A PSEUDO-CODE OF MOMENTUM TRACKING

The pseudo-code for Momentum Tracking is given in Alg. 1, where $\textbf{Transmit}_{i\to j}(\cdot)$ denotes that node i transmits parameters to node j and $\textbf{Receive}_{i\leftarrow j}(\cdot)$ denotes that node i receives parameters from node j.

Algorithm 1: Update rules of Momentum Tracking at node i.

Input: Step size $\eta > 0$, $\beta \in (0, 1]$, mixing matrix $\boldsymbol{W}$. Initialize $\boldsymbol{c}_i$ and $\boldsymbol{u}_i$ to $\frac{1}{1-\beta}(\nabla F_i(\boldsymbol{x}_i^{(0)}; \xi_i^{(0)}) - \frac{1}{N}\sum_j \nabla F_j(\boldsymbol{x}_j^{(0)}; \xi_j^{(0)}))$ for all $i \in V$ and $\boldsymbol{x}_i$ with the same parameter.
for $r = 0, \cdots, R$ **do**
  $\boldsymbol{u}_i^{(r+1)} \leftarrow \beta \boldsymbol{u}_i^{(r)} + \nabla F_i(\boldsymbol{x}_i^{(r)}; \xi_i^{(r)})$.
  for $j \in \mathcal{N}_i$ **do**
    $\textbf{Transmit}_{i\to j}(\boldsymbol{x}_i^{(r)})$ and $\textbf{Receive}_{i\leftarrow j}(\boldsymbol{x}_j^{(r)})$.
    $\textbf{Transmit}_{i\to j}(\boldsymbol{c}_i^{(r)} - \boldsymbol{u}_i^{(r+1)})$ and $\textbf{Receive}_{i\leftarrow j}(\boldsymbol{c}_j^{(r)} - \boldsymbol{u}_j^{(r+1)})$.
  end for
  $\boldsymbol{x}_i^{(r+1)} \leftarrow \sum_{j\in\mathcal{N}_i^+} W_{ij}\boldsymbol{x}_j^{(r)} - \eta\left(\boldsymbol{u}_i^{(r+1)} - \boldsymbol{c}_i^{(r)}\right)$.
  $\boldsymbol{c}_i^{(r+1)} \leftarrow \sum_{j\in\mathcal{N}_i^+} W_{ij}\left(\boldsymbol{c}_j^{(r)} - \boldsymbol{u}_j^{(r+1)}\right) + \boldsymbol{u}_i^{(r+1)}$.
end for

B ADDITIONAL DISCUSSION ABOUT CONVERGENCE RATE

Because Momentum Tracking is equivalent to Gradient Tracking when $\beta = 0$, Theorem 1 also provides the convergence rate of Gradient Tracking. In this section, we compare the convergence rate of Gradient Tracking provided in Theorem 1 to that provided by Koloskova et al. (2021).

From Theorem 1, we get the following statement.

Corollary 1. *Suppose that $\beta = 0$ and the assumptions of Theorem 1 hold. Then, for any $R \geq 1$, there exists a step size η such that the average parameter $\bar{\boldsymbol{x}} \coloneqq \frac{1}{N}\sum_i \boldsymbol{x}_i$ generated by Eqs. (7-9) satisfies*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{r_0\sigma^2 L}{NR}} + \left(\frac{r_0\sigma L}{p^2 R}\right)^{\frac{2}{3}} + \frac{Lr_0}{p^2 R}\right), \tag{11}$$

where $r_0 \coloneqq f(\bar{\boldsymbol{x}}^{(0)}) - f^\star$.

Then, under Assumptions 1, 2, 3, and 4, Koloskova et al. (2021) provided the convergence rate of Gradient Tracking as follows.

Theorem 3 ((Koloskova et al., 2021))**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Then, for any round $R > \frac{2}{p}\log(\frac{50}{p}(1+\log\frac{1}{p}))$, there exists a step size η that satisfies that the average parameter $\bar{\boldsymbol{x}} \coloneqq \frac{1}{N}\sum_i \boldsymbol{x}_i$ generated by Gradient Tracking satisfies*

$$\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{r_0\sigma^2 L}{NR}} + \left(\frac{r_0\sigma L}{(\sqrt{p}c + p\sqrt{N})R}\right)^{\frac{2}{3}} + \frac{Lr_0}{pcR}\right), \tag{12}$$

where $\tilde{O}(\cdot)$ hides the polylogarithmic factors, $c \coloneqq 1 - \min\{\lambda_{min}, 0\}^2$, and λ_{min} is the minimum eigenvalue of $\boldsymbol{W}$.

Comparing the convergence rates in Eqs. (11) and (12), the convergence rate in Eq. (12) is tighter than that in Eq. (11) because $c \geq p$ for any mixing matrix $\boldsymbol{W}$. However, because the convergence rate in Eq. (12) holds only when the number of round R is larger than $\frac{2}{p}\log(\frac{50}{p}(1+\log\frac{1}{p}))$, Theorem 3 can not describe the behavior of the convergence rate at the beginning of the training. In contrast, Corollary 1 provides the convergence rate for Gradient Tracking that holds for any $R \geq 1$.

C ADDITIONAL EXPERIMENTS

C.1 RESULTS WITH VARIOUS NETWORK TOPOLOGIES

We evaluated Momentum Tracking in more detail by changing the underlying network topology. Table 4 lists the test accuracy of all comparison methods when we set the underlying network topology to be a hypercube or a semantic exponential graph.

Table 4 indicates that when the data distributions held by each node are statistically homogeneous (i.e., 10-class), DSGDm, QG-DSGDm, DecentLaM, and Momentum Tracking are comparable and outperform DSGD and Gradient Tracking for all network topologies. When the data distributions are heterogeneous (i.e., 4-class), the results show that Momentum Tracking is more robust to data heterogeneity than DSGDm, QG-DSGDm, and DecentLaM and outperforms all comparison methods for all network topologies. Therefore, the results indicate that Momentum Tracking is robust to data heterogeneity regardless of the underlying network topology.

Table 4: Test accuracy on CIFAR-10 with different underlying network topologies.

	CIFAR-10			
	Hypercube		Semantic Exponential Graph	
	10-class	4-class	10-class	4-class
DSGD	63.3 ± 0.65	55.9 ± 4.11	64.0 ± 0.26	60.7 ± 1.82
Gradient Tracking	61.0 ± 1.34	60.2 ± 1.13	62.4 ± 0.53	62.4 ± 0.89
DSGDm	$\mathbf{73.2 \pm 0.09}$	45.0 ± 5.90	$\mathbf{73.4 \pm 0.13}$	51.5 ± 7.80
QG-DSGDm	73.0 ± 0.31	62.9 ± 3.68	$\mathbf{73.4 \pm 0.58}$	70.2 ± 1.09
DecentLaM	72.9 ± 0.24	69.1 ± 4.05	72.9 ± 0.73	71.2 ± 1.72
Momentum Tracking	72.8 ± 0.15	$\mathbf{72.7 \pm 0.28}$	72.7 ± 0.33	$\mathbf{72.9 \pm 0.07}$

C.2 INITIAL VALUE ANALYSIS

In this section, we discuss the initial values of $\boldsymbol{c}_i$ and $\boldsymbol{u}_i$. Table 5 lists the test accuracy for Momentum Tracking when we initialize $\boldsymbol{c}_i$ and $\boldsymbol{u}_i$ to zero and when we initialize $\boldsymbol{c}_i$ and $\boldsymbol{u}_i$ as in Theorem 1. The results indicate that the test accuracy are almost equivalent on both settings. Hence, Theorem 1 requires $\boldsymbol{c}_i$ and $\boldsymbol{u}_i$ to be initialized to $\frac{1}{1-\beta}(\nabla F_i(\boldsymbol{x}_i^{(0)};\xi_i^{(0)}) - \frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\boldsymbol{x}_j^{(0)};\xi_j^{(0)}))$. However, in practice, $\boldsymbol{c}_i$ and $\boldsymbol{u}_i$ can be initialized to zeros without any impact on accuracy.

Table 5: Test accuracy on FashionMNIST, SVHN, and CIFAR-10 with LeNet. "k-class" indicates that each node has only the data of randomly selected k classes.

	FashionMNIST			
	10-class	8-class	6-class	4-class
Momentum Tracking	89.5 ± 0.36	89.4 ± 0.05	88.9 ± 0.47	86.8 ± 1.56
Momentum Tracking ($\boldsymbol{c}_i^{(0)} = \boldsymbol{u}_i^{(0)} = \boldsymbol{0}$)	89.5 ± 0.38	89.4 ± 0.04	88.7 ± 0.63	85.8 ± 1.53

	SVHN			
	10-class	8-class	6-class	4-class
Momentum Tracking	92.6 ± 0.32	92.4 ± 0.40	92.3 ± 0.23	91.7 ± 0.53
Momentum Tracking ($\boldsymbol{c}_i^{(0)} = \boldsymbol{u}_i^{(0)} = \boldsymbol{0}$)	92.5 ± 0.34	92.3 ± 0.50	92.2 ± 0.29	92.0 ± 0.81

	CIFAR-10			
	10-class	8-class	6-class	4-class
Momentum Tracking	72.9 ± 0.59	73.0 ± 0.49	72.6 ± 0.41	70.7 ± 1.38
Momentum Tracking ($\boldsymbol{c}_i^{(0)} = \boldsymbol{u}_i^{(0)} = \boldsymbol{0}$)	72.8 ± 0.35	72.9 ± 0.32	73.0 ± 0.41	70.7 ± 2.00

C.3 LEARNING CURVES

In this section, we present the learning curves for the results whose final accuracy are presented in Tables 2 and 3. Figs. 3, 4, and 5 show the learning curves for FashionMNSIT, SVHN, and CIFAR-10, respectively, with LeNet. Figs. 6 and 7 show the learning curves for CIFAR-10 with VGG-11 and ResNet-34, respectively.

When the data distributions are statistically homogeneous (i.e., 10-class), the results indicate that DSGDm, QG-DSGDm, DecentLaM, and Momentum Tracking are comparable and can achieve high accuracy faster than DSGD and Gradient Tracking. When the data distributions are statistically heterogeneous (e.g., 2-class and 4-class), the results indicate that the learning curves for Momentum Tracking are more stable than those for DSGDm, QG-DSGDm, and DecentLaM, and Momentum Tracking outperforms all comparison methods. In particular, in Figs. 6 and 7, the accuracy of DSGD, DSGDm, QG-DSGDm, and DecentLaM continue to oscillate in the final training phase in the 2-class setting, whereas the accuracy of Momentum Tracking and Gradient Tracking converge in the 2-class setting as well as in the 10-class setting. Therefore, Momentum Tracking is more robust to data heterogeneity than DSGDm, QG-DSGDm, and DecentLaM.

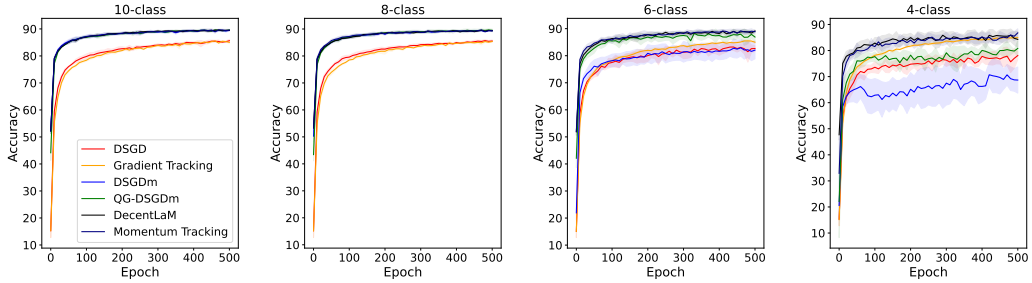

Figure 3: Learning curves on FashionMNIST. The accuracy is evaluated per 10 epochs.

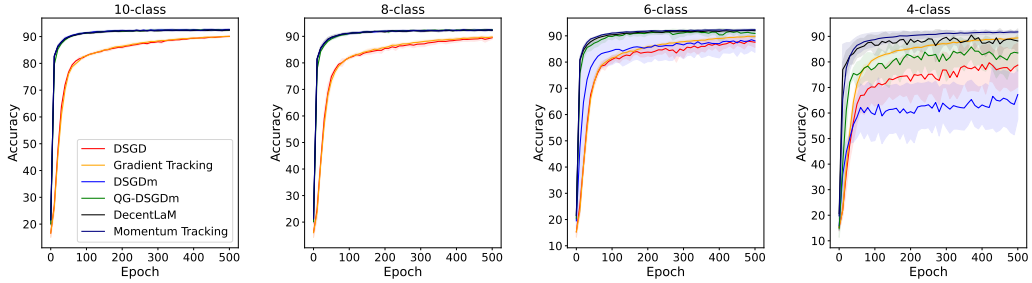

Figure 4: Learning curves on SVHN. The accuracy is evaluated per 10 epochs.

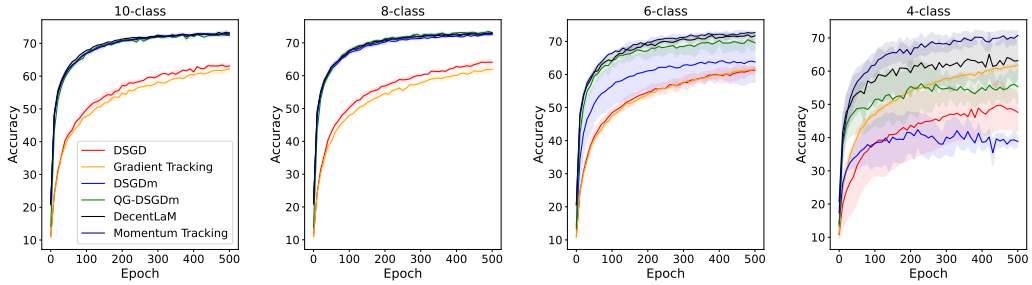

Figure 5: Learning curves on CIFAR-10. The accuracy is evaluated per 10 epochs.

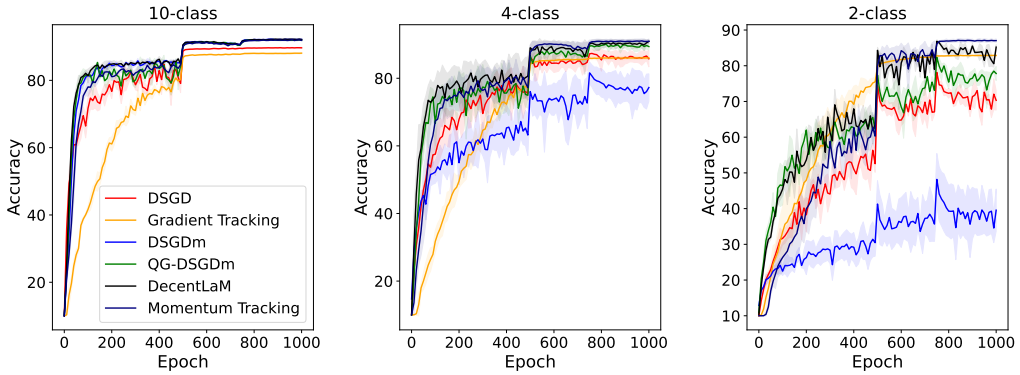

Figure 6: Learning curves on CIFAR-10 with VGG-11. The accuracy is evaluated per 10 epochs.

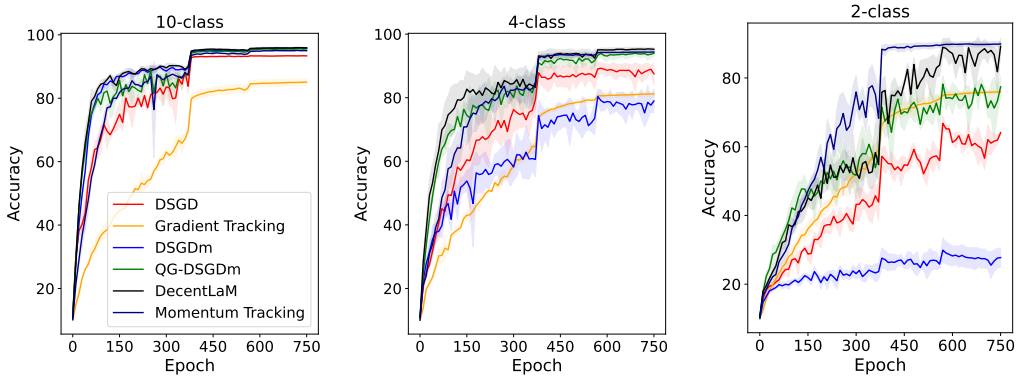

Figure 7: Learning curves on CIFAR-10 with ResNet-34. The accuracy is evaluated per 10 epochs.

C.4 Synthetic Experiment

In this section, we evaluate the convergence rate in more detail using a synthetic dataset. Following the previous work (Koloskova et al., 2020), we set the dimension of parameter d to 50, the number of nodes N to 25, and the network topology to a ring consisting of N nodes. We then defined the local objective function $f_i(\boldsymbol{x})$ to be $\frac{1}{2}\|\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{b}_i\|^2$ where $\boldsymbol{A}_i := i/\sqrt{N}$ and $\boldsymbol{b}_i$ are sampled from $\mathcal{N}(\boldsymbol{0}, \zeta^2/i^2\boldsymbol{1})$, and we defined the stochastic gradient $\nabla F_i(\boldsymbol{x};\xi_i)$ to be $\nabla f_i(\boldsymbol{x})+\epsilon$ where ϵ is drawn from $\mathcal{N}(\boldsymbol{0};\sigma^2/d\boldsymbol{1})$. For all comparison methods, we set the step size η to 1.0×10^{-4}.

Figs. 8 and 9 illustrate $\|\nabla f(\bar{\boldsymbol{x}})\|^2$ with respect to the number of rounds r when we vary data heterogeneity ζ^2 as $\{0, 25, 50\}$ and set σ^2 to one. The results show that Momentum Tracking converges in the same manner regardless of data heterogeneity ζ^2. On the other hand, for DSGDm, QG-DSGDm, and DecentLaM, $\|\nabla f(\bar{\boldsymbol{x}})\|^2$ increases as data heterogeneity ζ^2 increases. Hence, these results are consistent with our theoretical analysis that the convergence rate of Momentum Tracking is independent of data heterogeneity.

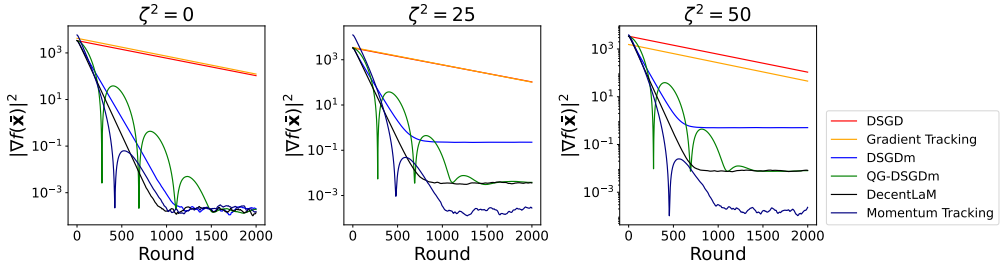

Figure 8: Comparison of the convergence in the initial training phase in the synthetic experiments.

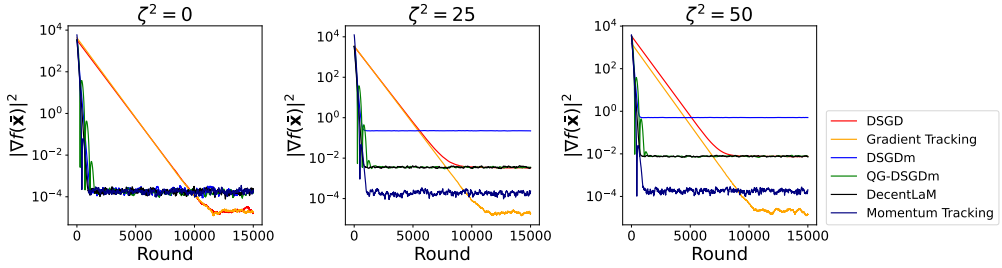

Figure 9: Comparison of the convergence in the synthetic experiments.

D PROOF OF THEOREM 1

D.1 TECHNICAL LEMMA

Lemma 1. *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\gamma > 0$, it holds that*

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 \leq (1+\gamma)\|\boldsymbol{x}\|^2 + (1+\gamma^{-1})\|\boldsymbol{y}\|^2. \tag{13}$$

Lemma 2. *For any $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N$, it holds that*

$$\left\|\sum_{i=1}^N \boldsymbol{a}_i\right\|^2 \leq N\sum_{i=1}^N \|\boldsymbol{a}_i\|^2. \tag{14}$$

Lemma 3. *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\gamma > 0$, it holds that*

$$2\langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq \gamma\|\boldsymbol{x}\|^2 + \gamma^{-1}\|\boldsymbol{y}\|^2. \tag{15}$$

D.2 MOMENTUM TRACKING IN MATRIX NOTATION

We define $\boldsymbol{U}^{(r)}$, $\boldsymbol{X}^{(r)}$, $\boldsymbol{C}^{(r)}$, $\nabla F(\boldsymbol{X}^{(r)};\xi^{(r)})$, and $\nabla f(\boldsymbol{X}^{(r)})$ as follows:

$$\boldsymbol{U}^{(r)} \coloneqq \left(\boldsymbol{u}_1^{(r)}, \cdots, \boldsymbol{u}_N^{(r)}\right), \quad \boldsymbol{X}^{(r)} \coloneqq \left(\boldsymbol{x}_1^{(r)}, \cdots, \boldsymbol{x}_N^{(r)}\right), \quad \boldsymbol{C}^{(r)} \coloneqq \left(\boldsymbol{c}_1^{(r)}, \cdots, \boldsymbol{c}_N^{(r)}\right),$$
$$\nabla F(\boldsymbol{X}^{(r)};\xi^{(r)}) \coloneqq \left(\nabla F_1(\boldsymbol{x}_1^{(r)};\xi_1^{(r)}), \cdots, \nabla F_N(\boldsymbol{x}_N^{(r)};\xi_N^{(r)})\right),$$
$$\nabla f(\boldsymbol{X}^{(r)}) \coloneqq \left(\nabla f_1(\boldsymbol{x}_1^{(r)}), \cdots, \nabla f_N(\boldsymbol{x}_N^{(r)})\right).$$

Then, the update rule of Momentum Tracking can then be rewritten as follows:

$$\boldsymbol{U}^{(r+1)} = \beta\boldsymbol{U}^{(r)} + \nabla F(\boldsymbol{X}^{(r)};\xi^{(r)}), \tag{16}$$
$$\boldsymbol{X}^{(r+1)} = \boldsymbol{X}^{(r)}\boldsymbol{W} - \eta(\boldsymbol{U}^{(r+1)} - \boldsymbol{C}^{(r)}), \tag{17}$$
$$\boldsymbol{C}^{(r+1)} = (\boldsymbol{C}^{(r)} - \boldsymbol{U}^{(r+1)})\boldsymbol{W} + \boldsymbol{U}^{(r+1)}, \tag{18}$$

where $\boldsymbol{U}^{(0)}$ and $\boldsymbol{C}^{(0)}$ are initialized as follows:

$$\boldsymbol{U}^{(0)} = \frac{1}{1-\beta}(\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)}) - \frac{1}{N}\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\boldsymbol{1}\boldsymbol{1}^\top),$$
$$\boldsymbol{C}^{(0)} = \frac{1}{1-\beta}(\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)}) - \frac{1}{N}\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\boldsymbol{1}\boldsymbol{1}^\top).$$

D.3 ADDITIONAL NOTATION

We define the update rules of $\boldsymbol{d}_i$ and $\boldsymbol{e}_i$ as follows:

$$\boldsymbol{d}_i^{(r+1)} = \beta\boldsymbol{d}_i^{(r)} + \nabla f_i(\bar{\boldsymbol{x}}^{(r)}), \tag{19}$$
$$\boldsymbol{e}_i^{(r+1)} = \beta\boldsymbol{e}_i^{(r)} + \nabla f(\bar{\boldsymbol{x}}^{(r)}), \tag{20}$$

where $\boldsymbol{d}_i^{(0)} = \frac{1}{1-\beta}(\nabla f_i(\bar{\boldsymbol{x}}^{(0)}) - \nabla f(\bar{\boldsymbol{x}}^{(0)}))$ and $\boldsymbol{e}_i^{(0)} = \boldsymbol{0}$. Note that it holds that $\bar{\boldsymbol{d}}^{(r)} = \bar{\boldsymbol{e}}^{(r)}$ for any round $r \geq 0$. Then, we define $\boldsymbol{D}$ and $\boldsymbol{E}$ as follows:

$$\boldsymbol{D}^{(r)} \coloneqq \left(\boldsymbol{d}_1^{(r)}, \cdots, \boldsymbol{d}_N^{(r)}\right), \quad \boldsymbol{E}^{(r)} \coloneqq \left(\boldsymbol{e}_1^{(r)}, \cdots, \boldsymbol{e}_N^{(r)}\right).$$

The update rules of $\boldsymbol{D}$ and $\boldsymbol{E}$ can be written as follows:

$$\boldsymbol{D}^{(r+1)} = \beta\boldsymbol{D}^{(r)} + \nabla f(\bar{\boldsymbol{X}}^{(r)}), \tag{21}$$
$$\boldsymbol{E}^{(r+1)} = \beta\boldsymbol{E}^{(r)} + \frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(r)})\boldsymbol{1}\boldsymbol{1}^\top, \tag{22}$$

where $\boldsymbol{D}^{(0)}$ and $\boldsymbol{E}^{(0)}$ are initialized as follows:

$$\boldsymbol{D}^{(0)} = \frac{1}{1-\beta}(\nabla f(\bar{\boldsymbol{X}}^{(0)}) - \frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(0)})\mathbf{1}\mathbf{1}^\top),$$
$$\boldsymbol{E}^{(0)} = \mathbf{0}.$$

Note that $\boldsymbol{d}_i$, $\boldsymbol{e}_i$, $\boldsymbol{D}$, and $\boldsymbol{E}$ are the only variables used in the proof that do not need to be computed in practice in Alg. 1. We define Ξ, $\mathcal{E}$, and $\mathcal{D}$ as follows:

$$\Xi^{(r)} \coloneqq \frac{1}{N}\mathbb{E}\left\|\boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r)}\right\|_F^2,$$
$$\mathcal{E}^{(r)} \coloneqq \frac{1}{N}\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)} - \boldsymbol{E}^{(r+1)}\right\|_F^2,$$
$$\mathcal{D}^{(r)} \coloneqq \frac{1}{N}\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{D}^{(r)} - \boldsymbol{E}^{(r+1)} + \boldsymbol{E}^{(r)}\right\|_F^2.$$

Inspired by Yu et al. (2019), we define $\bar{\boldsymbol{z}}$ as follows:

$$\bar{\boldsymbol{z}}^{(r)} \coloneqq \begin{cases} \bar{\boldsymbol{x}}^{(r)}, & \text{if } r = 0 \\ \frac{1}{1-\beta}\bar{\boldsymbol{x}}^{(r)} - \frac{\beta}{1-\beta}\bar{\boldsymbol{x}}^{(r-1)}, & \text{otherwise} \end{cases}.$$

In the following, we define $\pm a \coloneqq a - a = 0$ for any a and $\bar{a} \coloneqq \frac{1}{N}\sum_{i=1}^N a_i$ for any $a_1, \cdots, a_N$. Then, $\mathbb{E}[\cdot]$ denotes the expectation over all randomness that occurs during training (i.e., $\{\xi_i^{(r)}\}_{i,r}$), and $\mathbb{E}_r[\cdot]$ denotes the expectation over the randomness that occurs at round r (i.e., $\{\xi_i^{(r)}\}_i$).

D.4 USEFUL LEMMA

Lemma 4. *For any round $r \geq 0$, it holds that $\bar{\boldsymbol{c}}^{(r)} = \mathbf{0}$.*

Proof. For any round $r \geq 0$, we have

$$\begin{aligned}\sum_{i=1}^{N} \boldsymbol{c}_i^{(r+1)} &= \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\boldsymbol{c}_j^{(r)} - \boldsymbol{u}_j^{(r+1)}) + \sum_{i=1}^{N} \boldsymbol{u}_i^{(r+1)} \\ &= \sum_{j=1}^{N}(\boldsymbol{c}_j^{(r)} - \boldsymbol{u}_j^{(r+1)})\sum_{i=1}^{N} W_{ij} + \sum_{i=1}^{N} \boldsymbol{u}_i^{(r+1)}.\end{aligned}$$

Because $\boldsymbol{W}$ is a mixing matrix, we obtain

$$\sum_{i=1}^{N} \boldsymbol{c}_i^{(r+1)} = \sum_{j=1}^{N} \boldsymbol{c}_j^{(r)}.$$

Since we have

$$\sum_{i=1}^{N} \boldsymbol{c}_i^{(0)} = \frac{1}{1-\beta}\sum_{i=1}^{N}\left(\nabla F_i(\boldsymbol{x}_i^{(0)}; \xi_i^{(0)}) - \frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\boldsymbol{x}_j^{(0)}; \xi_j^{(0)})\right) = \mathbf{0},$$

we obtain the statement. □

Lemma 5. *For any round $r \geq 0$, it holds that*

$$\bar{\boldsymbol{x}}^{(r+1)} = \bar{\boldsymbol{x}}^{(r)} - \eta\bar{\boldsymbol{u}}^{(r+1)}.$$

Proof. We have

$$\begin{aligned}\bar{\boldsymbol{x}}^{(r+1)} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}\boldsymbol{x}_j^{(r)} - \eta(\bar{\boldsymbol{u}}^{(r+1)} - \bar{\boldsymbol{c}}^{(r)}) \\ &= \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}_j^{(r)}\sum_{i=1}^{N} W_{ij} - \eta(\bar{\boldsymbol{u}}^{(r+1)} - \bar{\boldsymbol{c}}^{(r)}).\end{aligned}$$

The fact that $\boldsymbol{W}$ is a mixing matrix gives us

$$\bar{\boldsymbol{x}}^{(r+1)} = \bar{\boldsymbol{x}}^{(r)} - \eta(\bar{\boldsymbol{u}}^{(r+1)} - \bar{\boldsymbol{c}}^{(r)}).$$

Then, using Lemma 4, we get the statement. □

Lemma 6. *For any round $r \geq 0$, it holds that*

$$\bar{\boldsymbol{z}}^{(r+1)} - \bar{\boldsymbol{z}}^{(r)} = -\frac{\eta}{1-\beta}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(r)}; \xi_i^{(r)}).$$

Proof. For any $r \geq 1$, we have

$$\begin{aligned}\bar{\boldsymbol{z}}^{(r+1)} - \bar{\boldsymbol{z}}^{(r)} &= \frac{1}{1-\beta}(\bar{\boldsymbol{x}}^{(r+1)} - \bar{\boldsymbol{x}}^{(r)}) - \frac{\beta}{1-\beta}(\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{x}}^{(r-1)}) \\ &= -\frac{\eta}{1-\beta}\bar{\boldsymbol{u}}^{(r+1)} + \frac{\eta\beta}{1-\beta}\bar{\boldsymbol{u}}^{(r)} \\ &= -\frac{\eta}{1-\beta}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(r)}; \xi_i^{(r)}),\end{aligned}$$

where we use Lemma 5. When $r = 0$, we have

$$\begin{aligned}\bar{\boldsymbol{z}}^{(1)} - \bar{\boldsymbol{z}}^{(0)} &= \frac{1}{1-\beta}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(0)}) \\ &= -\frac{\eta}{1-\beta}\bar{\boldsymbol{u}}^{(1)} \\ &= -\frac{\eta}{1-\beta}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(0)}; \xi_i^{(0)}),\end{aligned}$$

where we use $\bar{\boldsymbol{u}}^{(0)} = \mathbf{0}$. This concludes the proof. □

Lemma 7. *Suppose that Assumptions 1, 2, 3, and 4 hold. For any round $R \geq 0$, it holds that*

$$\sum_{r=0}^{R}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 \leq \frac{\beta^2\eta^2}{(1-\beta)^4}\sum_{r=0}^{R}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 + \frac{\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R.$$

Proof. From Lemma 5, we have

$$\begin{aligned}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 &= \mathbb{E}\left\|\frac{\eta\beta}{1-\beta}\bar{\boldsymbol{u}}^{(r)}\right\|^2 \\ &= \frac{\beta^2\eta^2}{(1-\beta)^2}\mathbb{E}\left\|\sum_{k=0}^{r-1}\beta^{r-k-1}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(k)}; \xi_i^{(k)})\right\|^2,\end{aligned}$$

for any $r \geq 1$. Defining $s^{(r)} \coloneqq \sum_{k=0}^{r}\beta^{r-k}$, we obtain

$$\begin{aligned}&\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 \\ &= \frac{\beta^2\eta^2}{(1-\beta)^2}{s^{(r-1)}}^2\mathbb{E}\left\|\sum_{k=0}^{r-1}\frac{\beta^{r-k-1}}{s^{(r-1)}}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(k)}; \xi_i^{(k)})\right\|^2 \\ &\overset{(a)}{\leq} \frac{\beta^2\eta^2}{(1-\beta)^2}s^{(r-1)}\sum_{k=0}^{r-1}\beta^{r-k-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(k)}; \xi_i^{(k)})\right\|^2 \\ &\overset{(6)}{\leq} \frac{\beta^2\eta^2}{(1-\beta)^2}s^{(r-1)}\sum_{k=0}^{r-1}\beta^{r-k-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(k)})\right\|^2 + \frac{\beta^2\sigma^2\eta^2}{N(1-\beta)^2}s^{(r-1)}\sum_{k=0}^{r-1}\beta^{r-k-1},\end{aligned}$$

where we use Jensen's inequality in (a). Using $s^{(r-1)} \leq \frac{1}{1-\beta}$, we obtain

$$\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 \leq \frac{\beta^2\eta^2}{(1-\beta)^3}\sum_{k=0}^{r-1}\beta^{r-k-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(k)})\right\|^2 + \frac{\beta^2\sigma^2\eta^2}{N(1-\beta)^4}.$$

Recursive addition yields

$$\begin{aligned}\sum_{r=1}^{R}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 &\leq \frac{\beta^2\eta^2}{(1-\beta)^3}\sum_{r=1}^{R}\sum_{k=0}^{r-1}\beta^{r-k-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(k)})\right\|^2 + \frac{\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R \\ &= \frac{\beta^2\eta^2}{(1-\beta)^3}\sum_{k=0}^{R-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(k)})\right\|^2\sum_{r=k+1}^{R}\beta^{r-k-1} + \frac{\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R \\ &\leq \frac{\beta^2\eta^2}{(1-\beta)^4}\sum_{k=0}^{R-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(k)})\right\|^2 + \frac{\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R,\end{aligned}$$

where we use $\sum_{r=k+1}^{R}\beta^{r-k-1} \leq \frac{1}{1-\beta}$ in the last inequality. From the definition of $\bar{\boldsymbol{z}}^{(0)}$, we have $\|\bar{\boldsymbol{x}}^{(0)} - \bar{\boldsymbol{z}}^{(0)}\|^2 = 0$. This yields the statement. □

Lemma 8. *Suppose that Assumptions 1, 2, 3, and 4 hold. For any round $r \geq 0$, it holds that*

$$\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r+1)} - \bar{\boldsymbol{x}}^{(r)}\right\|^2 \leq 4L^2\eta^2\Xi^{(r)} + 2\beta^2\eta^2\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 + 4\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{\sigma^2\eta^2}{N}.$$

Proof. From Lemma 5, we have

$$\begin{aligned}
\mathbb{E}_r\left\|\bar{\boldsymbol{x}}^{(r+1)} - \bar{\boldsymbol{x}}^{(r)}\right\|^2 &= \eta^2\mathbb{E}_r\left\|\beta\bar{\boldsymbol{u}}^{(r)} + \frac{1}{N}\sum_{i=1}^N \nabla F_i(\boldsymbol{x}_i^{(r)}; \xi_i^{(r)})\right\|^2 \\
&\overset{(6)}{\leq} \eta^2\left\|\beta\bar{\boldsymbol{u}}^{(r)} + \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 + \frac{\sigma^2\eta^2}{N} \\
&\overset{(13)}{\leq} 2\beta^2\eta^2\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 + 2\eta^2\underbrace{\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2}_{T} + \frac{\sigma^2\eta^2}{N}.
\end{aligned}$$

Then, T can be bounded from above as follows:

$$\begin{aligned}
T &= \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)}) \pm \nabla f_i(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
&\overset{(13)}{\leq} 2\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)}) - \nabla f_i(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + 2\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
&\overset{(14)}{\leq} \frac{2}{N}\sum_{i=1}^N \left\|\nabla f_i(\boldsymbol{x}_i^{(r)}) - \nabla f_i(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + 2\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
&\overset{(5)}{\leq} \frac{2L^2}{N}\sum_{i=1}^N \left\|\boldsymbol{x}_i^{(r)} - \bar{\boldsymbol{x}}^{(r)}\right\|^2 + 2\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2.
\end{aligned}$$

Then, we obtain the statement. □

Lemma 9. *For any round $r \geq 0$, it holds that*

$$\mathbb{E}\left\|\bar{\boldsymbol{e}}^{(r+1)}\right\|^2 \leq \frac{1}{1-\beta}\sum_{k=0}^r \beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2.$$

Proof. We have

$$\mathbb{E}\left\|\bar{\boldsymbol{e}}^{(r+1)}\right\|^2 = \mathbb{E}\left\|\sum_{k=0}^r \beta^{r-k}\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2,$$

where we use $\bar{\boldsymbol{e}}^{(0)} = \boldsymbol{0}$. Defining $s^{(r)} \coloneqq \sum_{k=0}^r \beta^{r-k}$, we obtain

$$\begin{aligned}
\mathbb{E}\left\|\bar{\boldsymbol{e}}^{(r+1)}\right\|^2 &= {s^{(r)}}^2\mathbb{E}\left\|\sum_{k=0}^r \frac{\beta^{r-k}}{s^{(r)}}\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\leq s^{(r)}\sum_{k=0}^r \beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2,
\end{aligned}$$

where we use Jensen's inequality. Using $s^{(r)} \leq \frac{1}{1-\beta}$, we obtain the statement. □

Lemma 10. *Suppose that Assumptions 1, 2, 3, and 4 hold. For any round $r \geq 0$, it holds that*

$$\frac{1}{N}\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{U}^{(r+1)}\right\|_F^2 \leq \frac{L^2}{1-\beta}\sum_{k=0}^r \beta^{r-k}\Xi^{(k)} + \frac{5\sigma^2}{(1-\beta)^3}.$$

Proof. We have

$$\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{U}^{(r+1)}\right\|_F^2 = \mathbb{E}\left\|\sum_{k=0}^{r}\beta^{r-k}(\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla F(\boldsymbol{X}^{(k)};\xi^{(k)}))+\beta^{r+1}(\boldsymbol{D}^{(0)}-\boldsymbol{U}^{(0)})\right\|_F^2.$$

Defining $s^{(r)} \coloneqq \sum_{k=0}^{r}\beta^{r-k}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{U}^{(r+1)}\right\|_F^2 \\
&= {s^{(r+1)}}^2\mathbb{E}\left\|\sum_{k=0}^{r}\frac{\beta^{r-k}}{s^{(r+1)}}(\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla F(\boldsymbol{X}^{(k)};\xi^{(k)}))+\frac{\beta^{r+1}}{s^{(r+1)}}(\boldsymbol{D}^{(0)}-\boldsymbol{U}^{(0)})\right\|_F^2 \\
&\overset{(a)}{\leq} s^{(r+1)}\sum_{k=0}^{r}\beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla F(\boldsymbol{X}^{(k)};\xi^{(k)})\right\|_F^2+s^{(r+1)}\beta^{r+1}\mathbb{E}\left\|\boldsymbol{D}^{(0)}-\boldsymbol{U}^{(0)}\right\|_F^2 \\
&\overset{(14)}{\leq} s^{(r+1)}\sum_{k=0}^{r}\beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla F(\boldsymbol{X}^{(k)};\xi^{(k)})\right\|_F^2 \\
&\quad+\frac{2s^{(r+1)}}{(1-\beta)^2}\beta^{r+1}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(0)})-\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\right\|_F^2 \\
&\quad+\frac{2s^{(r+1)}}{(1-\beta)^2}\beta^{r+1}\mathbb{E}\left\|\frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(0)})\mathbf{1}\mathbf{1}^\top-\frac{1}{N}\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\mathbf{1}\mathbf{1}^\top\right\|_F^2 \\
&\overset{(6)}{\leq} s^{(r+1)}\sum_{k=0}^{r}\beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla f(\boldsymbol{X}^{(k)})\right\|_F^2+s^{(r+1)}\sum_{k=0}^{r}\beta^{r-k}N\sigma^2+\frac{4s^{(r+1)}}{(1-\beta)^2}\beta^{r+1}N\sigma^2,
\end{aligned}$$

where we use Jensen's inequality for (a) and use $\boldsymbol{X}^{(0)}=\bar{\boldsymbol{X}}^{(0)}$ for the last inequality. Then, using $s^{(r)}\leq\frac{1}{1-\beta}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{U}^{(r+1)}\right\|_F^2 \\
&\leq\frac{1}{1-\beta}\sum_{k=0}^{r}\beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla f(\boldsymbol{X}^{(k)})\right\|_F^2+\frac{N\sigma^2}{(1-\beta)^2}+\frac{4N\sigma^2}{(1-\beta)^3}\beta^{r+1} \\
&\overset{\beta\in[0,1)}{\leq}\frac{1}{1-\beta}\sum_{k=0}^{r}\beta^{r-k}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(k)})-\nabla f(\boldsymbol{X}^{(k)})\right\|_F^2+\frac{5N\sigma^2}{(1-\beta)^3} \\
&\overset{(5)}{\leq}\frac{L^2}{1-\beta}\sum_{k=0}^{r}\beta^{r-k}\mathbb{E}\left\|\bar{\boldsymbol{X}}^{(k)}-\boldsymbol{X}^{(k)}\right\|_F^2+\frac{5N\sigma^2}{(1-\beta)^3}.
\end{aligned}$$

This concludes the proof. □

Lemma 11. *Suppose that Assumptions 1, 2, 3, and 4 hold. For any round $r\geq 0$, it holds that*

$$\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)}-\bar{\boldsymbol{d}}^{(r+1)}\right\|^2\leq\frac{L^2}{1-\beta}\sum_{k=0}^{r}\beta^{r-k}\Xi^{(k)}+\frac{\sigma^2}{N(1-\beta)^2}$$

Proof. We have

$$\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)}-\bar{\boldsymbol{d}}^{(r+1)}\right\|^2=\mathbb{E}\left\|\sum_{k=0}^{r}\beta^{r-k}(\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{x}_i^{(k)};\xi_i^{(k)})-\nabla f(\bar{\boldsymbol{x}}^{(k)}))\right\|^2,$$

where we use $\bar{\boldsymbol{u}}^{(0)} = \bar{\boldsymbol{d}}^{(0)} = \mathbf{0}$. Defining $s^{(r)} \coloneqq \sum_{k=0}^{r} \beta^{r-k}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)} - \bar{\boldsymbol{d}}^{(r+1)}\right\|^2 \\
&= {s^{(r)}}^2 \mathbb{E}\left\|\sum_{k=0}^{r} \frac{\beta^{r-k}}{s^{(r)}} \left(\frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\boldsymbol{x}_i^{(k)}; \xi_i^{(k)}) - \nabla f(\bar{\boldsymbol{x}}^{(k)})\right)\right\|^2 \\
&\overset{(a)}{\leq} s^{(r)} \sum_{k=0}^{r} \beta^{r-k} \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\boldsymbol{x}_i^{(k)}; \xi_i^{(k)}) - \nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\
&\overset{(6)}{\leq} s^{(r)} \sum_{k=0}^{r} \beta^{r-k} \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\boldsymbol{x}_i^{(k)}) - \nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 + s^{(r)} \sum_{k=0}^{r} \beta^{r-k} \frac{\sigma^2}{N} \\
&\overset{(14)}{\leq} s^{(r)} \sum_{k=0}^{r} \beta^{r-k} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\|\nabla f_i(\boldsymbol{x}_i^{(k)}) - \nabla f_i(\bar{\boldsymbol{x}}^{(k)})\right\|^2 + s^{(r)} \sum_{k=0}^{r} \beta^{r-k} \frac{\sigma^2}{N},
\end{aligned}
$$

where we use Jensen's inequality in (a). Then, using $s^{(r)} \leq \frac{1}{1-\beta}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)} - \bar{\boldsymbol{d}}^{(r+1)}\right\|^2 \\
&\leq \frac{1}{1-\beta} \sum_{k=0}^{r} \beta^{r-k} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\|\nabla f_i(\boldsymbol{x}_i^{(k)}) - \nabla f_i(\bar{\boldsymbol{x}}^{(k)})\right\|^2 + \frac{\sigma^2}{N(1-\beta)^2} \\
&\overset{(5)}{\leq} \frac{L^2}{1-\beta} \sum_{k=0}^{r} \beta^{r-k} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\|\boldsymbol{x}_i^{(k)} - \bar{\boldsymbol{x}}^{(k)}\right\|^2 + \frac{\sigma^2}{N(1-\beta)^2}.
\end{aligned}
$$

This concludes the proof. □

Lemma 12. *Suppose that Assumptions 1, 2, 3, and 4 hold. For any round $r \geq 0$, it holds that*

$$
\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)}\right\|^2 \leq \frac{2L^2}{1-\beta} \sum_{k=0}^{r} \beta^{r-k} \Xi^{(k)} + \frac{2}{1-\beta} \sum_{k=0}^{r} \beta^{r-k} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 + \frac{2\sigma^2}{N(1-\beta)^2}.
$$

Proof. We have

$$
\begin{aligned}
\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)}\right\|^2 &= \mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)} \pm \bar{\boldsymbol{d}}^{(r+1)}\right\|^2 \\
&\overset{(13)}{\leq} 2\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r+1)} - \bar{\boldsymbol{d}}^{(r+1)}\right\|^2 + 2\mathbb{E}\left\|\bar{\boldsymbol{d}}^{(r+1)}\right\|^2.
\end{aligned}
$$

From Lemmas 9 and 11, we obtain the statement. □

D.5 MAIN PROOF

Lemma 13 (Descent Lemma)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. If the step size η satisfies*

$$\eta \leq \frac{1-\beta}{4L},$$

then it holds that for any round $r \geq 0$,

$$\mathbb{E} f(\bar{\boldsymbol{z}}^{(r+1)}) \leq \mathbb{E} f(\bar{\boldsymbol{z}}^{(r)}) + \frac{L^2\eta}{1-\beta}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 + \frac{L^2\eta}{1-\beta}\Xi^{(r)} - \frac{\eta}{4(1-\beta)}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 - \frac{\eta}{4(1-\beta)}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{L\sigma^2\eta^2}{2N(1-\beta)^2}.$$

Proof. From Assumption 3 and Lemma 6, we have

$$\begin{aligned}
&\mathbb{E}_r f(\bar{\boldsymbol{z}}^{(r+1)}) \\
&\leq f(\bar{\boldsymbol{z}}^{(r)}) + \mathbb{E}_r\langle \nabla f(\bar{\boldsymbol{z}}^{(r)}), \bar{\boldsymbol{z}}^{(r+1)} - \bar{\boldsymbol{z}}^{(r)}\rangle + \frac{L}{2}\mathbb{E}_r\left\|\bar{\boldsymbol{z}}^{(r+1)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 \\
&= f(\bar{\boldsymbol{z}}^{(r)}) - \frac{\eta}{1-\beta}\left\langle \nabla f(\bar{\boldsymbol{z}}^{(r)}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\rangle + \frac{L\eta^2}{2(1-\beta)^2}\mathbb{E}_r\left\|\frac{1}{N}\sum_{i=1}^N \nabla F_i(\boldsymbol{x}_i^{(r)}; \xi_i^{(r)})\right\|^2 \\
&\overset{(6)}{\leq} f(\bar{\boldsymbol{z}}^{(r)}) - \frac{\eta}{1-\beta}\left\langle \nabla f(\bar{\boldsymbol{z}}^{(r)}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\rangle \\
&\quad + \frac{L\eta^2}{2(1-\beta)^2}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 + \frac{L\sigma^2\eta^2}{2N(1-\beta)^2} \\
&= f(\bar{\boldsymbol{z}}^{(r)}) + \frac{\eta}{1-\beta}\underbrace{\left\langle \nabla f(\bar{\boldsymbol{x}}^{(r)}) - \nabla f(\bar{\boldsymbol{z}}^{(r)}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\rangle}_{T_1} \\
&\quad - \frac{\eta}{1-\beta}\underbrace{\left\langle \nabla f(\bar{\boldsymbol{x}}^{(r)}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\rangle}_{T_2} + \frac{L\eta^2}{2(1-\beta)^2}\underbrace{\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2}_{T_3} + \frac{L\sigma^2\eta^2}{2N(1-\beta)^2}.
\end{aligned}$$

We can bound T_1 from above as follows:

$$\begin{aligned}
T_1 &\overset{(15),\gamma=2}{\leq} \left\|\nabla f(\bar{\boldsymbol{x}}^{(r)}) - \nabla f(\bar{\boldsymbol{z}}^{(r)})\right\|^2 + \frac{1}{4}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 \\
&\overset{(5)}{\leq} L^2\left\|\bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)}\right\|^2 + \frac{1}{4}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2.
\end{aligned}$$

We can bound $-T_2$ from above as follows:

$$\begin{aligned}
-T_2 &= \frac{1}{2}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)}) - \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 - \frac{1}{2}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 - \frac{1}{2}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 \\
&\overset{(14)}{\leq} \frac{1}{2}\frac{1}{N}\sum_{i=1}^N\left\|\nabla f_i(\bar{\boldsymbol{x}}^{(r)}) - \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 - \frac{1}{2}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 - \frac{1}{2}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2 \\
&\overset{(5)}{\leq} \frac{L^2}{2}\frac{1}{N}\sum_{i=1}^N\left\|\bar{\boldsymbol{x}}^{(r)} - \boldsymbol{x}_i^{(r)}\right\|^2 - \frac{1}{2}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 - \frac{1}{2}\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2.
\end{aligned}$$

Then, we can bound T_3 from above as follows:

$$\begin{aligned}
T_3 &= \left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)}) \pm \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 \\
&\overset{(13)}{\leq} 2\left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)}) - \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 + 2\left\| \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 \\
&\overset{(14)}{\leq} \frac{2}{N}\sum_{i=1}^N \left\| \nabla f_i(\boldsymbol{x}_i^{(r)}) - \nabla f_i(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 + 2\left\| \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 \\
&\overset{(5)}{\leq} \frac{2L^2}{N}\sum_{i=1}^N \left\| \boldsymbol{x}_i^{(r)} - \bar{\boldsymbol{x}}^{(r)} \right\|^2 + 2\left\| \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 .
\end{aligned}$$

By combining them, we obtain

$$\begin{aligned}
&\mathbb{E}_r f(\bar{\boldsymbol{z}}^{(r+1)}) \\
&\leq f(\bar{\boldsymbol{z}}^{(r)}) + \frac{L^2\eta}{1-\beta}\left\| \bar{\boldsymbol{x}}^{(r)} - \bar{\boldsymbol{z}}^{(r)} \right\|^2 - \frac{\eta}{4(1-\beta)}\left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\boldsymbol{x}_i^{(r)}) \right\|^2 \\
&\quad + \frac{L^2}{1-\beta}\left(\frac{1}{2} + \frac{L\eta}{1-\beta}\right)\eta\Xi^{(r)} - \frac{1}{1-\beta}\left(\frac{1}{2} - \frac{L\eta}{1-\beta}\right)\eta\left\| \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 + \frac{L\sigma^2\eta^2}{2N(1-\beta)^2}.
\end{aligned}$$

Using $\eta \leq \frac{1-\beta}{4L}$, we get the statement. □

Lemma 14 (Recursion for Ξ)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Then, it holds that for any round $r \geq 0$,*

$$\Xi^{(r+1)} \leq (1-\frac{p}{2})\Xi^{(r)} + \frac{9}{p}\eta^2\mathcal{E}^{(r)} + \frac{9}{Np}\eta^2\mathbb{E}\left\| \boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)} \right\|_F^2 + \frac{9}{Np}\eta^2\mathbb{E}\left\| \boldsymbol{E}^{(r+1)} \right\|_F^2 .$$

Proof. Because $\sum_{i=1}^N \|\boldsymbol{a}_i - \bar{\boldsymbol{a}}\|^2 \leq \sum_{i=1}^N \|\boldsymbol{a}_i\|^2$ for any $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N \in \mathbb{R}^d$, we have

$$N\Xi^{(r)} = \mathbb{E}\left\| (\boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r-1)}) + (\bar{\boldsymbol{X}}^{(r-1)} - \bar{\boldsymbol{X}}^{(r)}) \right\|_F^2 \leq \mathbb{E}\left\| \boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r-1)} \right\|_F^2 .$$

Then, we have

$$\begin{aligned}
\left\| \boldsymbol{X}^{(r+1)} - \bar{\boldsymbol{X}}^{(r+1)} \right\|_F^2 &\leq \left\| \boldsymbol{X}^{(r+1)} - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 \\
&= \left\| \boldsymbol{X}^{(r)}\boldsymbol{W} - \eta(\boldsymbol{U}^{(r+1)} - \boldsymbol{C}^{(r)}) - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 \\
&\overset{(13)}{\leq} (1+\gamma)\left\| \boldsymbol{X}^{(r)}\boldsymbol{W} - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 + (1+\gamma^{-1})\eta^2\left\| \boldsymbol{U}^{(r+1)} - \boldsymbol{C}^{(r)} \right\|_F^2 \\
&\overset{(4)}{\leq} (1+\gamma)(1-p)\left\| \boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 + (1+\gamma^{-1})\eta^2\left\| \boldsymbol{U}^{(r+1)} - \boldsymbol{C}^{(r)} \right\|_F^2 .
\end{aligned}$$

By substituting $\gamma = \frac{p}{2}$ and using $p \leq 1$, we obtain

$$\begin{aligned}
&\left\| \boldsymbol{X}^{(r+1)} - \bar{\boldsymbol{X}}^{(r+1)} \right\|_F^2 \\
&\leq (1-\frac{p}{2})\left\| \boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 + \frac{3}{p}\eta^2\left\| \boldsymbol{U}^{(r+1)} - \boldsymbol{C}^{(r)} \right\|_F^2 \\
&= (1-\frac{p}{2})\left\| \boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 + \frac{3}{p}\eta^2\left\| \boldsymbol{U}^{(r+1)} \pm \boldsymbol{D}^{(r+1)} \pm \boldsymbol{E}^{(r+1)} - \boldsymbol{C}^{(r)} \right\|_F^2 \\
&\overset{(14)}{\leq} (1-\frac{p}{2})\left\| \boldsymbol{X}^{(r)} - \bar{\boldsymbol{X}}^{(r)} \right\|_F^2 \\
&\quad + \frac{9}{p}\eta^2\left\| \boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)} \right\|_F^2 + \frac{9}{p}\eta^2\left\| \boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)} - \boldsymbol{E}^{(r+1)} \right\|_F^2 + \frac{9}{p}\eta^2\left\| \boldsymbol{E}^{(r+1)} \right\|_F^2 .
\end{aligned}$$

This concludes the proof. □

Lemma 15 (Recursion for $\mathcal{E}$)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Then, it holds that for any round $r \geq 0$,*

$$\mathcal{E}^{(r+1)} \leq (1-\frac{p}{2})\mathcal{E}^{(r)} + \frac{18\beta^2}{p}\mathcal{D}^{(r)} + \frac{24}{Np}\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 + \frac{144L^4}{p}\eta^2\Xi^{(r)} + \frac{72\beta^2L^2}{p}\eta^2\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 + \frac{144L^2}{p}\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{36L^2\sigma^2\eta^2}{Np}.$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{C}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2 \\
&= \mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - (\boldsymbol{C}^{(r)} - \boldsymbol{U}^{(r+1)})\boldsymbol{W} - \boldsymbol{U}^{(r+1)} - \boldsymbol{E}^{(r+2)} \pm \boldsymbol{D}^{(r+1)} \pm \boldsymbol{D}^{(r+1)}\boldsymbol{W} \pm \boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\overset{(13),(14)}{\leq} (1+\gamma)\mathbb{E}\left\|(\boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)})\boldsymbol{W} - \boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|(\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)})(\boldsymbol{W} - \boldsymbol{I})\right\|_F^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{D}^{(r+1)} + \boldsymbol{E}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2 \\
&\overset{(4)}{\leq} (1+\gamma)(1-p)\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)} - \boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|(\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)})(\boldsymbol{W} - \boldsymbol{I})\right\|_F^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{D}^{(r+1)} + \boldsymbol{E}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2,
\end{aligned}$$

where we use Lemma 4 and $\boldsymbol{E}^{(r+1)} = \frac{1}{N}\boldsymbol{D}^{(r+1)}\mathbf{1}\mathbf{1}^\top$ in the last inequality. Then, we have

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{C}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2 \\
&\overset{(a)}{\leq} (1+\gamma)(1-p)\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)} - \boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 \|\boldsymbol{W} - \boldsymbol{I}\|_{\mathrm{op}}^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{D}^{(r+1)} + \boldsymbol{E}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2 \\
&\overset{(b)}{\leq} (1+\gamma)(1-p)\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)} - \boldsymbol{E}^{(r+1)}\right\|_F^2 + 8(1+\gamma^{-1})\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 \\
&\quad + 2(1+\gamma^{-1})\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{D}^{(r+1)} + \boldsymbol{E}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2,
\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. In (a), we use the following definition of the operator norm: $\|\boldsymbol{W} - \boldsymbol{I}\|_{\mathrm{op}} \coloneqq \sup_{\hat{\boldsymbol{v}} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{\|(\boldsymbol{W}-\boldsymbol{I})\hat{\boldsymbol{v}}\|}{\|\hat{\boldsymbol{v}}\|} \geq \frac{\|(\boldsymbol{W}-\boldsymbol{I})\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$ for any $\boldsymbol{v} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. In (b), we use Gershgorin circle theorem and the fact that $\boldsymbol{W}$ is a mixing matrix. Substituting $\gamma = \frac{p}{2}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{C}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2 \\
&\leq (1-\frac{p}{2})\mathbb{E}\left\|\boldsymbol{D}^{(r+1)} - \boldsymbol{C}^{(r)} - \boldsymbol{E}^{(r+1)}\right\|_F^2 + \frac{24}{p}\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 \\
&\quad + \frac{6}{p}\underbrace{\mathbb{E}\left\|\boldsymbol{D}^{(r+2)} - \boldsymbol{D}^{(r+1)} + \boldsymbol{E}^{(r+1)} - \boldsymbol{E}^{(r+2)}\right\|_F^2}_{T}.
\end{aligned}$$

Then, we can bound T from above by expanding $\boldsymbol{D}^{(r+2)}$, $\boldsymbol{D}^{(r+1)}$, $\boldsymbol{E}^{(r+2)}$, and $\boldsymbol{E}^{(r+1)}$ as follows:

$$\begin{aligned}
T &\overset{(14)}{\leq} 3\beta^2\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2+3\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(r+1)})-\nabla f(\bar{\boldsymbol{X}}^{(r)})\right\|_F^2 \\
&\quad+3\mathbb{E}\left\|\frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(r)})\mathbf{1}\mathbf{1}^\top-\frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(r+1)})\mathbf{1}\mathbf{1}^\top\right\|_F^2 \\
&\overset{(5)}{\leq} 3\beta^2\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2+6L^2\mathbb{E}\left\|\bar{\boldsymbol{X}}^{(r+1)}-\bar{\boldsymbol{X}}^{(r)}\right\|_F^2.
\end{aligned}$$

Using Lemma 8, we obtain

$$\begin{aligned}
T &\leq 3\beta^2\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad+N(24L^4\eta^2\Xi^{(r)}+12\beta^2L^2\eta^2\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2+24L^2\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2+\frac{6L^2\sigma^2\eta^2}{N}).
\end{aligned}$$

This concludes the proof. □

Lemma 16 (Recursion for $\mathcal{D}$)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Then, it holds that for any round $r \geq 0$,*

$$\begin{aligned}
\mathcal{D}^{(r+1)} &\leq \frac{2\beta^2}{1+\beta^2}\mathcal{D}^{(r)}+\frac{32L^4\eta^2}{1-\beta^2}\Xi^{(r)}+\frac{16L^2\beta^2\eta^2}{1-\beta^2}\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 \\
&\quad+\frac{32L^2\eta^2}{1-\beta^2}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2+\frac{8L^2\sigma^2\eta^2}{N(1-\beta^2)}.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+2)}-\boldsymbol{D}^{(r+1)}-\boldsymbol{E}^{(r+2)}+\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\overset{(13),(14)}{\leq}(1+\gamma)\beta^2\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad+2(1+\gamma^{-1})\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(r+1)})-\nabla f(\bar{\boldsymbol{X}}^{(r)})\right\|_F^2 \\
&\quad+2(1+\gamma^{-1})\mathbb{E}\left\|\frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(r)})\mathbf{1}\mathbf{1}^\top-\frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(r+1)})\mathbf{1}\mathbf{1}^\top\right\|_F^2 \\
&\overset{(5)}{\leq}(1+\gamma)\beta^2\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2+4L^2(1+\gamma^{-1})\mathbb{E}\left\|\bar{\boldsymbol{X}}^{(r+1)}-\bar{\boldsymbol{X}}^{(r)}\right\|_F^2.
\end{aligned}$$

Substituting $\gamma=\frac{1-\beta^2}{1+\beta^2}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+2)}-\boldsymbol{D}^{(r+1)}-\boldsymbol{E}^{(r+2)}+\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\leq\frac{2\beta^2}{1+\beta^2}\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2+\frac{8L^2}{1-\beta^2}\mathbb{E}\left\|\bar{\boldsymbol{X}}^{(r+1)}-\bar{\boldsymbol{X}}^{(r)}\right\|_F^2.
\end{aligned}$$

Using Lemma 8, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{D}^{(r+2)}-\boldsymbol{D}^{(r+1)}-\boldsymbol{E}^{(r+2)}+\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\leq\frac{2\beta^2}{1+\beta^2}\mathbb{E}\left\|\boldsymbol{D}^{(r+1)}-\boldsymbol{D}^{(r)}+\boldsymbol{E}^{(r)}-\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad+N(\frac{32L^4\eta^2}{1-\beta^2}\Xi^{(r)}+\frac{16L^2\beta^2\eta^2}{1-\beta^2}\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2+\frac{32L^2\eta^2}{1-\beta^2}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2+\frac{8L^2\sigma^2\eta^2}{N(1-\beta^2)}).
\end{aligned}$$

This concludes the proof.

□

Lemma 17 (Recursion for Ξ, $\mathcal{E}$, and $\mathcal{D}$)**.** *We define $t \in \mathbb{R}$ and $A \in \mathbb{R}$ as follows:*

$$t \coloneqq \frac{2\beta^2 p}{1-\beta^2} + 4, \quad A \coloneqq \frac{648}{1 - \frac{p}{t} - \frac{2\beta^2}{1+\beta^2}}.$$

Note that it holds that $t \geq 4$ and $A > 0$. Suppose that Assumptions 1, 2, 3, and 4 hold, and step size η satisfies

$$\eta \leq \frac{p}{8L\sqrt{324 + \frac{2A\beta^2}{1-\beta^2}}}.$$

Then, it holds that

$$\begin{aligned}
&\Xi^{(r+1)} + \frac{36}{p^2}\eta^2\mathcal{E}^{(r+1)} + \frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r+1)} \\
&\leq (1 - \frac{p}{t})(\Xi^{(r)} + \frac{36}{p^2}\eta^2\mathcal{E}^{(r)} + \frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r)}) \\
&\quad + \frac{1}{p}\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{1}{Np}\left(9 + \frac{864}{p^2}\right)\eta^2\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 \\
&\quad + \frac{L^2}{p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 + \frac{9}{Np}\eta^2\mathbb{E}\left\|\boldsymbol{E}^{(r+1)}\right\|_F^2 + \frac{\sigma^2\eta^2}{p}.
\end{aligned}$$

Proof. From Lemmas 14 and 15, we have

$$\begin{aligned}
\Xi^{(r+1)} &\leq (1 - \frac{p}{2})\Xi^{(r)} + \frac{9}{p}\eta^2\mathcal{E}^{(r)} + \frac{9}{Np}\eta^2\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 + \frac{9}{Np}\eta^2\mathbb{E}\left\|\boldsymbol{E}^{(r+1)}\right\|_F^2, \\
\frac{36}{p^2}\eta^2\mathcal{E}^{(r+1)} &\leq (1 - \frac{p}{2})\frac{36}{p^2}\eta^2\mathcal{E}^{(r)} + \frac{648\beta^2}{p^3}\eta^2\mathcal{D}^{(r)} + \frac{5184L^4}{p^3}\eta^4\Xi^{(r)} + \frac{2592\beta^2L^2}{p^3}\eta^4\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 \\
&\quad + \frac{864}{Np^3}\eta^2\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 + \frac{5184L^2}{p^3}\eta^4\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{1296L^2\sigma^2\eta^4}{Np^3}.
\end{aligned}$$

Then, from Lemma 16, we have

$$\begin{aligned}
\frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r+1)} &\leq \frac{2A\beta^4}{(1+\beta^2)p^3}\eta^2\mathcal{D}^{(r)} + \frac{32A\beta^2L^4}{(1-\beta^2)p^3}\eta^4\Xi^{(r)} + \frac{16AL^2\beta^4}{(1-\beta^2)p^3}\eta^4\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 \\
&\quad + \frac{32A\beta^2L^2}{(1-\beta^2)p^3}\eta^4\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{8A\beta^2L^2\sigma^2}{N(1-\beta^2)p^3}\eta^4.
\end{aligned}$$

Using $\eta^2 \leq \frac{p^2}{L^2}$ and $\eta^2 \leq \frac{p^2}{128L^2(162 + \frac{A\beta^2}{1-\beta^2})}$, we have

$$\left((1 - \frac{p}{2}) + \frac{5184L^4}{p^3}\eta^4 + \frac{32A\beta^2L^4}{(1-\beta^2)p^3}\eta^4\right)\Xi^{(r)} \leq \left((1 - \frac{p}{2}) + \frac{L^2}{4p}\eta^2\right)\Xi^{(r)} \leq (1 - \frac{p}{4})\Xi^{(r)}.$$

In addition, we have

$$\begin{aligned}
\left(\frac{9}{p}\eta^2 + (1 - \frac{p}{2})\frac{36}{p^2}\eta^2\right)\mathcal{E}^{(r)} &= (1 - \frac{p}{4})\frac{36}{p^2}\eta^2\mathcal{E}^{(r)}, \\
\left(\frac{648\beta^2}{p^3}\eta^2 + \frac{2A\beta^4}{(1+\beta^2)p^3}\eta^2\right)\mathcal{D}^{(r)} &= \left(\frac{648}{A} + \frac{2\beta^2}{1+\beta^2}\right)\frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r)} = (1 - \frac{p}{t})\frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r)}.
\end{aligned}$$

Then, using $t \geq 4$, we obtain

$$\begin{aligned}
&\Xi^{(r+1)} + \frac{36}{p^2}\eta^2\mathcal{E}^{(r+1)} + \frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r+1)} \\
&\leq (1-\frac{p}{t})(\Xi^{(r)} + \frac{36}{p^2}\eta^2\mathcal{E}^{(r)} + \frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r)}) \\
&\quad + \frac{L^2}{p^3}\left(\frac{32A\beta^2}{1-\beta^2} + 5184\right)\eta^4\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
&\quad + \frac{1}{Np}\left(9 + \frac{864}{p^2}\right)\eta^2\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 \\
&\quad + \frac{9}{Np}\eta^2\mathbb{E}\left\|\boldsymbol{E}^{(r+1)}\right\|_F^2 \\
&\quad + \frac{L^2}{p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 \\
&\quad + \frac{L^2\sigma^2}{Np^3}\left(1296 + \frac{8A\beta^2}{1-\beta^2}\right)\eta^4.
\end{aligned}$$

Using $\eta^2 \leq \frac{p^2}{32L^2(162+\frac{A\beta^2}{1-\beta^2})}$, we have

$$\frac{L^2}{p^3}\left(\frac{32A\beta^2}{1-\beta^2} + 5184\right)\eta^4\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \leq \frac{1}{p}\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2.$$

Using $\eta^2 \leq \frac{p^2}{8L^2(162+\frac{A\beta^2}{1-\beta^2})}$, we obtain

$$\frac{L^2\sigma^2}{Np^3}\left(1296 + \frac{8A\beta^2}{1-\beta^2}\right)\eta^4 \leq \frac{\sigma^2\eta^2}{Np} \leq \frac{\sigma^2\eta^2}{p}.$$

This concludes the proof. □

Lemma 18. *We define $t \in \mathbb{R}$ and $A \in \mathbb{R}$ as follows:*

$$t \coloneqq \frac{2\beta^2 p}{1-\beta^2} + 4, \quad A \coloneqq \frac{648}{1 - \frac{p}{t} - \frac{2\beta^2}{1+\beta^2}}.$$

Under the same assumptions as those in Lemma 17, it holds that

$$\sum_{r=0}^{R}\Xi^{(r)} \leq \frac{t}{p}\sum_{k=0}^{R}\Psi^{(k)} + \frac{145t\sigma^2\eta^2}{(1-\beta)^2p^3}R.$$

where $\Psi^{(r)}$ is defined as follows:

$$\begin{aligned}
\Psi^{(r)} \coloneqq &\frac{1}{p}\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{1}{Np}\left(9 + \frac{864}{p^2}\right)\eta^2\mathbb{E}\left\|\boldsymbol{U}^{(r+1)} - \boldsymbol{D}^{(r+1)}\right\|_F^2 \\
&+ \frac{L^2}{p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 + \frac{9}{Np}\eta^2\mathbb{E}\left\|\boldsymbol{E}^{(r+1)}\right\|_F^2.
\end{aligned}$$

Proof. We define $\Theta^{(r)} \coloneqq \Xi^{(r)} + \frac{36}{p^2}\eta^2\mathcal{E}^{(r)} + \frac{A\beta^2}{p^3}\eta^2\mathcal{D}^{(r)}$. From Lemma 17, we obtain

$$\begin{aligned}
\Theta^{(r+1)} &\leq (1-\frac{p}{t})\Theta^{(r)} + \Psi^{(r)} + \frac{\sigma^2\eta^2}{p} \\
&\leq (1-\frac{p}{t})^{r+1}\Theta^{(0)} + \sum_{k=0}^{r}(1-\frac{p}{t})^{r-k}\Psi^{(k)} + \sum_{k=0}^{r}(1-\frac{p}{t})^{r-k}\frac{\sigma^2\eta^2}{p}.
\end{aligned}$$

Using $\sum_{k=0}^{r}(1-\frac{p}{t})^{r-k} \leq \frac{t}{p}$, we obtain

$$\Theta^{(r+1)} \leq (1-\frac{p}{t})^{r+1}\Theta^{(0)} + \sum_{k=0}^{r}(1-\frac{p}{t})^{r-k}\Psi^{(k)} + \frac{t\sigma^2\eta^2}{p^2}.$$

Then, for any $R \geq 1$, we obtain

$$\begin{aligned}
\sum_{r=1}^{R}\Theta^{(r)} &\leq \sum_{r=1}^{R}(1-\frac{p}{t})^{r}\Theta^{(0)} + \sum_{r=1}^{R}\sum_{k=0}^{r-1}(1-\frac{p}{t})^{r-k-1}\Psi^{(k)} + \frac{t\sigma^2\eta^2}{p^2}R \\
&= \sum_{r=1}^{R}(1-\frac{p}{t})^{r}\Theta^{(0)} + \sum_{k=0}^{R-1}\Psi^{(k)}\sum_{r=k+1}^{R}(1-\frac{p}{t})^{r-k-1} + \frac{t\sigma^2\eta^2}{p^2}R \\
&\leq \frac{t}{p}\Theta^{(0)} + \frac{t}{p}\sum_{k=0}^{R-1}\Psi^{(k)} + \frac{t\sigma^2\eta^2}{p^2}R,
\end{aligned}$$

where we use $\sum_{r=1}^{R}(1-\frac{p}{t})^{r} \leq \frac{t}{p}$ and $\sum_{r=k+1}^{R}(1-\frac{p}{t})^{r-k-1} \leq \frac{t}{p}$ in the last inequality. Then, from the definition of $\mathcal{E}^{(r)}$, we have

$$\begin{aligned}
\mathcal{E}^{(0)} &= \frac{1}{N}\mathbb{E}\left\|\boldsymbol{D}^{(1)} - \boldsymbol{C}^{(0)} - \boldsymbol{E}^{(1)}\right\|_F^2 \\
&= \frac{1}{(1-\beta)^2}\frac{1}{N}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(0)}) - \frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(0)})\mathbf{1}\mathbf{1}^\top - \nabla F(\boldsymbol{X}^{(0)};\xi^{(0)}) + \frac{1}{N}\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\mathbf{1}\mathbf{1}^\top\right\|_F^2 \\
&\overset{(14)}{\leq} \frac{2}{(1-\beta)^2}\frac{1}{N}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{X}}^{(0)}) - \nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\right\|_F^2 \\
&\quad + \frac{2}{(1-\beta)^2}\frac{1}{N}\mathbb{E}\left\|\frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(0)})\mathbf{1}\mathbf{1}^\top - \frac{1}{N}\nabla F(\boldsymbol{X}^{(0)};\xi^{(0)})\mathbf{1}\mathbf{1}^\top\right\|_F^2 \\
&\overset{(6)}{\leq} \frac{4}{(1-\beta)^2}\sigma^2,
\end{aligned}$$

where we use $\boldsymbol{X}^{(0)} = \bar{\boldsymbol{X}}^{(0)}$ in the last inequality. From the definition of $\mathcal{D}^{(r)}$, we have

$$\mathcal{D}^{(0)} = \frac{1}{N}\mathbb{E}\left\|(\beta-1)\boldsymbol{D}^{(0)} + \nabla f(\bar{\boldsymbol{X}}^{(0)}) - \frac{1}{N}\nabla f(\bar{\boldsymbol{X}}^{(0)})\mathbf{1}\mathbf{1}^\top\right\|_F^2 = 0.$$

Then using $\boldsymbol{X}^{(0)} = \bar{\boldsymbol{X}}^{(0)}$ (i.e., $\Xi^{(0)} = 0$), we have

$$\Theta^{(0)} \leq \frac{144\sigma^2\eta^2}{(1-\beta)^2p^2}.$$

Here, the above upper bounds of $\mathcal{E}^{(0)}$ and $\mathcal{D}^{(0)}$ are attributed to how we choose the initial values $\boldsymbol{u}_i^{(0)}$, $\boldsymbol{c}_i^{(0)}$, $\boldsymbol{d}_i^{(0)}$, and $\boldsymbol{e}_i^{(0)}$ for $i \in V$. Then, combining them, we obtain

$$\begin{aligned}
\sum_{r=1}^{R}\Theta^{(r)} &\leq \frac{t}{p}\sum_{k=0}^{R-1}\Psi^{(k)} + \frac{144t\sigma^2\eta^2}{(1-\beta)^2p^3} + \frac{t\sigma^2\eta^2}{p^2}R \\
&\leq \frac{t}{p}\sum_{k=0}^{R-1}\Psi^{(k)} + \frac{145t\sigma^2\eta^2}{(1-\beta)^2p^3}R,
\end{aligned}$$

where we use $p \in (0,1]$, $\beta \in [0,1)$, and $R \geq 1$ in the last inequality. Then, using $\Theta^{(r)} \geq \Xi^{(r)}$ and $\Xi^{(0)} = 0$, we obtain the statement. □

Lemma 19. *We define $t \in \mathbb{R}$ and $A \in \mathbb{R}$ as follows:*

$$t \coloneqq \frac{2\beta^2 p}{1-\beta^2} + 4, \quad A \coloneqq \frac{648}{1-\frac{p}{t}-\frac{2\beta^2}{1+\beta^2}}.$$

Note that it holds that $t \geq 4$ and $A > 0$. Suppose that the same assumptions as those in Lemma 17 hold. Then, if step size η satisfies

$$\eta \leq \min\{\frac{p}{4L\sqrt{324+\frac{2A\beta^2}{1-\beta^2}}}, \frac{(1-\beta)p}{2L\sqrt{t(5+\frac{432}{p^2})}}, \frac{(1-\beta)p}{8L\sqrt{5t}}\},$$

it holds that

$$4L^2\sum_{r=0}^{R}\Xi^{(r)} \leq \frac{1}{2}\sum_{k=0}^{R}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 + \frac{40L^2t}{(1-\beta)^3p^2}\left(10+\frac{29}{p}+\frac{864}{p^2}\right)\sigma^2\eta^2(R+1).$$

Proof. We define $\Psi^{(r)}$ as follows:

$$\begin{aligned}\Psi^{(r)} \coloneqq \frac{1}{p}\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 + \frac{1}{Np}\left(9+\frac{864}{p^2}\right)\eta^2\mathbb{E}\left\|\boldsymbol{U}^{(r+1)}-\boldsymbol{D}^{(r+1)}\right\|_F^2 \\ + \frac{L^2}{p^3}\left(2592\beta^2+\frac{16A\beta^4}{1-\beta^2}\right)\eta^4\mathbb{E}\left\|\bar{\boldsymbol{u}}^{(r)}\right\|^2 + \frac{9}{Np}\eta^2\mathbb{E}\left\|\boldsymbol{E}^{(r+1)}\right\|_F^2.\end{aligned}$$

Using $\bar{\boldsymbol{e}}^{(r)} = \boldsymbol{e}_i^{(r)}$ and Lemmas 9, 10, and 12, we obtain

$$\begin{aligned}\Psi^{(r)} \leq \frac{1}{p}\eta^2\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\ + \eta^2\frac{9}{(1-\beta)p}\sum_{k=0}^{r}\beta^{r-k}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\ + \frac{2L^2}{(1-\beta)p^3}\left(2592\beta^2+\frac{16A\beta^4}{1-\beta^2}\right)\eta^4\left(\sum_{k=0}^{r-1}\beta^{r-k-1}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2\right) \\ + \frac{L^2}{(1-\beta)p}\left(9+\frac{864}{p^2}\right)\eta^2\left(\sum_{k=0}^{r}\beta^{r-k}\Xi^{(k)}\right) \\ + \frac{2L^4}{(1-\beta)p^3}\left(2592\beta^2+\frac{16A\beta^4}{1-\beta^2}\right)\eta^4\left(\sum_{k=0}^{r-1}\beta^{r-k-1}\Xi^{(k)}\right) \\ + \frac{5}{(1-\beta)^3p}\left(9+\frac{864}{p^2}\right)\sigma^2\eta^2 \\ + \frac{2L^2}{N(1-\beta)^2p^3}\left(2592\beta^2+\frac{16A\beta^4}{1-\beta^2}\right)\sigma^2\eta^4,\end{aligned}$$

for any round $r \geq 1$. Then, we obtain

$$\begin{aligned}\sum_{r=1}^{R} \Psi^{(r)} \leq & \frac{1}{p}\eta^2 \sum_{r=1}^{R} \mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\ & + \eta^2 \frac{9}{(1-\beta)p} \sum_{r=1}^{R}\sum_{k=0}^{r} \beta^{r-k} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\ & + \frac{2L^2}{(1-\beta)p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4 \left(\sum_{r=1}^{R}\sum_{k=0}^{r-1} \beta^{r-k-1} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2\right) \\ & + \frac{L^2}{(1-\beta)p}\left(9 + \frac{864}{p^2}\right)\eta^2 \left(\sum_{r=1}^{R}\sum_{k=0}^{r} \beta^{r-k}\Xi^{(k)}\right) \\ & + \frac{2L^4}{(1-\beta)p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4 \left(\sum_{r=1}^{R}\sum_{k=0}^{r-1} \beta^{r-k-1}\Xi^{(k)}\right) \\ & + \frac{5}{(1-\beta)^3 p}\left(9 + \frac{864}{p^2}\right)\sigma^2\eta^2 R \\ & + \frac{2L^2}{N(1-\beta)^2 p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\sigma^2\eta^4 R.\end{aligned}$$

Then, we obtain

$$\begin{aligned}\sum_{r=1}^{R} \Psi^{(r)} \leq & \frac{1}{p}\eta^2 \sum_{r=1}^{R} \mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\ & + \eta^2 \frac{9}{(1-\beta)p} \sum_{k=0}^{R} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \sum_{r=\max\{1,k\}}^{R} \beta^{r-k} \\ & + \frac{2L^2}{(1-\beta)p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \sum_{r=k+1}^{R} \beta^{r-k-1}\right) \\ & + \frac{L^2}{(1-\beta)p}\left(9 + \frac{864}{p^2}\right)\eta^2 \left(\sum_{k=0}^{R} \Xi^{(k)} \sum_{r=\max\{1,k\}}^{R} \beta^{r-k}\right) \\ & + \frac{2L^4}{(1-\beta)p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \Xi^{(k)} \sum_{r=k+1}^{R} \beta^{r-k-1}\right) \\ & + \frac{5}{(1-\beta)^3 p}\left(9 + \frac{864}{p^2}\right)\sigma^2\eta^2 R \\ & + \frac{2L^2}{N(1-\beta)^2 p^3}\left(2592\beta^2 + \frac{16A\beta^4}{1-\beta^2}\right)\sigma^2\eta^4 R.\end{aligned}$$

Using $\sum_{r=k+1}^{R} \beta^{r-k-1} \leq \frac{1}{1-\beta}$ and $\sum_{r=\max\{1,k\}}^{R} \beta^{r-k} \leq \frac{1}{1-\beta}$, we obtain

$$\begin{aligned}
\sum_{r=1}^{R} \Psi^{(r)} \leq & \frac{1}{p}\eta^2 \sum_{r=1}^{R} \mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
& + \eta^2 \frac{9}{(1-\beta)^2 p} \sum_{k=0}^{R} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\
& + \frac{2L^2\beta^2}{(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2\right) \\
& + \frac{L^2}{(1-\beta)^2 p}\left(9 + \frac{864}{p^2}\right)\eta^2 \left(\sum_{k=0}^{R} \Xi^{(k)}\right) \\
& + \frac{2L^4\beta^2}{(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \Xi^{(k)}\right) \\
& + \frac{5}{(1-\beta)^3 p}\left(9 + \frac{864}{p^2}\right)\sigma^2\eta^2 R \\
& + \frac{2L^2\beta^2}{N(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\sigma^2\eta^4 R.
\end{aligned}$$

Then, using $\bar{\boldsymbol{u}}^{(0)} = \mathbf{0}$ and Lemmas 9 and 10, we have

$$\Psi^{(0)} \leq \frac{1}{p}\eta^2 \mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(0)})\right\|^2 + \frac{9}{(1-\beta)p}\eta^2 \left\|\nabla f(\bar{\boldsymbol{x}}^{(0)})\right\|^2 + \frac{5}{(1-\beta)^3 p}\left(9 + \frac{864}{p^2}\right)\sigma^2\eta^2.$$

Then, we obtain

$$\begin{aligned}
\sum_{r=0}^{R} \Psi^{(r)} \leq & \frac{1}{p}\eta^2 \sum_{r=0}^{R} \mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
& + \eta^2 \frac{18}{(1-\beta)^2 p} \sum_{k=0}^{R} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\
& + \frac{2L^2\beta^2}{(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2\right) \\
& + \frac{L^2}{(1-\beta)^2 p}\left(9 + \frac{864}{p^2}\right)\eta^2 \left(\sum_{k=0}^{R} \Xi^{(k)}\right) \\
& + \frac{2L^4\beta^2}{(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \Xi^{(k)}\right) \\
& + \frac{5}{(1-\beta)^3 p}\left(9 + \frac{864}{p^2}\right)\sigma^2\eta^2 (R+1) \\
& + \frac{2L^2\beta^2}{N(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\sigma^2\eta^4 R.
\end{aligned}$$

Using $\eta^2 \leq \frac{p^2}{32L^2(162+\frac{A\beta^2}{1-\beta^2})}$, we have

$$\frac{2L^2\beta^2}{(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2\right) \leq \frac{\beta^2}{(1-\beta)^2 p}\eta^2 \left(\sum_{k=0}^{R-1} \left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2\right).$$

Using $\eta^2 \leq \frac{p^2}{32L^2(162+\frac{A\beta^2}{1-\beta^2})}$, we have

$$\frac{2L^4\beta^2}{(1-\beta)^2 p^3}\left(2592 + \frac{16A\beta^2}{1-\beta^2}\right)\eta^4 \left(\sum_{k=0}^{R-1} \Xi^{(k)}\right) \leq \frac{L^2\beta^2}{(1-\beta)^2 p}\eta^2 \left(\sum_{k=0}^{R-1} \Xi^{(k)}\right).$$

Using $\eta^2 \leq \frac{p^2}{32L^2(162+\frac{A\beta^2}{1-\beta^2})}$, we have

$$\frac{2L^2\beta^2}{N(1-\beta)^2p^3}\left(2592+\frac{16A\beta^2}{1-\beta^2}\right)\sigma^2\eta^4R \leq \frac{\beta^2}{N(1-\beta)^2p}\sigma^2\eta^2R.$$

Then, using $\beta \in [0,1)$ and $N \geq 1$, we obtain

$$\begin{aligned}\sum_{r=0}^{R}\Psi^{(r)} \leq{}& \frac{1}{p}\eta^2\sum_{r=0}^{R}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\ &+\eta^2\frac{19}{(1-\beta)^2p}\sum_{k=0}^{R}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\ &+\frac{L^2}{(1-\beta)^2p}\left(10+\frac{864}{p^2}\right)\eta^2\left(\sum_{k=0}^{R}\Xi^{(k)}\right) \\ &+\frac{5}{(1-\beta)^3p}\left(10+\frac{864}{p^2}\right)\sigma^2\eta^2(R+1).\end{aligned}$$

Using $\beta \in [0,1)$ and Lemma 18, we obtain

$$\begin{aligned}\sum_{r=0}^{R}\Xi^{(r)} \leq{}& \frac{20t}{(1-\beta)^2p^2}\eta^2\sum_{k=0}^{R}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\ &+\frac{tL^2}{(1-\beta)^2p^2}\left(10+\frac{864}{p^2}\right)\eta^2\left(\sum_{k=0}^{R}\Xi^{(k)}\right) \\ &+\frac{5t}{(1-\beta)^3p^2}\left(10+\frac{864}{p^2}\right)\sigma^2\eta^2(R+1)+\frac{145t\sigma^2\eta^2}{(1-\beta)^2p^3}R \\ \leq{}& \frac{20t}{(1-\beta)^2p^2}\eta^2\sum_{k=0}^{R}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2 \\ &+\frac{tL^2}{(1-\beta)^2p^2}\left(10+\frac{864}{p^2}\right)\eta^2\left(\sum_{k=0}^{R}\Xi^{(k)}\right) \\ &+\frac{5t}{(1-\beta)^3p^2}\left(10+\frac{29}{p}+\frac{864}{p^2}\right)\sigma^2\eta^2(R+1).\end{aligned}$$

Then, using $\eta^2 \leq \frac{(1-\beta)^2p^2}{4tL^2(5+\frac{432}{p^2})}$, we obtain

$$\frac{1}{2}\sum_{r=0}^{R}\Xi^{(r)} \leq \frac{20t}{(1-\beta)^2p^2}\eta^2\sum_{k=0}^{R}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2+\frac{5t}{(1-\beta)^3p^2}\left(10+\frac{29}{p}+\frac{864}{p^2}\right)\sigma^2\eta^2(R+1).$$

Multiplying $8L^2$, we obtain

$$4L^2\sum_{r=0}^{R}\Xi^{(r)} \leq \frac{160L^2t}{(1-\beta)^2p^2}\eta^2\sum_{k=0}^{R}\left\|\nabla f(\bar{\boldsymbol{x}}^{(k)})\right\|^2+\frac{40L^2t}{(1-\beta)^3p^2}\left(10+\frac{29}{p}+\frac{864}{p^2}\right)\sigma^2\eta^2(R+1).$$

Using $\eta^2 \leq \frac{(1-\beta)^2p^2}{320L^2t}$, we obtain the statement. □

Lemma 20. *We define $t \in \mathbb{R}$ as follows:*

$$t \coloneqq \frac{2\beta^2p}{1-\beta^2}+4.$$

Suppose that the assumptions of Lemma 19 hold. Then, if step size η satisfies

$$\eta \leq \frac{(1-\beta)^2}{2\sqrt{2}L},$$

it holds that

$$\frac{1}{2(R+1)}\sum_{r=0}^{R}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \leq \frac{4(1-\beta)}{\eta(R+1)}\left(f(\bar{\boldsymbol{z}}^{(0)})-f^\star\right)+\frac{2L\sigma^2\eta}{N(1-\beta)} + \frac{L^2}{(1-\beta)^3}\left(\frac{40t}{p^2}\left(10+\frac{29}{p}+\frac{864}{p^2}\right)+\frac{4\beta^2}{N(1-\beta)}\right)\sigma^2\eta^2.$$

Proof. Using Lemma 13 and Assumption 1, we have

$$\sum_{r=0}^{R}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \leq \frac{4(1-\beta)}{\eta}\left(f(\bar{\boldsymbol{z}}^{(0)})-f^\star\right)+4L^2\sum_{r=0}^{R}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)}-\bar{\boldsymbol{z}}^{(r)}\right\|^2+4L^2\sum_{r=0}^{R}\Xi^{(r)} - \sum_{r=0}^{R}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2+\frac{2L\sigma^2\eta}{N(1-\beta)}(R+1).$$

From Lemma 7, we have

$$4L^2\sum_{r=0}^{R}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{(r)}-\bar{\boldsymbol{z}}^{(r)}\right\|^2 \leq \frac{4L^2\beta^2\eta^2}{(1-\beta)^4}\sum_{r=0}^{R}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2+\frac{4L^2\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R.$$

Combining them yields

$$\begin{aligned}&\sum_{r=0}^{R}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2\\ &\leq \frac{4(1-\beta)}{\eta}\left(f(\bar{\boldsymbol{z}}^{(0)})-f^\star\right)+4L^2\sum_{r=0}^{R}\Xi^{(r)}-\left(1-\frac{4L^2\beta^2\eta^2}{(1-\beta)^4}\right)\sum_{r=0}^{R}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^{(r)})\right\|^2\\ &\quad+\frac{2L\sigma^2\eta}{N(1-\beta)}(R+1)+\frac{4L^2\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R.\end{aligned}$$

Using $\eta^2 \leq \frac{(1-\beta)^4}{8L^2}$ and $\beta < 1$, we obtain

$$\begin{aligned}&\sum_{r=0}^{R}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2\\ &\leq \frac{4(1-\beta)}{\eta}\left(f(\bar{\boldsymbol{z}}^{(0)})-f^\star\right)+4L^2\sum_{r=0}^{R}\Xi^{(r)}+\frac{2L\sigma^2\eta}{N(1-\beta)}(R+1)+\frac{4L^2\beta^2\sigma^2\eta^2}{N(1-\beta)^4}R.\end{aligned}$$

Using Lemma 19, we obtain

$$\begin{aligned}&\frac{1}{2}\sum_{r=0}^{R}\mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2\\ &\leq \frac{4(1-\beta)}{\eta}\left(f(\bar{\boldsymbol{z}}^{(0)})-f^\star\right)+\frac{2L\sigma^2\eta}{N(1-\beta)}(R+1)\\ &\quad+\frac{L^2}{(1-\beta)^3}\left(\frac{40t}{p^2}\left(10+\frac{29}{p}+\frac{864}{p^2}\right)+\frac{4\beta^2}{N(1-\beta)}\right)\sigma^2\eta^2(R+1).\end{aligned}$$

This concludes the proof. □

Lemma 21. *We define $t \in \mathbb{R}$ and $A \in \mathbb{R}$ as follows:*

$$t \coloneqq \frac{2\beta^2 p}{1-\beta^2}+4, \quad A \coloneqq \frac{648}{1-\frac{p}{t}-\frac{2\beta^2}{1+\beta^2}}.$$

Then, it holds that

$$\frac{(1-\beta)^2p^2}{16L\sqrt{\frac{7836\beta^2}{(1-\beta^2)^3p}+282}} \leq \min\left\{\frac{1-\beta}{4L}, \frac{p}{8L\sqrt{324+\frac{2A\beta^2}{1-\beta^2}}}, \frac{(1-\beta)p^2}{2L\sqrt{t(5p^2+432)}}\frac{(1-\beta)p}{8L\sqrt{5t}}, \frac{(1-\beta)^2}{2\sqrt{2}L}\right\}.$$

Proof. Because $\sqrt{\frac{7836\beta^2}{(1-\beta^2)^3 p} + 282} > 1$, $p \in (0, 1]$, and $\beta \in [0, 1)$, we have

$$\frac{(1-\beta)^2 p^2}{16L\sqrt{\frac{7836\beta^2}{(1-\beta^2)^3 p} + 282}} \leq \min\left\{\frac{1-\beta}{4L}, \frac{(1-\beta)^2}{2\sqrt{2}L}\right\}.$$

From $p \leq 1$, we have

$$t - 3 = \frac{2\beta^2 p}{1-\beta^2} + 1 \geq \frac{1+\beta^2}{1-\beta^2} p = \frac{p}{1 - \frac{2\beta^2}{1+\beta^2}}.$$

Then, we obtain

$$1 - \frac{p}{t} - \frac{2\beta^2}{1+\beta^2} \geq \frac{p}{t-3} - \frac{p}{t} = \frac{3p}{t(t-3)} \geq \frac{3p}{t^2}.$$

Then, we obtain

$$A \leq \frac{216t^2}{p}.$$

Using the above inequality, we obtain

$$\frac{A\beta^2}{1-\beta^2} + 162 \leq \frac{216\beta^2 t^2}{p(1-\beta^2)} + 162.$$

From the definition of t, we obtain

$$\begin{aligned}
\frac{A\beta^2}{1-\beta^2} + 30t + 162 &\leq \frac{216\beta^2 t^2}{p(1-\beta^2)} + 30t + 162 \\
&= \frac{216\beta^2}{p(1-\beta^2)}\left(\frac{2\beta^2 p}{1-\beta^2} + 4\right)^2 + 30\left(\frac{2\beta^2 p}{1-\beta^2} + 4\right) + 162 \\
&= \frac{216\beta^2}{p(1-\beta^2)}\left(\frac{4\beta^4 p^2}{(1-\beta^2)^2} + \frac{16\beta^2 p}{1-\beta^2} + 16\right) + 30\left(\frac{2\beta^2 p}{1-\beta^2} + 4\right) + 162 \\
&\leq \frac{7836\beta^2}{p(1-\beta^2)^3} + 282,
\end{aligned}$$

where we use $\beta \in [0, 1)$ and $p \in (0, 1]$ in the last inequality. Then, we obtain

$$\begin{aligned}
\frac{(1-\beta)^2 p^2}{16L\sqrt{\frac{7836\beta^2}{(1-\beta^2)^3 p} + 282}} &\leq \frac{(1-\beta)^2 p^2}{16L\sqrt{\frac{A\beta^2}{1-\beta^2} + 30t + 162}} \\
&\leq \min\left\{\frac{p}{8L\sqrt{324 + \frac{2A\beta^2}{1-\beta^2}}}, \frac{(1-\beta)p^2}{2L\sqrt{t(5p^2+432)}}\frac{(1-\beta)p}{8L\sqrt{5t}}, \right\}.
\end{aligned}$$

This concludes the proof. □

Lemma 22 (Convergence Rate for Non-convex Case)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Then, for any $R \geq 1$, there exists a step size η such that it holds that*

$$\begin{aligned}
&\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left\|\nabla f(\bar{\boldsymbol{x}}^{(r)})\right\|^2 \\
&\leq \mathcal{O}\left(\sqrt{\frac{r_0\sigma^2 L}{NR}} + \left(\frac{r_0^2\sigma^2 L^2}{p^4 R^2(1-\beta)}\left(1 + \frac{p\beta^2}{1-\beta}\right)\right)^{\frac{1}{3}} + \frac{Lr_0}{(1-\beta)p^2 R}\sqrt{1 + \frac{\beta^2}{(1-\beta^2)^3 p}}\right),
\end{aligned}$$

where $r_0 \coloneqq f(\bar{\boldsymbol{x}}^{(0)}) - f^\star$.

Proof. From Lemmas 20 and 21, if the step size η satisfies the following:

$$\eta \leq \frac{(1-\beta)^2 p^2}{16L\sqrt{\frac{7836\beta^2}{(1-\beta^2)^3 p} + 282}},$$

then we have

$$\begin{aligned}
&\frac{1}{2(R+1)} \sum_{r=0}^{R} \mathbb{E} \left\| \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 \\
&\leq \frac{4}{\tilde{\eta}(R+1)} \left(f(\bar{\boldsymbol{z}}^{(0)}) - f^\star \right) + \frac{2L\sigma^2\tilde{\eta}}{N} + \underbrace{\frac{L^2}{1-\beta} \left(\frac{40t}{p^2} \left(10 + \frac{29}{p} + \frac{864}{p^2} \right) + \frac{4\beta^2}{N(1-\beta)} \right)}_{T} \sigma^2 \tilde{\eta}^2,
\end{aligned}$$

where we define $\tilde{\eta} \coloneqq \frac{\eta}{1-\beta}$. Then, we can bound T from above as follows:

$$\begin{aligned}
T &= \frac{L^2}{1-\beta} \left(\frac{40}{p^2} \left(\frac{2\beta^2 p}{1-\beta^2} + 4 \right) \left(10 + \frac{29}{p} + \frac{864}{p^2} \right) + \frac{4\beta^2}{N(1-\beta)} \right) \\
&\overset{p\in(0,1]}{\leq} \frac{L^2}{1-\beta} \left(\frac{36120}{p^4} \left(\frac{2\beta^2 p}{1-\beta^2} + 4 \right) + \frac{4\beta^2}{N(1-\beta)} \right) \\
&\overset{p\in(0,1],\beta\in[0,1)}{\leq} \frac{36120L^2}{(1-\beta)p^4} \left(\frac{3\beta^2 p}{1-\beta} + 4 \right).
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
&\frac{1}{2(R+1)} \sum_{r=0}^{R} \mathbb{E} \left\| \nabla f(\bar{\boldsymbol{x}}^{(r)}) \right\|^2 \\
&\leq \frac{4}{\tilde{\eta}(R+1)} \left(f(\bar{\boldsymbol{z}}^{(0)}) - f^\star \right) + \frac{2L\sigma^2\tilde{\eta}}{N} + \frac{36120L^2}{(1-\beta)p^4} \left(\frac{3\beta^2 p}{1-\beta} + 4 \right) \sigma^2 \tilde{\eta}^2.
\end{aligned}$$

Using Lemma 17 in the previous work (Koloskova et al., 2020), we obtain the statement. □

E HYPERPARAMETER SETTINGS

Tables 6, 7, 8, 9, and 10 list the hyperparameter settings for each dataset. We evaluated the performance of each comparison method for different step sizes and selected the step size that achieved the highest accuracy on the validation dataset.

Table 6: Experimental settings for FashionMNIST.

Neural network architecture	LeNet (LeCun et al., 1998)
Normalization	Group normalization (Wu & He, 2018)
Step size	$\{0.005, 0.001, 0.0005\}$
L2 penalty	0.001
Batch size	100
Data augmentation	RandomCrop
Total number of epochs	500

Table 7: Experimental settings for SVHN.

Neural network architecture	LeNet (LeCun et al., 1998)
Normalization	Group normalization (Wu & He, 2018)
Step size	$\{0.005, 0.001, 0.0005\}$
L2 penalty	0.001
Batch size	100
Data augmentation	RandomCrop
Total number of epochs	500

Table 8: Experimental settings for CIFAR-10.

Neural network architecture	LeNet (LeCun et al., 1998)
Normalization	Group normalization (Wu & He, 2018)
Step size	$\{0.005, 0.001, 0.0005\}$
L2 penalty	0.001
Batch size	100
Data augmentation	RandomCrop, RandomHorizontalFlip
Total number of epochs	500

Table 9: Experimental settings for CIFAR-10 with VGG-11.

Neural network architecture	VGG-11 (Simonyan & Zisserman, 2015)
Normalization	Group normalization (Wu & He, 2018)
Step size	$\{0.05, 0.01, 0.005\}$
Step size decay	$/10$ at epoch 500 and 750.
L2 penalty	0.001
Batch size	100
Data augmentation	RandomCrop, RandomHorizontalFlip, RandomErasing
Total number of epochs	1000

Table 10: Experimental settings for CIFAR-10 with ResNet-34.

Neural network architecture	ResNet-34 (He et al., 2016)
Normalization	Group normalization (Wu & He, 2018)
Step size	$\{0.05, 0.01, 0.005\}$
Step size decay	$/10$ at epoch 375 and 563.
L2 penalty	0.001
Batch size	100
Data augmentation	RandomCrop, RandomHorizontalFlip, RandomErasing
Total number of epochs	750