a)}{\leq} \frac{\eta^2 t^2}{N}\left[\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right]\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\\ &\overset{(b)}{=} \eta^2 t^2\left[\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right]\Phi\left(\underline{\boldsymbol{w}}^r\right), \end{aligned} \tag{20}$$

where (a) follows by choosing $\eta \leq \frac{2}{\mu_{min}}$ and (b) follows from the fact that $\frac{1}{N}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^r\right) = \Phi\left(\underline{\boldsymbol{w}}^r\right)$. Using recursion on equation 18, we get

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{4}\right)^{T}\Phi\left(\underline{\boldsymbol{w}}^r\right)+\frac{\eta L^2}{N}\sum_{\tau=0}^{T-1}\left(1-\frac{\eta\mu}{4}\right)^{\tau}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,T-1-\tau}-\underline{\boldsymbol{w}}^{r,T-1-\tau}\right\|^2.$$

It follows from the update step that $\frac{1}{N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,T-1-\tau}-\underline{\boldsymbol{w}}^{r,T-1-\tau}\right\|^2 = 0$ for $\tau = T-1$. Using $\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\| \leq \eta^2 t^2\left[\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right]\Phi\left(\underline{\boldsymbol{w}}^r\right)$ in the above results in

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq{}& \left(1-\frac{\eta\mu}{4}\right)^{T}\Phi\left(\underline{\boldsymbol{w}}^r\right)\\ &+\eta L^2\sum_{\tau=0}^{T-2}\left(1-\frac{\eta\mu}{4}\right)^{\tau}\eta^2(T-1)^2\left[\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right]\Phi\left(\underline{\boldsymbol{w}}^r\right).\end{aligned}$$

Setting $\eta \leq \frac{4}{\mu}$ gives

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left[\left(1-\frac{\eta\mu}{4}\right)^{T}+\eta^3 L^2(T-1)^3\left(\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right)\right]\Phi\left(\underline{\boldsymbol{w}}^r\right). \tag{21}$$

Using the fact $\left(1-\frac{\eta\mu}{4}\right)^{T} \leq \left(1-\frac{\eta\mu}{4}\right)$, and choosing $\eta \leq \left[\frac{\mu}{8L^2T^3\left(\frac{2l_{max}}{b}+\frac{2b(b-1)L_{max}}{b^2}\right)}\right]^{\frac{1}{2}}$ results in the following exponential bound

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{8}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right]. \quad \square$$

A.5 Proof of Theorem 2

In this section, we first present the overview of the proof. Then, we will state and prove Lemmas required to prove the Theorem. The proof mainly consists of three intermediate steps, namely bounding i) the local loss, ii) the loss in terms of future iterates, and iii) the global drift. In the Lemma 6, we bound the local loss. We use L_k smoothness (see definition 1) and local PL inequality to show loss at local weight is bounded in terms loss at global average weight and the drift.

A.5.1 Proof of Theorem 2

We simplify the presentation of the proof by using the following matrix notations. Let the local average weights be denoted by $\underline{W}_l^r := \left[\underline{\boldsymbol{w}}_1^r, \underline{\boldsymbol{w}}_2^r, \ldots, \underline{\boldsymbol{w}}_N^r\right]^T \in \mathbb{R}^{N\times d}$, where $\underline{\boldsymbol{w}}_k^r \in \mathbb{R}^d$. The `Aggregation` step of Algorithm 2 can be compactly written in matrix form as

$$\underline{\boldsymbol{w}}_k^{r+1} = \sum_{i\in\mathcal{N}_k} p_{k,i}\boldsymbol{w}_k^{r,T} \quad \equiv \quad \underline{W}_l^{r+1} = PW^r, \tag{22}$$

where $\mathcal{N}_k := \{i : p_{k,i} > 0\}$. Further, we define the global average as

$$\underline{\boldsymbol{w}}^r := \frac{1}{N}\sum_{k=1}^{N}\underline{\boldsymbol{w}}_k^r \quad \equiv \quad \underline{W}^r = Q\underline{W}_l^r, \tag{23}$$

where the average matrix $Q := \frac{1}{N}\mathbf{1}\mathbf{1}^T$. Now, let us represent the gradients compactly in the matrix form as

$$\partial\hat{\Phi}\left(W^{r,t}\right) = \left[\frac{1}{b}\sum_{j\in B_1^{r,t}} G_{1,j}^{(r,t)}, \frac{1}{b}\sum_{j\in B_2^{r,t}} G_{2,j}^{(r,t)}, \ldots, \frac{1}{b}\sum_{j\in B_N^{r,t}} G_{N,j}^{(r,t)}\right], \tag{24}$$

where $G_{l,j}^{(r,t)} := \nabla\Phi_{l,j}\left(\boldsymbol{w}_l^{r,t}\right)$. The mixing matrix P also preserves the average, and hence $QP = P$.

We start by proving an upper bound on the average loss $\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right]$ in terms of the loss $\Phi\left(\underline{\boldsymbol{w}}^{r}\right)$ in the r-th communication round, and the drift $\mathcal{D}_{r,0}$, as shown in the following Lemma.

Lemma 3. *The average loss is bounded in terms of the drift as follows*

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1 - \frac{\eta\mu}{8}\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right) + \frac{6\eta^2 L}{N}\mathcal{D}_{r,0}, \tag{25}$$

where the drift $\mathcal{D}_{r,0} := \left\|\underline{W}_l^{r,0} - \underline{W}^{r,0}\right\|_F^2$, *and* η *is chosen according to equation 2.*

Proof: The proof is provided in Appendix A.6. □

It is easy to see from Lemma 3 that we can obtain the convergence result provided in theorem 2 provided the drift term on the right hand side of equation 25 is bounded in terms of loss. More specifically, if $\mathcal{D}_{r,0} \leq \texttt{constant} \times \Phi\left(\underline{\boldsymbol{w}}^{r}\right)$, then the linear convergence stated in Theorem 2 can be easily proved by substitution. Before proving this, in the following lemma, we provide a recursion of the drift in terms of the average loss and the past drift.

Lemma 4. *The drift is bounded in terms of* $\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)$ *as follows*

$$\mathcal{D}_{r,0} \leq \eta^2\beta T^2 N L_m\left(\sum_{\tau=0}^{r-1}\lambda^{r-\tau}D_{\tau,0} + \sum_{\tau=0}^{r-1}\lambda^{r-\tau}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau+1,0}\right)\right]\right), \tag{26}$$

where $L_m := \max\left\{L_{max}^2, 2L_{max}N\right\}$, $\beta := \frac{4l_{max}\psi^2 N}{(1+\psi)\mu_{min}}$, $\lambda \triangleq \left(1 + \frac{1}{\psi}\right)\lambda_2^2$.

Proof: The proof is provided in Appendix A.6. □

Next, our task is to show that the recursion in equation 26 satisfies a bound of the form $\mathcal{D}_{r,0} \leq \texttt{constant}^r \times \Phi\left(\underline{\boldsymbol{w}}^{0}\right)$, which is the desired result. Here, the `constant` is less than one. We use induction along with carefully choosing η to achieve this goal. The following lemma provides the desired result.

Lemma 5. *Using equation 26 and equation 25 and by induction on* $\mathcal{D}_{r+1,0}$ *we get*

$$\mathcal{D}_{r+1,0} \leq (2r+3)\eta^2\beta T^2 L_m N\lambda^2\Lambda^{r+1}\Phi\left(\underline{\boldsymbol{w}}^{0}\right), \tag{27}$$

where $L_m := \max\left\{L_{max}^2, 2L_{max}N\right\}$ *and* $\beta := \frac{4l_{max}\psi^2 N}{(1+\psi)\mu_{min}}$.

Proof: The proof is provided in Appendix A.6.1. □

First, note that if the network is fully connected or centralized, i.e., $\lambda_2 = 0$, then the drift term becomes zero, as expected. Further, the drift increases with the number of clients N and the number of local rounds T. Nevertheless, it goes down with Λ exponentially provided $\Lambda < 1$. This ensures that the exponential bound in our main result holds good. Finally, the proof of Theorem 2 is complete by using equation 26 and equation 27 in equation 25. In the next subsection, we state and prove some useful Lemmas that are required to prove the main result.

A.5.2 Useful Lemmas to Prove Theorem 2

Lemma 6. *The function $\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)$ satisfies local PL inequality and can be bounded in terms of global average weight i.e., $\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$ as follows*

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2, \tag{28}$$

where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$.

Proof: From assumption 1, the function $\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)$ is written as

$$\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right) \leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) + \left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \boldsymbol{w}_k^{r,\tau} - \boldsymbol{w}_k^{r,\tau-1}\right\rangle + \frac{L_k}{2}\left\|\boldsymbol{w}_k^{r,\tau} - \boldsymbol{w}_k^{r,\tau-1}\right\|_2^2. \tag{29}$$

We know from step 7 of **Algorithm** 2, $\boldsymbol{w}_k^{r,\tau} - \boldsymbol{w}_k^{r,\tau-1} = -\frac{\eta}{b}\sum_{j\in B_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)$. Using this in equation 29, we get

$$\begin{aligned}
\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right) &\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \frac{1}{b}\sum_{j\in B_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle + \frac{\eta^2 L_k}{2}G_k(r,\tau).\\
&\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \frac{1}{b}\sum_{j\in B_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1)}\right)\right\rangle\\
&+ \frac{\eta^2 L_k}{2b^2}\sum_{j\in B_k^{r,\tau-1}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_k}{2b^2}\sum_{j\neq j'}\left\langle\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right), \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle.\\
&\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \frac{1}{b}\sum_{j\in B_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1)}\right)\right\rangle\\
&+ \frac{\eta^2 L_{max}}{2b^2}\sum_{j\in B_k^{r,\tau-1}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_{max}}{2b^2}\sum_{j\neq j'}\left\langle\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right), \nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle.
\end{aligned}$$

where $G_k(r,\tau) := \left\|\frac{1}{b}\sum_{j\in B_k^{r,\tau-1}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2$, and $L_{max} := \max_k L_k$. Taking expectation with respect to $\boldsymbol{w}_k^{r,\tau-1}$ in the above, gives us

$$\begin{aligned}
\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] &\leq \mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\langle\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right), \nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\rangle + \frac{\eta^2 L_{max}}{2b}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2\right.\\
&+ \left.\frac{\eta^2 L_{max} b(b-1)}{2b^2}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2\right].
\end{aligned}$$

Applying smoothness assumption of each sample, i.e., $\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 \leq 2l_{k,j}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)$, we have

$$\begin{aligned}
\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] &\leq \mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_{max} l_{k,j}}{b}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right.\\
&+ \left.\frac{\eta^2 L_{max} b(b-1)L_k}{b^2}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right]\right].\\
&\leq \Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right) - \eta\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 + \frac{\eta^2 L_{max} l_{max}}{b}\mathbb{E}\left[\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right]\\
&+ \frac{\eta^2 L_{max} b(b-1)L_{max}}{b^2}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].
\end{aligned} \tag{30}$$

where $l_{max} := \max_k L_k$. From the local PL inequality (see definition 2), it follows that $\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right\|_2^2 \geq \mu_{min}\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)$ for $k = \{1,2,\ldots,N\}$, where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$.

Using this in equation 30 results in

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left[1-\eta\mu_{min}+\eta^2\left(\frac{l_{max}L_{max}}{b}+\frac{L_{max}^2 b(b-1)}{b^2}\right)\right]\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].$$

By setting $\eta \leq \frac{\mu_{min}}{2\left[\frac{l_{max}L_{max}}{b}+\frac{L_{max}^2 b(b-1)}{b^2}\right]}$, the above can be further bounded as

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau-1}\right)\right].$$

Since $\boldsymbol{w}_k^{r,0}=\underline{\boldsymbol{w}}_k^r$, the above can be written as

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)\right]. \tag{31}$$

Using the local PL inequality, i.e., $\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right) \leq \frac{1}{\mu_{min}}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)\right\|_2^2$ in equation 31, we have

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\frac{1}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)\right\|_2. \tag{32}$$

Now, adding and subtracting the term $\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$ in the above, and using the fact that $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$, we get

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\frac{2}{\mu_{min}}\mathbb{E}\left(\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}_k^r\right)-\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|_2^2+\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|_2^2\right).$$

Using L_k smoothness assumption (see Assumption 3), we have

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \left(1-\frac{\eta\mu_{min}}{2}\right)^{\tau}\mathbb{E}\left(\frac{2L_k^2}{\mu_{min}}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{2}{\mu_{min}}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|_2^2\right).$$

Choosing $\eta \leq \frac{2}{\mu_{min}}$ and using the fact that $L_{max}=\max_k L_k$, we get

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2. \quad \square \tag{33}$$

Corollary 3. *The function $\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)$ satisfies local PL inequality and can be bounded in terms of global average weight i.e., $\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$ as follows*

$$\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{4L_{max}}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right], \tag{34}$$

where $\mu_{min} := \min_{k\in[N]}\{\mu_k\}$ and $L_{max} := \max_k L_k$.

Proof: The proof directly follows from Lemma 6 by using the smoothness assumption, i.e., $\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2 \leq 2L_{max}\Phi_k\left(\underline{\boldsymbol{w}}^r\right)$. This completes the proof. $\square$

Next, we show that the loss can be bounded in terms of the future iterates as follows.

Lemma 7. *The function $\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)$ is bounded in terms of the future value of the function as given below*

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right] \leq 2\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right)+\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r-1}-\underline{\boldsymbol{w}}^{r-1}\right\|_2^2\right].$$

Proof: It follows from the smoothness assumption that

$$\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) \geq \Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)+\left\langle\nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right),\underline{\boldsymbol{w}}^{r,0}-\underline{\boldsymbol{w}}^{r-1,0}\right\rangle-\frac{L}{2}\left\|\underline{\boldsymbol{w}}^{r,0}-\underline{\boldsymbol{w}}^{r-1,0}\right\|^2. \tag{35}$$

Telescoping the update in step 7 of **Algorithm 2**, we get $\boldsymbol{w}_i^{r-1,T} = \boldsymbol{w}_i^{r-1,0} - \frac{\eta}{b}\sum_{\tau=0}^{T-1}\sum_{j\in\mathcal{B}_k^{r-1,\tau}} \nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r-1,\tau}\right)$. Averaging over all neighboring nodes $i \in \mathcal{N}_k$, we get

$$\underline{\boldsymbol{w}}_k^{r,0} = \sum_{i\in\mathcal{N}_k} p_{k,i}\boldsymbol{w}_i^{r-1,T} = \sum_{i\in\mathcal{N}_k} p_{k,i}\boldsymbol{w}_i^{r-1,0} - \frac{\eta}{b}\sum_{\tau=0}^{T-1}\sum_{i\in\mathcal{N}_k} p_{k,i}\sum_{j\in\mathcal{B}_k^{r-1,\tau}} \nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r-1,\tau}\right).$$

Averaging over $k \in [N]$ leads to

$$\underline{\boldsymbol{w}}^{r,0} = \underline{\boldsymbol{w}}^{r-1,0} - \frac{\eta}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in\mathcal{B}_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right).$$

Using the above update in equation 35, we get

$$\begin{aligned}\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) \quad &\geq \quad \Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right) - \eta\underbrace{\left\langle \nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right), \frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in\mathcal{B}_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\rangle}_{:=\mathcal{A}_1} - \\ &\frac{\eta^2 L}{2}\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in\mathcal{B}_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|^2. \end{aligned} \tag{36}$$

The term $\mathcal{A}_1$ in equation 36 can be bounded as

$$\begin{aligned}\mathcal{A}_1 \quad &\overset{(a)}{=} \quad \frac{1}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right\|^2 + \frac{1}{2}\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in\mathcal{B}_k^{r,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|^2 \\ &- \quad \frac{1}{2}\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in B_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right) - \nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right\|^2 \\ &\overset{(b)}{\leq} \quad \frac{1}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right\|^2 + \frac{1}{2}\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in\mathcal{B}_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|^2. \end{aligned} \tag{37}$$

where (a) follows from the inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a-b\|^2$, and (b) follows from the fact that the term $\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in B_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right) - \nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1}\right)\right\|^2 > 0$. Next, using equation 37 in equation 36, we get

$$\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) \geq \Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right) - \frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right\|^2 - \frac{\eta}{2}(1+L\eta)\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in B_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|^2.$$

Using the smoothness assumption in the above, we get

$$\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) \geq \Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right) - \eta L\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right) - \frac{\eta}{2}(1+L\eta)\underbrace{\left\|\frac{1}{bN}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in B_k^{r-1,\tau}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|^2}_{:=\mathcal{A}_2}. \tag{38}$$

A part of the third term in the above can be bounded as follows

$$\begin{aligned}\mathcal{A}_2 &\overset{(a)}{\leq} \frac{T}{N}\sum_{K=1}^{N}\sum_{\tau=0}^{T-1}\frac{1}{b}\sum_{j\in B_k^{r-1,\tau}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|^2\\ &\overset{(b)}{\leq} \frac{T}{N}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\frac{1}{b}\sum_{j\in B_k^{r-1,\tau}}2l_{k,j}\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right),\\ &\overset{(c)}{\leq} \frac{Tl_{max}}{Nb}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\sum_{j\in B_k^{r-1,\tau}}2\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right),\end{aligned}$$

where (a) follows from the fact that for any vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$, $\left(\sum_{i=1}^{N}\boldsymbol{z}_i\right)^2 \leq N\sum_{i=1}^{N}(\boldsymbol{z}_i)^2$, (b) follows from smoothness assumption, and (c) follows from the fact that $l_{max} := \max_{k,j} l_{k,j}$. Next, taking the expectation

$$\mathbb{E}\left[\mathcal{A}_2\right] \leq \frac{2l_{max}T}{N}\sum_{k=1}^{N}\sum_{\tau=0}^{T-1}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right].$$

Using $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_k^2 + \frac{4L_{max}}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^{r,0}\right)\right]$ from Corollary 3, the above can be further bounded as

$$\mathbb{E}\left[\mathcal{A}_2\right] \leq \frac{2l_{max}T^2}{N}\sum_{k=1}^{N}\left(\frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r-1} - \underline{\boldsymbol{w}}^{r-1}\right\|_2^2 + \frac{4L_{max}}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right]\right).$$

Using the above result in equation 38 and rearranging, we obtain

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right)\right] \geq\ & \mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right) - \frac{\eta\left(1+\eta L\right)2l_{max}L_{max}^2T^2}{N\mu_{min}}\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r-1} - \underline{\boldsymbol{w}}^{r-1}\right\|_2^2\right.\\ &\left.-\eta\left(\frac{\left(1+\eta L\right)4l_{max}T^2L_{max}}{\mu_{min}} + L\right)\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right].\end{aligned}$$

Choosing $\eta \leq \frac{1}{L}$ and rearranging the terms, we get

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right] \leq \frac{1}{\left(1-\eta\left(\frac{8l_{max}T^2L_{max}}{\mu_{min}} + L\right)\right)}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) + \frac{\eta 4l_{max}L_{max}^2T^2}{N\mu_{min}}\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r-1} - \underline{\boldsymbol{w}}^{r-1}\right\|_2^2\right].$$

Further, choosing $\eta \leq \frac{1}{2\left(\frac{8l_{max}T^2L_{max}}{\mu_{min}}+L\right)}$, the above can be bounded as

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right] \leq 2\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) + \frac{\eta 4l_{max}L_{max}^2T^2}{N\mu_{min}}\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r-1} - \underline{\boldsymbol{w}}^{r-1}\right\|_2^2\right]. \tag{39}$$

The following bound can be obtained by using $\eta \leq \frac{N\mu_{min}}{4l_{max}L_{max}^2T^2}$ in equation 39:

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r-1,0}\right)\right] \leq 2\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,0}\right) + \sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r-1} - \underline{\boldsymbol{w}}^{r-1}\right\|_2^2\right]. \quad \square$$

Now, it suffices to bound the drift term in terms of the loss to obtain the linear convergence.

Lemma 8. *The consensus term, i.e.,* $\mathcal{D}_{r,0} := \left\|\underline{W}_l^{r,0} - \underline{W}^{r,0}\right\|_F^2$ *satisfies the following bound*

$$\mathcal{D}_{r,0} \leq \eta^2\beta L_m T^2 N\left(\sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathcal{D}_{\tau,0} + \sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)\right]\right). \tag{40}$$

where $\beta := \frac{4l\psi^2 N}{(1+\psi)\mu_{min}}$, $\lambda := \left(1+\frac{1}{\psi}\right)\lambda_2^2$, $L_m := \max\left\{L_{max}^2, 2L_{max}N\right\}$, *and* $\psi > \frac{1}{\frac{1}{\lambda_2^2}-1}$. *Here,* λ_2 *is the second largest eigenvalue of the mixing matrix* P.

Proof: Let, $\mathcal{D}_{r,0} = \mathbb{E}\left\|\underline{W}_l^{r,0} - \underline{W}^{r,0}\right\|_F^2 = \sum_{k=1}^N \mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r,0} - \underline{\boldsymbol{w}}^{r,0}\right\|^2$. Using equation 22 and equation 23, the consensus term can be written as

$$\begin{aligned}\mathcal{D}_{r,0} &= \mathbb{E}\left\|QPW^{r,0} - PW^{r,0}\right\|_F^2 \\ &= \mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2. \end{aligned} \tag{41}$$

Recall that $Q = \frac{1}{N}\mathbf{1}\mathbf{1}^T$ is the average matrix, P is the mixing matrix and $QP = Q$. Using $\underline{W}_l^{r,0} = PW^{r-1,T}$ (see equation 22), substituting for the update in $W^{r-1,T}$ and taking the telescopic sum, we get

$$W^{r,0} = \underline{W}_l^{r,0} = P\left(W^{r-1,0} - \eta\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right).$$

Plugging the above in equation 41, and using the generalized Cauchy's inequality, i.e., $\|a+b\|^2 \leq \left(1+\frac{1}{\psi}\right)\|a\|^2 + (1+\psi)\|b\|^2$ for any $\psi \geq 0$, the consensus term can be upper bounded as

$$\begin{aligned}\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\Xi + (1+\psi)\eta^2\mathbb{E}\left\|\left(Q-P^2\right)\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 \\ &\overset{(a)}{\leq} \left(1+\frac{1}{\psi}\right)\Xi + (1+\psi)\eta^2 N\left\|\left(Q-P^2\right)\right\|_{op}^2\mathbb{E}\left\|\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 \\ &\overset{(b)}{\leq} \left(1+\frac{1}{\psi}\right)\Xi + (1+\psi)\eta^2\lambda_2^4 NT\sum_{\tau=0}^{T-1}\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2, \end{aligned} \tag{42}$$

where $\Xi := \mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2$, and (a) follows from Lemma 1 and (b) follows from Lemma 2. Next, consider bounding the following

$$\begin{aligned}\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 &= \mathbb{E}\sum_{k=1}^N\left\|\frac{1}{b}\sum_{j\in B_k^{r-1,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|_2^2 \\ &\leq \mathbb{E}\sum_{k=1}^N\frac{1}{b}\sum_{j\in B_k^{r-1,\tau}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right\|_2^2 \\ &\overset{(a)}{\leq} 2l_{max}\sum_{k=1}^N\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right], \end{aligned} \tag{43}$$

where (a) follows from the smoothness assumption and $l_{max} := \max_{k,j} l_{k,j}$. Substituting the bound in equation 28 of Lemma 7, i.e., $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r-1,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^{r-1} - \underline{\boldsymbol{w}}^{r-1}\right\|_2^2 + \frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r-1}\right)\right\|^2$ in the above, and writing it in the matrix form, we get

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r-1,\tau}\right)\right\|_F^2 = \frac{4l_{max}L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{W}_l^{r-1,0} - \underline{W}^{r-1,0}\right\|_F^2 + \frac{4l_{max}}{\mu_{min}}\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2.$$

Using the above in equation 42

$$\begin{aligned}\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2 + \eta^2\lambda_2^4\alpha NT^2L_{max}^2\mathcal{D}_{r-1,0} \\ &\quad + \eta^2\lambda_2^4\alpha NT^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2, \end{aligned} \tag{44}$$

where $\alpha := \frac{4l_{max}(1+\psi)}{\mu_{min}}$. First, let us consider bounding $\mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2$. Using the update step $W^{r-1,0} = \underline{W}_l^{r-1,0} = P\left(W^{r-2,0} - \eta\sum_{\tau=0}^{T-1}\partial\hat{\Phi}\left(W^{r-2,\tau}\right)\right)$ and following a similar approach as used in steps equation 42 to equation 44, we get the following bound

$$\begin{aligned}\mathbb{E}\left\|\left(Q-P^2\right)W^{r-1,0}\right\|_F^2 &\leq \left(1+\frac{1}{\psi}\right)\mathbb{E}\left\|\left(Q-P^3\right)W^{r-2,0}\right\|_F^2 + \eta^2\lambda_2^6\alpha L_{max}^2NT^2\mathcal{D}_{r-2,0} \\ &\quad +\eta^2\lambda_2^6\alpha NT^2\mathbb{E}\left\|\partial\Phi\left(W^{r-2,0}\right)\right\|.\end{aligned}$$

Using the above result in equation 44

$$\begin{aligned}\mathcal{D}_{r,0} &\leq \left(1+\frac{1}{\psi}\right)^2\mathbb{E}\left\|\left(Q-P^3\right)W^{r-2,0}\right\|_F^2 + \left(1+\frac{1}{\psi}\right)\eta^2\lambda_2^6\alpha L_{max}^2NT^2\mathcal{D}_{r-2,0} + \\ &\quad \left(1+\frac{1}{\psi}\right)\eta^2\lambda_2^6\alpha NT^2\mathbb{E}\left\|\partial\Phi\left(W^{r-2,0}\right)\right\| + \eta^2\lambda_2^4\alpha NT^2L_{max}^2\mathcal{D}_{r-1,0} + \\ &\quad \eta^2\lambda_2^4\alpha NT^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r-1,0}\right)\right\|_F^2.\end{aligned}$$

Proceeding further in a similar manner as above, we get

$$\begin{aligned}\mathcal{D}_{r,0} &\leq \left(1+\frac{1}{\psi}\right)^r\mathbb{E}\left\|\left(Q-P^{r+1}\right)W^{0,0}\right\|_F^2 + \eta^2\alpha L_{max}^2NT^2\sum_{\tau=0}^{r-1}\lambda_2^{2(r+1-\tau)}\left(1+\frac{1}{\psi}\right)^{(r-1-\tau)}\mathcal{D}_{\tau,0} \\ &+ \eta^2\alpha NT^2\sum_{\tau=0}^{r-1}\lambda_2^{2(r+1-\tau)}\left(1+\frac{1}{\psi}\right)^{(r-1-\tau)}\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{\tau,0}\right)\right\|_F^2.\end{aligned}$$

We initialize $W^{0,0} = 0$. Further, multiplying and dividing by $\left(1+\frac{1}{\psi}\right)$ to the second and the third term in the above, we get

$$\mathcal{D}_{r,0} \leq \frac{\eta^2\psi^2\alpha L_{max}^2NT^2}{(1+\psi)^2}\sum_{\tau=0}^{r-1}\lambda^{(r+1-\tau)}\mathcal{D}_{\tau,0} + \frac{\eta^2\psi^2\alpha NT^2}{(1+\psi)^2}\sum_{\tau=0}^{r-1}\lambda^{(r+1-\tau)}\delta^{r,0}. \tag{45}$$

where $\delta^{r,0} := \mathbb{E}\left\|\partial\Phi\left(\underline{W}^{\tau,0}\right)\right\|_F^2$ and $\lambda := \left(1+\frac{1}{\psi}\right)\lambda_2^2$. Using $\alpha = \frac{4l_{max}(1+\psi)}{\mu_{min}}$ in equation 45, we have

$$\mathcal{D}_{r,0} \leq \frac{\eta^2 4l_{max}\psi^2L_{max}^2NT^2}{(1+\psi)\mu_{min}}\sum_{\tau=0}^{r-1}\lambda^{(r+1-\tau)}\mathcal{D}_{\tau,0} + \frac{\eta^2 4l_{max}\psi^2NT^2}{(1+\psi)\mu_{min}}\sum_{\tau=0}^{r-1}\lambda^{(r+1-\tau)}\delta^{r,0}. \tag{46}$$

The term, $\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{\tau,0}\right)\right\|_F^2$ in the above, is bounded as follows

$$\begin{aligned}\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{\tau,0}\right)\right\|_F^2 &= \mathbb{E}\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{\tau,0}\right)\right\|_2^2 \\ &\overset{(a)}{\leq} 2L_{max}N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)\right],\end{aligned} \tag{47}$$

where (a) follows from smoothness assumption and using the fact that $\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right) = \frac{1}{N}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^{\tau,0}\right)$, and $L_{max} = \max_k L_k$. Using equation 47 in equation 46, we get

$$\mathcal{D}_{r,0} \leq \frac{\eta^2 4l_{max}\psi^2NL_{max}^2T^2}{(1+\psi)\mu_{min}}\sum_{\tau=0}^{r-1}\lambda^{(r+1-\tau)}\mathcal{D}_{\tau,0} + \frac{\eta^2 4l_{max}\psi^2T^2N2L_{max}N}{(1+\psi)\mu_{min}}\sum_{\tau=0}^{r-1}\lambda^{(r+1-\tau)}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau}\right)\right].$$

Let $L_m := \max\left\{L_{max}^2, 2L_{max}N\right\}$ and $\beta := \frac{4l_{max}\psi^2N}{(1+\psi)\mu_{min}}$. Therefore, the drift term results in

$$\mathcal{D}_{r,0} \leq \eta^2\beta L_mT^2N\left(\sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathcal{D}_{\tau,0} + \sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)\right]\right).$$

□

A.6 COMPLETING THE PROOF OF THEOREM 2

From L-smoothness assumption (see 1) of $\Phi\left(\boldsymbol{w}\right)$, we have

$$\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right) \;\leq\; \Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) + \langle \nabla\Phi(\underline{\boldsymbol{w}}^{r,t}), \underline{\boldsymbol{w}}^{r,t+1} - \underline{\boldsymbol{w}}^{r,t}\rangle + \frac{L}{2}\left\|\underline{\boldsymbol{w}}^{r,t+1} - \underline{\boldsymbol{w}}^{r,t}\right\|^2. \tag{48}$$

Using step 7 of **Algorithm 2** we have, $\boldsymbol{w}_i^{r,t+1} = \boldsymbol{w}_i^{r,t} - \frac{\eta}{b}\sum_{j\in B_i^{r,t}} \nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,t}\right)$. Multiplying both sides by $p_{k,i}$ and summing over $i \in \mathcal{N}_k$, we get

$$\underline{\boldsymbol{w}}_k^{r,t+1} \;=\; \underline{\boldsymbol{w}}_k^{r,t} - \frac{\eta}{b}\sum_{i\in\mathcal{N}_k} p_{k,i} \sum_{j\in B_i^{r,t}} \nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,t}\right). \tag{49}$$

Averaging on both sides over $k \in [N]$, we get

$$\underline{\boldsymbol{w}}^{r,t+1} \;=\; \underline{\boldsymbol{w}}^{r,t} - \frac{\eta}{bN}\sum_{k=1}^{N}\sum_{j\in B_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right).$$

Using the above update, i.e., $\underline{\boldsymbol{w}}^{r,t+1} - \underline{\boldsymbol{w}}^{r,t}$ in equation 48, we get

$$\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right) \;\leq\; \Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \eta\left\langle \nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right), \frac{1}{bN}\sum_{k=1}^{N}\sum_{j\in B_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle + \frac{\eta^2 L}{2b^2N^2}\left\|\mathcal{G}^{r,t}\right\|^2.$$

where $\mathcal{G}^{r,t} := \sum_{k=1}^{N}\sum_{j\in B_k^{r,t}} \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)$. Taking expectation conditioning on $\boldsymbol{w}_k^{r,t}$ and past, we get

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \;\leq\; \mathbb{E}\Bigg[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \eta\underbrace{\left\langle \nabla\Phi(\underline{\boldsymbol{w}}^{r,t}), \frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle}_{:=\mathcal{A}_1} + \frac{\eta^2 L\mathcal{A}_2}{2} \\ + \underbrace{\frac{1}{b^2N^2}\sum_{k\neq k'}\left\langle \sum_{j\in B_k^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right), \sum_{i\in B_k^{r,t}}\nabla\Phi_{k',i}\left(\boldsymbol{w}_{k'}^{r,t}\right)\right\rangle}_{:=\mathcal{A}_3}\Bigg)\Bigg],\end{aligned} \tag{50}$$

where $\mathcal{A}_2 := \frac{1}{b^2N^2}\sum_{k=1}^{N}\left\|\sum_{j\in B_k^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2$. This term can be bounded as follows

$$\mathcal{A}_2 \;=\; \frac{1}{b^2N^2}\sum_{k=1}^{N}\sum_{j\in B_k^{r,t}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 + \frac{1}{b^2N^2}\sum_{k=1}^{N}\sum_{j\neq j'}\left\langle \nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right), \nabla\Phi_{k,j'}\left(\boldsymbol{w}_k^{r,t}\right)\right\rangle.$$

Taking expectation, we get

$$\mathbb{E}[\mathcal{A}_2] \;=\; \frac{1}{bN^2}\sum_{k=1}^{N}\mathbb{E}\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \frac{b(b-1)}{b^2N^2}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2, \tag{51}$$

Similarly the term $\mathcal{A}_3$ in equation 50 can be bounded by taking expectation as follows

$$\begin{aligned}\mathbb{E}[\mathcal{A}_3] \;&=\; \frac{1}{b^2N^2}\sum_{k\neq k'}\left\langle \nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right), \nabla\Phi_{k'}\left(\boldsymbol{w}_{k'}^{r,t}\right)\right\rangle \\ &\overset{(a)}{\leq}\; \frac{1}{2b^2N^2}\sum_{k\neq k'}\left[\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 + \|\nabla\Phi_{k'}\left(\boldsymbol{w}_{k'}^{r,t}\right)\|^2\right] \\ &=\; \frac{2(N-1)}{2b^2N^2}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 \\ &\leq\; \frac{1}{b^2N}\sum_{k=1}^{N}\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2,\end{aligned} \tag{52}$$

where (a) follows from $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. Next, we lower bound the term $\mathcal{A}_1$ in equation 50 as

$$\begin{aligned}\mathcal{A}_1 &= \frac{1}{2}\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 + \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 - \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 \\ &\geq \frac{1}{2}\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\|^2 + \frac{1}{2}\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\|^2 - \frac{L^2}{2N}\sum_{k=1}^{N}\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\|^2. \end{aligned} \tag{53}$$

Substituting equation 51, equation 52 and equation 53 in equation 48, we get the following

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq \mathbb{E}\Bigg[&\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 - \frac{\eta}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 + \frac{\eta L^2}{2N}\sum_{k=1}^{N}\left\|\boldsymbol{\Delta}_k^{r,t}\right\|^2 \\ &+ \frac{\eta^2 L}{2bN^2}\underbrace{\sum_{k=1}^{N}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2}_{:=\mathcal{A}_4} + \left(\frac{\eta^2 Lb(b-1)}{2b^2N^2} + \frac{\eta^2 L}{2N}\right)\mathcal{A}_5\Bigg], \end{aligned} \tag{54}$$

where $\boldsymbol{\Delta}_k^{r,t} := \boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}$ and $\mathcal{A}_5 := \sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2$. The term $\mathcal{A}_4$ in equation 54 is bounded as follows

$$\begin{aligned}\mathcal{A}_4 &\overset{(a)}{\leq} \sum_{k=1}^{N} 2\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 + \sum_{k=1}^{N} 2\left\|\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\ &\overset{(b)}{\leq} 2\sum_{k=1}^{N} l_{k,j}^2\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4\sum_{k=1}^{N} l_{k,j}\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right) \\ &\overset{(c)}{\leq} 2l_{max}^2\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4l_{max}\sum_{k=1}^{N}\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right), \end{aligned}$$

where (a) follows by adding and subtracting the term $\nabla\Phi_{k,j}\left(\underline{\boldsymbol{w}}^{r,t}\right)$ and using the fact that, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (b) follows from Assumption 3, and (c) follows from the fact that $l_{max} := \max_{k,j} l_{k,j}$. Taking expectation, we get

$$\mathbb{E}\left[\mathcal{A}_4\right] \leq 2l_{max}^2\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4l_{max}\sum_{k=1}^{N}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]. \tag{55}$$

The term $\mathcal{A}_5$ in equation 54 is bounded as

$$\begin{aligned}\mathcal{A}_5 &\overset{(a)}{\leq} 2\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right) - \nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 + 2\sum_{k=1}^{N}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 \\ &\overset{(b)}{\leq} 2\sum_{k=1}^{N} L_k^2\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4\sum_{k=1}^{N} L_k\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right) \\ &\overset{(c)}{\leq} 2L_{max}^2\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 + 4L_{max}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right), \end{aligned} \tag{56}$$

where (a) follows by adding and subtracting $\nabla\Phi_k\left(\underline{\boldsymbol{w}}^{r,t}\right)$, and (b) follows from assumption 3 and (c) follows from $L_{max} := \max_k L_k$. Substituting upper bounds from equation 55 and equation 56 in equation 67, we get

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] \leq \mathbb{E}\Bigg[&\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right) - \frac{\eta}{2}\left\|\nabla\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right\|^2 - \frac{\eta}{2}\left\|\frac{1}{N}\sum_{k=1}^{N}\nabla\Phi_k\left(\boldsymbol{w}_k^{r,t}\right)\right\|^2 \\ &+ \left(\frac{\eta L^2}{2N} + \frac{\eta^2 Ll_{max}^2}{bN^2} + \frac{\eta^2 LL_{max}^2}{N^2} + \frac{\eta^2 LL_{max}^2}{N}\right)\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t} - \underline{\boldsymbol{w}}^{r,t}\right\|^2 \\ &+ \left(\frac{2\eta^2 Ll_{max}}{bN} + \frac{2\eta^2 LL_{max}}{N} + 2\eta^2 LL_{max}\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\Bigg]. \end{aligned} \tag{57}$$

Now, using PL inequality (see definition 2), i.e., $\|\nabla\Phi(\boldsymbol{w})\|^2 \geq \mu\Phi(\boldsymbol{w}),\ \forall \boldsymbol{w} \in \mathbb{R}^d$ and rearranging, we get

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] &\leq \mathbb{E}\Bigg[\left(1-\frac{\eta\mu}{2}+\left(\frac{2\eta^2 L l_{max}}{bN}+\frac{2\eta^2 L L_{max}}{N}+2\eta^2 L L_{max}\right)\right)\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\\ &\quad+\left(\frac{\eta L^2}{2N}+\frac{\eta^2 L l_{max}^2}{bN^2}+\frac{\eta^2 L L_{max}^2}{N^2}+\frac{\eta^2 L L_{max}^2}{N}\right)\frac{1}{N}\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2\Bigg].\end{aligned}$$

Choosing $\eta \leq \min\left\{\frac{\mu}{4\left(\frac{2Ll_{max}}{bN}+\frac{2LL_{max}}{N}+2LL_{max}\right)}, \frac{L^2}{2\left(\frac{Ll_{max}^2}{bN}+\frac{LL_{max}^2}{N}+LL_{max}^2\right)}\right\}$, the above can be further bounded as

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t+1}\right)\right] &\leq \left(1-\frac{\eta\mu}{4}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]+\frac{\eta L^2}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|^2 \quad (58)\\ &\overset{(a)}{\leq} \left(1-\frac{\eta\mu}{4}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r,t}\right)\right]+\frac{2\eta L^2}{N}\sum_{k=1}^{N}\mathbb{E}\left(\left\|\underline{\boldsymbol{\Delta}}_k^{r,t}\right\|^2+\left\|\bar{\boldsymbol{\Delta}}_k^{r,t}\right\|^2\right), \quad (59)\end{aligned}$$

where $\underline{\boldsymbol{\Delta}}_k^{r,t} := \boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}$ and $\bar{\boldsymbol{\Delta}}_k^{r,t} := \underline{\boldsymbol{w}}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}$. In the above, (a) follows by adding and subtracting the term $\underline{\boldsymbol{w}}_k^{r,t}$ and using the fact that, $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$. First, let us consider the local drift term i.e., $\sum_{k=1}^{N}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|$ in equation 59. Telescoping the update from step 7 of **Algorithm** 2 we have,

$$\boldsymbol{w}_k^{r,t} = \boldsymbol{w}_k^{r,0}-\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right). \quad (60)$$

Further, consider the local average at node k, i.e., $\underline{\boldsymbol{w}}_k^{r,t}$

$$\underline{\boldsymbol{w}}_k^{r,t}=\sum_{i\in\mathcal{N}_k}p_{k,i}\boldsymbol{w}_i^{r,t}=\underline{\boldsymbol{w}}_k^{r,0}-\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\sum_{j\in\mathcal{B}_i^{r,\tau}}\nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right). \quad (61)$$

Now noting the fact that $\boldsymbol{w}_k^{r,0}=\underline{\boldsymbol{w}}_k^{r,0}$ and using equation 60 and equation 61, we can bound the drift term as

$$\begin{aligned}\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 &= \sum_{k=1}^{N}\mathbb{E}\left\|\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)-\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\sum_{j\in\mathcal{B}_i^{r,\tau}}\nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right)\right\|^2\\ &\overset{(a)}{\leq} 2\sum_{k=1}^{N}\mathbb{E}\left[\left\|\frac{\eta}{b}\sum_{\tau=0}^{t-1}G_{kj}(r,\tau)\right\|^2+\left\|\frac{\eta}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}G_{ij}(r,\tau)\right\|^2\right]\\ &\overset{(b)}{\leq} 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\|G_{kj}(r,\tau)\|^2+\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\left\|\sum_{i\in\mathcal{N}_k}p_{k,i}G_{ij}(r,\tau)\right\|^2\right],\end{aligned}$$

where $G_{ij}(r,\tau) := \sum_{j\in\mathcal{B}_i^{r,\tau}}\nabla\Phi_{i,j}\left(\boldsymbol{w}_i^{r,\tau}\right)$. In the above, (a) follows from the fact that, $\|a+b\|^2 \leq \|a\|^2+\|b\|^2$, and (b) follows from the fact that for any vector $\boldsymbol{z}_i$, $\left(\sum_{i=1}^{N}\boldsymbol{z}_i\right)^2 \leq N\sum_{i=1}^{N}(\boldsymbol{z}_i)^2$.

The second term in (b) can be further bounded using Jensen's inequality as follows

$$\begin{aligned}
\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 &\leq 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\left\|G_{kj}(r,\tau)\right\|^2+\frac{\eta^2 t}{b^2}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\left\|G_{ij}(r,\tau)\right\|^2\right]\\
&\leq 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\left\|g_{kj}^{r,\tau}\right\|^2+\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{i\in\mathcal{N}_k}p_{k,i}\sum_{j\in\mathcal{B}_i^{r,\tau}}\left\|g_{ij}^{r,\tau}\right\|^2\right]\\
&\overset{(a)}{\leq} 2\sum_{k=1}^{N}\mathbb{E}\left[\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}2l_{k,j}\mathbb{L}_{kj}^{r,\tau}+\frac{\eta^2 t}{b}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\sum_{i\in\mathcal{N}_k}p_{k,i}2l_{i,j}\mathbb{L}_{ij}^{r,\tau}\right]\\
&\overset{(b)}{=} \mathbb{E}\left[\frac{2\eta^2 t}{b}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}2l_{max}\mathbb{L}_{kj}^{r,\tau}+\frac{2\eta^2 t}{b}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\sum_{i\in\mathcal{N}_k}p_{k,i}2l_{max}\mathbb{L}_{ij}^{r,\tau}\right],
\end{aligned}$$

where $g_{kj}^{r,\tau}:=\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)$, $\mathbb{L}_{kj}^{r,\tau}:=\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)$ and (a) follows from smoothness assumption and (b) follows from the fact that mixing matrix P preserves the average and $l_{max}:=\max_{k,j}l_{k,j}$. Simplifying the above results in

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \leq \mathbb{E}\left[\frac{8\eta^2 t l_{max}}{b}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\sum_{j\in\mathcal{B}_k^{r,\tau}}\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right].$$

Taking expectation, we get

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \leq 8\eta^2 t l_{max}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right]. \tag{62}$$

According to equation 34 of Corollary 3 we have, $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{4L_{max}}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]$. Using this in equation 62, we get

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \leq 8\eta^2 t l_{max}\sum_{k=1}^{N}\sum_{\tau=0}^{t-1}\left(\frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2+\frac{4L_N}{\mu_{min}}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]\right).$$

Simplifying the above results in

$$\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{w}_k^{r,t}-\underline{\boldsymbol{w}}_k^{r,t}\right\|^2 \leq 16\eta^2 t^2 l_{max}L_{max}^2\sum_{k=1}^{N}\frac{\bar{\boldsymbol{\Delta}}^r}{\mu_{min}}+32\eta^2 t^2 l_{max}L_{max}\sum_{k=1}^{N}\frac{\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right]}{\mu_{min}}. \tag{63}$$

where $\bar{\boldsymbol{\Delta}}^r:=\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r-\underline{\boldsymbol{w}}^r\right\|_2^2$. Next, let us consider the global drift term i.e., $\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{r,t}-\underline{\boldsymbol{w}}^{r,t}\right\|_2^2$ in equation 59, which can be rewritten in matrix notation as $\mathcal{D}_{r,t}:=\left\|\underline{W}_l^{r,t}-\underline{W}^{r,t}\right\|_F^2$. This term is bounded as

$$\begin{aligned}
\mathcal{D}_{r,t} &\overset{(a)}{=} \mathbb{E}\left\|QPW^{r,t}-PW^{r,t}\right\|_F^2\\
&\overset{(b)}{=} \mathbb{E}\left\|(Q-P)W^{r,t}\right\|_F^2\\
&\overset{(c)}{=} \mathbb{E}\left\|(Q-P)\left(W^{r,0}-\eta\sum_{\tau=0}^{t-1}\partial\hat{\Phi}\left(W^{r,\tau}\right)\right)\right\|_F^2,
\end{aligned}$$

where (a) follows since $QPW^{r,t}=\underline{W}^{r,t}$ and $PW^{r,t}=\underline{W}_l^{r,t}$, (b) follows from $QP=Q$, and (c) follows from the update $W^{r,t}=W^{r,0}-\eta\sum_{\tau=0}^{t-1}\partial\hat{\Phi}\left(W^{r,\tau}\right)$. Using the fact that $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$ in the above, we get

$$\begin{aligned}
\mathcal{D}_{r,t} &\leq 2\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2+2\eta^2 t\sum_{\tau=0}^{t-1}\mathbb{E}\left\|(Q-P)\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2\\
&\leq 2\mathbb{E}\left\|(Q-P)W^{r,0}\right\|+2\eta^2 t\sum_{\tau=0}^{t-1}N\lambda_2^2\mathbb{E}\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\|_F^2,
\end{aligned} \tag{64}$$

The term $\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2$ in the above can be bounded as

$$\begin{aligned}\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2 &= \mathbb{E}\sum_{k=1}^{N}\left\|\frac{1}{b}\sum_{j\in B_k^{r,t}}\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right\|_2^2 \\ &\leq \mathbb{E}\sum_{k=1}^{N}\frac{1}{b}\sum_{j\in B_k^{r,t}}\left\|\nabla\Phi_{k,j}\left(\boldsymbol{w}_k^{r,\tau}\right)\right\|_2^2 \\ &\overset{(a)}{\leq} 2l_{max}\sum_{k=1}^{N}\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right],\end{aligned}$$

where (a) follows from the smoothness assumption and the fact that $l_{max} := \max_{k,j} l_{k,j}$. Using equation 28 of Lemma 6, i.e., $\mathbb{E}\left[\Phi_k\left(\boldsymbol{w}_k^{r,\tau}\right)\right] \leq \frac{2L_{max}^2}{\mu_{min}}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_k^2 + \frac{2}{\mu_{min}}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2$ in the above, we get

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2 \leq \frac{4L_{max}^2 l_{max}}{\mu_{min}}\sum_{k=1}^{N}\mathbb{E}\left\|\underline{\boldsymbol{w}}_k^r - \underline{\boldsymbol{w}}^r\right\|_2^2 + \frac{4l_{max}}{\mu_{min}}\sum_{k=1}^{N}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2.$$

The result above can be written in matrix form as,

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(W^{r,\tau}\right)\right\|_F^2 = \frac{4L_{max}^2 l_{max}}{\mu_{min}}\mathcal{D}_{r,0} + \frac{4l_{max}}{\mu_{min}}\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2,$$

Substituting the above result in equation 64, we get

$$\mathcal{D}_{r,t} \leq 2\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 + 4\eta^2 L_{max}^2\lambda_2^2\gamma t^2\mathcal{D}_{r,0} + 4\eta^2\lambda_2^2\gamma t^2\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2, \quad (65)$$

where $\gamma := \frac{2l_{max}N}{\mu_{min}}$.

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] &\leq \left(1-\frac{\eta\mu}{4}\right)^T\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right] + \frac{2\eta L^2}{N}\sum_{\tau=0}^{T-1}\left(1-\frac{\eta\mu}{4}\right)^\tau\sum_{k=1}^{N}\mathbb{E}\left(\left\|\underline{\boldsymbol{\Delta}}_k^{r,T-1-\tau}\right\|^2 + \left\|\bar{\boldsymbol{\Delta}}_k^{r,T-1-\tau}\right\|^2\right) \\ &\overset{(a)}{\leq} \left(1-\frac{\eta\mu}{4}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right] + \frac{2\eta L^2}{N}\sum_{\tau=0}^{T-2}\left(1-\frac{\eta\mu}{4}\right)^\tau\sum_{k=1}^{N}\mathbb{E}\left(\left\|\underline{\boldsymbol{\Delta}}_k^{r,T-1-\tau}\right\|^2 + \left\|\bar{\boldsymbol{\Delta}}_k^{r,T-1-\tau}\right\|^2\right) \quad (66)\end{aligned}$$

where (a) follows from the fact that $\left(1-\frac{\eta\mu}{4}\right)^T \leq \left(1-\frac{\eta\mu}{4}\right)$, $\left\|\underline{\boldsymbol{\Delta}}_k^{r,T-1-\tau}\right\|^2 = 0$ and $\left\|\bar{\boldsymbol{\Delta}}_k^{r,T-1-\tau}\right\|^2 = 0$ for $\tau = T-1$. Now choosing $\eta < \frac{4}{\mu}$ and substituting equation 63 and equation 65 in equation 66, we get

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq &\mathbb{E}\left[\left(1-\frac{\eta\mu}{4}+\frac{64\eta^4T^3lLL_{max}}{\mu_{min}}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right) + \frac{2\eta^2TL}{N}\left[\left(\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}} + 4\lambda_2^2\eta^2\gamma L_{max}^2T^2\right)\mathcal{D}_{r,0}\right.\right. \\ &\left.\left. + 2\left\|(Q-P)W^{r,0}\right\|_F^2 + 4\eta^2\gamma T^2\lambda_2^2\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2\right]\right]. \quad (67)\end{aligned}$$

The term $\mathbb{E}\left\|\partial\Phi\left(\underline{W}^{r,0}\right)\right\|_F^2$ can be bounded as

$$\mathbb{E}\left\|\partial\hat{\Phi}\left(\underline{W}^{r,0}\right)\right\|_F^2 = \sum_{k=1}^{N}\mathbb{E}\left\|\nabla\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right\|^2 \overset{(a)}{\leq} \sum_{k=1}^{N}2L_{max}\mathbb{E}\left[\Phi_k\left(\underline{\boldsymbol{w}}^r\right)\right] = 2L_{max}N\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^r\right)\right],$$

where (a) follows from smoothness assumption and (b) follows from the fact that $\Phi\left(\underline{\boldsymbol{w}}^{r}\right) = \frac{1}{N}\sum_{k=1}^{N}\Phi_k\left(\underline{\boldsymbol{w}}^{r}\right)$. Using the above result in equation 67, we get

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] &\leq \mathbb{E}\left[\left(1-\frac{\eta\mu}{4}+\frac{64\eta^4T^3l_{max}LL_{max}}{\mu_{min}}\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right)+\frac{2\eta^2TL}{N}\left[2\left\|(Q-P)W^{r,0}\right\|_F^2+\right.\right.\\
&\quad\left.\left.\left(\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}}+4\lambda_2^2\eta^2\gamma L^2T^2\right)\mathcal{D}_{r,0}+8\eta^2\lambda_2^2\gamma T^2L_{max}N\Phi\left(\underline{\boldsymbol{w}}^{r}\right)\right]\right]\\
&\leq \mathbb{E}\left[\left(1-\frac{\eta\mu}{4}+\frac{64\eta^4T^3l_{max}LL_{max}}{\mu_{min}}+16\eta^4\gamma T^3\lambda_2^2LL_{max}\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right)+\right.\\
&\quad\left.\frac{2\eta^2TL}{N}\left(\left[\frac{16l_{max}\eta^2T^2L_{max}^2}{\mu_{min}}+4\lambda_2^2\eta^2\gamma L^2T^2\right]\mathcal{D}_{r,0}+2\left\|(Q-P)W^{r,0}\right\|_F^2\right)\right].
\end{aligned}$$

Choosing $\eta \leq \frac{1}{8}\left(\frac{\mu}{\frac{64T^3l_{max}LL_{max}}{\mu_{min}}+16\gamma T^3L\lambda_2^2L_{max}}\right)^{1/3}$ in the above result in

$$\begin{aligned}
\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] &\leq \mathbb{E}\left[\left(1-\frac{\eta\mu}{8}\right)\Phi\left(\underline{\boldsymbol{w}}^{r}\right)+\frac{2\eta^4TL}{N}\left[\frac{16T^2L_{max}^2l_{max}}{\mu_{min}}+4\lambda_2^2\gamma L^2T^2\right]\mathcal{D}_{r,0}\right.\\
&\quad\left.+\frac{4\eta^2L}{N}\left\|(Q-P)W^{r,0}\right\|_F^2\right].
\end{aligned}$$

Again choosing $\eta \leq \left[\left(\frac{1}{\frac{16T^2l_{max}L_{max}^2}{\mu_{min}}+4\lambda_2^2\gamma T^2L^2}\right)\right]^{\frac{1}{2}}$, the above results in

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{8}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r}\right)\right]+\frac{2\eta^2TL}{N}\mathcal{D}_{r,0}+\frac{4\eta^2TL}{N}\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2.$$

It is easy to see that $\mathbb{E}\left\|(Q-P)W^{r,0}\right\|_F^2 = \mathbb{E}\left\|\underline{W}_l^{r,0}-\underline{W}^{r,0}\right\|_F^2 = \mathcal{D}_{r,0}$. Using this above, gives us

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)\right] \leq \left(1-\frac{\eta\mu}{8}\right)\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{r}\right)\right]+\frac{6\eta^2TL}{N}\mathcal{D}_{r,0}. \tag{68}$$

From Lemma 8, we have

$$\mathcal{D}_{r,0} \leq \eta^2\beta L_mT^2N\left(\sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathcal{D}_{\tau,0}+\sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)\right)\right]. \tag{69}$$

From Lemma 7, we know that

$$\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)\right] \leq 2\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau+1,0}\right)+\sum_{k=1}^{N}\left\|\underline{\boldsymbol{w}}_k^{\tau}-\underline{\boldsymbol{w}}^{\tau}\right\|_2^2\right].$$

Using the above result on $\Phi\left(\underline{\boldsymbol{w}}^{\tau,0}\right)$ in equation 69, we get

$$\mathcal{D}_{r,0} \leq 3\eta^2\beta L_mT^2N\sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathcal{D}_{\tau,0}+2\eta^2\beta L_mT^2N\sum_{\tau=0}^{r-1}\lambda^{r+1-\tau}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau+1,0}\right)\right].$$

Let $L_m = \max\{2L_m, 3L_m\}$. The above can be further bounded as

$$\mathcal{D}_{r,0} \leq \eta^2\beta T^2NL_m\left(\sum_{\tau=0}^{r-1}\lambda^{r-\tau}D_{\tau,0}+\sum_{\tau=0}^{r-1}\lambda^{r-\tau}\mathbb{E}\left[\Phi\left(\underline{\boldsymbol{w}}^{\tau+1,0}\right)\right]\right). \tag{70}$$

This completes the proof. □

A.6.1 PROOF OF PROPOSITION 4

Note that we need to prove the following set of inequalities hold good for all r

$$\mathcal{D}_{r,0} \leq (2r+3)\eta^2\beta T^2 L_m N\lambda^2\Lambda^r\Phi\left(\underline{\boldsymbol{w}}^0\right) \tag{71}$$

$$\Phi\left(\underline{\boldsymbol{w}}^r\right) \leq \Lambda^{r-1}\Big(\Lambda + 4\eta^4 L L_m\beta T^3\lambda^2 r^2\Big)\Phi\left(\underline{\boldsymbol{w}}^0\right), r = \{1,2,\ldots,R\} \tag{72}$$

where $\lambda = \left(1+\frac{1}{\psi}\right)\lambda_2^2$, $\Phi\left(\underline{\boldsymbol{w}}^0\right) = \Phi\left(\underline{\boldsymbol{w}}^{0,0}\right)$ and $\Lambda = \max\left(\left(1-\frac{\eta\mu}{8}\right),\lambda\right)$. We use induction method to prove that the above set of inequalities hold good for all r. Since $\mathcal{D}_{0,0} = 0$, the inequalities hold good for $r = 0$. Next, assuming that the above inequalities hold good for every communication rounds in $\{1,2,\ldots,r\}$, we need to prove that the respective inequalities hold for $\mathcal{D}_{r+1,0}$ and $\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)$. Towards this, consider the following

$$\begin{aligned}
\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right) &\leq \left(1-\frac{\eta\mu}{8}\right)\Phi\left(\underline{\boldsymbol{w}}^r\right) + \frac{4\eta^2 LT}{N}\mathcal{D}_{r,0} \\
&\overset{(a)}{\leq} \Lambda\left[\Lambda + 4\eta^4 L L_m\beta T^3\lambda^2 r^2\right]\Lambda^{r-1}\Phi\left(\underline{\boldsymbol{w}}^0\right) + 4\eta^4\beta T^3 L_m L(2r+1)\lambda^2\Lambda^r\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&= \left[\Lambda + 4\eta^4 L L_m\beta T^3\lambda^2 r^2\right]\Lambda^r\Phi\left(\underline{\boldsymbol{w}}^0\right) + 4\eta^4\beta T^3 L_m L(2r+1)\lambda^2\Lambda^r\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&= \left[\Lambda^{r+1} + 4\eta^4 L L_m\beta T^3\lambda^2\Lambda^r\left(r^2+2r+1\right)\right]\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&= \Lambda^r\left[\Lambda + 4\eta^4 L L_m\beta T^3\lambda^2\left(r+1\right)^2\right]\Phi\left(\underline{\boldsymbol{w}}^0\right),
\end{aligned}$$

where (a) follows by substituting equation 71, equation 72 and using $\Lambda := \left(1-\frac{\eta\mu}{8}\right)$. Let us recall from equation 40 of Lemma 8 that

$$\mathcal{D}_{r+1,0} \leq \eta^2\beta T^2 N L_m\Big(\sum_{\tau=0}^{r}\lambda^{r+2-\tau}\mathcal{D}_{\tau,0} + \sum_{\tau=0}^{r}\lambda^{r+2-\tau}\Phi\left(\underline{\boldsymbol{w}}^{\tau+1,0}\right)\Big). \tag{73}$$

Substituting for $\mathcal{D}_{\tau,0}$ from equation 71 in the first term of equation 73, we get

$$\begin{aligned}
\sum_{\tau=0}^{r}\lambda^{r+2-\tau}\mathcal{D}_{\tau,0} &= \eta^2\beta T^2 L_m N\sum_{\tau=0}^{r}\lambda^{r+2-\tau}(2\tau+1)\lambda^2\Lambda^\tau\Phi\left(\underline{\boldsymbol{w}}^0\right). \\
&\leq \eta^2\beta T^2 L_m N\lambda^2\lambda\sum_{\tau=0}^{r}(2\tau+1)\lambda^{r-\tau}\Lambda^{\tau+1}\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&\overset{(a)}{\leq} \eta^2\beta T^2 L_m N\lambda^2\lambda\Lambda^{r+1}\left(\sum_{\tau=0}^{r}2\tau+1\right)\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&\leq \eta^2\beta T^2 L_m N\lambda^2\lambda\Lambda^{r+1}r(r+1)\Phi\left(\underline{\boldsymbol{w}}^0\right),
\end{aligned}$$

where (a) follows from the fact that $\lambda \leq \Lambda$. Now picking $\eta^2 \leq \frac{1}{\beta T^2 L_m N R\lambda}$ results in

$$\sum_{\tau=0}^{r}\lambda^{r+2-\tau}\mathcal{D}_{\tau,0} \leq (r+1)\lambda^2\Lambda^{r+1}\Phi\left(\underline{\boldsymbol{w}}^0\right). \tag{74}$$

Next, substituting for $\Phi\left(\underline{\boldsymbol{w}}^{\tau+1}\right)$ from equation 72 in the second term of equation 73, we get

$$\begin{aligned}
\sum_{\tau=0}^{r}\lambda^{r+2-\tau}\Phi\left(\underline{\boldsymbol{w}}^{\tau+1}\right) &= \sum_{\tau=0}^{r}\lambda^{r+2-\tau}\left[\Lambda + 4\eta^4 L L_m\beta T^3\lambda^2(\tau+1)^2\right]\Lambda^\tau\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&\leq \sum_{\tau=0}^{r}\lambda^{r+2-\tau}\Lambda^{\tau+1} + (r+1)^2\sum_{\tau=0}^{r}\lambda^2\lambda^{r+\tau}\Lambda^\tau 4\eta^4 L L_m\beta T^3\lambda\Phi\left(\boldsymbol{w}^0\right).
\end{aligned}$$

The last inequality follows from the fact that $\tau \leq r$, and $\lambda \leq \Lambda$. By choosing $\eta^4 \leq \frac{1}{4LL_m\beta T^3(r+1)^3\lambda}$, we get

$$\begin{aligned}
\sum_{\tau=0}^{r}\lambda^{r+2-\tau}\Phi\left(\underline{\boldsymbol{w}}^{\tau+1}\right) &\leq \left[\sum_{\tau=0}^{r}\lambda^2\Lambda^{r+1} + \sum_{\tau=0}^{r}\frac{1}{(r+1)}\lambda^2\Lambda^{r+1}\right]\Phi\left(\underline{\boldsymbol{w}}^0\right) \\
&= (r+2)\lambda^2\Lambda^{r+1}\Phi\left(\underline{\boldsymbol{w}}^0\right).
\end{aligned} \tag{75}$$

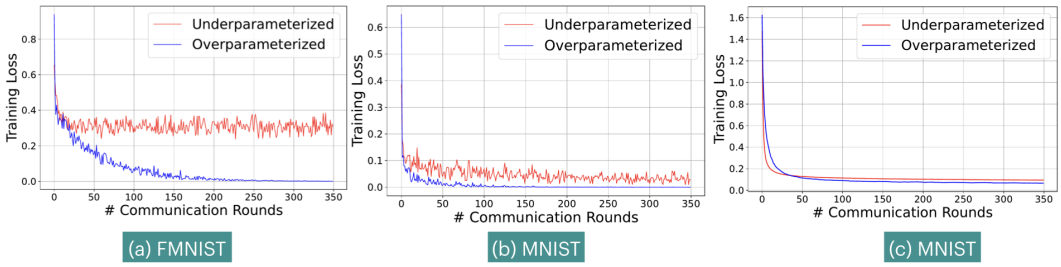

Figure 5: Training loss for *server* FedAvg (see (a) FMNIST and (b) MNIST) and *decentralized* FedAvg (see (c) MNIST) versus communication rounds.

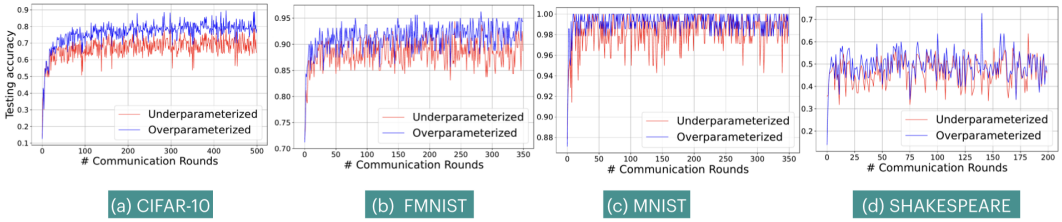

Figure 6: Testing accuracy on different datasets versus the communication rounds for FedAvg in the *Server* setting.

Using equation 74 and equation 75 in equation 73, and after some algebraic manipulations, we get the following desired result

$$\mathcal{D}_{r+1,0} \quad \leq \quad (2r+3)\eta^2\beta T^2 L_m N \lambda^2 \Lambda^{r+1} \Phi\left(\underline{\boldsymbol{w}}^0\right).$$

Using the above result in the upper bound for $\Phi\left(\underline{\boldsymbol{w}}^{r+1}\right)$, we get the desired bound on . □

A.7 Additional Experiments

In this section, we provide the details of the experimental setup and some additional results for experiments carried on different datasets for both *Server* and *Decentralized* setting. We have used NVIDIA $DGX\ A100$ to implement all our experiments. The experimental setup consists of the following model and data set:

Overparameterized regression: We consider a model with 3 linear layers and no activation function with 231490 trainable parameters. Note that this formulation models a simple regression problem. We condsider a image classification task on MNIST dataset and evaluate the performance of FedAvg under different settings.

Deep neural network: In this case, we consider an image classification task on CIFAR–10 dataset. Each edge device implements a three hidden layer convolutional neural network (CNN) followed by two linear layers with 1046426 trainable model parameters. In the overparameterized setting, for the

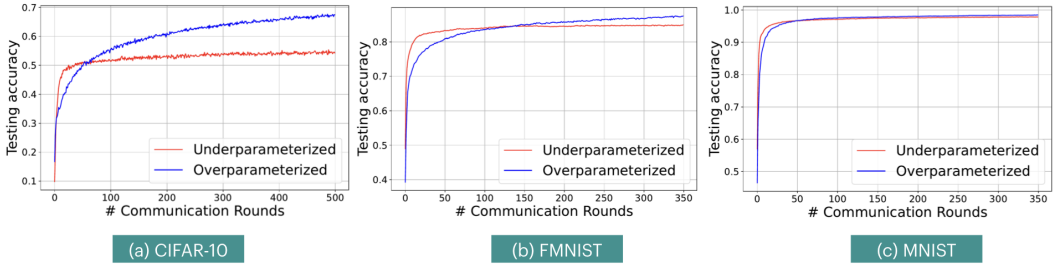

Figure 7: Testing accuracy on different datasets versus the communication rounds for FedAvg in the *Decentralized* setting.

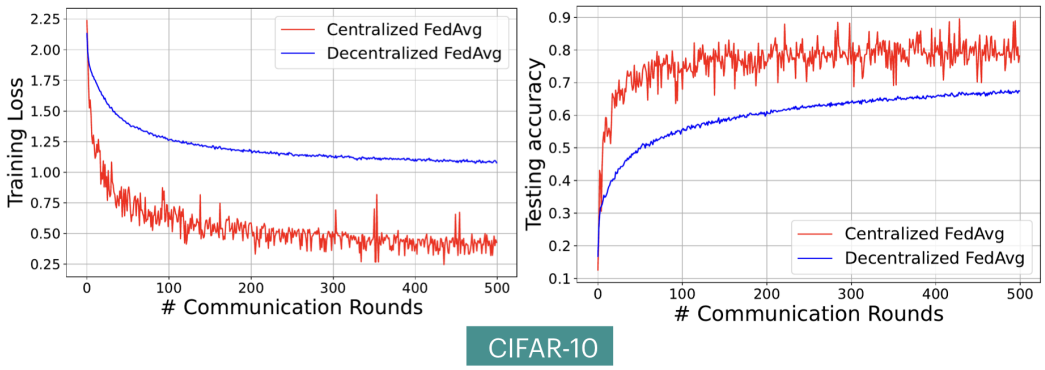

Figure 8: Training loss and Testing accuracy for centralized ($\lambda_2 = 0$) and *decentralized* FedAvg algorithm with ring topology ($\lambda_2 = 0.33$) on CIFAR-10 dataset versus communication rounds.

CIFAR-10, MNIST and FMNIST, each edge device implements a three hidden layer convolutional neural network (CNN) with 256, 128 and 64 filters followed by three linear layers having 1642849 trainable parameters for CIFAR-10 and two linear layers for MNIST and FMNIST with 1046426 trainable parameters. For Shakespeare dataset, LSTM models are used at each edge device. We consider an embedding layer with embedding size of 10 followed by 2 LSTM layers with 256 hidden neurons and one linear layer. On the other hand, in the underparameterized setting, we consider a comparatively smaller neural network. For the CIFAR-10, MNIST, FMNIST datasets each device implements two hidden layer CNN network with 25 and 52 filters followed by two linear layers for CIFAR-10 and one linear layer for MNIST and FMNIST datasets. For the Shakespeare dataset each device has embedding layer followed by one LSTM layer with 56 hidden neurons and a linear layer. For the experiments, we chose $T = 10$ and tune for the learning rate in the range $\eta \in [0.001 : 0.01]$ for CIFAR-10, MNIST, FMNIST datasets whereas we choose $\eta = 0.8$ for the Shakespeare dataset. Each device has access to 490 training samples and 90 test samples for CIFAR-10 whereas for MNIST and FMNIST datasets, 540 samples are used for training and 80 samples are used for testing.

Figure 5 show the training loss on FMNIST and MNIST dataset for *server* and *decentralized* FedAvg settings. Figure 6 show the testing accuracy for FedAvg in the server setting for four different datasets. As expected the convergence speed of underparameterized case is slower than the overparameterized case. Similarly, figure 7 show plots for testing accuracy for FedAvg in the *decentralized* setting.

Finally, in Figure 8, we compare the training loss and testing accuracy for *server* and *decentralized* FedAvg algorithm against the communication rounds for classification task on CIFAR-10 dataset. It is clear from the figures that the centralized case achieves a very good performance at a faster rate as opposed to the *decentralized* case, i.e., the ring topology.