

A Appendix

A.1 Compact Notations and Useful Lemmas

We simplify the presentation of the proof by using the following matrix notations. Let the local average of the parameters 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 $\underline{w}_k^r \in \mathbb{R}^d$ is the parameter vector at node k . The **Aggregation** step of Algorithm 1 can be compactly written in matrix form as

$$\underline{w}_k^{r+1} = \sum_{i \in \mathcal{N}_k} p_{k,i} \underline{w}_k^{r,T} \quad \equiv \quad \underline{W}_l^{r+1} = P \underline{W}_l^{r,T}, \quad (7)$$

where $\mathcal{N}_k := \{i : p_{k,i} > 0\}$, and the symbol \equiv means “equivalent to”. Further, we define the global average as

$$\underline{w}^r := \frac{1}{N} \sum_{k=1}^N \underline{w}_k^r \quad \equiv \quad \underline{W}^r = Q \underline{W}_l^r, \quad (8)$$

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

$$\partial \hat{\Phi}(\underline{W}^{r,t}) := \left[\frac{1}{b} \sum_{j \in B_1^{r,t}} G_{1,j}^{(r,t)}, \frac{1}{b} \sum_{j \in B_2^{r,t}} G_{2,j}^{(r,t)}, \dots, \frac{1}{b} \sum_{j \in B_N^{r,t}} G_{N,j}^{(r,t)} \right], \quad (9)$$

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 $QP = P$. In the following subsection, we provide a Lemma that relates the local average with the drift. Next, we present two Lemmas that will be used in proving the convergence result of *Decentralized FedAvg* algorithm, in particular, while bounding the global drift.

Lemma 4. (See (Horn & Johnson, 2012)) For any matrices $A \in \mathbb{C}^{N \times N}$ and $B \in \mathbb{C}^{N \times d}$, we have $\|AB\|_F^2 \leq \|A\|_{op}^2 \|B\|_F^2$, where $\|A\|_{op}$ denotes the operator norm of A .

Lemma 5. (See Lemma 1 in (Sun et al., 2021)) Suppose Assumption 5 holds, then 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 $Q := \frac{1}{N} \mathbf{1}\mathbf{1}^T$.

The above Lemmas are standard which come in handy while bounding the consensus error (Koloskova et al., 2020; Liu et al., 2022b; Wang & Joshi, 2021). See Liu et al. (2022b); Wang & Joshi (2021); Sun et al. (2021) for the detailed proofs.

B Proof of Theorem 1

In this section, we will prove the main theorem by proving Lemmas 1 and 2. The proof mainly consists of two intermediate steps, namely bounding i) the local loss (see Lemma 1) using L_k smoothness (see Definition 1) and local PL inequality to show that the loss at local parameter is bounded in terms of the loss at the global average parameter and the drift and ii) the global drift (see Lemma 2).

B.1 Useful Lemma to Prove Theorem 1

Lemma 1. The expected local loss function $\Phi_k(\underline{w}_k^{r,\tau})$ satisfies the following bound

$$\mathbb{E}[\Phi_k(\underline{w}_k^{r,\tau})] \leq \frac{2L_{max}^2}{\mu_{min}} \mathbb{E}\|\underline{w}_k^r - \underline{w}^r\|_2^2 + \frac{2}{\mu_{min}} \mathbb{E}\|\nabla \Phi_k(\underline{w}^r)\|^2, \quad (10)$$

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

Proof: Using Assumption 1, we have

$$\Phi_k(\mathbf{w}_k^{r,\tau}) \leq \Phi_k(\mathbf{w}_k^{r,\tau-1}) + \left\langle \nabla \Phi_k(\mathbf{w}_k^{r,\tau-1}), \mathbf{w}_k^{r,\tau} - \mathbf{w}_k^{r,\tau-1} \right\rangle + \frac{L_k}{2} \|\mathbf{w}_k^{r,\tau} - \mathbf{w}_k^{r,\tau-1}\|_2^2. \quad (11)$$

We know from Step 7 of the **Algorithm 1** that $\mathbf{w}_k^{r,\tau} - \mathbf{w}_k^{r,\tau-1} = -\frac{\eta}{b} \sum_{j \in B_k^{r,\tau-1}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1})$. Using this in equation 11, we get

$$\begin{aligned} \Phi_k(\mathbf{w}_k^{r,\tau}) &\leq \Phi_k(\mathbf{w}_k^{r,\tau-1}) - \eta \left\langle \nabla \Phi_k(\mathbf{w}_k^{r,\tau-1}), \frac{1}{b} \sum_{j \in B_k^{r,\tau-1}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}) \right\rangle + \frac{\eta^2 L_k}{2} G_k(r, \tau). \\ &= \Phi_k(\mathbf{w}_k^{r,\tau-1}) - \eta \left\langle \nabla \Phi_k(\mathbf{w}_k^{r,\tau-1}), \frac{1}{b} \sum_{j \in B_k^{r,\tau-1}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}) \right\rangle \\ &\quad + \frac{\eta^2 L_k}{2b^2} \sum_{j \in B_k^{r,\tau-1}} \|\nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1})\|_2^2 + \frac{\eta^2 L_k}{2b^2} \sum_{j \neq j'} \left\langle \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}), \nabla \Phi_{k,j'}(\mathbf{w}_k^{r,\tau-1}) \right\rangle. \\ &\leq \Phi_k(\mathbf{w}_k^{r,\tau-1}) - \eta \left\langle \nabla \Phi_k(\mathbf{w}_k^{r,\tau-1}), \frac{1}{b} \sum_{j \in B_k^{r,\tau-1}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}) \right\rangle \\ &\quad + \frac{\eta^2 L_{max}}{2b^2} \sum_{j \in B_k^{r,\tau-1}} \|\nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1})\|_2^2 + \frac{\eta^2 L_{max}}{2b^2} \sum_{j \neq j'} \left\langle \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}), \nabla \Phi_{k,j'}(\mathbf{w}_k^{r,\tau-1}) \right\rangle \end{aligned} \quad (12)$$

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 of the above leads to

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

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

$$\begin{aligned} \mathbb{E}[\Phi_k(\mathbf{w}_k^{r,\tau})] &\leq \mathbb{E} \left[\Phi_k(\mathbf{w}_k^{r,\tau-1}) - \eta \|\nabla \Phi_k(\mathbf{w}_k^{r,\tau-1})\|_2^2 + \frac{\eta^2 L_{max} l_{k,j}}{b} \Phi_{k,j}(\mathbf{w}_k^{r,\tau-1}) \right. \\ &\quad \left. + \frac{\eta^2 L_{max} b(b-1) L_k}{b^2} [\Phi_k(\mathbf{w}_k^{r,\tau-1})] \right]. \\ &\leq \Phi_k(\mathbf{w}_k^{r,\tau-1}) - \eta \|\nabla \Phi_k(\mathbf{w}_k^{r,\tau-1})\|_2^2 + \frac{\eta^2 L_{max} l_{max}}{b} \mathbb{E}[\Phi_{k,j}(\mathbf{w}_k^{r,\tau-1})] \\ &\quad + \frac{\eta^2 L_{max}^2 b(b-1)}{b^2} [\Phi_k(\mathbf{w}_k^{r,\tau-1})], \end{aligned} \quad (13)$$

where $l_{max} := \max_k L_k$. From the local PL inequality (see definition 2), it follows that $\|\nabla \Phi_k(\mathbf{w}_k^{r,\tau-1})\|_2^2 \geq \mu_{min} \Phi_k(\mathbf{w}_k^{r,\tau-1})$ for $k = \{1, 2, \dots, N\}$, where $\mu_{min} := \min_{k \in [N]} \{\mu_k\}$. Using this in equation 13 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

$$\mathbb{E} [\Phi_k(\mathbf{w}_k^{r,\tau})] \leq \left(1 - \frac{\eta\mu_{\min}}{2}\right)^\tau \mathbb{E} [\Phi_k(\underline{\mathbf{w}}_k^r)]. \quad (14)$$

Using the local PL inequality, i.e., $\Phi_k(\underline{\mathbf{w}}_k^r) \leq \frac{1}{\mu_{\min}} \|\nabla \Phi_k(\underline{\mathbf{w}}_k^r)\|_2^2$ in equation 14, we have

$$\mathbb{E} [\Phi_k(\mathbf{w}_k^{r,\tau})] \leq \left(1 - \frac{\eta\mu_{\min}}{2}\right)^\tau \frac{1}{\mu_{\min}} \mathbb{E} \|\nabla \Phi_k(\underline{\mathbf{w}}_k^r)\|_2^2. \quad (15)$$

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

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

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

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

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

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

□

Using smoothness condition, the above leads to the following corollary. The below result comes in handy while proving the main result.

Corollary 1. *The function $\Phi_k(\mathbf{w}_k^{r,\tau})$ satisfies local PL inequality and can be bounded in terms of global average parameter i.e., $\Phi_k(\underline{\mathbf{w}}^r)$ as follows*

$$\mathbb{E} [\Phi_k(\mathbf{w}_k^{r,\tau})] \leq \frac{2L_{\max}^2}{\mu_{\min}} \mathbb{E} \|\underline{\mathbf{w}}_k^r - \underline{\mathbf{w}}^r\|_2^2 + \frac{4L_{\max}}{\mu_{\min}} \mathbb{E} [\Phi_k(\underline{\mathbf{w}}^r)], \quad (16)$$

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

B.2 Proof of Lemma 1

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

$$\Phi(\underline{\mathbf{w}}^{r,t+1}) \leq \Phi(\underline{\mathbf{w}}^{r,t}) + \langle \nabla \Phi(\underline{\mathbf{w}}^{r,t}), \underline{\mathbf{w}}^{r,t+1} - \underline{\mathbf{w}}^{r,t} \rangle + \frac{L}{2} \|\underline{\mathbf{w}}^{r,t+1} - \underline{\mathbf{w}}^{r,t}\|^2. \quad (17)$$

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

$$\underline{\mathbf{w}}_k^{r,t+1} = \underline{\mathbf{w}}_k^{r,t} - \frac{\eta}{b} \sum_{i \in \mathcal{N}_k} p_{k,i} \sum_{j \in B_i^{r,t}} \nabla \Phi_{i,j}(\mathbf{w}_i^{r,t}). \quad (18)$$

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

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

Substituting for $\underline{\mathbf{w}}^{r,t+1} - \underline{\mathbf{w}}^{r,t}$ from the above update in equation 17, 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

$$\begin{aligned} \mathbb{E}[\Phi(\underline{\mathbf{w}}^{r,t+1})] &\leq \mathbb{E} \left[\underbrace{\Phi(\underline{\mathbf{w}}^{r,t}) - \eta \left\langle \nabla \Phi(\underline{\mathbf{w}}^{r,t}), \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) \right\rangle}_{:=\mathcal{A}_1} + \underbrace{\frac{\eta^2 L}{2} \left(\frac{1}{b^2 N^2} \sum_{k=1}^N \left\| \sum_{j \in B_k^{r,t}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,t}) \right\|^2 \right)}_{:=\mathcal{A}_2} \right. \\ &\quad \left. + \underbrace{\frac{1}{b^2 N^2} \sum_{k \neq k'} \left\langle \sum_{j \in B_k^{r,t}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,t}), \sum_{i \in B_{k'}^{r,t}} \nabla \Phi_{k',i}(\mathbf{w}_{k'}^{r,t}) \right\rangle}_{:=\mathcal{A}_3} \right], \end{aligned} \quad (19)$$

First, consider the second term above, i.e., \mathcal{A}_2

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

Taking expectation, we get

$$\mathbb{E}[\mathcal{A}_2] = \frac{1}{bN^2} \sum_{k=1}^N \mathbb{E} \|\nabla \Phi_{k,j}(\mathbf{w}_k^{r,t})\|^2 + \frac{b(b-1)}{b^2 N^2} \sum_{k=1}^N \|\nabla \Phi_k(\mathbf{w}_k^{r,t})\|^2. \quad (20)$$

Similarly the term \mathcal{A}_3 in equation 19 can be bounded by taking expectation as follows

$$\begin{aligned} \mathbb{E}[\mathcal{A}_3] &= \frac{1}{N^2} \sum_{k \neq k'} \left\langle \nabla \Phi_k(\mathbf{w}_k^{r,t}), \nabla \Phi_{k'}(\mathbf{w}_{k'}^{r,t}) \right\rangle \\ &\stackrel{(a)}{\leq} \frac{1}{2N^2} \sum_{k \neq k'} \left[\|\nabla \Phi_k(\mathbf{w}_k^{r,t})\|^2 + \|\nabla \Phi_{k'}(\mathbf{w}_{k'}^{r,t})\|^2 \right] \\ &= \frac{2(N-1)}{2N^2} \sum_{k=1}^N \|\nabla \Phi_k(\mathbf{w}_k^{r,t})\|^2 \\ &\leq \frac{1}{N} \sum_{k=1}^N \|\nabla \Phi_k(\mathbf{w}_k^{r,t})\|^2, \end{aligned} \quad (21)$$

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 19 as

$$\begin{aligned} \mathcal{A}_1 &= \frac{1}{2} \|\nabla \Phi(\underline{\mathbf{w}}^{r,t})\|^2 + \frac{1}{2} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) \right\|^2 - \frac{1}{2} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) - \nabla \Phi(\underline{\mathbf{w}}^{r,t}) \right\|^2 \\ &\stackrel{\text{Jensen} + \text{smoothness}}{\geq} \frac{1}{2} \|\nabla \Phi(\underline{\mathbf{w}}^{r,t})\|^2 + \frac{1}{2} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) \right\|^2 - \frac{L^2}{2N} \sum_{k=1}^N \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2. \end{aligned} \quad (22)$$

Substituting equation 20, equation 21 and equation 22 in equation 17, we get the following

$$\begin{aligned} \mathbb{E}[\Phi(\underline{\mathbf{w}}^{r,t+1})] &\leq \mathbb{E} \left[\Phi(\underline{\mathbf{w}}^{r,t}) - \frac{\eta}{2} \|\nabla \Phi(\underline{\mathbf{w}}^{r,t})\|^2 - \frac{\eta}{2} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) \right\|^2 + \frac{\eta L^2}{2N} \sum_{k=1}^N \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2 \right. \\ &\quad \left. + \underbrace{\frac{\eta^2 L}{2bN^2} \sum_{k=1}^N \|\nabla \Phi_{k,j}(\mathbf{w}_k^{r,t})\|^2}_{:=\mathcal{A}_4} + \left(\frac{\eta^2 L b(b-1)}{2b^2 N^2} + \frac{\eta^2 L}{2N} \right) \underbrace{\sum_{k=1}^N \|\nabla \Phi_k(\mathbf{w}_k^{r,t})\|^2}_{:=\mathcal{A}_5} \right]. \end{aligned} \quad (23)$$

The term \mathcal{A}_4 in equation 23 is bounded as follows

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

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

$$\mathbb{E}[\mathcal{A}_4] \leq 2l_{max}^2 \sum_{k=1}^N \mathbb{E} \left\| \mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t} \right\|^2 + 4l_{max} \sum_{k=1}^N \mathbb{E}[\Phi_k(\underline{\mathbf{w}}^{r,t})]. \quad (24)$$

The term \mathcal{A}_5 in equation 23 is bounded as

$$\begin{aligned}
\mathcal{A}_5 &\stackrel{(a)}{\leq} 2 \sum_{k=1}^N \left\| \nabla \Phi_k(\mathbf{w}_k^{r,t}) - \nabla \Phi_k(\underline{\mathbf{w}}^{r,t}) \right\|^2 + 2 \sum_{k=1}^N \left\| \nabla \Phi_k(\underline{\mathbf{w}}^{r,t}) \right\|^2 \\
&\stackrel{(b)}{\leq} 2 \sum_{k=1}^N L_k^2 \left\| \mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t} \right\|^2 + 4 \sum_{k=1}^N L_k \Phi_k(\underline{\mathbf{w}}^{r,t}) \\
&\stackrel{(c)}{\leq} 2L_{max}^2 \sum_{k=1}^N \left\| \mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t} \right\|^2 + 4L_{max} \sum_{k=1}^N \Phi_k(\underline{\mathbf{w}}^{r,t}),
\end{aligned} \quad (25)$$

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 24 and equation 25 in equation 23, we get

$$\begin{aligned}
\mathbb{E}[\Phi(\underline{\mathbf{w}}^{r,t+1})] &\leq \mathbb{E} \left[\Phi(\underline{\mathbf{w}}^{r,t}) - \frac{\eta}{2} \left\| \nabla \Phi(\underline{\mathbf{w}}^{r,t}) \right\|^2 - \frac{\eta}{2} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \Phi_k(\mathbf{w}_k^{r,t}) \right\|^2 \right. \\
&\quad + \left(\frac{\eta L^2}{2N} + \frac{\eta^2 L l_{max}^2}{b N^2} + \frac{\eta^2 L L_{max}^2}{N^2} + \frac{\eta^2 L L_{max}^2}{N} \right) \sum_{k=1}^N \left\| \mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t} \right\|^2 \\
&\quad \left. + \left(\frac{2\eta^2 L l_{max}}{b N} + \frac{2\eta^2 L L_{max}}{N} + 2\eta^2 L L_{max} \right) \Phi(\underline{\mathbf{w}}^{r,t}) \right].
\end{aligned} \quad (26)$$

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

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

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

$$\mathbb{E} [\Phi(\underline{\mathbf{w}}^{r,t+1})] \leq \left(1 - \frac{\eta\mu}{4}\right) \mathbb{E} [\Phi(\underline{\mathbf{w}}^{r,t})] + \frac{\eta L^2}{N} \sum_{k=1}^N \mathbb{E} \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2 \quad (27)$$

$$\stackrel{(a)}{\leq} \left(1 - \frac{\eta\mu}{4}\right) \mathbb{E} [\Phi(\underline{\mathbf{w}}^{r,t})] + \frac{2\eta L^2}{N} \sum_{k=1}^N \mathbb{E} \left(\|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}_k^{r,t}\|^2 + \|\underline{\mathbf{w}}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|^2 \right), \quad (28)$$

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 28. Telescoping the update from step 7 of **Algorithm 1** we have,

$$\mathbf{w}_k^{r,t} = \mathbf{w}_k^{r,0} - \frac{\eta}{b} \sum_{\tau=0}^{t-1} \sum_{j \in \mathcal{B}_k^{r,\tau}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau}). \quad (29)$$

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

$$\underline{\mathbf{w}}_k^{r,t} = \sum_{i \in \mathcal{N}_k} p_{k,i} \mathbf{w}_i^{r,t} = \underline{\mathbf{w}}_k^{r,0} - \frac{\eta}{b} \sum_{\tau=0}^{t-1} \sum_{i \in \mathcal{N}_k} p_{k,i} \sum_{j \in \mathcal{B}_i^{r,\tau}} \nabla \Phi_{i,j}(\mathbf{w}_i^{r,\tau}). \quad (30)$$

Now noting the fact that $\mathbf{w}_k^{r,0} = \underline{\mathbf{w}}_k^{r,0}$ and using equation 29 and equation 30, 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, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, 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

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

where $g_{k,j}^{r,\tau} := \nabla \Phi_{k,j}(\mathbf{w}_k^{r,\tau})$. In the above, (a) follows from smoothness assumption and (b) follows from the fact that the 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

$$\sum_{k=1}^N \mathbb{E} \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}_k^{r,t}\|^2 \leq 8\eta^2 t l_{max} \sum_{k=1}^N \sum_{\tau=0}^{t-1} \mathbb{E} [\Phi_k(\mathbf{w}_k^{r,\tau})]. \quad (31)$$

From equation 16 of Corollary 1, we have, $\mathbb{E} [\Phi_k(\mathbf{w}_k^{r,\tau})] \leq \frac{2L_{max}^2}{\mu_{min}} \mathbb{E} \|\underline{\mathbf{w}}_k^r - \underline{\mathbf{w}}^r\|_2^2 + \frac{4L_{max}}{\mu_{min}} \mathbb{E} [\Phi_k(\underline{\mathbf{w}}^r)]$. Using this in equation 31, we get

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

Simplifying the above results in

$$\sum_{k=1}^N \mathbb{E} \|\mathbf{w}_k^{r,t} - \underline{\mathbf{w}}_k^{r,t}\|^2 \leq 16\eta^2 t^2 l_{max} L_{max}^2 \sum_{k=1}^N \frac{\mathbb{E} \|\underline{\mathbf{w}}_k^r - \underline{\mathbf{w}}^r\|_2^2}{\mu_{min}} + 32\eta^2 t^2 l_{max} L_{max} \sum_{k=1}^N \frac{\mathbb{E} [\Phi_k(\underline{\mathbf{w}}^r)]}{\mu_{min}}. \quad (32)$$

Next, let us consider the global drift term i.e., $\sum_{k=1}^N \|\underline{\mathbf{w}}_k^{r,t} - \underline{\mathbf{w}}^{r,t}\|_2^2$ in equation 28, 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

$$\begin{aligned} \mathcal{D}_{r,t} &\leq 2\mathbb{E} \|(Q - P)W^{r,0}\|_F^2 + 2\eta^2 t \sum_{\tau=0}^{t-1} \mathbb{E} \|(Q - P)\partial \hat{\Phi}(W^{r,\tau})\|_F^2 \\ &\leq 2\mathbb{E} \|(Q - P)W^{r,0}\|_F^2 + 2\eta^2 t \sum_{\tau=0}^{t-1} \lambda_2^2 \mathbb{E} \|\partial \hat{\Phi}(W^{r,\tau})\|_F^2. \end{aligned} \quad (33)$$

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

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

where (a) follows from the smoothness assumption and the fact that $l_{max} := \max_{k,j} l_{k,j}$. Using equation 10 of Lemma 1, i.e., $\mathbb{E}[\Phi_k(\mathbf{w}_k^{r,\tau})] \leq \frac{2L_{max}^2}{\mu_{min}} \mathbb{E}\|\mathbf{w}_k^r - \underline{\mathbf{w}}^r\|_2^2 + \frac{2}{\mu_{min}} \mathbb{E}\|\nabla\Phi_k(\underline{\mathbf{w}}^r)\|^2$ in the above, we get

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

The result above can be written in the matrix form as,

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

Substituting the above result in equation 33, we get

$$\mathcal{D}_{r,t} \leq 2\mathbb{E}\|(Q-P)W^{r,0}\|_F^2 + 4\eta^2 L_{max}^2 \lambda_2^2 \gamma t^2 \mathcal{D}_{r,0} + 4\eta^2 \lambda_2^2 \gamma t^2 \mathbb{E}\left\|\partial\Phi(W^{r,0})\right\|_F^2, \quad (34)$$

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

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

where (a) follows from the fact that $\left\|\mathbf{w}_k^{r,T-1-\tau} - \underline{\mathbf{w}}_k^{r,T-1-\tau}\right\|^2 = 0$ and $\left\|\underline{\mathbf{w}}_k^{r,T-1-\tau} - \underline{\mathbf{w}}^{r,T-1-\tau}\right\|^2 = 0$ for $\tau = T-1$. Now choosing $\eta < \frac{4}{\mu}$ and using equation 32 and equation 34, the average loss becomes

$$\begin{aligned} \mathbb{E}[\Phi(\underline{\mathbf{w}}^{r+1})] &\leq \mathbb{E}\left[\left(\left(1 - \frac{\eta\mu}{4}\right)^T + \frac{64\eta^3 T^3 l_{max} L^2 L_{max}}{\mu_{min}}\right) \Phi(\underline{\mathbf{w}}^r) + \frac{4\eta T L^2}{N} \|(Q-P)W^{r,0}\|_F^2 + \right. \\ &\quad \left. \frac{2\eta T L^2}{N} \left[\left(\frac{16l_{max}\eta^2 T^2 L_{max}^2}{\mu_{min}} + 4\lambda_2^2 \eta^2 \gamma L_{max}^2 T^2\right) \mathcal{D}_{r,0} + 4\eta^2 \gamma T^2 \lambda_2^2 \mathbb{E}\left\|\partial\Phi(W^{r,0})\right\|_F^2\right]\right]. \quad (35) \end{aligned}$$

Using the fact that $\left(1 - \frac{\eta\mu}{4}\right)^T \leq \left(1 - \frac{\eta\mu}{4}\right)$, the above can be further bounded as

$$\begin{aligned} \mathbb{E}[\Phi(\underline{\mathbf{w}}^{r+1})] &\leq \mathbb{E}\left[\left(1 - \frac{\eta\mu}{4} + \frac{64\eta^3 T^3 l_{max} L^2 L_{max}}{\mu_{min}}\right) \Phi(\underline{\mathbf{w}}^r) + \frac{2\eta T L^2}{N} \left[2\|(Q-P)W^{r,0}\|_F^2 + \right. \right. \\ &\quad \left. \left. \left(\frac{16l_{max}\eta^2 T^2 L_{max}^2}{\mu_{min}} + 4\lambda_2^2 \eta^2 \gamma L_{max}^2 T^2\right) \mathcal{D}_{r,0} + 4\eta^2 \gamma T^2 \lambda_2^2 \mathbb{E}\left\|\partial\Phi(W^{r,0})\right\|_F^2\right]\right]. \quad (36) \end{aligned}$$

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

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

where (a) follows from the 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 36, we get

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

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

$$\begin{aligned} \mathbb{E} [\Phi(\underline{\mathbf{w}}^{r+1})] &\leq \mathbb{E} \left[\left(1 - \frac{\eta\mu}{8} \right) \Phi(\underline{\mathbf{w}}^r) + \frac{2\eta^3 T L^2}{N} \left[\frac{16T^2 L_{max}^2 l_{max}}{\mu_{min}} + 4\lambda_2^2 \gamma L_{max}^2 T^2 \right] \mathcal{D}_{r,0} \right. \\ &\quad \left. + \frac{4\eta T L^2}{N} \|(Q - P)W^{r,0}\|_F^2 \right]. \end{aligned} \quad (37)$$

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

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

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

$$\mathbb{E} [\Phi(\underline{\mathbf{w}}^{r+1})] \leq \left(1 - \frac{\eta\mu}{8} \right) \mathbb{E} [\Phi(\underline{\mathbf{w}}^r)] + \frac{6\eta T L^2}{N} \mathcal{D}_{r,0}. \quad (38)$$

This completes the proof. \square

B.3 Proof of Lemma 2

Let $\mathcal{D}_{r,0} = \mathbb{E} \|\underline{W}_l^{r,0} - \underline{W}^{r,0}\|_F^2 = \sum_{k=1}^N \mathbb{E} \|\underline{\mathbf{w}}_k^{r,0} - \underline{\mathbf{w}}^{r,0}\|^2$. Using compact notations for the updates in equation 7 and equation 8, the global drift term can be written as

$$\begin{aligned} \mathcal{D}_{r,0} &= \mathbb{E} \|QPW^{r,0} - PW^{r,0}\|_F^2 \\ &= \mathbb{E} \|(Q - P)W^{r,0}\|_F^2. \end{aligned} \quad (39)$$

Recall that $Q = \frac{1}{N} \mathbf{1}\mathbf{1}^T$ is the average matrix, P is the mixing matrix and $QP = Q$. Using $\underline{W}_l^{r,0} = PW^{r-1,T}$ (see equation 7), substituting for the update in $W^{r-1,T}$ and taking the telescopic sum, we get

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

Plugging the above in equation 39, 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 global drift term can be upper bounded as

$$\begin{aligned} \mathbb{E} \|(Q - P)W^{r,0}\|_F^2 &\leq \left(1 + \frac{1}{\psi} \right) \Xi + (1 + \psi) \eta^2 \mathbb{E} \left\| (Q - P^2) \sum_{\tau=0}^{T-1} \partial \hat{\Phi}(W^{r-1,\tau}) \right\|_F^2 \\ &\stackrel{(a)}{\leq} \left(1 + \frac{1}{\psi} \right) \Xi + (1 + \psi) \eta^2 \|(Q - P^2)\|_{op}^2 \mathbb{E} \left\| \sum_{\tau=0}^{T-1} \partial \hat{\Phi}(W^{r-1,\tau}) \right\|_F^2 \\ &\stackrel{(b)}{\leq} \left(1 + \frac{1}{\psi} \right) \Xi + (1 + \psi) \eta^2 \lambda_2^4 T \sum_{\tau=0}^{T-1} \mathbb{E} \left\| \partial \hat{\Phi}(W^{r-1,\tau}) \right\|_F^2, \end{aligned} \quad (40)$$

where λ_2 is the second largest eigenvalue of the mixing matrix P and $\Xi := \mathbb{E} \|(Q - P^2) W^{r-1,0}\|_F^2$. In the above, (a) follows from Lemma 4 and (b) follows from Lemma 5. Next, consider bounding the following

$$\begin{aligned} \mathbb{E} \|\partial \hat{\Phi}(W^{r-1,\tau})\|_F^2 &= \mathbb{E} \sum_{k=1}^N \left\| \frac{1}{b} \sum_{j \in B_k^{r-1,\tau}} \nabla \Phi_{k,j}(\mathbf{w}_k^{r-1,\tau}) \right\|_2^2 \\ &\stackrel{\text{Jensen's}}{\leq} \mathbb{E} \sum_{k=1}^N \frac{1}{b} \sum_{j \in B_k^{r-1,\tau}} \|\nabla \Phi_{k,j}(\mathbf{w}_k^{r-1,\tau})\|_2^2 \\ &\stackrel{(a)}{\leq} 2l_{max} \sum_{k=1}^N \mathbb{E} [\Phi_k(\mathbf{w}_k^{r-1,\tau})], \end{aligned} \quad (41)$$

where (a) follows from the smoothness assumption and $l_{max} := \max_{k,j} l_{k,j}$. Recall from Lemma 1 that

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

Substituting this in equation 41, and writing it in the matrix form, we get

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

Using the above in equation 40

$$\begin{aligned} \mathbb{E} \|(Q - P)W^{r,0}\|_F^2 &\leq \left(1 + \frac{1}{\psi}\right) \mathbb{E} \|(Q - P^2)W^{r-1,0}\|_F^2 + \eta^2 \lambda_2^4 \alpha T^2 L_{max}^2 \mathcal{D}_{r-1,0} \\ &\quad + \eta^2 \lambda_2^4 \beta T^2 \mathbb{E} \|\partial \Phi(\underline{W}^{r-1,0})\|_F^2, \end{aligned} \quad (42)$$

where $\beta := \frac{4l_{max}(1+\psi)}{\mu_{min}}$. First, let us consider bounding a part of the first term above, i.e., $\mathbb{E} \|(Q - P^2)W^{r-1,0}\|_F^2$. Using the fact that $QP = Q$ and $Q^2 = Q$, it follows that $P^2 - Q = (Q - P)^2$. Using this in equation 42, we get

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

Applying the results of Lemma 4 and 5, the first term in the above can further be bounded as,

$$\begin{aligned} \mathbb{E} \|(Q - P)W^{r,0}\|_F^2 &\leq \left(1 + \frac{1}{\psi}\right) \|(Q - P)\|^2 \mathbb{E} \|(Q - P)W^{r-1,0}\|_F^2 + \eta^2 \lambda_2^4 \beta T^2 L_{max}^2 \mathcal{D}_{r-1,0} \\ &\quad + \eta^2 \lambda_2^4 \beta T^2 \mathbb{E} \|\partial \Phi(\underline{W}^{r-1,0})\|_F^2 \\ &\stackrel{(a)}{\leq} \left(1 + \frac{1}{\psi}\right) \lambda_2^2 \mathcal{D}_{r-1,0} + \eta^2 \lambda_2^4 \beta T^2 L_{max}^2 \mathcal{D}_{r-1,0} + \eta^2 \lambda_2^4 \beta T^2 \mathbb{E} \|\partial \Phi(\underline{W}^{r-1,0})\|_F^2, \end{aligned} \quad (43)$$

where (a) follows by substituting the results from Lemma 5. The term $\mathbb{E} \|\partial \Phi(\underline{W}^{r-1,0})\|_F^2$ in the above is bounded as follows

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

where (a) follows from smoothness assumption and using the fact that $\Phi(\underline{\mathbf{w}}^{r-1,0}) = \frac{1}{N} \sum_{k=1}^N \Phi_k(\underline{\mathbf{w}}^{r-1,0})$, and $L_{max} = \max_k L_k$. Using the above result in equation 43, we get

$$\mathcal{D}_{r,0} \leq \left(\left(1 + \frac{1}{\psi}\right) \lambda_2^2 + \eta^2 \lambda_2^4 \beta T^2 L_{max}^2 \right) \mathcal{D}_{r-1,0} + 2\eta^2 \lambda_2^4 \beta T^2 L_{max} N \mathbb{E} [\Phi(\underline{\mathbf{w}}^{r-1,0})].$$

This completes the proof. \square

B.4 Proof of Lemma 3 (Completing the proof of Theorem 1)

Let us recall the equations for $\mathcal{D}_{r+1,0}$ and $\Phi(\underline{\mathbf{w}}^{r+1})$ from Lemma 1 and Lemma 2

$$\Phi(\underline{\mathbf{w}}^{r+1}) \leq \alpha \Phi(\underline{\mathbf{w}}^r) + \rho \mathcal{D}_{r,0}, \quad (44)$$

$$\mathcal{D}_{r+1,0} \leq \nu \mathcal{D}_{r,0} + \chi \Phi(\underline{\mathbf{w}}^r), \quad (45)$$

where $\alpha := (1 - \frac{\eta\mu}{8})$, $\rho := \frac{6\eta L^2 T}{N}$, $\nu := (1 + \frac{1}{\psi}) \lambda_2^2 + \eta^2 \lambda_2^4 \beta T^2 L_{max}^2$ and $\chi := 2\eta^2 \lambda_2^4 \beta T^2 L_{max} N$. To ensure $\nu < 1$, we choose $\psi = \frac{2\lambda_2^2}{1-\lambda_2^2}$ and any $\eta \leq \sqrt{\frac{1-\lambda_2^2}{4\lambda_2^4 \beta T^2 L_{max}^2}}$. Further, to ensure $\chi < 1$, we choose $\eta \leq \sqrt{\frac{1}{2\lambda_2^4 \beta T^2 L_{max} N}}$. Now consider the following Lyapunov function for some constant $\theta > 0$

$$\begin{aligned} \Phi(\underline{\mathbf{w}}^{r+1}) + \theta \mathcal{D}_{r+1,0} &\leq \alpha \Phi(\underline{\mathbf{w}}^r) + \rho \mathcal{D}_{r,0} + \theta (\nu \mathcal{D}_{r,0} + \chi \Phi(\underline{\mathbf{w}}^r)) \\ &\stackrel{(a)}{\leq} (\alpha + \theta \chi) \Phi(\underline{\mathbf{w}}^r) + (\rho + \theta \nu) \mathcal{D}_{r,0}, \end{aligned} \quad (46)$$

where (a) follows from equation 44 and equation 45. To show linear convergence we want the coefficients of the first and second terms in equation 46 to satisfy the following inequalities

$$\alpha + \theta \chi \leq \left(1 - \frac{\eta\mu}{16}\right) \text{ and } (\rho + \theta \nu) \leq \theta \left(1 - \frac{\eta\mu}{16}\right). \quad (47)$$

Now, consider the first inequality above. Substituting for α and χ and choosing

$$\eta \leq \frac{\mu}{32\theta \lambda_2^4 \beta T^2 L_{max} N},$$

ensures that the first inequality in equation 47 is satisfied. Next, substituting the values for ρ and ν in the second inequality in equation 47, and simplifying results in

$$\frac{6\eta L^2 T}{N} + \theta \eta^2 \lambda_2^4 \beta T^2 L_{max}^2 + \frac{\theta \eta \mu}{16} \leq \theta \left(1 - \lambda_2^2 - \frac{\lambda_2^2}{\psi}\right),$$

where the above quantity is non-negative by choosing $\psi > \frac{\lambda_2^2}{1-\lambda_2^2}$. Now, picking $\eta \leq \frac{1}{\theta \lambda_2^4 \beta T^2 L_{max}^2}$ leads to

$$\eta \left(1 + \frac{\theta \mu}{16} + \frac{6L^2 T}{N}\right) \leq \theta \left(1 - \lambda_2^2 - \frac{\lambda_2^2}{\psi}\right).$$

Choosing $\eta \leq \frac{\theta \left(1 - \lambda_2^2 - \frac{\lambda_2^2}{\psi}\right)}{\left(1 + \frac{\theta \mu}{16} + \frac{6L^2 T}{N}\right)}$ ensures that $(\rho + \theta \nu) \leq \theta \left(1 - \frac{\eta\mu}{16}\right)$. Finally, choosing

$$\begin{aligned} \eta \leq \min &\left\{ \frac{4}{\mu}, \frac{2}{\mu_{min}}, \frac{\mu}{4\zeta_1}, \frac{L^2}{2\zeta_2}, \frac{1}{8} \left(\frac{\mu}{\zeta_3 T^3} \right)^{\frac{1}{3}}, \left(\frac{1}{\zeta_4 T^2} \right)^{\frac{1}{2}}, \frac{\mu_{min}}{\zeta_5}, \right. \\ &\left. \sqrt{\frac{1-\lambda_2^2}{\zeta_6 T^2}}, \sqrt{\frac{1}{\zeta_7 T^2}}, \frac{\mu}{\zeta_8 T^2}, \frac{1}{\zeta_9 T^2}, \frac{\theta(1-\lambda_2^2-\frac{\lambda_2^2}{\psi})}{\left(1 + \frac{\theta \mu}{16} + \frac{6L^2 T}{N}\right)} \right\}, \end{aligned} \quad (48)$$

Substituting the conditions in equation 47 in equation 46, we get

$$\Phi(\underline{\mathbf{w}}^{r+1}) + \theta \mathcal{D}_{r+1,0} \leq \left(1 - \frac{\eta\mu}{16}\right) (\Phi(\underline{\mathbf{w}}^r) + \theta \mathcal{D}_{r,0}).$$

for any constant $\theta > 0$. Here, $\mu_{\min} := \min_{k \in [N]} \{\mu_k\}$, $l_{\max} := \max_{k,j} l_{k,j}$ and $L_{\max} := \max_k L_k$, $\psi = \frac{2\lambda_2^2}{1-\lambda_2^2}$, $\gamma := \frac{2l_{\max}N}{\mu_{\min}}$, $\beta := \frac{4l_{\max}(1+\psi)}{\mu_{\min}}$, and $\mathcal{D}_{r,0} := \sum_{k=1}^N \mathbb{E} \left\| \mathbf{w}_k^{r,0} - \mathbf{w}^{r,0} \right\|^2$. Moreover, in the above $\zeta_1 := 4 \left(\frac{2Ll_{\max}}{bN} + \frac{2LL_{\max}}{N} + 2LL_{\max} \right)$, $\zeta_2 := 2 \left(\frac{Ll_{\max}^2}{bN} + \frac{LL_{\max}^2}{N} + LL_{\max}^2 \right)$, $\zeta_3 := \frac{64l_{\max}LL_{\max}}{\mu_{\min}} + 16\gamma L\lambda_2^2 L_{\max}$, $\zeta_4 := \frac{16l_{\max}L_{\max}^2}{\mu_{\min}} + 4\lambda_2^2 \gamma L^2$, $\zeta_5 := 2 \left[\frac{l_{\max}L_{\max}}{b} + \frac{L_{\max}^2 b(b-1)}{b^2} \right]$, $\zeta_6 := 4\lambda_2^4 \beta L_{\max}^2$, $\zeta_7 := 2\lambda_2^4 \beta L_{\max} N$, $\zeta_8 := 32\theta\lambda_2^4 \beta L_{\max} N$ and $\zeta_9 := \theta\lambda_2^4 \beta L_{\max}^2$.

□

C Differences between strongly convex and our setting

In the following, we provide a very simple 1-D examples for strongly convex ($f_1(x) = \log(x - 0.5 + \sqrt{1 + (x - 0.5)^2}) + (x - 0.5)^2 + 0.25 + \text{constant}$ and $f_2(x) = 2x^2$), and non-convex settings ($f_1(x) = x^2 + 2\sin^2(x)$ and $f_2(x) = 0 \times \mathbf{1}\{f_1(x) \leq 4\} + f_1(x)\mathbf{1}\{f_1(x) > 4\}$), as shown in Fig. 8. Note that in the strongly convex setting, both clients share the unique minima $x^* = 0$ due to the interpolation assumption. In this case, both clients do not need to communicate since each client can run local rounds to reach the global minima that minimize the average, and hence making decentralized or collaborative learning vacuous! On the other hand, for the PL setting, multiple local rounds lead to different optimal points. For example, running multiple rounds of GD results in client 1 reaching $x_1^* = 0$ while client 2 reaches ≈ -1.3 or $\approx +1.3$ depending on the initialization. However, the optimal point $x^* = 0$. This simple scenario suggests challenges while proving the results. For example, in the strongly convex setting ((Koloskova et al., 2020)), one can start with the difference between global optimum and the local/global update while we cannot use this to prove our results.

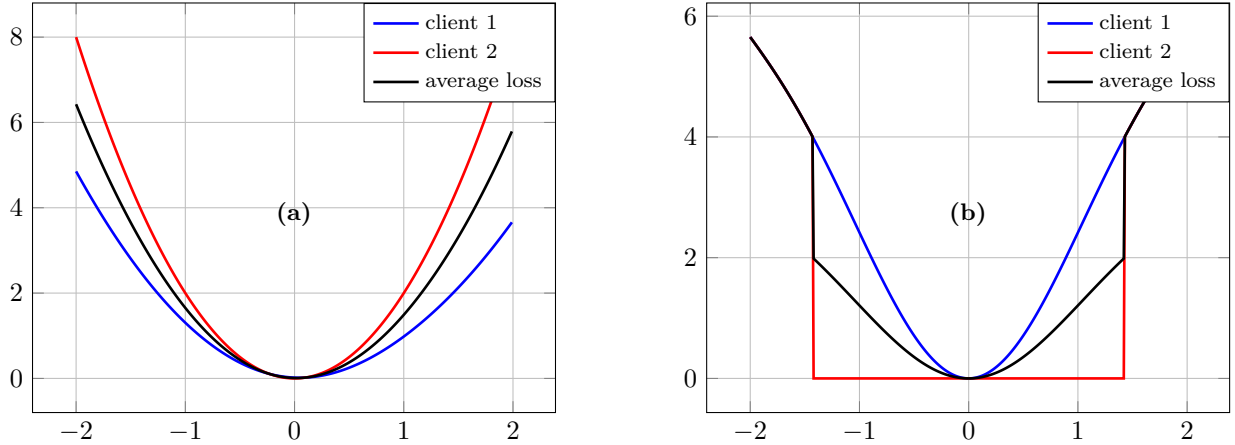


Figure 8: Strongly convex losses in the overparameterized regime (see (a)) and Losses satisfying PL inequality (see (b)).