

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, 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 Konečný, 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 Agüera 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 Quanguan 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 $\|\cdot\|_2$ and the Frobenius norm by $\|\cdot\|_F$. Also $\langle \cdot, \cdot \rangle$ 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 $\Phi(\mathbf{w})$, the gradient is denoted by $\nabla\Phi(\mathbf{w})$, and the Hessian is denoted by $\nabla^2\Phi(\mathbf{w})$. We use $\mathbf{1}$ to represent a column vector with all ones.

A.2 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.

A.3 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. 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 $\mathcal{O}(1/N\epsilon^2)$. 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 non-convex objectives. 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. Similarly, the authors in (Woodworth et al., 2020b) extended their analyses of (Woodworth et al., 2020a) to the heterogeneous data settings. 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; 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 as well. 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 conver-

gence 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.

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.4 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.5 PROOF OF THEOREM 1

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

A.5.1 USEFUL LEMMAS TO PROVE 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.5.2. 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 $\frac{1}{N} \sum_{k=1}^N \mathbb{E} \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2$ is bounded in terms of local weight i.e., $\Phi_k(\mathbf{w}_k^{r,\tau})$ as follows

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\mathbb{E} [\Phi_k(\mathbf{w}_k^{r,\tau})]$ is bounded in terms of global average weight i.e., $\Phi_k(\underline{\mathbf{w}}^r)$ as follows

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

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

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

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

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

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

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

where $\mathcal{F}_k^{r,\tau-1} := \left\langle \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}), \nabla \Phi_{k,j'}(\mathbf{w}_k^{r,\tau-1}) \right\rangle$, 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

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

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

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

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

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

It is easy to see that the above implies

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

□

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

A.5.2 COMPLETING THE PROOF OF THEOREM 1

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

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

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

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

where $\mathcal{G}^{r,t} := \frac{1}{bN} \sum_{k=1}^N \sum_{j \in \mathcal{B}^{r,t}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,t})$. Taking the expectation conditioning on $\mathbf{w}_k^{r,t}$, we get²

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \mathbb{E} [\Phi(\underline{\mathbf{w}}^{r,t+1})] &\leq \mathbb{E} \left[\Phi(\underline{\mathbf{w}}^{r,t}) - \frac{\eta}{2} \|\nabla \Phi(\underline{\mathbf{w}}^{r,t})\|^2 - \frac{\eta}{2} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) \right\|^2 \right. \\ &\quad \left. + \frac{\eta L^2}{2N} \sum_{k=1}^N \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2 + \frac{\eta^2 L}{2bN^2} \sum_{k=1}^N \|\nabla \Phi_{k,j}(\mathbf{w}_k^{r,t})\|^2 \right. \\ &\quad \left. + \left(\frac{\eta^2 Lb(b-1)}{2b^2 N^2} + \frac{\eta^2 L}{2N} \right) \sum_{k=1}^N \|\nabla \Phi_k(\mathbf{w}_k^{r,t})\|^2 \right]. \end{aligned} \quad (14)$$

$:= \mathcal{A}_4$
 $:= \mathcal{A}_5$

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

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

where (a) follows by adding and subtracting $\nabla \Phi_{k,j}(\underline{\mathbf{w}}^{r,t})$ 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

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

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

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

In the above, (a) follows by adding and subtracting $\nabla \Phi_k(\underline{\mathbf{w}}^{r,t})$ 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

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

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

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

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

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

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

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

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

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

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

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

It follows from the update step that $\frac{1}{N} \sum_{k=1}^N \|\mathbf{w}_k^{r,T-1-\tau} - \mathbf{w}^{r,T-1-\tau}\|^2 = 0$ for $\tau = T-1$. Using

$$\frac{1}{N} \sum_{k=1}^N \mathbb{E} \|\mathbf{w}_k^{r,t} - \mathbf{w}^{r,t}\|^2 \leq \eta^2 t^2 \left[\frac{2l_{max}}{b} + \frac{2b(b-1)L_{max}}{b^2} \right] \Phi(\mathbf{w}^r) \text{ in the above results in}$$

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

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

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

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

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

A.6 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.6.1 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 $\underline{W}_l^r := [\underline{w}_1^r, \underline{w}_2^r, \dots, \underline{w}_N^r]^T \in \mathbb{R}^{N \times d}$, where $\underline{w}_k^r \in \mathbb{R}^d$. The Aggregation step of Algorithm 2 can be compactly written in matrix form as

$$\underline{w}_k^{r+1} = \sum_{i \in \mathcal{N}_k} p_{k,i} \underline{w}_i^{r,T} \equiv \underline{W}_l^{r+1} = P \underline{W}_l^r, \quad (22)$$

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

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

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

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

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

We start by proving an upper bound on the average loss $\mathbb{E}[\Phi(\underline{w}^{r+1})]$ in terms of the loss $\Phi(\underline{w}^r)$ 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*

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

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.7. \square

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 $\mathcal{D}_{r,0} \leq \text{constant} \times \Phi(\underline{w}^r)$, 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 $\Phi(\underline{w}^{\tau,0})$ as follows*

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

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

Proof: The proof is provided in Appendix A.7. \square

Next, our task is to show that the recursion in equation 26 satisfies a bound of the form $\mathcal{D}_{r,0} \leq \text{constant}^r \times \Phi(\underline{w}^0)$, 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

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

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

Proof: The proof is provided in Appendix A.7.1. \square

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.6.2 USEFUL LEMMAS TO PROVE THEOREM 2

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

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

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

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

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

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

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

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

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

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

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

where $l_{max} := \max_k L_k$. From the local PL inequality (see definition 2), it follows that $\left\| \nabla \Phi_k \left(\mathbf{w}_k^{r,\tau-1} \right) \right\|_2^2 \geq \mu_{min} \Phi_k \left(\mathbf{w}_k^{r,\tau-1} \right)$ for $k = \{1, 2, \dots, N\}$, where $\mu_{min} := \min_{k \in [N]} \{\mu_k\}$. Using this in equation 30 results in

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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., $\left\| \nabla \Phi_k (\underline{\mathbf{w}}^r) \right\|_2^2 \leq 2L_{max} \Phi_k (\underline{\mathbf{w}}^r)$. 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 $\Phi(\underline{\mathbf{w}}^{r-1,0})$ is bounded in terms of the future value of the function as given below

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

Proof: It follows from the smoothness assumption that

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

Telescoping the update in step 7 of **Algorithm 2**, we get

$\mathbf{w}_i^{r-1,T} = \mathbf{w}_i^{r-1,0} - \frac{\eta}{b} \sum_{\tau=0}^{T-1} \sum_{j \in \mathcal{B}_k^{r-1,\tau}} \nabla \Phi_{i,j}(\mathbf{w}_i^{r-1,\tau})$. Averaging over all neighboring nodes $i \in \mathcal{N}_k$, we get

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

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

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

Using the above update in equation 35, we get

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

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

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

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 $\left\| \frac{1}{bN} \sum_{k=1}^N \sum_{\tau=0}^{T-1} \sum_{j \in \mathcal{B}_k^{r-1,\tau}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r-1,\tau}) - \nabla \Phi(\underline{\mathbf{w}}^{r-1,0}) \right\|^2 > 0$. Next, using equation 37 in equation 36, we get

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

Using the smoothness assumption in the above, we get

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

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

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

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

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

Using $\mathbb{E}[\Phi_k(\mathbf{w}_k^{r,\tau})] \leq \frac{2L_{max}^2}{\mu_{min}} \mathbb{E}\|\underline{\mathbf{w}}_k^r - \underline{\mathbf{w}}^r\|_2^2 + \frac{4L_{max}}{\mu_{min}} \mathbb{E}[\Phi_k(\underline{\mathbf{w}}^{r,0})]$ from Corollary 3, the above can be further bounded as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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., $\mathbb{E} [\Phi_k(\underline{w}_k^{r-1,\tau})] \leq \frac{2L_{\max}^2}{\mu_{\min}} \mathbb{E} \left\| \underline{w}_k^{r-1} - \underline{w}^{r-1} \right\|_2^2 + \frac{2}{\mu_{\min}} \mathbb{E} \left\| \nabla \Phi_k(\underline{w}^{r-1}) \right\|^2$ in the above, and writing it in the matrix form, we get

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

Using the above in equation 42

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

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

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

Using the above result in equation 44

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

Proceeding further in a similar manner as above, we get

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

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

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

where $\delta^{r,0} := \mathbb{E} \|\partial \Phi(\underline{W}^{\tau,0})\|_F^2$ 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

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

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

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

where (a) follows from smoothness assumption and using the fact that $\Phi(\underline{\mathbf{w}}^{\tau,0}) = \frac{1}{N} \sum_{k=1}^N \Phi_k(\underline{\mathbf{w}}^{\tau,0})$, and $L_{max} = \max_k L_k$. Using equation 65 in equation 46, we get

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

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

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

□

A.7 COMPLETING THE PROOF OF THEOREM 2

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

$$\Phi(\underline{\mathbf{w}}^{r,t+1}) \leq \Phi(\underline{\mathbf{w}}^{r,t}) + \langle \nabla \Phi(\underline{\mathbf{w}}^{r,t}), \underline{\mathbf{w}}^{r,t+1} - \underline{\mathbf{w}}^{r,t} \rangle + \frac{L}{2} \|\underline{\mathbf{w}}^{r,t+1} - \underline{\mathbf{w}}^{r,t}\|^2. \quad (48)$$

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

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

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

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

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

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

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

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

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

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

Taking expectation, we get

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

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

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

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

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

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

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

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

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

where (a) follows by adding and subtracting the term $\nabla \Phi_{k,j}(\underline{\mathbf{w}}^{r,t})$ 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

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

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

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

where (a) follows by adding and subtracting $\nabla \Phi_k(\underline{\mathbf{w}}^{r,t})$, 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 68, we get

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

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

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

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

$$\mathbb{E}[\Phi(\underline{\mathbf{w}}^{r,t+1})] \leq \left(1 - \frac{\eta\mu}{4}\right) \mathbb{E}[\Phi(\underline{\mathbf{w}}^{r,t})] + \frac{\eta L^2}{N} \sum_{k=1}^N \mathbb{E}\|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2 \quad (58)$$

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

where $\underline{\Delta}_k^{r,t} := \mathbf{w}_k^{r,t} - \underline{\mathbf{w}}_k^{r,t}$ and $\bar{\Delta}_k^{r,t} := \underline{\mathbf{w}}_k^{r,t} - \underline{\mathbf{w}}^{r,t}$. In the above, (a) follows by adding and subtracting the term $\underline{\mathbf{w}}_k^{r,t}$ 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., $\sum_{k=1}^N \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}_k^{r,t}\|^2$ in equation 59. Telescoping the update from step 7 of **Algorithm 2** we have,

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

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

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

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

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

where $G_{ij}(r, \tau) := \sum_{j \in \mathcal{B}_i^{r,\tau}} \nabla\Phi_{i,j}(\mathbf{w}_i^{r,\tau})$. 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 \mathbf{z}_i , $\left(\sum_{i=1}^N \mathbf{z}_i\right)^2 \leq N \sum_{i=1}^N (\mathbf{z}_i)^2$.

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

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

where $g_{kj}^{r,\tau} := \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau})$, $\mathbb{L}_{kj}^{r,\tau} := \Phi_{k,j}(\mathbf{w}_k^{r,\tau})$ 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

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

Taking expectation, we get

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

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

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

Simplifying the above results in

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

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

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

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 $W^{r,t} = W^{r,0} - \eta \sum_{\tau=0}^{t-1} \partial \hat{\Phi}(W^{r,\tau})$. Using the fact that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ in the above, we get

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

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

$$\begin{aligned} \mathbb{E} \left\| \partial \hat{\Phi} (W^{r,\tau}) \right\|_F^2 &= \mathbb{E} \sum_{k=1}^N \left\| \frac{1}{b} \sum_{j \in B_k^{r,t}} \nabla \Phi_{k,j} (\mathbf{w}_k^{r,\tau}) \right\|_2^2 \\ &\leq \mathbb{E} \sum_{k=1}^N \frac{1}{b} \sum_{j \in B_k^{r,t}} \|\nabla \Phi_{k,j} (\mathbf{w}_k^{r,\tau})\|_2^2 \\ &\stackrel{(a)}{\leq} 2l_{max} \sum_{k=1}^N \mathbb{E} [\Phi_k (\mathbf{w}_k^{r,\tau})], \end{aligned} \quad (65)$$

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., $\mathbb{E} [\Phi_k (\mathbf{w}_k^{r,\tau})] \leq \frac{2L_{max}^2}{\mu_{min}} \mathbb{E} \|\underline{\mathbf{w}}_k^r - \underline{\mathbf{w}}^r\|_2^2 + \frac{2}{\mu_{min}} \mathbb{E} \|\nabla \Phi_k (\underline{\mathbf{w}}^r)\|^2$ in the above, we get

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

The result above can be written in matrix form as,

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

Substituting the above result in equation 64, we get

$$\mathcal{D}_{r,t} \leq 2\mathbb{E} \|(Q - P)W^{r,0}\|_F^2 + 4\eta^2 L_{max}^2 \lambda_2^2 \gamma t^2 \mathcal{D}_{r,0} + 4\eta^2 \lambda_2^2 \gamma t^2 \mathbb{E} \|\partial \Phi (\underline{W}^{r,0})\|_F^2, \quad (66)$$

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

$$\begin{aligned} \mathbb{E} [\Phi (\underline{\mathbf{w}}^{r+1})] &\leq \left(1 - \frac{\eta\mu}{4}\right)^T \mathbb{E} [\Phi (\underline{\mathbf{w}}^r)] + \frac{2\eta L^2}{N} \sum_{\tau=0}^{T-1} \left(1 - \frac{\eta\mu}{4}\right)^\tau \sum_{k=1}^N \mathbb{E} \left(\left\| \underline{\Delta}_k^{r,T-1-\tau} \right\|^2 + \left\| \bar{\Delta}_k^{r,T-1-\tau} \right\|^2 \right) \\ &\stackrel{(a)}{\leq} \left(1 - \frac{\eta\mu}{4}\right) \mathbb{E} [\Phi (\underline{\mathbf{w}}^r)] + \frac{2\eta L^2}{N} \sum_{\tau=0}^{T-2} \left(1 - \frac{\eta\mu}{4}\right)^\tau \sum_{k=1}^N \mathbb{E} \left(\left\| \underline{\Delta}_k^{r,T-1-\tau} \right\|^2 + \left\| \bar{\Delta}_k^{r,T-1-\tau} \right\|^2 \right) \end{aligned} \quad (67)$$

where (a) follows from the fact that $\left(1 - \frac{\eta\mu}{4}\right)^T \leq \left(1 - \frac{\eta\mu}{4}\right)$, $\left\| \underline{\Delta}_k^{r,T-1-\tau} \right\|^2 = 0$ and $\left\| \bar{\Delta}_k^{r,T-1-\tau} \right\|^2 = 0$ for $\tau = T - 1$. Now choosing $\eta < \frac{4}{\mu}$ and substituting equation 63 and equation 66 in equation 67, we get

$$\begin{aligned} \mathbb{E} [\Phi (\underline{\mathbf{w}}^{r+1})] &\leq \mathbb{E} \left[\left(1 - \frac{\eta\mu}{4} + \frac{64\eta^4 T^3 l L_{max}}{\mu_{min}}\right) \Phi (\underline{\mathbf{w}}^r) + \frac{2\eta^2 T L}{N} \left[\left(\frac{16l_{max} \eta^2 T^2 L_{max}^2}{\mu_{min}} + 4\lambda_2^2 \eta^2 \gamma L_{max}^2 T^2 \right) \mathcal{D}_{r,0} \right. \right. \\ &\quad \left. \left. + 2 \|(Q - P)W^{r,0}\|_F^2 + 4\eta^2 \gamma T^2 \lambda_2^2 \|\partial \Phi (\underline{W}^{r,0})\|_F^2 \right] \right]. \end{aligned} \quad (68)$$

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

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

where (a) follows from smoothness assumption and (b) follows from the fact that $\Phi(\underline{\mathbf{w}}^r) = \frac{1}{N} \sum_{k=1}^N \Phi_k(\underline{\mathbf{w}}^r)$. Using the above result in equation 68, we get

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

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

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

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

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

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

$$\mathbb{E}[\Phi(\underline{\mathbf{w}}^{r+1})] \leq \left(1 - \frac{\eta\mu}{8}\right) \mathbb{E}[\Phi(\underline{\mathbf{w}}^r)] + \frac{6\eta^2 T L}{N} \mathcal{D}_{r,0}. \quad (69)$$

From Lemma 8, we have

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

From Lemma 7, we know that

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

Using the above result on $\Phi(\underline{\mathbf{w}}^{\tau,0})$ in equation 70, we get

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

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

$$\mathcal{D}_{r,0} \leq \eta^2 \beta T^2 N L_m \left(\sum_{\tau=0}^{r-1} \lambda^{r-\tau} \mathcal{D}_{\tau,0} + \sum_{\tau=0}^{r-1} \lambda^{r-\tau} \mathbb{E}[\Phi(\underline{\mathbf{w}}^{\tau+1,0})]\right). \quad (71)$$

This completes the proof. \square

A.7.1 PROOF OF PROPOSITION 5

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

$$\mathcal{D}_{r,0} \leq (2r+3)\eta^2\beta T^2 L_m N \lambda^2 \Lambda^r \Phi(\underline{\mathbf{w}}^0) \quad (72)$$

$$\Phi(\underline{\mathbf{w}}^r) \leq \Lambda^{r-1} \left(\Lambda + 4\eta^4 L L_m \beta T^3 \lambda^2 r^2 \right) \Phi(\underline{\mathbf{w}}^0), r = \{1, 2, \dots, R\} \quad (73)$$

where $\lambda = \left(1 + \frac{1}{\psi}\right) \lambda_2^2$, $\Phi(\underline{\mathbf{w}}^0) = \Phi(\underline{\mathbf{w}}^{0,0})$ 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 $\{1, 2, \dots, r\}$, we need to prove that the respective inequalities hold for $\mathcal{D}_{r+1,0}$ and $\Phi(\underline{\mathbf{w}}^{r+1})$. Towards this, consider the following

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

where (a) follows by substituting equation 72, equation 73 and using $\Lambda := \left(1 - \frac{\eta\mu}{8}\right)$. Let us recall from equation 40 of Lemma 8 that

$$\mathcal{D}_{r+1,0} \leq \eta^2 \beta T^2 N L_m \left(\sum_{\tau=0}^r \lambda^{r+2-\tau} \mathcal{D}_{\tau,0} + \sum_{\tau=0}^r \lambda^{r+2-\tau} \Phi(\underline{\mathbf{w}}^{\tau+1,0}) \right). \quad (74)$$

Substituting for $\mathcal{D}_{\tau,0}$ from equation 72 in the first term of equation 74, we get

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

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

$$\sum_{\tau=0}^r \lambda^{r+2-\tau} \mathcal{D}_{\tau,0} \leq (r+1) \lambda^2 \Lambda^{r+1} \Phi(\underline{\mathbf{w}}^0). \quad (75)$$

Next, substituting for $\Phi(\underline{\mathbf{w}}^{\tau+1})$ from equation 73 in the second term of equation 74, we get

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

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

$$\begin{aligned} \sum_{\tau=0}^r \lambda^{r+2-\tau} \Phi(\underline{\mathbf{w}}^{\tau+1}) &\leq \left[\sum_{\tau=0}^r \lambda^2 \Lambda^{r+1} + \sum_{\tau=0}^r \frac{1}{(r+1)} \lambda^2 \Lambda^{r+1} \right] \Phi(\underline{\mathbf{w}}^0) \\ &= (r+2) \lambda^2 \Lambda^{r+1} \Phi(\underline{\mathbf{w}}^0). \end{aligned} \quad (76)$$

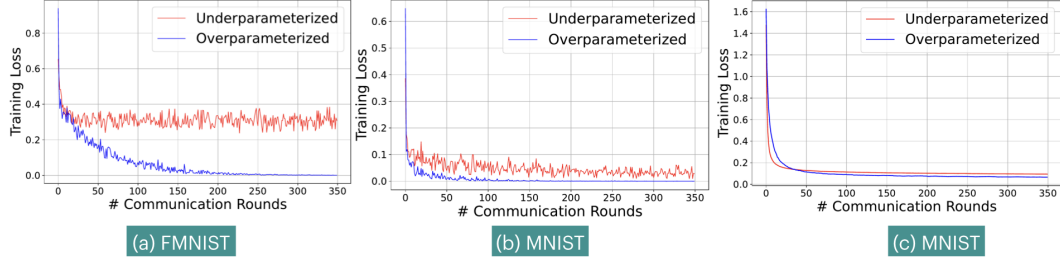


Figure 5: Training loss for *server* FedAvg (see (a) FMNIST and (b) MNIST) and *decentralized* FedAvg (see (c) MNIST) versus communication rounds.

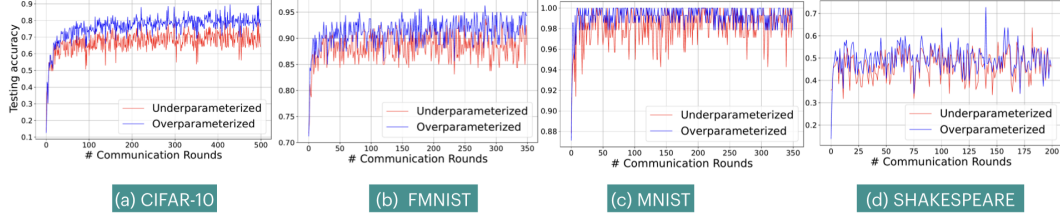


Figure 6: Testing accuracy on different datasets versus the communication rounds for FedAvg in the *Server* setting.

Using equation 75 and equation 76 in equation 74, and after some algebraic manipulations, we get the following desired result

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

Using the above result in the upper bound for $\Phi(\mathbf{w}^{r+1})$, we get the desired bound on . \square

A.8 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 consider 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

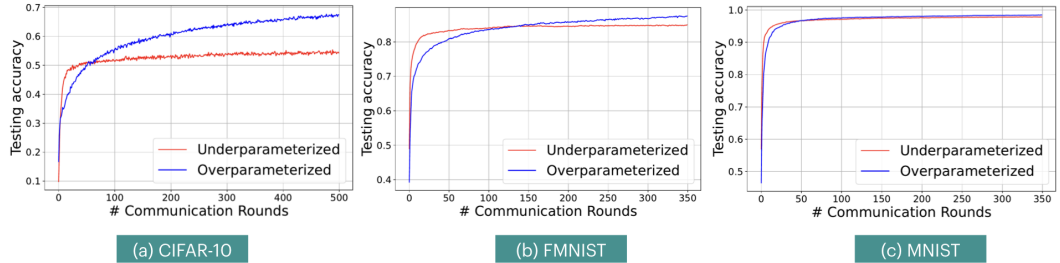


Figure 7: Testing accuracy on different datasets versus the communication rounds for FedAvg in the *Decentralized* setting.

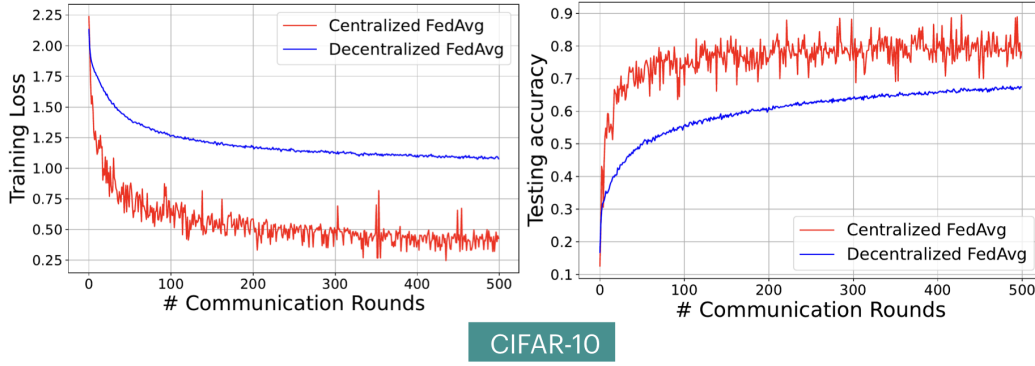


Figure 8: Training loss and Testing accuracy for centralized ($\lambda_2 = 0$) and decentralized FedAvg algorithm with ring topology ($\lambda_2 = 0.33$) on CIFAR-10 dataset versus communication rounds.

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 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 centralized and decentralized FedAvg algorithm against the communications 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.