Faster Adaptive Decentralized Learning Algorithms

Feihu Huang [1 2] **Jianyu Zhao** [1]

Abstract

Decentralized learning recently has received increasing attention in machine learning due to its advantages in implementation simplicity and system robustness, data privacy. Meanwhile, the adaptive gradient methods show superior performances in many machine learning tasks such as training neural networks. Although some works focus on studying decentralized optimization algorithms with adaptive learning rates, these adaptive decentralized algorithms still suffer from high sample complexity. To fill these gaps, we propose a class of faster adaptive decentralized algorithms (i.e., AdaMDOS and AdaMDOF) for distributed nonconvex stochastic and finite-sum optimization, respectively. Moreover, we provide a solid convergence analysis framework for our methods. In particular, we prove that our AdaMDOS obtains a near-optimal sample complexity of $\tilde{O}(\epsilon^{-3})$ for finding an ϵ-stationary solution of nonconvex stochastic optimization. Meanwhile, our AdaMDOF obtains a near-optimal sample complexity of $O(\sqrt{n}\epsilon^{-2})$ for finding an ϵ-stationary solution of for nonconvex finite-sum optimization, where n denotes the sample size. To the best of our knowledge, our AdaMDOF algorithm is the first adaptive decentralized algorithm for nonconvex finite-sum optimization. Some experimental results demonstrate efficiency of our algorithms.

1. Introduction

With the rapidly increasing dataset sizes and the high dimensionality of the machine learning problems, training large-scale machine learning models has been increasingly concerned. Clearly, training large-scale models by a single centralized machine has become inefficient and unscalable. Due to addressing the efficiency and scalability challenges, recently distributed machine learning optimization is widely studied. In particular, decentralized optimization (Lian et al., 2017) has received increasing attention in recent years in machine learning due to liberating the centralized agent with large communication load and privacy risk. In the paper, we study decentralized learning algorithms to solve the distributed **stochastic** problem over a communication network $G = (V, E)$, defined as

$$\min_{x\in\mathbb{R}^d} F(x) \equiv \frac{1}{m}\sum_{i=1}^m f^i(x),\ f^i(x) = \mathbb{E}_{\xi^i}\big[f^i\big(x;\xi^i\big)\big] \quad (1)$$

where for any $i \in [m]$, $f^i(x)$ denotes the objective function in i-th client, which is a differentiable and possibly nonconvex function. Here ξ^i for any $i \in [m]$ is an independent random variable following an unknown distribution $\mathcal{D}^i$, and for any $i, j \in [m]$ possibly $\mathcal{D}^i \neq \mathcal{D}^j$. $G = (V, E)$ is a communication network including m computing agents, where any agents $i, j \in V$ can communicate only if $(i, j) \in E$. Meanwhile, we also consider decentralized learning algorithms for solving the distributed **finite-sum** problem over a communication network $G = (V, E)$, defined as

$$\min_{x\in\mathbb{R}^d} F(x) \equiv \frac{1}{m}\sum_{i=1}^m f^i(x),\ f^i(x) = \frac{1}{n}\sum_{k=1}^n f^i_k(x) \quad (2)$$

where $f^i_k(x) = f^i(x;\xi^i_k)$ for $k = 1, 2\cdots, n$. Here $\{\xi^i_k\}_{k=1}^n$ can be seen as n samples drawn from distribution $\mathcal{D}^i$ for $i = 1, 2, \cdots, m$. In fact, Problems (1) and (2) frequently appear many machine learning applications such as training Deep Neural Networks (DNNs) (Lian et al., 2017) and reinforcement learning (Chen et al., 2022).

Many decentralized stochastic gradient-based algorithms recently have been developed to solve the above stochastic Problem (1). For example, (Lian et al., 2017) proposed an efficient decentralized stochastic gradient descent (D-PSGD) algorithm, which integrates average consensus with local-SGD steps and outperforms the standard centralized SGD methods. Due to the presence of inconsistency under non-i.i.d. setting, some variants of D-PSGD (Tang et al., 2018; Xin et al., 2021b) are studied to handle the data heterogeneity issue, e.g., D^2 method (Tang et al., 2018) by storing previous status, and GT-DSGD method (Xin et al., 2021a)

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China [2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China. Correspondence to: Feihu Huang <huangfeihu2018@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1: **Sample** and **Communication** complexities comparison of the representative **adaptive** decentralized stochastic algorithms for finding an ϵ-stationary point of Problem (1) or (2), i.e., $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$ or its equivalent variants. **Note that** the AdaRWGD (Sun et al., 2022) relies on the random walk instead of parallel framework used in other algorithms. For fair comparison, here we do not consider some specific cases such as sparse stochastic gradients.

Problem	Algorithm	Reference	Sample Complexity	Communication Complexity
Stochastic	DADAM	(Nazari et al., 2022)	$O(\epsilon^{-4})$	$O(\epsilon^{-4})$
	AdaRWGD	(Sun et al., 2022)	$O(\epsilon^{-4})$	$O(\epsilon^{-4})$
	DAMSGrad/DAdaGrad	(Chen et al., 2023)	$O(\epsilon^{-4})$	$O(\epsilon^{-4})$
	AdaMDOS	Ours	$\tilde{O}(\epsilon^{-3})$	$O(\epsilon^{-3})$
Finite-Sum	AdaMDOF	Ours	$O(\sqrt{n}\epsilon^{-2})$	$O(\epsilon^{-2})$

by using gradient tracking technique (Xu et al., 2015). Subsequently, (Sun et al., 2020; Pan et al., 2020) proposed some accelerated decentralized SGD algorithms (i.e., D-GET and D-SPIDER-SFO) by using variance reduced gradient estimator of SARAH/SPIDER (Nguyen et al., 2017; Fang et al., 2018), which obtain a near-optimal sample complexity of $O(\epsilon^{-3})$ for finding the stationary solution of stochastic optimization problems. To reduce large batch-size at each iteration, (Zhang et al., 2021; Xin et al., 2021a) proposed a class of efficient momentum-based decentralized SGD algorithms (i.e., GT-HSGD and GT-STORM) based on momentum-based variance reduced gradient estimator of ProxHSGD/STORM (Cutkosky & Orabona, 2019; Tran-Dinh et al., 2022), which also obtain a near-optimal sample complexity of $\tilde{O}(\epsilon^{-3})$.

Meanwhile, some decentralized stochastic gradient-based algorithms have been developed to solve the above finite-sum Problem (2). (Sun et al., 2020; Xin et al., 2022) proposed a class of efficient decentralized algorithms for nonconvex finite-sum optimization based on variance reduced gradient estimator of SARAH (Nguyen et al., 2017). Subsequently, (Zhan et al., 2022) presented a fast decentralized algorithm for nonconvex finite-sum optimization based on variance reduced gradient estimator of ZeroSARAH (Li et al., 2021) without computing multiple full gradients.

It well known that the adaptive gradient methods show superior performances n many machine learning tasks such as training DNNs. More recently, (Nazari et al., 2022; Sun et al., 2022; Chen et al., 2023) proposed some adaptive decentralized algorithms for stochastic optimization based on the existing Adam algorithm (Kingma & Ba, 2014) or its variants. However, these adaptive decentralized algorithms still suffer high sample and communication complexities in finding the stationary solution of Problem (1) (Please see Table 1). Naturally, there still exists an open question:

> *Could we design adaptive decentralized algorithms with lower sample and communication complexities to solve Problems (1) and (2) ?*

In the paper, to fill this gap, we affirmatively answer to this question, and propose a class of faster adaptive decentralized algorithms to solve Problems (1) and (2), respectively, based on the momentum-based variance-reduced and gradient tracking techniques. In particular, our methods use a unified adaptive matrix to flexibly incorporate various adaptive learning rates. Our main contributions are several folds:

(1) We propose a class of efficient adaptive decentralized optimization algorithms (i.e., AdaMDOS and AdaMDOF) to solve Problems (1) and (2), respectively, based on the momentum-based variance-reduced and gradient tracking techniques simultaneously. Moreover, we provide a convergence analysis framework for our methods.

(2) We prove that our AdaMDOS algorithm reaches the near optimal sample complexity of $\tilde{O}(\epsilon^{-3})$ for finding an ϵ-stationary solution of Problem (1), which matches the lower bound of smooth nonconvex stochastic optimization (Arjevani et al., 2023).

(3) We prove that our AdaMDOF algorithm reaches the near optimal sample complexity of $O(\sqrt{n}\epsilon^{-2})$ for finding an ϵ-stationary solution of Problem (2), which matches the lower bound of smooth nonconvex finite-sum optimization (Fang et al., 2018).

(4) We conduct some numerical experiments on training nonconvex machine learning tasks to verify the efficiency of our proposed algorithms.

Since our algorithms use a unified adaptive matrix including various adaptive learning rates, our convergence analysis does not consider some specific cases such as sparse stochastic gradients. Despite this, our adaptive algorithms still obtain lower sample and communication complexities compared to the existing adaptive decentralized algorithms.

2. Related Works

In this section, we overview some representative decentralized optimization algorithms and adaptive gradient algorithms, respectively.

2.1. Decentralized Optimization

Decentralized optimization is an efficient framework to collaboratively solve distributed problems by multiple worker nodes, where a worker node only needs to communicate with its neighbors at each iteration. The traditional decentralized optimization methods include some popular algorithms such as Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011), Dual Averaging (Duchi et al., 2011b). Subsequently, some efficient decentralized optimization algorithms have been developed, e.g., Extra (Shi et al., 2015), Next (Di Lorenzo & Scutari, 2016), Prox-PDA (Hong et al., 2017). Meanwhile, (Lian et al., 2017) proposed an efficient decentralized stochastic gradient descent algorithm (i.e., D-PSGD), which shows that the decentralized SGD can outperform the parameter server-based SGD algorithms relying on high communication cost. From this, decentralized algorithms have begun to shine in machine learning such as training DNNs. Subsequently, (Tang et al., 2018) proposed an accelerated D-PSGD (i.e., D^2) by using previous status. Meanwhile, (Xin et al., 2021a) further proposed an efficient decentralized SGD algorithm (i.e., GT-DSGD) by using gradient tracking technique (Xu et al., 2015). By using the variance-reduced techniques, some other accelerated decentralized SGD algorithms (Sun et al., 2020; Pan et al., 2020; Cutkosky & Orabona, 2019; Tran-Dinh et al., 2022) have been proposed, include D-SPIDER-SFO (Pan et al., 2020) and GT-HSGD (Xin et al., 2021a). Meanwhile, (Nazari et al., 2022) studied the decentralized version of AMSGrad (Reddi et al., 2019) for online optimization. Moreover, (Sun et al., 2022; Chen et al., 2023) developed adaptive decentralized algorithms for stochastic optimization by using local adaptive learning rates.

2.2. Adaptive Gradient Methods

Adaptive gradient methods (Duchi et al., 2011a; Kingma & Ba, 2014; Loshchilov & Hutter, 2017) recently have been successfully applied in machine learning tasks such as training DNNs. Adam (Kingma & Ba, 2014) is one of popular adaptive gradient methods by using a coordinate-wise adaptive learning rate and momentum technique to accelerate algorithm, which is the default optimization tool for training attention models (Zhang et al., 2020). Subsequently, some variants of Adam (Reddi et al., 2019; Chen et al., 2018; Guo et al., 2021) have been presented to obtain a convergence guarantee under the nonconvex setting. Due to using the coordinate-wise type of adaptive learning rates, Adam frequently shows a bad generalization performance in training DNNs. To improve the generalization performances, recently some adaptive gradient methods such as AdamW (Loshchilov & Hutter, 2017), AdaGrad (Li & Orabona, 2019) and AdaBelief (Zhuang et al., 2020) have been proposed. More recently, some accelerated adaptive gradient methods (Cutkosky & Orabona, 2019; Huang et al., 2021; Levy et al., 2021; Kavis et al., 2022) have been proposed based on the variance-reduced techniques.

3. Preliminaries

3.1. Notations

$[m]$ denotes the set $\{1, 2, \cdots, m\}$. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices. $\langle x, y\rangle$ denotes the inner product of two vectors x and y. For vectors x and y, x^r $(r > 0)$ denotes the element-wise power operation, x/y denotes the element-wise division and $\max(x, y)$ denotes the element-wise maximum. I_d denotes a d-dimensional identity matrix. $a_t = O(b_t)$ denotes that $a_t \le cb_t$ for some constant $c > 0$. The notation $\tilde{O}(\cdot)$ hides logarithmic terms. $\text{ones}(d, 1)$ denotes an all-one d-dimensional vector.

3.2. Assumptions

In this subsection, we give some mild assumptions on the Problems (1) and (2).

Assumption 3.1. (**Smoothness**) For any $i \in [m]$, each component loss function $f^i(x; \xi^i)$ is L-smooth, such that for all $x_1, x_2 \in \mathbb{R}^d$

$$\|\nabla f^i(x_1; \xi^i) - \nabla f^i(x_2; \xi^i)\| \le L\|x_1 - x_2\|. \tag{3}$$

Clearly, based on Assumptions 3.1, we have

$$\begin{aligned}
&\|\nabla F(x_1) - \nabla F(x_2)\| \\
&= \|\frac{1}{m}\sum_{i=1}^m \big(\mathbb{E}_{\xi^i}[\nabla f^i(x_1; \xi^i)] - \mathbb{E}_{\xi^i}[\nabla f^i(x_2; \xi^i)]\big)\| \\
&\le \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi^i}\|\nabla f^i(x_1; \xi^i) - \nabla f^i(x_2; \xi^i)\| \le L\|x_1 - x_2\|,
\end{aligned}$$

i.e., the global function $F(x)$ is L-smooth as well.

Assumption 3.2. (**Sampling Oracle**) Stochastic function $f^i(x; \xi^i)$ has an unbiased stochastic gradient with bounded variance for any $i \in [m]$, i.e.,

$$\mathbb{E}[\nabla f^i(x; \xi^i)] = \nabla f^i(x), \ \mathbb{E}\|\nabla f^i(x; \xi^i) - \nabla f^i(x)\|^2 \le \sigma^2.$$

Assumption 3.3. (**Lower Bounded**) The objective function $F(x)$ is lower bounded, i.e., $F^* = \inf_{x\in\mathbb{R}^d} F(x)$.

Assumption 3.4. (**Network Protocol**) The graph $G = (V, E)$ is connected and undirected, which can be represented by a mixing matrix $W \in \mathbb{R}^{m\times m}$: 1) $W_{i,j} > 0$ if $W_{i,j} \in E$ and $W_{i,j} = 0$ otherwise; 2) W is doubly stochastic such that $W = W^T$, $\sum_{i=1}^m W_{i,j} = 1$ and $\sum_{j=1}^m W_{i,j} = 1$; 3) the eigenvalues of W satisfy $\lambda_m \le \cdots \le \lambda_2 < \lambda_1 = 1$ and $\nu = \max(|\lambda_2|, |\lambda_m|) < 1$.

Assumption 3.5. In our algorithms, the local adaptive matrices $A_t^i \succeq \rho I_d \succ 0$ for all $i \in [m]$, $t \ge 1$ for updating the variables x, where $\rho > 0$ is an appropriate positive number.

Assumptions 3.1 and 3.2 are commonly used in stochastic smooth nonconvex optimization (Sun et al., 2020; Pan

Algorithm 1 Adaptive Momentum-Based Decentralized Optimization (**AdaMDOS**) Algorithm for **Stochastic** Optimization

1: **Input:** $T > 0$, tuning parameters $\{\gamma, \eta_t, \beta_t\}$, initial inputs $x_1^i \in \mathbb{R}^d$ for all $i \in [m]$;
2: **initialize:** Set $x_0^i = \tilde{x}_0^i$ for $i \in [m]$, and draw one sample ξ_0^i and then compute $u_0^i = \nabla f^i(x_0^i; \xi_0^i)$ and $w_0^i = \sum_{j \in \mathcal{N}_i} W_{i,j} u_0^j$ for all $i \in [m]$.
3: **for** $t = 0$ **to** $T-1$ **do**
4:   **for** $i = 1, \cdots, m$ **(in parallel) do**
5:     Generate the adaptive matrix $A_t^i \in \mathbb{R}^{d \times d}$;
    One example of A_t^i by using update rule ($a_0^i = 0$, $0 < \varrho < 1$, $\rho > 0$.)
    Compute $a_t^i = \varrho a_{t-1}^i + (1-\varrho)(\nabla f^i(x_t^i; \xi_t^i))^2$, $A_t^i = \mathrm{diag}(\sqrt{a_t^i} + \rho I_d)$;
6:     $\tilde{x}_{t+1}^i = \sum_{j \in \mathcal{N}_i} W_{i,j} x_t^j - \gamma (A_t^i)^{-1} w_t^i$;
7:     $x_{t+1}^i = x_t^i + \eta_t(\tilde{x}_{t+1}^i - x_t^i)$;
8:     Randomly draw a sample $\xi_{t+1}^i \sim \mathcal{D}^i$;
9:     $u_{t+1}^i = \nabla f^i(x_{t+1}^i; \xi_{t+1}^i) + (1-\beta_{t+1})(u_t^i - \nabla f^i(x_t^i; \xi_{t+1}^i))$;
10:     $w_{t+1}^i = \sum_{j \in \mathcal{N}_i} W_{i,j}\big(w_t^j + u_{t+1}^j - u_t^j\big)$;
11:   **end for**
12: **end for**
13: **Output:** Chosen uniformly random from $\{x_{t\geq 1}^i\}_{i=1}^m$.

Algorithm 2 Adaptive Momentum-Based Decentralized Optimization (**AdaMDOF**) Algorithm for **Finite-Sum** Optimization

Input: $T > 0$, tuning parameters $\{\gamma, \eta_t, \beta_t\}$, initial inputs $x_1^i \in \mathbb{R}^d$ for all $i \in [m]$;
initialize: Set $x_0^i = x_1^i = \tilde{x}_1^i$, $z_{1,0}^i = z_{2,0}^i = \cdots = z_{n,0}^i = 0$ and $u_0^i = w_0^i = 0$ for any $i \in [m]$.
for $t = 1$ **to** T **do**
  for $i = 1, \cdots, m$ **(in parallel) do**
    Randomly draw a minibatch samples $\mathcal{I}_t^i$ with $|\mathcal{I}_t^i| = b$;
    $u_t^i = \frac{1}{b}\sum_{k \in \mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big) + (1-\beta_t)u_{t-1}^i + \beta_t\Big(\frac{1}{b}\sum_{k \in \mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\big) + \frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i\Big)$;
    $w_t^i = \sum_{j \in \mathcal{N}_i} W_{i,j}\big(w_{t-1}^j + u_t^j - u_{t-1}^j\big)$;
    Generate the adaptive matrix $A_t^i \in \mathbb{R}^{d \times d}$;
    One example of A_t^i by using update rule ($a_0^i = 0$, $0 < \varrho < 1$, $\rho > 0$.)
    Compute $a_t^i = \varrho a_{t-1}^i + (1-\varrho)\big(\frac{1}{b}\sum_{k \in \mathcal{I}_t^i} \nabla f_k^i(x_t^i)\big)^2$, $A_t^i = \mathrm{diag}(\sqrt{a_t^i} + \rho I_d)$;
    $\tilde{x}_{t+1}^i = \sum_{j \in \mathcal{N}_i} W_{i,j} x_t^j - \gamma (A_t^i)^{-1} w_t^i$;
    $x_{t+1}^i = x_t^i + \eta_t(\tilde{x}_{t+1}^i - x_t^i)$;
    $z_{k,t}^i = \nabla f_k^i(x_t)$ for $k \in \mathcal{I}_t^i$ and $z_{k,t}^i = z_{k,t-1}^i$ for $k \notin \mathcal{I}_t^i$.
  end for
end for
Output: Chosen uniformly random from $\{x_{t\geq 1}^i\}_{i=1}^m$.

et al., 2020; Cutkosky & Orabona, 2019; Tran-Dinh et al., 2022). Assumption 3.3 ensures the feasibility of Problems (1) and (2). Assumption 3.4 shows the protocol properties of network $G = (V, E)$, which is very common in the decentralized distributed optimization (Lian et al., 2017; Xin et al., 2021a). Assumption 3.5 imposes that each local adaptive matrix is positive definite, which is commonly used in many adaptive gradient methods for non-distributed optimization (Huang et al., 2021; Yun et al., 2021).

4. Adaptive Momentum-Based Decentralized Algorithms

In this section, we propose a class of efficient adaptive momentum-based decentralized algorithms to solve Problems (1) and (2), respectively, which build on the momentum-based and gradient tracking techniques.

4.1. AdaMDOS Algorithm for Stochastic Optimization

In this subsection, we propose a faster adaptive momentum-based decentralized (AdaMDOS) algorithm for the stochastic Problem (1) over a network, which builds on the variance-reduced momentum technique of STORM (Cutkosky & Orabona, 2019; Tran-Dinh et al., 2022) and gradient tracking technique (Xu et al., 2015). In particular, our AdaMDOS algorithm also uses the momentum iteration and unified adaptive learning rate to update variable. Algorithm 1 provides the algorithmic framework of our AdaMDOS algorithm.

At the line 5 of Algorithm 1, we generate an adaptive matrix based on the historical stochastic gradients $\big\{\nabla f^i(x_l^i; \xi_l^i)\big\}_{1 \leq l \leq t}$. And we give an Adam-like adaptive learning rate, defined as

$$
\begin{aligned}
a_t^i &= \varrho a_{t-1}^i + (1-\varrho)(\nabla f^i(x_t^i; \xi_t^i))^2 \\
&= \sum_{l=1}^{t} (1-\varrho)\varrho^{t-l}(\nabla f^i(x_l^i; \xi_l^i))^2, \\
A_t^i &= \mathrm{diag}(\sqrt{a_t^i} + \rho I_d), \qquad (4)
\end{aligned}
$$

clearly, we have $A_t^i \succeq \rho I_d$, which satisfies Assumption 3.5. Besides one example (4), we can also generate many adaptive matrices satisfying the above Assumption 3.5. e.g., the Barzilai-Borwein-like adaptive matrix, defined as

$$
\begin{aligned}
a_t^i &= \frac{|\langle x_t^i - x_{t-1}^i, \nabla f^i(x_t^i; \xi_t^i) - \nabla f^i(x_{t-1}^i; \xi_t^i)\rangle|}{\|x_t^i - x_{t-1}^i\|^2} + \rho, \\
A_t^i &= a_t^i I_d \succeq \rho I_d. \qquad (5)
\end{aligned}
$$

At the line 6 of Algorithm 1, each client updates the local variable x^i based on adaptive matrix A_t^i and momentum-based gradient estimator w_t^i:

$$\tilde{x}_{t+1}^i = \sum_{j\in\mathcal{N}_i} W_{i,j}x_t^j - \gamma(A_t^i)^{-1}w_t^i, \tag{6}$$

where the constant $\gamma > 0$. Here $\mathcal{N}_i = \{j \in V \mid (i,j) \in E, j = i\}$ denotes the neighborhood of the i-th client. Here each client communicates with its neighbors to update the variable x. Then we further use the momentum iteration to update the variable x at the line 7 of Algorithm 1:

$$x_{t+1}^i = x_t^i + \eta_t(\tilde{x}_{t+1}^i - x_t^i), \tag{7}$$

where $\eta_t \in (0,1)$.

At lines 8-9 of Algorithm 1, each client uses the variance-reduced momentum-based technique (Tran-Dinh et al., 2022; Cutkosky & Orabona, 2019) to update the stochastic gradients by using local data ξ_{t+1}^i: for $i \in [m]$

$$\begin{aligned} u_{t+1}^i &= \nabla f^i(x_{t+1}^i;\xi_{t+1}^i) \\ &\quad + (1-\beta_{t+1})(u_t^i - \nabla f^i(x_t^i;\xi_{t+1}^i)), \end{aligned} \tag{8}$$

where $\beta_{t+1} \in (0,1)$. At the line 10 of our Algorithm 1, then each client communicates with its neighbors to compute gradient estimators w_{t+1}^i, defined as

$$w_{t+1}^i = \sum_{j\in\mathcal{N}_i} W_{i,j}\big(w_t^j + u_{t+1}^j - u_t^j\big), \tag{9}$$

which uses gradient tracking technique (Xu et al., 2015; Di Lorenzo & Scutari, 2016) to reduce the consensus error. Thus, the local stochastic gradient estimator w_{t+1}^i can track the directions of global gradients.

4.2. AdaMDOF Algorithm for Finite-Sum Optimization

In this subsection, we propose a faster adaptive momentum-based decentralized (AdaMDOF) algorithm for distributed finite-sum problem (2) over a network, which builds on the variance-reduced momentum technique of ZeroSARAH (Li et al., 2021) and gradient tracking technique. Algorithm 2 provides the algorithmic framework of our AdaMDOF algorithm.

Algorithm 2 is fundamentally similar to Algorithm 1, differing primarily in its application of the variance-reduced momentum technique from ZeroSARAH (Li et al., 2021) to update the stochastic gradients by using local data: for $i \in [m]$

$$\begin{aligned} u_t^i &= \frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big) + (1-\beta_t)u_{t-1}^i \\ &\quad + \beta_t\Big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\big) + \frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i\Big), \end{aligned}$$

where $\beta_{t+1} \in (0,1)$. Here the ZeroSARAH technique can be seen as the combination of SARAH (Nguyen et al., 2017) and SAGA (Defazio et al., 2014) techniques.

5. Convergence Analysis

In this section, under some mild assumptions, we provide the convergence properties of our AdaMDOS and AdaMDOF algorithms for Problems (1) and (2), respectively. All related proofs are provided in the following Appendix. For notational simplicity, let $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_t^i$ for all $t \geq 1$.

5.1. Convergence Properties of our AdaMDOS Algorithm

For AdaMDOS algorithm, we define a useful Lyapunov function, for any $t \geq 1$

$$\begin{aligned} \Omega_t &= \mathbb{E}\Big[F(\bar{x}_t) + (\lambda_{t-1} - \frac{9\gamma\eta}{2\rho})\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 \\ &\quad + \chi_{t-1}\frac{1}{m}\sum_{i=1}^m\|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\ &\quad + (\theta_{t-1} - \frac{19\gamma\eta L^2}{4\rho})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i - \bar{x}_{t-1}\|^2 \\ &\quad + (\vartheta_{t-1} - \frac{\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m\|w_{t-1}^i - \bar{w}_{t-1}\|^2 \\ &\quad + \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2\Big], \end{aligned}$$

where $g_t^i = (A_t^i)^{-1}w_t^i$, $\overline{\nabla f(x_t)} = \frac{1}{m}\sum_{i=1}^m \nabla f^i(x_t^i)$, $\chi_t \geq 0$, $\lambda_t \geq \frac{9\gamma\eta}{2\rho}$, $\theta_t \geq \frac{29\gamma\eta L^2}{6\rho}$, $\vartheta_t \geq \frac{\gamma\eta}{4\rho}$ and $\eta_t = \eta$ for all $t \geq 0$.

Theorem 5.1. *Suppose the sequences $\big\{\{x_t^i\}_{i=1}^m\big\}_{t=1}^T$ be generated from Algorithm 1. Under the above Assumptions 3.1-3.5, and let $\eta_t = \eta$, $0 < \beta_t \leq 1$ for all $t \geq 0$, $\gamma \leq \min\big(\frac{\rho(1-\nu^2)}{48\theta_t}, \frac{3\rho(1-\nu^2)\theta_t}{58L^2}\big)$, $\eta \leq \min\big(\frac{\rho\sqrt{1-\nu^2}}{4L\gamma\sqrt{3(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta_t}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)$ with $H_t = \frac{9}{2\beta_t} + \frac{8\nu^2}{(1-\nu)^2}$ for all $t \geq 1$, we have*

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T\mathbb{E}\|\nabla F(\bar{x}_t)\| \\ &\leq \frac{1}{T}\sum_{t=1}^T\frac{1}{m}\sum_{i=1}^m\mathbb{E}[\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\|] \\ &\leq \Big(\frac{6\sqrt{G}}{\sqrt{T}} + \frac{12\sigma}{\rho}\sqrt{\frac{1}{T}\sum_{t=1}^T H_t\beta_t^2}\Big)\sqrt{\frac{1}{T}\sum_{t=1}^T\frac{1}{m}\sum_{i=1}^m\mathbb{E}\|A_t^i\|^2}, \end{aligned} \tag{10}$$

where $G = \frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta} + \big(\frac{4\nu^2}{\rho^2(1-\nu)} + \frac{9}{2\rho^2\beta_0} - \frac{9}{2\rho^2}\big)\sigma^2$.

Remark 5.2. Based on Assumption 3.1, if using Barzilai-Borwein-like adaptive matrix (5), we have $\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} \leq L+\rho$.

Let $\beta_t = \frac{1}{T^{2/3}}$ for all $t \geq 1$, we have

$$H_t = \frac{9}{2\beta_t} + \frac{8\nu^2}{(1-\nu)^2} = O(T^{2/3}), \tag{11}$$

and then we can obtain $\sqrt{\frac{1}{T}\sum_{t=1}^T H_t\beta_t^2} = O(\frac{1}{T^{1/3}})$. We set $\theta_t = \theta \geq \frac{29\gamma\eta L^2}{6\rho}$ for all $t \geq 1$. Meanwhile, we can set $\rho = O(1)$, $\gamma = O(1)$ and $\eta = O(\frac{1}{T^{1/3}})$. Then we obtain $G = O(T^{1/3})$ and

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \\ &\leq O(\frac{1}{T^{1/3}} + \frac{\sigma}{T^{1/3}})\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2}. \end{aligned} \tag{12}$$

Since $\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} = O(1)$, let

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \leq O(\frac{1}{T^{1/3}} + \frac{\sigma}{T^{1/3}}) \leq \epsilon, \tag{13}$$

we have $T = O(\epsilon^{-3})$. Since our AdaMDOS algorithm only use one sample at each iteration, it obtains a near-optimal sample complexity of $1 \cdot T = O(\epsilon^{-3})$ for finding an ϵ-stationary point of Problem (1), which matches the lower bound of smooth nonconvex stochastic optimization (Arjevani et al., 2023).

Assumption 5.3. (**Lipschitz Continuity**) For any $i \in [m]$, each component loss function $f^i(x;\xi^i)$ is M-Lipschitz continuity, such that for all $x \in \mathbb{R}^d$

$$\|\nabla f^i(x;\xi^i)\| \leq M. \tag{14}$$

Assumption 5.3 is commonly used in the adaptive gradient algorithms (Reddi et al., 2019; Chen et al., 2018; Guo et al., 2021; Sun et al., 2022; Chen et al., 2023).

Remark 5.4. Based on Assumption 5.3, if using Adam-like adaptive matrix given in Algorithm 1, we have $\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} \leq M+\rho$.

Based on the above Theorem 5.1, our AdaMDOS algorithm still obtains a near-optimal sample (gradient) complexity of $O(\epsilon^{-3})$ for finding an ϵ-stationary point of Problem (1).

5.2. Convergence Properties of our AdaMDOF Algorithm

For AdaMDOF algorithm, we first define a useful Lyapunov function for any $t \geq 1$

$$\begin{aligned} \Phi_t &= \mathbb{E}\Big[F(\bar{x}_t) + (\lambda_{t-1} - \frac{9\gamma\eta}{2\rho})\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 \\ &+ \chi_{t-1}\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 + \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2 \\ &+ \alpha_{t-1}\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\ &+ (\theta_{t-1} - \frac{19\gamma\eta L^2}{4\rho})\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 \\ &+ (\vartheta_{t-1} - \frac{3\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2\Big], \end{aligned}$$

where $\alpha_t \geq 0$, $\chi_t \geq 0$, $\lambda_t \geq \frac{9\gamma\eta}{2\rho}$, $\theta_t \geq \frac{29\gamma\eta L^2}{6\rho}$, $\vartheta_t \geq \frac{3\gamma\eta}{4\rho}$ and $\eta_t = \eta$ for all $t \geq 0$.

Theorem 5.5. *Suppose the sequences $\big\{\{x_t^i\}_{i=1}^m\big\}_{t=1}^T$ be generated from Algorithm 2. Under the above Assumptions 3.1-3.5, and let $\eta_t = \eta$, $0 < \beta_t \leq 1$ for all $t \geq 0$, $\gamma \leq \min\big(\frac{\rho(1-\nu^2)}{48\theta_t}, \frac{3\rho(1-\nu^2)\theta_t}{58L^2}\big)$, $\eta \leq \min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta_t}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)$ with $H_t = \frac{9}{b\beta_t} + \frac{6\nu^2}{b(1-\nu)^2} + \frac{4n^2\beta_t^2}{b^3}\big(\frac{9}{\beta_t} + \frac{9\nu^2}{(1-\nu)^2}\big) + \frac{3\nu^2}{(1-\nu)^2}$ for all $t \geq 0$, we have*

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \\ &\leq \frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\|] \\ &\leq \frac{6\sqrt{G}}{\sqrt{T}}\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2}, \end{aligned} \tag{15}$$

where $G = \frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta} + \big(\frac{18\beta_0}{\rho^2} + \frac{18\beta_0^2\nu^2}{\rho^2(1-\nu)^2} + \frac{3\nu^2}{\rho^2(1-\nu)^2} + \frac{9}{2\rho^2\beta_0} - \frac{9}{2\rho^2}\big)\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_0^i)\|^2$ is independent on T, b and n.

Remark 5.6. Based on Assumption 3.1, if using Barzilai-Borwein-like adaptive matrix (5), we have $\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} \leq L+\rho$. Based on Assumption 5.3, we have $\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} \leq M+\rho$.

Let $\theta_t = \theta \geq \frac{29\gamma\eta L^2}{6\rho}$, $b = \sqrt{n}$ and $\beta_t = \frac{b}{n}$ for all $t \geq 1$, we have

$$H_t = \frac{9}{b\beta_t} + \frac{6\nu^2}{b(1-\nu)^2} + \frac{4n^2\beta_t^2}{b^3}\big(\frac{9}{\beta_t} + \frac{9\nu^2}{(1-\nu)^2}\big) + \frac{3\nu^2}{(1-\nu)^2} \leq 45 + \frac{45\nu^2}{(1-\nu)^2}. \quad (16)$$

Then we have $\frac{1}{\sqrt{H_t}} \geq \frac{1}{\sqrt{45+\frac{45\nu^2}{(1-\nu)^2}}}$ and

$$\min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big) \geq \min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}}, \frac{\sqrt{\rho(1-\nu^2)\theta}}{2L\sqrt{\gamma(3+\nu^2)}}\big)\sqrt{45+\frac{45\nu^2}{(1-\nu)^2}}.$$

Thus we can let $\eta = \min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}}, \frac{\sqrt{\rho(1-\nu^2)\theta}}{2L\sqrt{\gamma(3+\nu^2)}}\big)\sqrt{45+\frac{45\nu^2}{(1-\nu)^2}}$ and $\gamma = \min\big(\frac{\rho(1-\nu^2)}{48\theta}, \frac{3\rho(1-\nu^2)\theta}{58L^2}\big)$. **Note that** we set $\beta_t = \frac{b}{n}$ for all $t \geq 1$, while we can set $\beta_0 \in (0,1)$, which is independent on T, b and n. Let $\rho = O(1)$, $\eta = O(1)$ and $\gamma = O(1)$, we have $G = \frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta} + \big(\frac{18\beta_0}{\rho^2} + \frac{18\beta_0^2\nu^2}{\rho^2(1-\nu)^2} + \frac{3\nu^2}{\rho^2(1-\nu)^2} + \frac{9}{2\rho^2\beta_0} - \frac{9}{2\rho^2}\big)\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_0^i)\|^2 = O(1)$ is independent on T, b and n. Since $\sqrt{\frac{1}{T}\sum_{t=1}^T\frac{1}{m}\sum_{i=1}^m\mathbb{E}\|A_t^i\|^2} = O(1)$, set

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq O(\frac{1}{\sqrt{T}}) \leq \epsilon,$$

we have $T = O(\epsilon^{-2})$. Since our AdaMDOF algorithm requires b samples, we can obtain a near-optimal sample complexity of $T\cdot b = O(\sqrt{n}\epsilon^{-2})$, which matches the lower bound of smooth nonconvex finite-sum optimization (Fang et al., 2018).

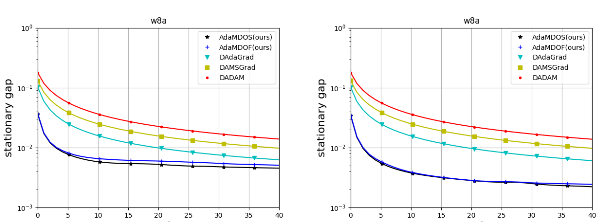


Figure 1: Stationary gap vs epoch at **w8a** dataset under the ring network (Left) and the 3-regular network (Right).

6. Numerical Experiments

In this section, we apply some numerical experiments to demonstrate efficiency of our AdaMDOS and AdaMDOF

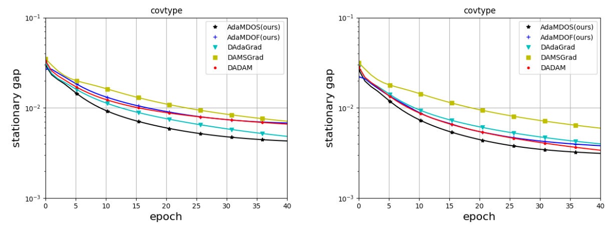


Figure 2: Stationary gap vs epoch at **covertype** dataset under the ring network (Left) and the 3-regular network (Right).

algorithms. Since the AdaRWGD (Sun et al., 2022) relies on the random walk instead of parallel framework used in other algorithms, our adaptive decentralized algorithms only compare to the existing adaptive decentralized methods (i.e, DADAM (Nazari et al., 2022), DAMSGrad (Chen et al., 2023), DAdaGrad (Chen et al., 2023)) given in Tabel 1. In decentralized algorithms, we consider two classical undirected networks that connect all clients, i.e., the *ring* and *3-regular* expander networks (Hoory et al., 2006), described in the following Appendix B.

6.1. Training Logistic Model

In this subsection, we consider learning a non-convex logistic model (Allen-Zhu & Hazan, 2016) for binary classification over a decentralized network of m nodes with n data samples at each node:

$$\min_{x\in\mathbb{R}^d}\frac{1}{m}\sum_{i=1}^m\Big(\frac{1}{n}\sum_{k=1}^n f^i(x;\xi_k^i) + \lambda\|x\|^2\Big), \quad (17)$$

where $\lambda > 0$ and $f^i(x;\xi_k^i) = \frac{1}{1+\exp(l_k^i\langle a_k^i,x\rangle)}$ is a nonconvex sigmoid loss function. Here $\xi_k^i = (a_k^i, l_k^i)$ denotes the k-th sample at i-th node, where $a_k^i \in \mathbb{R}^d$ denotes the features and $l_k^i \in \{-1,1\}$ is a label. In the experiment, we set the regularization parameter $\lambda = 10^{-5}$, and use the same initial solution $x_0 = x_0^i = 0.01\cdot\text{ones}(d,1)$ for all $i \in [m]$ for all algorithms. We use public w8a and covertype datasets[1]. The w8a dataset includes 60,000 training examples, where we partitioned into 5 clients each containing 12,000 training examples. The covertype dataset includes 100,000 training examples, where we partitioned into 5 clients each containing 20,000 training examples.

In the experiment, we characterize performance of the algorithms in comparison in terms of the decrease of stationary gap versus epochs, where the *stationary gap* is defined as $\|\nabla F(\bar{x}_t)\| + \frac{1}{m}\sum_{i=1}^m\|\bar{x}_t - x_t^i\|$, where x_t^i is the estimate of the stationary solution at the i-th node and $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_t^i$, and each epoch represents n component gradient computa-

[1]available at https://www.openml.org/

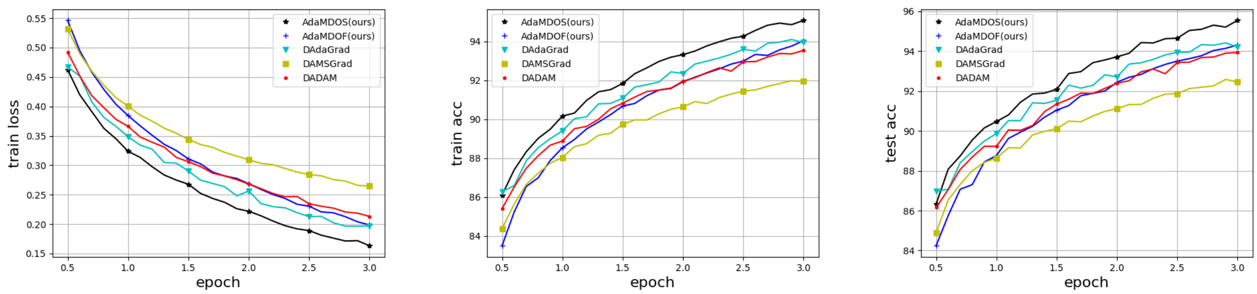


Figure 3: Training CNN on MNIST dataset: training loss vs epoch (Left), training accuracy (%) vs epoch (Middle), and test accuracy (%) vs epoch (Right) under the *ring* network.

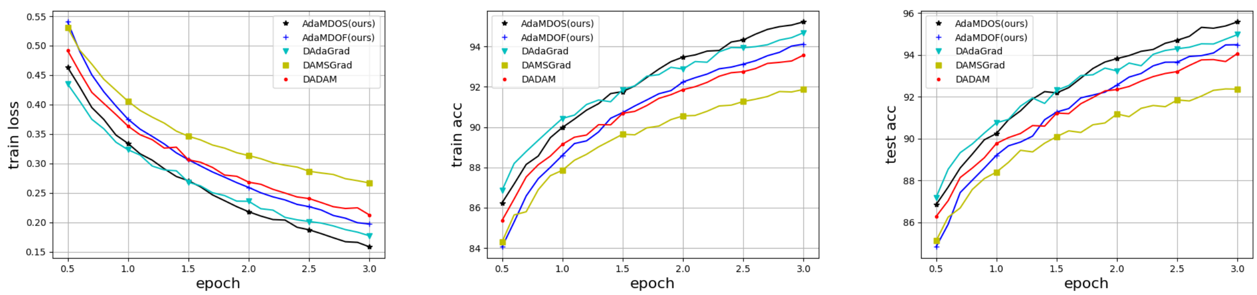


Figure 4: Training CNN on MNIST dataset: training loss vs epoch (Left), training accuracy (%) vs epoch (Middle), and test accuracy (%) vs epoch (Right) under the *3-regular* network.

tions at each node. In the experiment, for fair comparison, we use the batch size $b = 10$ in all algorithms, and set $\beta_1 = \beta_2 = 0.9$ in the DADAM (Nazari et al., 2022) and DAMSGrad (Chen et al., 2023), and set $\beta_1 = 0.9$ in the DAdaGrad (Chen et al., 2023), and set $\varrho = \beta_t = \eta_t = 0.9$ for all $t \geq 1$ in our algorithms.

Figures 1 and 2 show that our AdaMDOS method outperforms all comparisons, while due to requiring large batch-size, our AdaMDOF is comparable with the existing adaptive methods.

6.2. Training Convolutional Neural Network

In this subsection, we consider training a Convolutional Neural Network (CNN) for MNIST classification over a decentralized network. Here we use the same CNN architecture as in (McMahan et al., 2017). This CNN includes two 5×5 convolution layers (the first with 32 channels, the second with 64, each followed with 2×2 max pooling), a fully connected layer with 512 units and ReLu activation, and a final softmax output layer (1,663,370 total parameters). The MNIST dataset (LeCun et al., 2010) consists of 10 classes of 28×28 grayscale images, which includes 60,000 training examples and 10,000 testing examples, which we partitioned into 5 clients each containing 12000 training and 2000 testing examples.

In the experiment, we characterize performance of the algorithms in comparison in terms of the decrease of training loss versus epochs, where the loss denotes the objective function value in training CNN. Meanwhile, we also use the training accuracy and test accuracy, where the accuracy denotes the classification accuracy. For fair comparison, we use the batch size $b = 10$ in all algorithms, and set $\beta_1 = \beta_2 = 0.9$ in the DADAM (Nazari et al., 2022) and DAMSGrad (Chen et al., 2023), and set $\beta_1 = 0.9$ in the DAdaGrad (Chen et al., 2023), and set $\varrho = \beta_t = \eta_t = 0.9$ for all $t \geq 1$ in our algorithms. Figures 3 and 4 also show that our AdaMDOS method outperforms all comparisons, while due to requiring large batch size, our AdaMDOF is comparable with the existing adaptive methods.

6.3. Training Residual Network

In this subsection, we consider training a residual neural network for Tiny-ImageNet classification over a decentralized network. Here we use the ResNet-18 as in (He et al., 2016), which includes a 3×3 convolution layer followed with batch-norm and ReLU activation, eight residual blocks (start with 64 channels, channel number doubled at third, fifth and seventh block, end with 512 channels), a 4×4 max pooling, a fully connected layer with 512 units and ReLU activation, and a final softmax output layer. Each residual block contains a shortcut and two 3×3 convolution layers, the first followed with batch-norm and ReLU activation, the second followed with batch-norm.

The Tiny-ImageNet dataset (Le & Yang, 2015) consists of 200 classes of 64 x 64 RGB images, which includes 100,000 training examples and 10,000 testing examples, respectively. Here we partitioned into 5 clients, where each client contains 20,000 training and 2000 testing examples, respectively.

For fair comparison, we use the batch size $b = 10$ in all algorithms, and set $\beta_1 = \beta_2 = 0.9$ in the DADAM and DAMSGrad, and set $\beta_1 = 0.9$ in the DAdaGrad, and set $\varrho = \beta_t = \eta_t = 0.9$ for all $t \geq 1$ in our algorithms. In this experiment, we add two basic non-adaptive decentralized algorithms, i.e., D-PSGD (Lian et al., 2017) and D^2 (Tang et al., 2018), as the comparisons. From Figure 5, we find that although DADAM and D^2 methods outperform our AdaMDOS method at the beginning of the iteration, while our AdaMDOS then outperforms all comparisons.

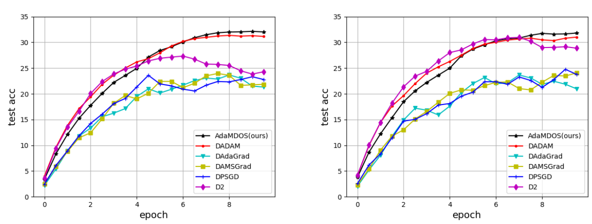


Figure 5: Training ResNet-18 on Tiny-ImageNet dataset: test accuracy (%) vs epoch under the *ring* network (Left) and the *3-regular* network (Right).

7. Conclusion

In the paper, we studied the distributed nonconvex stochastic and finite-sum optimization problems over a network. Moreover, we proposed a faster adaptive momentum-based decentralized optimization algorithm (i.e., AdaMDOS) to solve the stochastic problems, which reaches a near-optimal sample complexity of $\tilde{O}(\epsilon^{-3})$ for nonconvex stochastic optimization. Meanwhile, we proposed a faster adaptive momentum-based decentralized optimization algorithm (i.e., AdaMDOF) to solve the finite-sum problems, which obtains a near-optimal sample complexity of $O(\sqrt{n}\epsilon^{-2})$ for nonconvex finite-sum optimization. In particular, our methods use a unified adaptive matrix including various types of adaptive learning rate.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This paper was partially supported by NSFC under Grant No. 62376125. It was also partially supported by the Fundamental Research Funds for the Central Universities NO.NJ2023032.

Impact Statement

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

References

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pp. 699–707. PMLR, 2016.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1-2):165–214, 2023.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

Chen, X., Karimi, B., Zhao, W., and Li, P. On the convergence of decentralized adaptive gradient methods. In *Asian Conference on Machine Learning*, pp. 217–232. PMLR, 2023.

Chen, Z., Zhou, Y., Chen, R.-R., and Zou, S. Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. In *International Conference on Machine Learning*, pp. 3794–3834. PMLR, 2022.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Di Lorenzo, P. and Scutari, G. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011a.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011b.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hong, M., Hajinezhad, D., and Zhao, M.-M. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, pp. 1529–1538. PMLR, 2017.

Hoory, S., Linial, N., and Wigderson, A. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

Huang, F., Li, J., and Huang, H. Super-adam: faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34:9074–9085, 2021.

Kavis, A., Skoulakis, S., Antonakopoulos, K., Dadi, L. T., and Cevher, V. Adaptive stochastic variance reduction for non-convex finite-sum minimization. *Advances in Neural Information Processing Systems*, 35:23524–23538, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Le, Y. and Yang, X. S. Tiny imagenet visual recognition challenge. 2015.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Levy, K., Kavis, A., and Cevher, V. Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. *Advances in Neural Information Processing Systems*, 34:20571–20582, 2021.

Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992. PMLR, 2019.

Li, Z., Hanzely, S., and Richtárik, P. Zerosarah: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Nazari, P., Tarzanagh, D. A., and Michailidis, G. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *IEEE Transactions on Signal Processing*, 70:6065–6079, 2022.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Pan, T., Liu, J., and Wang, J. D-spider-sfo: A decentralized optimization algorithm with faster convergence rate for nonconvex problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1619–1626, 2020.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Sun, H., Lu, S., and Hong, M. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International conference on machine learning*, pp. 9217–9228. PMLR, 2020.

Sun, T., Li, D., and Wang, B. Adaptive random walk gradient descent for decentralized optimization. In *International Conference on Machine Learning*, pp. 20790–20809. PMLR, 2022.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. d^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071, 2022.

Xin, R., Khan, U., and Kar, S. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *International Conference on Machine Learning*, pp. 11459–11469. PMLR, 2021a.

Xin, R., Khan, U. A., and Kar, S. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 2021b.

Xin, R., Khan, U. A., and Kar, S. Fast decentralized nonconvex finite-sum optimization with recursive variance reduction. *SIAM Journal on Optimization*, 32(1):1–28, 2022.

Xu, J., Zhu, S., Soh, Y. C., and Xie, L. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060. IEEE, 2015.

Yun, J., Lozano, A. C., and Yang, E. Adaptive proximal gradient methods for structured neural networks. *Advances in Neural Information Processing Systems*, 34: 24365–24378, 2021.

Zhan, W., Wu, G., and Gao, H. Efficient decentralized stochastic gradient descent method for nonconvex finite-sum optimization problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9006–9013, 2022.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Zhang, X., Liu, J., Zhu, Z., and Bentley, E. S. Gt-storm: Taming sample, communication, and memory complexities in decentralized non-convex learning. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 271–280, 2021.

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S. C., Dvornek, N., Papademetris, X., and Duncan, J. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33: 18795–18806, 2020.

A. Appendix

In this section, we provide the detailed convergence analysis of our algorithms. We first give some notations and review some useful lemmas.

For the **stochastic** problem (1). Let $\mathbb{E}_t = \mathbb{E}_{\xi_t,\xi_{t-1},\cdots,\xi_1}$ with $\xi_t \in \{\xi_t^i\}_{i=1}^m$. Let $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_t^i$, $\bar{w}_t = \frac{1}{m}\sum_{i=1}^m w_t^i$, $\nabla f^i(x_t^i) = \mathbb{E}[\nabla f^i(x_t^i;\xi_t^i)]$ for all $i \in [m],\ t \geq 1$ and

$$\overline{\nabla f(x_t)} = \frac{1}{m}\sum_{i=1}^m \nabla f^i(x_t^i),\ \nabla F(\bar{x}_t) = \frac{1}{m}\sum_{i=1}^m \nabla f^i(\bar{x}_t). \tag{18}$$

For the **finite-sum** problem (2). Let $\mathbb{E}_t = \mathbb{E}_{\mathcal{I}_t,\mathcal{I}_{t-1},\cdots,\mathcal{I}_1}$ with $\mathcal{I}_t \in \{\mathcal{I}_t^i\}_{i=1}^m$.Let $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_t^i$, $\bar{w}_t = \frac{1}{m}\sum_{i=1}^m w_t^i$, $\nabla f^i(x_t^i) = \frac{1}{n}\sum_{k=1}^n \nabla f_k^i(x_t^i)$ for all $i \in [m]$ and

$$\overline{\nabla f(x_t)} = \frac{1}{m}\sum_{i=1}^m (\frac{1}{n}\sum_{k=1}^n \nabla f_k^i(x_t^i)) = \frac{1}{m}\sum_{i=1}^m \nabla f^i(x_t^i),\ \nabla F(\bar{x}_t) = \frac{1}{m}\sum_{i=1}^m \nabla f^i(\bar{x}_t). \tag{19}$$

Lemma A.1. *Given m vectors $\{u^i\}_{i=1}^m$, the following inequalities satisfy: $||u^i + u^j||^2 \leq (1+c)||u^i||^2 + (1+\frac{1}{c})||u^j||^2$ for any $c > 0$, and $||\frac{1}{m}\sum_{i=1}^m u^i||^2 \leq \frac{1}{m}\sum_{i=1}^m ||u^i||^2$.*

Lemma A.2. *Given a finite sequence $\{u^i\}_{i=1}^m$, and $\bar{u} = \frac{1}{m}\sum_{i=1}^m u^i$, the following inequality satisfies $\sum_{i=1}^m \|u^i - \bar{u}\|^2 \leq \sum_{i=1}^m \|u^i\|^2$.*

Lemma A.3. *The sequences $\{u_{t\geq 1}^i, w_{t\geq 1}^i\}_{i=1}^m$ be generated from our Algorithm 1 or 2, we have for all $t \geq 1$,*

$$\frac{1}{m}\sum_{i=1}^m u_t^i = \bar{u}_t = \bar{w}_t = \frac{1}{m}\sum_{i=1}^m w_t^i. \tag{20}$$

Proof. We proceed by induction. From our Algorithm 1, since $w_1^i = \sum_{j\in\mathcal{N}_i} W_{i,j}u_1^j$, we have

$$\bar{w}_1 = \frac{1}{m}\sum_{i=1}^m w_1^i = \frac{1}{m}\sum_{i=1}^m\sum_{j\in\mathcal{N}_i} W_{i,j}u_1^j = \frac{1}{m}\sum_{j=1}^m u_1^j\sum_{i=1}^m W_{i,j} = \frac{1}{m}\sum_{j=1}^m u_1^j = \bar{u}_1, \tag{21}$$

where the second last equality is due to $\sum_{i=1}^m W_{i,j} = 1$ from Assumption 3.4.

From the line 10 of Algorithm 1, we have for all $t \geq 1$

$$\begin{aligned}\bar{w}_{t+1} &= \frac{1}{m}\sum_{i=1}^m w_{t+1}^i = \frac{1}{m}\sum_{i=1}^m\sum_{j\in\mathcal{N}_i} W_{i,j}\big(w_t^j + u_{t+1}^j - u_t^j\big) \\ &= \frac{1}{m}\sum_{j=1}^m \big(w_t^j + u_{t+1}^j - u_t^j\big)\sum_{i=1}^m W_{i,j} = \frac{1}{m}\sum_{j=1}^m \big(w_t^j + u_{t+1}^j - u_t^j\big) = \bar{w}_t + \bar{u}_{t+1} - \bar{u}_t = \bar{u}_{t+1}, \end{aligned} \tag{22}$$

where the second last equality is due to $\mathcal{N}_i = \{j \in V \mid (i,j) \in E, j = i\}$ and $\sum_{i=1}^m W_{i,j} = 1$, and the last equality holds by the inductive hypothesis, i.e., $\bar{w}_t = \bar{u}_t$.

□

Lemma A.4. *Suppose the sequence $\left\{x_{t\geq 1}^i, \tilde{x}_{t\geq 1}^i\right\}_{i=1}^m$ be generated from Algorithm 1 or 2. Let $0 < \gamma \leq \frac{\rho}{4L\eta_t}$ for all $t \geq 0$, then we have*

$$F(\bar{x}_{t+1}) \leq F(\bar{x}_t) + \frac{2\gamma\eta_t}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\gamma\eta_t}{2\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta_t}{4}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2, \tag{23}$$

where $g_t^i = (A_t^i)^{-1}w_t^i$ for any $i \in [m]$.

Proof. Let $g_t^i = (A_t^i)^{-1} w_t^i$, we have $\tilde{x}_{t+1}^i = \sum_{j\in\mathcal{N}_i} W_{i,j} x_t^j - \gamma (A_t^i)^{-1} w_t^i$. Then we have

$$\begin{aligned}\bar{\tilde{x}}_{t+1} &= \frac{1}{m}\sum_{i=1}^m \tilde{x}_{t+1}^i = \frac{1}{m}\sum_{i=1}^m \sum_{j\in\mathcal{N}_i} W_{i,j} x_t^j - \gamma \frac{1}{m}\sum_{i=1}^m (A_t^i)^{-1} w_t^i \\ &= \frac{1}{m}\sum_{j\in\mathcal{N}_i} x_t^j \sum_{i=1}^m W_{i,j} - \gamma \frac{1}{m}\sum_{i=1}^m g_t^i = \bar{x}_t - \gamma \bar{g}_t, \end{aligned} \tag{24}$$

where the last equality is due to $\sum_{i=1}^m W_{i,j} = 1$. Since $A_t^i \succeq \rho I_d$ for all $i \in [m]$, we have

$$\begin{aligned}\rho\|g_t^i\|^2 &\le \langle A_t^i g_t^i, g_t^i\rangle = \langle w_t^i, g_t^i\rangle \overset{(i)}{=} \langle w_t^i - \bar{w}_t, g_t^i\rangle + \langle \bar{u}_t, g_t^i\rangle \\ &\le \frac{1}{2\rho}\|w_t^i - \bar{w}_t\|^2 + \frac{\rho}{2}\|g_t^i\|^2 + \langle \bar{u}_t, g_t^i\rangle, \end{aligned} \tag{25}$$

where the equality (i) is due to $\bar{u}_t = \bar{w}_t$. So we can obtain

$$0 \le \frac{1}{2\rho}\|w_t^i - \bar{w}_t\|^2 - \frac{\rho}{2}\|g_t^i\|^2 + \langle \bar{u}_t, g_t^i\rangle, \tag{26}$$

Then we have for any $i \in [m]$

$$0 \le \frac{\gamma\eta_t}{2\rho}\|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta_t}{2}\|g_t^i\|^2 + \gamma\eta_t\langle \bar{u}_t, g_t^i\rangle. \tag{27}$$

Then taking an average over i from 1 to m yields that

$$\begin{aligned} 0 &\le \frac{\gamma\eta_t}{2\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta_t}{2}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 + \gamma\eta_t \frac{1}{m}\sum_{i=1}^m \langle \bar{u}_t, g_t^i\rangle \\ &= \frac{\gamma\eta_t}{2\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta_t}{2}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 + \gamma\eta_t\langle \bar{u}_t, \bar{g}_t\rangle. \end{aligned} \tag{28}$$

According to Assumption 3.1, i.e., the function $F(x)$ is L-smooth, we have

$$\begin{aligned} F(\bar{x}_{t+1}) &\le F(\bar{x}_t) + \langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\ &= F(\bar{x}_t) + \eta_t\langle \nabla F(\bar{x}_t), \bar{\tilde{x}}_{t+1} - \bar{x}_t\rangle + \frac{L\eta_t^2}{2}\|\bar{\tilde{x}}_{t+1} - \bar{x}_t\|^2 \\ &= F(\bar{x}_t) + \eta_t\langle \nabla F(\bar{x}_t) - \bar{u}_t + \bar{u}_t, -\gamma\bar{g}_t\rangle + \frac{L\gamma^2\eta_t^2}{2}\|\bar{g}_t\|^2 \\ &\le F(\bar{x}_t) + \frac{2\gamma\eta_t}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\rho\gamma\eta_t}{8}\|\bar{g}_t\|^2 - \gamma\eta_t\langle \bar{u}_t, \bar{g}_t\rangle + \frac{L\gamma^2\eta_t^2}{2}\|\bar{g}_t\|^2 \\ &\le F(\bar{x}_t) + \frac{2\gamma\eta_t}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\rho\gamma\eta_t}{8}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 - \gamma\eta_t\langle \bar{u}_t, \bar{g}_t\rangle + \frac{L\gamma^2\eta_t^2}{2}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2, \end{aligned} \tag{29}$$

where the second equality is due to $\bar{x}_{t+1} = \bar{x}_t + \eta_t(\bar{\tilde{x}}_{t+1} - \bar{x}_t)$.

By summing the above inequalities (28) with (29), we can obtain

$$\begin{aligned} F(\bar{x}_{t+1}) &\le F(\bar{x}_t) + \frac{2\gamma\eta_t}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\gamma\eta_t}{2\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 - \Big(\frac{3\rho\gamma\eta_t}{8} - \frac{L\gamma^2\eta_t^2}{2}\Big)\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 \\ &\le F(\bar{x}_t) + \frac{2\gamma\eta_t}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\gamma\eta_t}{2\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta_t}{4}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2, \end{aligned} \tag{30}$$

where the last inequality is due to $\gamma \le \frac{\rho}{4L\eta_t}$ for all $t \ge 1$.

□

A.1. Convergence Analysis of AdaMDOS Algorithm

In this subsection, we provide the convergence analysis of our AdaMDOS Algorithm for **stochastic** optimization.

Lemma A.5. *Under the above assumptions, and assume the stochastic gradient estimators* $\{u^i_{t\geq 1}\}_{i=1}^m$ *be generated from Algorithm 1, we have*

$$\mathbb{E}\|u^i_{t+1} - \nabla f^i(x_{t+1})\|^2 \leq (1-\beta_{t+1})\mathbb{E}\|u^i_t - \nabla f^i(x_t)\|^2 + 2\beta_{t+1}^2\sigma^2 + 2L^2\eta_t^2\mathbb{E}\|\tilde{x}^i_{t+1} - x^i_t\|^2, \ \forall i \in [m] \tag{31}$$

$$\mathbb{E}\|\bar{u}_{t+1} - \overline{\nabla f(x_{t+1})}\|^2 \leq (1-\beta_{t+1})\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{m} + \frac{2L^2\eta_t^2}{m^2}\sum_{i=1}^m \mathbb{E}\|\tilde{x}^i_{t+1} - x^i_t\|^2, \tag{32}$$

where $\overline{\nabla f(x_t)} = \frac{1}{m}\sum_{i=1}^m \nabla f^i(x^i_t)$.

Proof. Since $u^i_{t+1} = \nabla f^i(x^i_{t+1};\xi^i_{t+1}) + (1-\beta_{t+1})(u^i_t - \nabla f^i(x^i_t;\xi^i_{t+1}))$ for any $i \in [m]$, we have

$$\begin{aligned}\bar{u}_{t+1} &= \frac{1}{m}\sum_{i=1}^m \big(\nabla f^i(x^i_{t+1};\xi^i_{t+1}) + (1-\beta_{t+1})(u^i_t - \nabla f^i(x^i_t;\xi^i_{t+1}))\big) \\ &= \overline{\nabla f(x_{t+1};\xi_{t+1})} + (1-\beta_{t+1})(\bar{u}_t - \overline{\nabla f(x_t;\xi_{t+1})}).\end{aligned} \tag{33}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\|\bar{u}_{t+1} - \overline{\nabla f(x_{t+1})}\|^2 \\
&= \mathbb{E}\|\frac{1}{m}\sum_{i=1}^m \big(u^i_{t+1} - \nabla f^i(x^i_{t+1})\big)\|^2 \\
&= \mathbb{E}\|\frac{1}{m}\sum_{i=1}^m \big(\nabla f^i(x^i_{t+1};\xi^i_{t+1}) + (1-\beta_{t+1})\big(u^i_t - \nabla f^i(x^i_t;\xi^i_{t+1})\big) - \nabla f^i(x^i_{t+1})\big)\|^2 \\
&= \mathbb{E}\|\frac{1}{m}\sum_{i=1}^m \Big(\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_{t+1}) - (1-\beta_{t+1})\big(\nabla f^i(x^i_t;\xi^i_{t+1}) - \nabla f^i(x^i_t)\big)\Big) \\
&\quad + (1-\beta_{t+1})\frac{1}{m}\sum_{i=1}^m \big(u^i_t - \nabla f^i(x^i_t)\big)\big\|^2 \\
&= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\big\|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_{t+1}) - (1-\beta_{t+1})\big(\nabla f^i(x^i_t;\xi^i_{t+1}) - \nabla f^i(x^i_t)\big)\big\|^2 \\
&\quad + (1-\beta_{t+1})^2\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\big\|^2 \\
&\leq \frac{2(1-\beta_{t+1})^2}{m^2}\sum_{i=1}^m \mathbb{E}\|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_{t+1}) - \nabla f^i(x^i_t;\xi^i_{t+1}) + \nabla f^i(x^i_t)\big)\|^2 \\
&\quad + \frac{2\beta_{t+1}^2}{m^2}\sum_{i=1}^m \mathbb{E}\|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_{t+1})\|^2 + (1-\beta_{t+1})^2\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \\
&\leq \frac{2(1-\beta_{t+1})^2}{m^2}\sum_{i=1}^m \mathbb{E}\|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_t;\xi^i_{t+1})\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{m} + (1-\beta_{t+1})^2\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \\
&\leq (1-\beta_{t+1})^2\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{m} + \frac{2(1-\beta_{t+1})^2L^2}{m^2}\sum_{i=1}^m \mathbb{E}\|x^i_{t+1} - x^i_t\|^2 \\
&\leq (1-\beta_{t+1})\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{m} + \frac{2L^2\eta_t^2}{m^2}\sum_{i=1}^m \mathbb{E}\|\tilde{x}^i_{t+1} - x^i_t\|^2,
\end{aligned}$$

where the forth equality holds by the following fact: for any $i \in [m]$,

$$\mathbb{E}_{\xi^i_{t+1}}\big[\nabla f^i(x^i_{t+1};\xi^i_{t+1})\big] = \nabla f^i(x^i_{t+1}), \ \mathbb{E}_{\xi^i_{t+1}}\big[\nabla f^i(x^i_t;\xi^i_{t+1}))\big] = \nabla f^i(x^i_t),$$

and for any $i \neq j \in [m]$, ξ_{t+1}^i and ξ_{t+1}^j are independent; the second inequality holds by the inequality $\mathbb{E}\|\zeta-\mathbb{E}[\zeta]\|^2 \leq \mathbb{E}\|\zeta\|^2$ and Assumption 3.4; the second last inequality is due to Assumption 3.2; the last inequality holds by $0 < \beta_{t+1} \leq 1$ and $x_{t+1}^i = x_t^i + \eta_t(\tilde{x}_{t+1}^i - x_t^i)$.

Similarly, we have

$$\mathbb{E}\|u_{t+1}^i - \nabla f^i(x_{t+1})\|^2 \leq (1-\beta_{t+1})\mathbb{E}\|u_t^i - \nabla f^i(x_t)\|^2 + 2\beta_{t+1}^2\sigma^2 + 2L^2\eta_t^2\mathbb{E}\|\tilde{x}_{t+1}^i - x_t^i\|^2, \ \forall i \in [m]. \tag{34}$$

□

Lemma A.6. *Given the sequence $\{x_{t\geq 1}^i, w_{t\geq 1}^i\}_{i=1}^m$ be generated from Algorithm 1. We have*

$$\sum_{i=1}^m \|x_{t+1}^i - \bar{x}_{t+1}\|^2 \leq (1 - \frac{(1-\nu^2)\eta_t}{2})\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + \frac{2\eta_t\gamma^2}{1-\nu^2}\sum_{i=1}^m \|g_t^i - \bar{g}_t\|^2,$$
$$\sum_{i=1}^m \|\tilde{x}_{t+1}^i - x_t^i\|^2 \leq (3+\nu^2)\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + \frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\sum_{i=1}^m \|g_t^i\|^2$$
$$\sum_{i=1}^m \|w_{t+1}^i - \bar{w}_{t+1}\|^2 \leq \nu\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 + \frac{\nu^2}{1-\nu}\big(4\beta_{t+1}^2\sum_{i=1}^m \|u_t^i - \nabla f^i(x_t^i)\|^2 + 4\beta_{t+1}^2 m\sigma^2 + 8\eta_t^2L^2\sum_{i=1}^m \|\tilde{x}_{t+1}^i - x_t^i\|^2\big).$$

Proof. For notational simplicity, let $x_t = [(x_t^1)^T, \cdots, (x_t^m)^T]^T \in \mathbb{R}^{md}$, $\tilde{x}_t = [(\tilde{x}_t^1)^T, \cdots, (\tilde{x}_t^m)^T]^T \in \mathbb{R}^{md}$ and $g_t = [(g_t^1)^T, \cdots, (g_t^m)^T]^T \in \mathbb{R}^{md}$ for all $t \geq 1$. By using Assumption 3.4, since $W\mathbf{1} = \mathbf{1}$ and $\tilde{W} = W \otimes I_d$, we have $\tilde{W}(\mathbf{1}\otimes\bar{x}_t) = \mathbf{1}\otimes\bar{x}_t$. Meanwhile, we have $\mathbf{1}^T(x_t - \mathbf{1}\otimes\bar{x}) = 0$ and $\tilde{W}\mathbf{1} = \mathbf{1}$. Thus, we have

$$\|\tilde{W}x_t - \mathbf{1}\otimes\bar{x}_t\|^2 = \|\tilde{W}(x_t - \mathbf{1}\otimes\bar{x}_t)\|^2 \leq \nu^2\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2, \tag{35}$$

where the above inequality holds by $x_t - \mathbf{1}\otimes\bar{x}_t$ is orthogonal to $\mathbf{1}$ that is the eigenvector corresponding to the largest eigenvalue of $\tilde{W}$, and ν denotes the second largest eigenvalue of $\tilde{W}$.

Since $\tilde{x}_{t+1}^i = \sum_{j\in\mathcal{N}_i} W_{i,j}x_t^j - \gamma(A_t^i)^{-1}w_t^i = \sum_{j\in\mathcal{N}_i} W_{i,j}x_t^j - \gamma g_t^i$ for all $i \in [m]$, we have $\tilde{x}_{t+1} = \tilde{W}x_t - \gamma g_t$ and $\bar{\tilde{x}}_{t+1} = \bar{x}_t - \gamma\bar{g}_t$. Since $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$ and $\bar{x}_{t+1} = \bar{x}_t + \eta_t(\bar{\tilde{x}}_{t+1} - \bar{x}_t)$, we have

$$\begin{aligned}
\sum_{i=1}^m \|x_{t+1}^i - \bar{x}_{t+1}\|^2 &= \big\|x_{t+1} - \mathbf{1}\otimes\bar{x}_{t+1}\big\|^2 \qquad (36)\\
&= \big\|x_t + \eta_t(\tilde{x}_{t+1} - x_t) - \mathbf{1}\otimes\big((\bar{x}_t + \eta_t(\bar{\tilde{x}}_{t+1} - \bar{x}_t)\big)\big\|^2\\
&\leq (1+\alpha_1)(1-\eta_t)^2\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + (1+\frac{1}{\alpha_1})\eta_t^2\|\tilde{x}_{t+1} - \mathbf{1}\otimes\bar{\tilde{x}}_{t+1}\|^2\\
&\overset{(i)}{=}(1-\eta_t)\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + \eta_t\|\tilde{x}_{t+1} - \mathbf{1}\otimes\bar{\tilde{x}}_{t+1}\|^2\\
&= (1-\eta_t)\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + \eta_t\|\tilde{W}x_t - \gamma g_t - \mathbf{1}\otimes\big(\bar{x}_t - \gamma\bar{g}_t\big)\|^2\\
&\leq (1-\eta_t)\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + (1+\alpha_2)\eta_t\|\tilde{W}x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + (1+\frac{1}{\alpha_2})\eta_t\gamma^2\|g_t - \mathbf{1}\otimes\bar{g}_t\|^2\\
&\overset{(ii)}{\leq}(1-\eta_t)\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + \frac{(1+\nu^2)\eta_t}{2}\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + \frac{\eta_t\gamma^2(1+\nu^2)}{1-\nu^2}\|g_t - \mathbf{1}\otimes\bar{g}_t\|^2\\
&\overset{(iii)}{\leq}(1 - \frac{(1-\nu^2)\eta_t}{2})\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + \frac{2\eta_t\gamma^2}{1-\nu^2}\|g_t - \mathbf{1}\otimes\bar{g}_t\|^2\\
&= (1 - \frac{(1-\nu^2)\eta_t}{2})\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + \frac{2\eta_t\gamma^2}{1-\nu^2}\sum_{i=1}^m \|g_t^i - \bar{g}_t\|^2, \qquad (37)
\end{aligned}$$

where the above equality (i) is due to $\alpha_1 = \frac{\eta_t}{1-\eta_t}$, and the second inequality (ii) holds by $\alpha_2 = \frac{1-\nu^2}{2\nu^2}$ and $\|\tilde{W}x_t - \mathbf{1}\otimes\bar{x}_t\|^2 \leq \nu^2\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2$, and the above inequality (ii) is due to $0 < \nu < 1$. Meanwhile, we have

$$
\begin{aligned}
\sum_{i=1}^m \|\tilde{x}_{t+1}^i - \bar{x}_t\|^2 &= \|\tilde{x}_{t+1} - \mathbf{1}\otimes\bar{x}_t\|^2 \\
&= \|\tilde{W}x_t - \gamma g_t - \mathbf{1}\otimes\bar{x}_t\|^2 \\
&\leq (1+\alpha_2)\nu^2\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 + (1+\frac{1}{\alpha_2})\gamma^2\|g_t\|^2 \\
&\overset{(i)}{=} \frac{1+\nu^2}{2}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 + \frac{1+\nu^2}{1-\nu^2}\gamma^2\sum_{i=1}^m\|g_t^i\|^2, \qquad (38)
\end{aligned}
$$

where the last equality (i) holds by $\alpha_2 = \frac{1-\nu^2}{2\nu^2}$. Then we have

$$
\begin{aligned}
\sum_{i=1}^m \|\tilde{x}_{t+1}^i - x_t^i\|^2 &= \|\tilde{x}_{t+1} - x_t\|^2 \\
&= \|\tilde{x}_{t+1} - \mathbf{1}\otimes\bar{x}_t + \mathbf{1}\otimes\bar{x}_t - x_t\|^2 \\
&\leq 2\|\tilde{x}_{t+1} - \mathbf{1}\otimes\bar{x}_t\|^2 + 2\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2 \\
&= (3+\nu^2)\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 + \frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\sum_{i=1}^m\|g_t^i\|^2. \qquad (39)
\end{aligned}
$$

Let $w_t = [(w_t^1)^T, (w_t^2)^T, \cdots, (w_t^m)^T]^T$, $u_t = [(u_t^1)^T, (u_t^2)^T, \cdots, (u_t^m)^T]^T$ and $\bar{w}_t = \frac{1}{m}\sum_{i=1}^m w_t^i$ and $\bar{u}_t = \frac{1}{m}\sum_{i=1}^m u_t^i$. Then we have for any $t \geq 1$,

$$
w_{t+1} = \tilde{W}\big(w_t + u_{t+1} - u_t\big).
$$

According to the above proof of Lemma A.3, we have $\bar{w}_{t+1} = \bar{w}_t + \bar{u}_{t+1} - \bar{u}_t$ for all $t \geq 1$. Thus we have

$$
\begin{aligned}
\sum_{i=1}^m \|w_{t+1}^i - \bar{w}_{t+1}\|^2 &= \|w_{t+1} - \mathbf{1}\otimes\bar{w}_{t+1}\|^2 \\
&= \big\|\tilde{W}\big(w_t + u_{t+1} - u_t\big) - \mathbf{1}\otimes\big(\bar{w}_t + \bar{u}_{t+1} - \bar{u}_t\big)\big\|^2 \\
&\leq (1+c)\|\tilde{W}w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\big\|\tilde{W}\big(u_{t+1} - u_t\big) - \mathbf{1}\otimes\big(\bar{u}_{t+1} - \bar{u}_t\big)\big\|^2 \\
&\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big\|u_{t+1} - u_t - \mathbf{1}\otimes\big(\bar{u}_{t+1} - \bar{u}_t\big)\big\|^2 \\
&\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big\|u_{t+1} - u_t\big\|^2, \qquad (40)
\end{aligned}
$$

where the last inequality holds by Lemma A.2.

Since $u^i_{t+1} = \nabla f^i(x^i_{t+1};\xi^i_{t+1}) + (1-\beta_{t+1})(u^i_t - \nabla f^i(x^i_t;\xi^i_{t+1}))$ for any $i\in[m]$ and $t\geq 1$, we have

$$\begin{aligned}
\big\|u_{t+1}-u_t\big\|^2 &= \sum_{i=1}^m \|u^i_{t+1}-u^i_t\|^2 \\
&= \sum_{i=1}^m \|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \beta_{t+1}u^i_t - (1-\beta_{t+1})\nabla f^i(x^i_t;\xi^i_{t+1})\|^2 \\
&= \sum_{i=1}^m \|\beta_{t+1}(\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_{t+1})) + \beta_{t+1}(\nabla f^i(x^i_{t+1}) - \nabla f^i(x^i_t)) + \beta_{t+1}(\nabla f^i(x^i_t) - u^i_t) \\
&\quad + (1-\beta_{t+1})(\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_t;\xi^i_{t+1}))\|^2 \\
&= 4\beta^2_{t+1}\sum_{i=1}^m \|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_{t+1})\|^2 + 4\beta^2_{t+1}\sum_{i=1}^m \|\nabla f^i(x^i_{t+1}) - \nabla f^i(x^i_t)\|^2 \\
&\quad + 4\beta^2_{t+1}\sum_{i=1}^m \|\nabla f^i(x^i_t) - u^i_t\|^2 + 4(1-\beta_{t+1})^2\sum_{i=1}^m \|\nabla f^i(x^i_{t+1};\xi^i_{t+1}) - \nabla f^i(x^i_t;\xi^i_{t+1})\|^2 \\
&\leq 4\beta^2_{t+1}\sum_{i=1}^m \|u^i_t - \nabla f^i(x^i_t)\|^2 + 4\beta^2_{t+1}m\sigma^2 + 4\beta^2_{t+1}L^2\sum_{i=1}^m \|x^i_{t+1}-x^i_t\|^2 \\
&\quad + 4(1-\beta_{t+1})^2L^2\sum_{i=1}^m \|x^i_{t+1}-x^i_t\|^2 \\
&\leq 4\beta^2_{t+1}\sum_{i=1}^m \|u^i_t - \nabla f^i(x^i_t)\|^2 + 4\beta^2_{t+1}m\sigma^2 + 8\eta_t^2L^2\sum_{i=1}^m \|\tilde{x}^i_{t+1}-x^i_t\|^2,
\end{aligned} \tag{41}$$

where the last inequality holds by $0<\beta_t<1$ and $x^i_{t+1} = x^i_t + \eta_t(\tilde{x}^i_{t+1}-x^i_t)$.

Plugging the above inequalities (41) into (40), we have

$$\begin{aligned}
\sum_{i=1}^m \|w^i_{t+1}-\bar{w}_{t+1}\|^2 &\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big\|u_{t+1}-u_t\big\|^2 \\
&\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big(4\beta^2_{t+1}\sum_{i=1}^m \|u^i_t - \nabla f^i(x^i_t)\|^2 + 4\beta^2_{t+1}m\sigma^2 \\
&\quad + 8\eta_t^2L^2\sum_{i=1}^m \|\tilde{x}^i_{t+1}-x^i_t\|^2\big).
\end{aligned} \tag{42}$$

Let $c=\frac{1}{\nu}-1$, we have

$$\begin{aligned}
\sum_{i=1}^m \|w^i_{t+1}-\bar{w}_{t+1}\|^2 \leq \nu\sum_{i=1}^m \|w^i_t-\bar{w}_t\|^2 + \frac{\nu^2}{1-\nu}\big(4\beta^2_{t+1}\sum_{i=1}^m \|u^i_t - \nabla f^i(x^i_t)\|^2 \\
+ 4\beta^2_{t+1}m\sigma^2 + 8\eta_t^2L^2\sum_{i=1}^m \|\tilde{x}^i_{t+1}-x^i_t\|^2\big).
\end{aligned} \tag{43}$$

□

Theorem A.7. *(Restatement of Theorem 5.1) Suppose the sequences* $\big\{\{x^i_t\}^m_{i=1}\big\}^T_{t=1}$ *be generated from Algorithm 1. Under the above Assumptions 3.1-3.5, and let* $\eta_t=\eta$, $0<\beta_t\leq 1$ *for all* $t\geq 0$, $\gamma\leq\min\big(\frac{\rho(1-\nu^2)}{48\theta_t}, \frac{3\rho(1-\nu^2)\theta_t}{58L^2}\big)$,

$\eta \leq \min\big(\frac{\rho\sqrt{1-\nu^2}}{4L\gamma\sqrt{3(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta_t}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)$ *with* $H_t = \frac{9}{2\beta_t} + \frac{8\nu^2}{(1-\nu)^2}$ *for all* $t \geq 1$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\|]$$
$$\leq \Big(\frac{6\sqrt{G}}{\sqrt{T}} + \frac{12\sigma}{\rho}\sqrt{\frac{1}{T}\sum_{t=1}^{T}H_t\beta_t^2}\Big)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{44}$$

where $G = \frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta} + \big(\frac{4\nu^2}{\rho^2(1-\nu)} + \frac{9}{2\rho^2\beta_0} - \frac{9}{2\rho^2}\big)\sigma^2$.

Proof. Without loss of generality, let $\eta = \eta_1 = \cdots = \eta_T$. According to Lemma A.4, we have

$$F(\bar{x}_{t+1}) \leq F(\bar{x}_t) + \frac{2\gamma\eta}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\gamma\eta}{2\rho}\frac{1}{m}\sum_{i=1}^{m}\|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta}{4}\frac{1}{m}\sum_{i=1}^{m}\|g_t^i\|^2. \tag{45}$$

According to the Lemma A.5, we have

$$\mathbb{E}\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \leq (1-\beta_t)\mathbb{E}\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2\beta_t^2\sigma^2}{m} + \frac{2L^2\eta^2}{m^2}\sum_{i=1}^{m}\mathbb{E}\|\tilde{x}_t^i - x_{t-1}^i\|^2, \tag{46}$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|u_t^i - \nabla f^i(x_t)\|^2 \leq (1-\beta_t)\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|u_{t-1}^i - \nabla f^i(x_{t-1})\|^2 + 2\beta_t^2\sigma^2 + 2L^2\eta^2\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\tilde{x}_t^i - x_{t-1}^i\|^2. \tag{47}$$

According to Lemma A.6, we have

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|w_t^i - \bar{w}_t\|^2 \leq \nu\frac{1}{m}\sum_{i=1}^{m}\|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \frac{4\nu^2}{1-\nu}\frac{1}{m}\sum_{i=1}^{m}\Big(2L^2\eta^2\|\tilde{x}_t^i - x_{t-1}^i\|^2 + \beta_t^2\|\nabla f^i(x_{t-1}^i) - u_{t-1}^i\|^2$$
$$+ \beta_t^2 m\sigma^2\Big). \tag{48}$$

Meanwhile, we also have

$$\frac{1}{m}\sum_{i=1}^{m}\|x_t^i - \bar{x}_t\|^2 \leq (1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^{m}\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \frac{2\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^{m}\|g_{t-1}^i - \bar{g}_{t-1}\|^2$$
$$\leq (1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^{m}\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \frac{2\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^{m}(\|g_{t-1}^i\|^2 + \|\bar{g}_{t-1}\|^2)$$
$$\leq (1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^{m}\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^{m}\|g_{t-1}^i\|^2. \tag{49}$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\|\tilde{x}_t^i - x_{t-1}^i\|^2 \leq (3+\nu^2)\frac{1}{m}\sum_{i=1}^{m}\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\frac{1}{m}\sum_{i=1}^{m}\|g_{t-1}^i\|^2. \tag{50}$$

Next considering the term $\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2$, we have

$$\begin{aligned}
\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 &= \|\bar{u}_t - \overline{\nabla f(x_t)} + \overline{\nabla f(x_t)} - \nabla F(\bar{x}_t)\|^2 \\
&\leq 2\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + 2\|\overline{\nabla f(x_t)} - \nabla F(\bar{x}_t)\|^2 \\
&\leq 2\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + 2\|\frac{1}{m}\sum_{i=1}^m \nabla f^i(x_t^i) - \frac{1}{m}\sum_{i=1}^m \nabla f^i(\bar{x}_t)\|^2 \\
&\leq 2\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + \frac{2L^2}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2,
\end{aligned}$$

where the last inequality is due to Assumption 3.1. Then we can obtain

$$-\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \leq -\frac{1}{2}\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 + \frac{L^2}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2. \tag{51}$$

Since $\bar{u}_t = \bar{w}_t$ for all $t \geq 1$, we have

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \|w_t^i - \nabla F(x_t^i)\|^2 &= \frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t + \bar{u}_t - \nabla F(\bar{x}_t) + \nabla F(\bar{x}_t) - \nabla F(x_t^i)\|^2 \\
&\leq 3\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 + 3\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 + 3\frac{1}{m}\sum_{i=1}^m \|\nabla F(\bar{x}_t) - \nabla F(x_t^i)\|^2 \\
&\leq 3\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 + 3\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 + 3L^2\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2.
\end{aligned}$$

Then we have

$$-\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 \leq -\frac{1}{3m}\sum_{i=1}^m \|w_t^i - \nabla F(x_t^i)\|^2 + \frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 + \frac{L^2}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2. \tag{52}$$

We define a useful Lyapunov function (i.e., potential function), for any $t \geq 1$

$$\begin{aligned}
\Omega_t = \mathbb{E}_t\big[F(\bar{x}_t) &+ (\lambda_{t-1} - \frac{9\gamma\eta}{2\rho})\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \chi_{t-1}\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\
&+ (\theta_{t-1} - \frac{19\gamma\eta L^2}{4\rho})\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 + (\vartheta_{t-1} - \frac{\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2 \\
&+ \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2\big],
\end{aligned} \tag{53}$$

where $\chi_{t-1} \geq 0$, $\lambda_{t-1} \geq \frac{9\gamma\eta}{2\rho}$, $\theta_{t-1} \geq \frac{29\gamma\eta L^2}{6\rho}$ and $\vartheta_{t-1} \geq \frac{\gamma\eta}{4\rho}$ for all $t \geq 1$. Then we have

$$\begin{aligned}
\Omega_{t+1} &= \mathbb{E}_{t+1}\big[F(\bar{x}_{t+1}) + (\lambda_t - \frac{9\gamma\eta}{2\rho})\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + \chi_t \frac{1}{m}\sum_{i=1}^m \|u_t^i - \nabla f^i(x_t^i)\|^2 \\
&\quad + (\theta_t - \frac{29\gamma\eta L^2}{6\rho})\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + (\vartheta_t - \frac{3\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 + \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2\big] \\
&\overset{(i)}{\leq} \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{4\rho}\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 + \frac{9\gamma\eta L^2}{2\rho}\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + \lambda_t\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \\
&\quad + \chi_t \frac{1}{m}\sum_{i=1}^m \|u_t^i - \nabla f^i(x_t^i)\|^2 + (\theta_t - \frac{29\gamma\eta L^2}{6\rho})\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + (\vartheta_t - \frac{\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2\Big] \\
&\overset{(ii)}{\leq} \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 \\
&\quad + \lambda_t(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2\lambda_t\beta_t^2\sigma^2}{m} + \frac{2\lambda_t L^2\eta^2}{m^2}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 \\
&\quad + \chi_t(1-\beta_t)\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1})\|^2 + 2\chi_t\beta_t^2\sigma^2 + 2\chi_t L^2\eta^2\frac{1}{m}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 \\
&\quad + \theta_t(1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \theta_t\frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2 \\
&\quad + \vartheta_t\nu\frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \frac{4\vartheta_t\nu^2}{1-\nu}\frac{1}{m}\sum_{i=1}^m \Big(2L^2\eta^2\|\tilde{x}_t^i - x_{t-1}^i\|^2 + \beta_t^2\|\nabla f^i(x_{t-1}^i) - u_{t-1}^i\|^2 \\
&\quad + \beta_t^2 m\sigma^2\Big)\Big],
\end{aligned} \tag{54}$$

where the inequality (i) is due to the above inequalities (45) and (51); and the inequality (ii) holds by the above inequalities (46), (47), (49) and (52).

Then we have

$$
\begin{aligned}
\Omega_{t+1} &\leq \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 \\
&\quad + \lambda_t(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \big(\chi_t - \chi_t\beta_t + \frac{4\vartheta_t\beta_t^2\nu^2}{1-\nu}\big)\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\
&\quad + 2L^2\eta^2\big(\lambda_t + \chi_t + \frac{4\vartheta_t\nu^2}{1-\nu}\big)\frac{1}{m}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 + \theta_t(1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 \\
&\quad + \theta_t\frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2 + \vartheta_t\nu\frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \big(\frac{2\lambda_t}{m} + 2\chi_t + \frac{4\vartheta_t\nu^2}{1-\nu}\big)\beta_t^2\sigma^2\Big] \\
&\overset{(i)}{\leq} \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 \\
&\quad + \lambda_t(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \big(\chi_t - \chi_t\beta_t + \frac{4\vartheta_t\beta_t^2\nu^2}{1-\nu}\big)\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\
&\quad + 2L^2\eta^2\big(\lambda_t + \chi_t + \frac{4\vartheta_t\nu^2}{1-\nu}\big)\Big((3+\nu^2)\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2\Big) \\
&\quad + \theta_t(1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \theta_t\frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2 \\
&\quad + \vartheta_t\nu\frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \big(\frac{2\lambda_t}{m} + 2\chi_t + \frac{4\vartheta_t\nu^2}{1-\nu}\big)\beta_t^2\sigma^2\Big], \qquad (55)
\end{aligned}
$$

where the inequality (i) is due to the above inequality (50).

Since $0 < \beta_t < 1$ for all $t \geq 1$, $\lambda_t = \frac{9\gamma\eta}{2\rho\beta_t} \geq \frac{9\gamma\eta}{2\rho}$ and $\lambda_t \leq \lambda_{t-1}$, then we have $\lambda_t(1-\beta_t) \leq \lambda_{t-1} - \frac{9\gamma\eta}{2\rho}$. Let $\chi_t = \frac{4\vartheta_t\nu^2}{1-\nu} \geq \frac{4\vartheta_t\beta_t\nu^2}{1-\nu}$ and $\chi_t \leq \chi_{t-1}$, we have $\chi_t - \chi_t\beta_t + \frac{4\vartheta_t\beta_t^2\nu^2}{1-\nu} \leq \chi_{t-1}$. Let $\vartheta_t = \frac{\gamma\eta}{\rho(1-\nu)}$ for all $t \geq 1$, since $0 < \nu < 1$, we have $\vartheta_t\nu = \vartheta_t - (1-\nu)\vartheta_t \leq \vartheta_t - \frac{3\gamma\eta}{4\rho} = \vartheta_{t-1} - \frac{3\gamma\eta}{4\rho}$. Meanwhile, let $\theta_t \leq \theta_{t-1}$ for all $t \geq 1$, $\gamma \leq \min\big(\frac{\rho(1-\nu^2)}{48\theta_t}, \frac{3\rho(1-\nu^2)\theta_t}{58L^2}\big)$, $\eta \leq \min\big(\frac{\rho\sqrt{1-\nu^2}}{4L\gamma\sqrt{3(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta_t}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)$ with $H_t = \frac{9}{2\beta_t} + \frac{8\nu^2}{(1-\nu)^2}$ for all $t \geq 1$, we can obtain

$$
\theta_t(1 - \frac{(1-\nu^2)\eta}{2}) + 2L^2\eta^2(3+\nu^2)\big(\lambda_t + \chi_t + \frac{4\vartheta_t\nu^2}{1-\nu}\big) \leq \theta_{t-1} - \frac{29\gamma\eta L^2}{6\rho}, \qquad (56)
$$

$$
\theta_t\frac{4\eta\gamma^2}{1-\nu^2} + \frac{4(1+\nu^2)}{1-\nu^2}L^2\eta^2\gamma^2\big(\lambda_t + \chi_t + \frac{4\vartheta_t\nu^2}{1-\nu}\big) \leq \frac{\rho\gamma\eta}{6}. \qquad (57)
$$

Based on the choice of these parameters and the above inequality (55), we have

$$\begin{aligned}
\Omega_{t+1} &\leq \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 \\
&\quad + (\lambda_{t-1} - \frac{9\gamma\eta}{2\rho})\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \chi_{t-1}\frac{1}{m}\sum_{i=1}^m\|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\
&\quad + (\theta_{t-1} - \frac{29\gamma\eta L^2}{6\rho})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + (\vartheta_{t-1} - \frac{3\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m\|w_{t-1}^i - \bar{w}_{t-1}\|^2 \\
&\quad + \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2 + \frac{2\gamma\eta H_t}{\rho}\beta_t^2\sigma^2\Big] \\
&= \Omega_t - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 + \frac{2\gamma\eta H_t}{\rho}\beta_t^2\sigma^2. \qquad (58)
\end{aligned}$$

Then we can obtain

$$\frac{1}{\rho^2}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 + \frac{L^2}{\rho^2}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 + \frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 \leq \frac{12(\Omega_t - \Omega_{t+1})}{\gamma\rho\eta} + \frac{48H_t}{\rho^2}\beta_t^2\sigma^2. \qquad (59)$$

Since $x_0^1 = \tilde{x}_0^1 = \cdots = x_0^m = \tilde{x}_0^m$, $u_0^1 = \cdots = u_0^m$ and $w_0^1 = \cdots = w_0^m = 0$, we have

$$\begin{aligned}
\Omega_1 &= \mathbb{E}\big[F(\bar{x}_1) + (\lambda_0 - \frac{9\gamma\eta}{2\rho})\|\bar{u}_0 - \overline{\nabla f(x_0)}\|^2 + \chi_0\frac{1}{m}\sum_{i=1}^m\|u_0^i - \nabla f^i(x_0^i)\|^2 \\
&\leq F(\bar{x}_1) + \big(\chi_0 + \lambda_0 - \frac{9\gamma\eta}{2\rho}\big)\sigma^2 \\
&= F(\bar{x}_1) + \big(\frac{4\gamma\eta\nu^2}{\rho(1-\nu)^2} + \frac{9\gamma\eta}{2\rho\beta_0} - \frac{9\gamma\eta}{2\rho}\big)\sigma^2. \qquad (60)
\end{aligned}$$

Let $\mathcal{M}_t^i = \|g_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\|$. Then we have

$$\begin{aligned}
\mathcal{M}_t^i &= \|g_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\
&\overset{(i)}{=} \|(A_t^i)^{-1}w_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\
&= \frac{1}{\|A_t^i\|}\|A_t^i\|\|(A_t^i)^{-1}w_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\
&\geq \frac{1}{\|A_t^i\|}\|w_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\
&\overset{(ii)}{\geq} \frac{1}{\|A_t^i\|}\|w_t^i\| + \frac{1}{\|A_t^i\|}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\|A_t^i\|}\|\bar{x}_t - x_t^i\| \\
&\geq \frac{1}{\|A_t^i\|}\big(\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\|\big), \qquad (61)
\end{aligned}$$

where the equality (i) holds by $g_t^i = (A_t^i)^{-1}w_t^i$, and the inequality (ii) holds by $\|A_t^i\| \geq \rho$ for all $t \geq 1$ due to Assumption 3.4. Then we have

$$\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\| \leq \mathcal{M}_t^i\|A_t^i\|. \qquad (62)$$

According to the above inequality (59), we have

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\mathcal{M}_t^i]^2 \leq \frac{1}{T}\sum_{t=1}^{T}\big[\frac{3}{\rho^2}\frac{1}{m}\sum_{i=1}^{m}\|w_t^i-\nabla F(x_t^i)\|^2+\frac{3L^2}{\rho^2}\frac{1}{m}\sum_{i=1}^{m}\|x_t^i-\bar{x}_t\|^2+\frac{3}{m}\sum_{i=1}^{m}\|g_t^i\|^2\big]$$
$$\leq \frac{1}{T}\sum_{t=1}^{T}\frac{36(\Omega_t-\Omega_{t+1})}{\gamma\rho\eta} \leq \frac{36(\Omega_1-F^*)}{T\gamma\rho\eta}+\frac{1}{T}\sum_{t=1}^{T}\frac{144H_t}{\rho^2}\beta_t^2\sigma^2. \tag{63}$$

By using Cauchy-Schwarz inequality to the above inequality (62), we have

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|] \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\big[\mathcal{M}_t^i\|A_t^i\|\big]$$
$$\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\mathcal{M}_t^i]^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{64}$$

By plugging the above inequalities (64) into (63), we can obtain

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|] \leq \big(\frac{6\sqrt{\Omega_1-F^*}}{\sqrt{T\gamma\rho\eta}}+\frac{12\sigma\sqrt{\frac{1}{T}\sum_{t=1}^{T}H_t\beta_t^2}}{\rho}\big)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{65}$$

Since $F(x)$ is L-smooth, we have

$$\|\nabla F(\bar{x}_t)\| = \|\nabla F(\bar{x}_t)-\nabla F(x_t^i)+\nabla F(x_t^i)\| \leq \|\nabla F(x_t^i)\|+L\|x_t^i-\bar{x}_t\|. \tag{66}$$

Meanwhile, let $G=\frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta}+\big(\frac{4\nu^2}{\rho^2(1-\nu)}+\frac{9}{2\rho^2\beta_0}-\frac{9}{2\rho^2}\big)\sigma^2$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|]$$
$$\leq \big(\frac{6\sqrt{\Omega_1-F^*}}{\sqrt{T\gamma\rho\eta}}+\frac{12\sigma\sqrt{\frac{1}{T}\sum_{t=1}^{T}H_t\beta_t^2}}{\rho}\big)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}$$
$$= \big(\frac{6\sqrt{G}}{\sqrt{T}}+\frac{12\sigma\sqrt{\frac{1}{T}\sum_{t=1}^{T}H_t\beta_t^2}}{\rho}\big)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{67}$$

Let $\beta_t=\frac{1}{T^{2/3}}$ for all $t\geq 1$, we have

$$H_t=\frac{9}{2\beta_t}+\frac{8\nu^2}{(1-\nu)^2}=O(T^{2/3}). \tag{68}$$

We set $\theta_t=\theta\geq\frac{29\gamma\eta L^2}{6\rho}$ for all $t\geq 1$. Then we can set $\gamma=O(1)$ and $\eta=O(\frac{1}{T^{1/3}})$. Thus we have $G=\frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta}+\big(\frac{4\nu^2}{\rho^2(1-\nu)}+\frac{9}{2\rho^2\beta_0}-\frac{9}{2\rho^2}\big)\sigma^2=O(T^{1/3})$, and then we can obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|]$$
$$\leq O(\frac{1}{T^{1/3}}+\frac{\sigma}{T^{1/3}})\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{69}$$

Given $\sqrt{\frac{1}{T}\sum_{t=1}^T\frac{1}{m}\sum_{i=1}^m\mathbb{E}\|A_t^i\|^2} = O(1)$, set

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq O(\frac{1}{T^{1/3}} + \frac{\sigma}{T^{1/3}})\sqrt{\frac{1}{T}\sum_{t=1}^T\frac{1}{m}\sum_{i=1}^m\mathbb{E}\|A_t^i\|^2} \leq \epsilon, \tag{70}$$

we have $T = O(\epsilon^{-3})$. Since our Algorithm 1, it reaches a near-optimal sample complexity of $1 \cdot T = O(\epsilon^{-3})$ for finding an ϵ-stationary solution of Problem (1).

□

A.2. Convergence Analysis of AdaMDOF Algorithm

In this subsection, we provide the convergence analysis of our AdaMDOF Algorithm for **finite-sum** optimization.

Lemma A.8. *Under the above Assumptions 3.1-3.2, and assume the gradient estimators* $\{u^i_{t\geq 1}\}_{i=1}^m$ *be generated from Algorithm 2, we have*

$$\begin{aligned}\mathbb{E}_t\|u_t^i - \nabla f^i(x_t^i)\|^2 \leq (1-\beta_t)\|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 + \frac{2\beta_t^2}{b}\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\ + \frac{2L^2\eta_{t-1}^2}{b}\|\tilde{x}_t^i - x_{t-1}^i\|^2, \ \forall i \in [m] \end{aligned} \tag{71}$$

$$\begin{aligned}\mathbb{E}_t\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \leq (1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2L^2\eta_t^2}{bm}\sum_{i=1}^m\|\tilde{x}_t^i - x_{t-1}^i\|^2 \\ + \frac{2\beta_t^2}{bm}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2, \end{aligned} \tag{72}$$

where $\beta_t \in (0,1)$ *for all* $t \geq 1$.

Proof. According the line 6 of Algorithm 2, i.e, $u_t^i = \frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big) + (1-\beta_t)u_{t-1}^i +$

$\beta_t\Big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\big)+\frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i\Big)$, we have

$$
\begin{aligned}
&\mathbb{E}_t\|u_t^i-\nabla f^i(x_t^i)\|^2 \qquad (73)\\
&=\mathbb{E}_t\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)\big)+(1-\beta_t)u_{t-1}^i+\beta_t\Big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\big)+\frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i\Big)-\nabla f^i(x_t^i)\Big\|^2\\
&=\mathbb{E}_t\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)\big)-\nabla f^i(x_t^i)+\nabla f^i(x_{t-1}^i)+(1-\beta_t)(u_{t-1}^i-\nabla f^i(x_{t-1}^i))\\
&\quad+\beta_t\big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\big)+\frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i-\nabla f^i(x_{t-1}^i)\big)\Big\|^2\\
&=\mathbb{E}_t\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)\big)-\nabla f^i(x_t^i)+\nabla f^i(x_{t-1}^i)+\beta_t\big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\big)\\
&\quad+\frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i-\nabla f^i(x_{t-1}^i)\big)\Big\|^2+(1-\beta_t)^2\|u_{t-1}^i-\nabla f^i(x_{t-1}^i)\|^2\\
&\le 2\mathbb{E}_t\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)\big)-\nabla f^i(x_t^i)+\nabla f^i(x_{t-1}^i)\Big\|^2\\
&\quad+2\beta_t^2\mathbb{E}_t\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\big)+\frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i-\nabla f^i(x_{t-1}^i)\Big\|^2+(1-\beta_t)^2\|u_{t-1}^i-\nabla f^i(x_{t-1}^i)\|^2\\
&\le\frac{2L^2}{b}\|x_t^i-x_{t-1}^i\|^2+\frac{2\beta_t^2}{b}\frac{1}{n}\sum_{k=1}^n\|z_{k,t-1}^i-\nabla f_k^i(x_{t-1}^i)\|^2+(1-\beta_t)^2\|u_{t-1}^i-\nabla f^i(x_{t-1}^i)\|^2\\
&\le\frac{2L^2\eta_{t-1}^2}{b}\|\tilde{x}_t^i-x_{t-1}^i\|^2+\frac{2\beta_t^2}{b}\frac{1}{n}\sum_{k=1}^n\|z_{k,t-1}^i-\nabla f_k^i(x_{t-1}^i)\|^2+(1-\beta_t)\|u_{t-1}^i-\nabla f^i(x_{t-1}^i)\|^2, \qquad (74)
\end{aligned}
$$

where the third equality holds by

$$
\begin{aligned}
&\mathbb{E}_t\big[\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)\big)\big]=\nabla f^i(x_t^i)-\nabla f^i(x_{t-1}^i),\\
&\mathbb{E}_t\big[\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\big)\big]=\nabla f^i(x_{t-1}^i)-\frac{1}{n}\sum_{j=1}^n z_{t-1}^j,
\end{aligned}
$$

and the second last inequality holds by the inequality $\mathbb{E}\|\zeta-\mathbb{E}[\zeta]\|^2\le\mathbb{E}\|\zeta\|^2$ and Assumption 3.1; the last inequality holds by $0<\beta_t\le 1$ and $x_t^i=x_{t-1}^i+\eta_{t-1}(\tilde{x}_t-x_{t-1}^i)$.

Let $\bar{u}_t=\frac{1}{m}\sum_{i=1}^m u_t^i$, $\nabla f^i(x_t^i)=\frac{1}{n}\sum_{k=1}^n\nabla f_k^i(x_t^i)$ for all $i\in[m]$ and $\overline{\nabla f(x_t)}=\frac{1}{m}\sum_{i=1}^m(\frac{1}{n}\sum_{k=1}^n\nabla f_k^i(x_t^i))=$

$\frac{1}{m}\sum_{i=1}^m \nabla f^i(x_t^i)$. We can obtain

$$
\begin{aligned}
&\mathbb{E}_t\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 = \mathbb{E}_t\big\|\frac{1}{m}\sum_{i=1}^m (u_t^i - \nabla f^i(x_t^i))\big\|^2 \\
&= \mathbb{E}_t\big\|\frac{1}{m}\sum_{i=1}^m \Big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big) - \nabla f^i(x_t^i) + \nabla f^i(x_{t-1}^i) + (1-\beta_t)(u_{t-1}^i - \nabla f^i(x_{t-1}^i)) \\
&\quad + \beta_t\big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\big) + \frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i - \nabla f^i(x_{t-1}^i)\big)\Big)\big\|^2 \\
&\overset{(i)}{=}(1-\beta_t)^2\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \mathbb{E}_t\big\|\frac{1}{m}\sum_{i=1}^m \big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big) - \nabla f^i(x_t^i) + \nabla f^i(x_{t-1}^i) \\
&\quad + \beta_t\big(\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\big) + \frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i - \nabla f^i(x_{t-1}^i)\big)\big)\big\|^2 \\
&\leq (1-\beta_t)^2\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2}{m}\sum_{i=1}^m \mathbb{E}_t\big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big) - \nabla f^i(x_t^i) + \nabla f^i(x_{t-1}^i)\big\|^2 \\
&\quad + \frac{2\beta_t^2}{m}\sum_{i=1}^m \mathbb{E}_t\big\|\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\big) + \frac{1}{n}\sum_{j=1}^n z_{j,t-1}^i - \nabla f^i(x_{t-1}^i)\big\|^2 \\
&\leq (1-\beta_t)^2\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2L^2}{mb}\sum_{i=1}^m \|x_t^i - x_{t-1}^i\|^2 + \frac{2\beta_t^2}{mb}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\overset{(ii)}{\leq}(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2L^2\eta_t^2}{bm}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 + \frac{2\beta_t^2}{bm}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2,
\end{aligned} \tag{75}
$$

where the above equality (i) is due to for all $i \in [m]$

$$
\begin{aligned}
&\mathbb{E}_t\big[\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_t^i) - \nabla f_k^i(x_{t-1}^i)\big)\big] = \nabla f^i(x_t^i) - \nabla f^i(x_{t-1}^i), \\
&\mathbb{E}_t\big[\frac{1}{b}\sum_{k\in\mathcal{I}_t^i}\big(\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\big)\big] = \nabla f^i(x_{t-1}^i) - \frac{1}{n}\sum_{j=1}^n z_{t-1}^j,
\end{aligned}
$$

and $\{\mathcal{I}_t^i\}_{i=1}^m$ are independent, and the above inequality (ii) holds by $0 < \beta_t < 1$ and $x_t^i = x_{t-1}^i + \eta_t(\tilde{x}_t^i - x_{t-1}^i)$.

□

Lemma A.9. *Under Assumption 3.1, the sequence $\{z_{k,t}^i\}$ is defined in Line 11 of Algorithm 2, then we have*

$$
\begin{aligned}
\mathbb{E}_t\big[\frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_t^i) - z_{k,t}^i\|^2\big] &\leq (1-\frac{b}{2n})\frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\quad + (\frac{2n}{b} - \frac{b}{n} - 1)L^2\eta_{t-1}^2\|\tilde{x}_t^i - x_{t-1}^i\|^2, \ \forall i \in [m].
\end{aligned} \tag{76}
$$

Proof. According to the Line 11 of Algorithm 2, we have

$$\begin{aligned}\mathbb{E}_t\big[\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_t^i)-z_{k,t}^i\|^2\big] &= (1-\frac{b}{n})\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_t^i)-z_{k,t-1}^i\|^2 \\ &= (1-\frac{b}{n})\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)+\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\|^2 \\ &\le (1+\frac{1}{\alpha})(1-\frac{b}{n})\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_t^i)-\nabla f_k^i(x_{t-1}^i)\|^2 \\ &\quad + (1+\alpha)(1-\frac{b}{n})\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\|^2 \\ &\le (1+\frac{1}{\alpha})(1-\frac{b}{n})L^2\|x_t^i-x_{t-1}^i\|^2+(1+\alpha)(1-\frac{b}{n})\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\|^2,\end{aligned} \tag{77}$$

where the last inequality is due to Assumption 3.1.

Let $\alpha=\frac{b}{2n}$, then we have

$$\begin{aligned}\mathbb{E}_t\big[\frac{1}{n}\sum_{i=1}^n\|\nabla f^i(x_t^i)-z_t^i\|^2\big] &\le (1-\frac{b}{2n})\frac{1}{n}\sum_{i=1}^n\|\nabla f^i(x_{t-1}^i)-z_{t-1}^i\|^2+(\frac{2n}{b}-\frac{b}{n}-1)L^2\|x_t^i-x_{t-1}^i\|^2 \\ &= (1-\frac{b}{2n})\frac{1}{n}\sum_{i=1}^n\|\nabla f^i(x_{t-1}^i)-z_{t-1}^i\|^2+(\frac{2n}{b}-\frac{b}{n}-1)L^2\eta_{t-1}^2\|\tilde{x}_t-x_{t-1}^i\|^2,\end{aligned} \tag{78}$$

where the above equality holds by $x_t^i-x_{t-1}^i=\eta_{t-1}(\tilde{x}_t-x_{t-1}^i)$.

□

Lemma A.10. *Given the sequence* $\{x_{t\ge1}^i,\tilde{x}_{t\ge1}^i,w_{t\ge1}^i\}_{i=1}^m$ *be generated from Algorithm 2. We have*

$$\begin{aligned}\sum_{i=1}^m\|x_{t+1}^i-\bar{x}_{t+1}\|^2 &\le (1-\frac{(1-\nu^2)\eta_t}{2})\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2+\frac{2\eta_t\gamma^2}{1-\nu^2}\sum_{i=1}^m\|g_t^i-\bar{g}_t\|^2, \\ \sum_{i=1}^m\|\tilde{x}_{t+1}^i-x_t^i\|^2 &\le (3+\nu^2)\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2+\frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\sum_{i=1}^m\|g_t^i\|^2, \\ \sum_{i=1}^m\mathbb{E}_{t+1}\|w_{t+1}^i-\bar{w}_{t+1}\|^2 &\le \nu\sum_{i=1}^m\|w_t^i-\bar{w}_t\|^2+\frac{3\nu^2}{1-\nu}\sum_{i=1}^m\big(L^2\eta_t^2\|\tilde{x}_{t+1}^i-x_t^i\|^2+\beta_{t+1}^2\|\nabla f^i(x_t^i)-u_t^i\|^2 \\ &\quad +\frac{\beta_{t+1}^2}{bn}\sum_{k=1}^n\|z_{k,t}^i-\nabla f^i(x_t^i)\|^2\big).\end{aligned}$$

Proof. For notational simplicity, let $x_t=[(x_t^1)^T,\cdots,(x_t^m)^T]^T\in\mathbb{R}^{md}$, $\tilde{x}_t=[(\tilde{x}_t^1)^T,\cdots,(\tilde{x}_t^m)^T]^T\in\mathbb{R}^{md}$ and $g_t=[(g_t^1)^T,\cdots,(g_t^m)^T]^T\in\mathbb{R}^{md}$ for all $t\ge1$. By using Assumption 3.4, since $W\mathbf{1}=\mathbf{1}$ and $\tilde{W}=W\otimes I_d$, we have $\tilde{W}(\mathbf{1}\otimes\bar{x}_t)=\mathbf{1}\otimes\bar{x}_t$. Meanwhile, we have $\mathbf{1}^T(x_t-\mathbf{1}\otimes\bar{x})=0$ and $\tilde{W}\mathbf{1}=\mathbf{1}$. Thus, we have

$$\|\tilde{W}x_t-\mathbf{1}\otimes\bar{x}_t\|^2=\|\tilde{W}(x_t-\mathbf{1}\otimes\bar{x}_t)\|^2\le\nu^2\|x_t-\mathbf{1}\otimes\bar{x}_t\|^2, \tag{79}$$

where the above inequality holds by $x_t-\mathbf{1}\otimes\bar{x}_t$ is orthogonal to $\mathbf{1}$ that is the eigenvector corresponding to the largest eigenvalue of $\tilde{W}$, and ν denotes the second largest eigenvalue of $\tilde{W}$.

Since $\tilde{x}_{t+1}^i=\sum_{j\in\mathcal{N}_i}W_{i,j}x_t^j-\gamma(A_t^i)^{-1}w_t^i=\sum_{j\in\mathcal{N}_i}W_{i,j}x_t^j-\gamma g_t^i$ for all $i\in[m]$, we have $\tilde{x}_{t+1}=\tilde{W}x_t-\gamma g_t$ and

$\bar{\tilde{x}}_{t+1} = \bar{x}_t - \gamma \bar{g}_t$. Since $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$ and $\bar{x}_{t+1} = \bar{x}_t + \eta_t(\bar{\tilde{x}}_{t+1} - \bar{x}_t)$, we have

$$\begin{aligned}
\sum_{i=1}^m \|x^i_{t+1} - \bar{x}_{t+1}\|^2 &= \big\|x_{t+1} - \mathbf{1} \otimes \bar{x}_{t+1}\big\|^2 \qquad (80) \\
&= \big\|x_t + \eta_t(\tilde{x}_{t+1} - x_t) - \mathbf{1} \otimes \big((\bar{x}_t + \eta_t(\bar{\tilde{x}}_{t+1} - \bar{x}_t)\big)\big\|^2 \\
&\leq (1+\alpha_1)(1-\eta_t)^2 \|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + (1 + \frac{1}{\alpha_1})\eta_t^2 \|\tilde{x}_{t+1} - \mathbf{1} \otimes \bar{\tilde{x}}_{t+1}\|^2 \\
&\overset{(i)}{=} (1-\eta_t)\|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + \eta_t \|\tilde{x}_{t+1} - \mathbf{1} \otimes \bar{\tilde{x}}_{t+1}\|^2 \\
&= (1-\eta_t)\|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + \eta_t \|\tilde{W} x_t - \gamma g_t - \mathbf{1} \otimes \big(\bar{x}_t - \gamma \bar{g}_t\big)\|^2 \\
&\leq (1-\eta_t)\|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + (1+\alpha_2)\eta_t \|\tilde{W} x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + (1 + \frac{1}{\alpha_2})\eta_t \gamma^2 \|g_t - \mathbf{1} \otimes \bar{g}_t\|^2 \\
&\overset{(ii)}{\leq} (1-\eta_t)\|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + \frac{(1+\nu^2)\eta_t}{2} \|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + \frac{\eta_t \gamma^2 (1+\nu^2)}{1-\nu^2} \|g_t - \mathbf{1} \otimes \bar{g}_t\|^2 \\
&\overset{(iii)}{\leq} (1 - \frac{(1-\nu^2)\eta_t}{2})\|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + \frac{2\eta_t\gamma^2}{1-\nu^2}\|g_t - \mathbf{1} \otimes \bar{g}_t\|^2 \\
&= (1 - \frac{(1-\nu^2)\eta_t}{2}) \sum_{i=1}^m \|x^i_t - \bar{x}_t\|^2 + \frac{2\eta_t\gamma^2}{1-\nu^2} \sum_{i=1}^m \|g^i_t - \bar{g}_t\|^2, \qquad (81)
\end{aligned}$$

where the above equality (i) is due to $\alpha_1 = \frac{\eta_t}{1-\eta_t}$, and the second inequality (ii) holds by $\alpha_2 = \frac{1-\nu^2}{2\nu^2}$ and $\|\tilde{W}x_t - \mathbf{1} \otimes \bar{x}_t\|^2 \leq \nu^2 \|x_t - \mathbf{1} \otimes \bar{x}_t\|^2$, and the above inequality (ii) is due to $0 < \nu < 1$. Meanwhile, we have

$$\begin{aligned}
\sum_{i=1}^m \|\tilde{x}^i_{t+1} - \bar{x}_t\|^2 &= \|\tilde{x}_{t+1} - \mathbf{1} \otimes \bar{x}_t\|^2 \\
&= \|\tilde{W} x_t - \gamma g_t - \mathbf{1} \otimes \bar{x}_t\|^2 \\
&\leq (1+\alpha_2)\nu^2 \|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 + (1 + \frac{1}{\alpha_2})\gamma^2 \|g_t\|^2 \\
&\overset{(i)}{=} \frac{1+\nu^2}{2} \sum_{i=1}^m \|x^i_t - \bar{x}_t\|^2 + \frac{1+\nu^2}{1-\nu^2}\gamma^2 \sum_{i=1}^m \|g^i_t\|^2, \qquad (82)
\end{aligned}$$

where the last equality (i) holds by $\alpha_2 = \frac{1-\nu^2}{2\nu^2}$. Then we have

$$\begin{aligned}
\sum_{i=1}^m \|\tilde{x}^i_{t+1} - x^i_t\|^2 &= \|\tilde{x}_{t+1} - x_t\|^2 \\
&= \|\tilde{x}_{t+1} - \mathbf{1} \otimes \bar{x}_t + \mathbf{1} \otimes \bar{x}_t - x_t\|^2 \\
&\leq 2\|\tilde{x}_{t+1} - \mathbf{1} \otimes \bar{x}_t\|^2 + 2\|x_t - \mathbf{1} \otimes \bar{x}_t\|^2 \\
&= (3+\nu^2) \sum_{i=1}^m \|x^i_t - \bar{x}_t\|^2 + \frac{2(1+\nu^2)}{1-\nu^2}\gamma^2 \sum_{i=1}^m \|g^i_t\|^2. \qquad (83)
\end{aligned}$$

Let $w_t = [(w^1_t)^T, (w^2_t)^T, \cdots, (w^m_t)^T]^T$, $u_t = [(u^1_t)^T, (u^2_t)^T, \cdots, (u^m_t)^T]^T$ and $\bar{w}_t = \frac{1}{m}\sum_{i=1}^m w^i_t$ and $\bar{u}_t = \frac{1}{m}\sum_{i=1}^m u^i_t$. Then we have for any $t \geq 1$,

$$w_{t+1} = \tilde{W}\big(w_t + u_{t+1} - u_t\big).$$

According to the above proof of Lemma A.3, we have $\bar{w}_{t+1} = \bar{w}_t + \bar{u}_{t+1} - \bar{u}_t$ for all $t \geq 1$. Thus we have

$$\begin{aligned}\sum_{i=1}^m \|w^i_{t+1} - \bar{w}_{t+1}\|^2 &= \|w_{t+1} - \mathbf{1}\otimes \bar{w}_{t+1}\|^2 \\ &= \big\|\tilde{W}\big(w_t + u_{t+1} - u_t\big) - \mathbf{1}\otimes\big(\bar{w}_t + \bar{u}_{t+1} - \bar{u}_t\big)\big\|^2 \\ &\leq (1+c)\|\tilde{W}w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\big\|\tilde{W}\big(u_{t+1}-u_t\big) - \mathbf{1}\otimes\big(\bar{u}_{t+1}-\bar{u}_t\big)\big\|^2 \\ &\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big\|u_{t+1} - u_t - \mathbf{1}\otimes\big(\bar{u}_{t+1}-\bar{u}_t\big)\big\|^2 \\ &\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big\|u_{t+1}-u_t\big\|^2, \end{aligned} \tag{84}$$

where the last inequality holds by Lemma A.2.

Since $u^i_{t+1} = \frac{1}{b}\sum_{k\in\mathcal{I}^i_{t+1}}\big(\nabla f^i_k(x_{t+1}) - \nabla f^i_k(x_t)\big) + (1-\beta_{t+1})u^i_t + \beta_{t+1}\Big(\frac{1}{b}\sum_{k\in\mathcal{I}^i_{t+1}}\big(\nabla f^i_k(x_t) - z^i_{k,t}\big) + \frac{1}{n}\sum_{j=1}^n z^i_{j,t}\Big)$ for any $i\in[m]$ and $t\geq 1$, we have

$$\begin{aligned}&\mathbb{E}_{t+1}\big\|u_{t+1}-u_t\big\|^2 \\ &= \sum_{i=1}^m \mathbb{E}_{t+1}\|u^i_{t+1} - u^i_t\|^2 \\ &= \sum_{i=1}^m \mathbb{E}_{t+1}\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}^i_{t+1}}\big(\nabla f^i_k(x^i_{t+1}) - \nabla f^i_k(x^i_t)\big) - \beta_{t+1}u^i_t + \beta_{t+1}\Big(\frac{1}{b}\sum_{k\in\mathcal{I}^i_{t+1}}\big(\nabla f^i_k(x^i_t) - z^i_{k,t}\big) + \frac{1}{n}\sum_{j=1}^n z^i_{j,t}\Big)\Big\|^2 \\ &= \sum_{i=1}^m \mathbb{E}_{t+1}\Big\|\frac{1}{b}\sum_{k\in\mathcal{I}^i_{t+1}}\big(\nabla f^i_k(x^i_{t+1}) - \nabla f^i_k(x^i_t)\big) - \beta_{t+1}u^i_t + \beta_{t+1}\nabla f^i(x^i_t) \\ &\quad + \beta_{t+1}\Big(\frac{1}{b}\sum_{k\in\mathcal{I}^i_{t+1}}\big(\nabla f^i_k(x^i_t) - z^i_{k,t}\big) + \frac{1}{n}\sum_{j=1}^n z^i_{j,t} - \nabla f^i(x^i_t)\Big)\Big\|^2 \\ &\leq 3L^2\sum_{i=1}^m\|x^i_{t+1} - x^i_t\|^2 + 3\beta^2_{t+1}\sum_{i=1}^m\|\nabla f^i(x^i_t) - u^i_t\|^2 + \frac{3\beta^2_{t+1}}{bn}\sum_{i=1}^m\sum_{k=1}^n\|z^i_{k,t} - \nabla f^i(x^i_t)\|^2 \\ &= 3L^2\eta_t^2\sum_{i=1}^m\|\tilde{x}^i_{t+1} - x^i_t\|^2 + 3\beta^2_{t+1}\sum_{i=1}^m\|\nabla f^i(x^i_t) - u^i_t\|^2 + \frac{3\beta^2_{t+1}}{bn}\sum_{i=1}^m\sum_{k=1}^n\|z^i_{k,t} - \nabla f^i(x^i_t)\|^2, \end{aligned} \tag{85}$$

where the second last inequality holds by Assumption 3.1 and the last equality is due to $x^i_{t+1} = x^i_t\eta_t(\tilde{x}^i_{t+1} - x^i_t)$.

Plugging the above inequalities (85) into (84), we have

$$\begin{aligned}\sum_{i=1}^m\|w^i_{t+1} - \bar{w}_{t+1}\|^2 &\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + (1+\frac{1}{c})\nu^2\big\|u_{t+1}-u_t\big\|^2 \\ &\leq (1+c)\nu^2\|w_t - \mathbf{1}\otimes\bar{w}_t\|^2 + 3(1+\frac{1}{c})\nu^2\sum_{i=1}^m\big(L^2\eta_t^2\|\tilde{x}^i_{t+1} - x^i_t\|^2 + \beta^2_{t+1}\|\nabla f^i(x^i_t) - u^i_t\|^2 \\ &\quad + \frac{\beta^2_{t+1}}{bn}\sum_{k=1}^n\|z^i_{k,t} - \nabla f^i(x^i_t)\|^2\big). \end{aligned} \tag{86}$$

Let $c = \frac{1}{\nu} - 1$, we have

$$\begin{aligned}\sum_{i=1}^m\mathbb{E}_{t+1}\|w^i_{t+1} - \bar{w}_{t+1}\|^2 \leq \nu\sum_{i=1}^m\|w^i_t - \bar{w}_t\|^2 + \frac{3\nu^2}{1-\nu}\sum_{i=1}^m\big(L^2\eta_t^2\|\tilde{x}^i_{t+1} - x^i_t\|^2 + \beta^2_{t+1}\|\nabla f^i(x^i_t) - u^i_t\|^2 \\ + \frac{\beta^2_{t+1}}{bn}\sum_{k=1}^n\|z^i_{k,t} - \nabla f^i(x^i_t)\|^2\big). \end{aligned} \tag{87}$$

□

Theorem A.11. *(Restatement of Theorem 5.5) Suppose the sequences* $\big\{\{x_t^i\}_{i=1}^m\big\}_{t=1}^T$ *be generated from Algorithm 2. Under the above Assumptions 3.1-3.5, and let* $\eta_t = \eta$, $0 < \beta_t \leq 1$ *for all* $t \geq 0$, $\gamma \leq \min\big(\frac{\rho(1-\nu^2)}{48\theta_t}, \frac{3\rho(1-\nu^2)\theta_t}{58L^2}\big)$, $\eta \leq \min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta_t}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)$ *with* $H_t = \frac{9}{b\beta_t} + \frac{6\nu^2}{b(1-\nu)^2} + \frac{4n^2\beta_t^2}{b^3}\big(\frac{9}{\beta_t} + \frac{9\nu^2}{(1-\nu)^2}\big) + \frac{3\nu^2}{(1-\nu)^2}$ *for all* $t \geq 0$*, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\|] \leq \frac{6\sqrt{G}}{\sqrt{T}}\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2}, \tag{88}$$

where $G = \frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta} + \big(\frac{18\beta_0}{\rho^2} + \frac{18\beta_0^2\nu^2}{\rho^2(1-\nu)^2} + \frac{3\nu^2}{\rho^2(1-\nu)^2} + \frac{9}{2\rho^2\beta_0} - \frac{9}{2\rho^2}\big)\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_0^i)\|^2$ *is independent on* T*,* b *and* n*.*

Proof. Without loss of generality, let $\eta = \eta_1 = \cdots = \eta_T$. According to Lemma A.4, we have

$$F(\bar{x}_{t+1}) \leq F(\bar{x}_t) + \frac{2\gamma\eta}{\rho}\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2 + \frac{\gamma\eta}{2\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 - \frac{\rho\gamma\eta}{4}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2. \tag{89}$$

According to the Lemma A.8, we have

$$\begin{aligned}\mathbb{E}_t\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 &\leq (1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2L^2\eta^2}{bm}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 \\ &\quad + \frac{2\beta_t^2}{bm}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2,\end{aligned} \tag{90}$$

and

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \mathbb{E}_t\|u_t^i - \nabla f^i(x_t^i)\|^2 &\leq (1-\beta_t)\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 + \frac{2L^2\eta^2}{bm}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 \\ &\quad + \frac{2\beta_t^2}{b}\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2.\end{aligned} \tag{91}$$

According to Lemma A.9, we have for any $i \in [m]$

$$\begin{aligned}\mathbb{E}_t\big[\frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_t) - z_{k,t}^i\|^2\big] &\leq (1 - \frac{b}{2n})\frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 + (\frac{2n}{b} - \frac{b}{n} - 1)L^2\eta^2\|\tilde{x}_t^i - x_{t-1}^i\|^2 \\ &\leq (1 - \frac{b}{2n})\frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 + \frac{2n}{b}L^2\eta^2\|\tilde{x}_t^i - x_{t-1}^i\|^2.\end{aligned} \tag{92}$$

According to Lemma A.10, we have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \mathbb{E}_t\|w_t^i - \bar{w}_t\|^2 &\leq \nu\frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \frac{3\nu^2}{1-\nu}\frac{1}{m}\sum_{i=1}^m \Big(L^2\eta^2\|\tilde{x}_t^i - x_{t-1}^i\|^2 + \beta_t^2\|\nabla f^i(x_{t-1}^i) - u_{t-1}^i\|^2 \\ &\quad + \frac{\beta_t^2}{bn}\sum_{k=1}^n \|z_{k,t-1}^i - \nabla f^i(x_{t-1}^i)\|^2\Big).\end{aligned} \tag{93}$$

Meanwhile, we also have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2 &\le (1-\frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i-\bar{x}_{t-1}\|^2+\frac{2\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i-\bar{g}_{t-1}\|^2 \\ &\le (1-\frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i-\bar{x}_{t-1}\|^2+\frac{2\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m(\|g_{t-1}^i\|^2+\|\bar{g}_{t-1}\|^2) \\ &\le (1-\frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i-\bar{x}_{t-1}\|^2+\frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2. \end{aligned} \tag{94}$$

and

$$\frac{1}{m}\sum_{i=1}^m\|\tilde{x}_t^i-x_{t-1}^i\|^2 \le (3+\nu^2)\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i-\bar{x}_{t-1}\|^2+\frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2. \tag{95}$$

Next considering the term $\|\bar{u}_t-\nabla F(\bar{x}_t)\|^2$, we have

$$\begin{aligned}\|\bar{u}_t-\nabla F(\bar{x}_t)\|^2 &= \|\bar{u}_t-\overline{\nabla f(x_t)}+\overline{\nabla f(x_t)}-\nabla F(\bar{x}_t)\|^2 \\ &\le 2\|\bar{u}_t-\overline{\nabla f(x_t)}\|^2+2\|\overline{\nabla f(x_t)}-\nabla F(\bar{x}_t)\|^2 \\ &\le 2\|\bar{u}_t-\overline{\nabla f(x_t)}\|^2+2\|\frac{1}{m}\sum_{i=1}^m\nabla f^i(x_t^i)-\frac{1}{m}\sum_{i=1}^m\nabla f^i(\bar{x}_t)\|^2 \\ &\le 2\|\bar{u}_t-\overline{\nabla f(x_t)}\|^2+\frac{2L^2}{m}\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2, \end{aligned}$$

where the last inequality is due to Assumption 3.1. Then we can obtain

$$-\|\bar{u}_t-\overline{\nabla f(x_t)}\|^2 \le -\frac{1}{2}\|\bar{u}_t-\nabla F(\bar{x}_t)\|^2+\frac{L^2}{m}\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2. \tag{96}$$

Since $\bar{u}_t=\bar{w}_t$ for all $t\ge 1$, we have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m\|w_t^i-\nabla F(x_t^i)\|^2 &= \frac{1}{m}\sum_{i=1}^m\|w_t^i-\bar{w}_t+\bar{u}_t-\nabla F(\bar{x}_t)+\nabla F(\bar{x}_t)-\nabla F(x_t^i)\|^2 \\ &\le 3\frac{1}{m}\sum_{i=1}^m\|w_t^i-\bar{w}_t\|^2+3\|\bar{u}_t-\nabla F(\bar{x}_t)\|^2+3\frac{1}{m}\sum_{i=1}^m\|\nabla F(\bar{x}_t)-\nabla F(x_t^i)\|^2 \\ &\le 3\frac{1}{m}\sum_{i=1}^m\|w_t^i-\bar{w}_t\|^2+3\|\bar{u}_t-\nabla F(\bar{x}_t)\|^2+3L^2\frac{1}{m}\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2. \end{aligned}$$

Then we have

$$-\|\bar{u}_t-\nabla F(\bar{x}_t)\|^2 \le -\frac{1}{3m}\sum_{i=1}^m\|w_t^i-\nabla F(x_t^i)\|^2+\frac{1}{m}\sum_{i=1}^m\|w_t^i-\bar{w}_t\|^2+\frac{L^2}{m}\sum_{i=1}^m\|x_t^i-\bar{x}_t\|^2. \tag{97}$$

We define a useful Lyapunov function (i.e., potential function), for any $t\ge 1$

$$\begin{aligned}\Phi_t = \mathbb{E}_t\big[&F(\bar{x}_t)+(\lambda_{t-1}-\frac{9\gamma\eta}{2\rho})\|\bar{u}_{t-1}-\overline{\nabla f(x_{t-1})}\|^2+\chi_{t-1}\frac{1}{m}\sum_{i=1}^m\|u_{t-1}^i-\nabla f^i(x_{t-1}^i)\|^2 \\ &+\alpha_{t-1}\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i)-z_{k,t-1}^i\|^2+(\theta_{t-1}-\frac{19\gamma\eta L^2}{4\rho})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i-\bar{x}_{t-1}\|^2 \\ &+(\vartheta_{t-1}-\frac{3\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m\|w_{t-1}^i-\bar{w}_{t-1}\|^2+\frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2\big], \end{aligned} \tag{98}$$

where $\alpha_{t-1} \geq 0$, $\chi_{t-1} \geq 0$, $\lambda_{t-1} \geq \frac{9\gamma\eta}{2\rho}$, $\theta_{t-1} \geq \frac{29\gamma\eta L^2}{6\rho}$ and $\vartheta_{t-1} \geq \frac{3\gamma\eta}{4\rho}$ for all $t \geq 1$. Then we have

$$
\begin{aligned}
\Phi_{t+1} &= \mathbb{E}_{t+1}\Big[F(\bar{x}_{t+1}) + (\lambda_t - \frac{9\gamma\eta}{2\rho})\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 + \chi_t \frac{1}{m}\sum_{i=1}^m \|u_t^i - \nabla f^i(x_t^i)\|^2 \\
&\quad + \alpha_t \frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_t^i) - z_{k,t}^i\|^2 + (\theta_t - \frac{29\gamma\eta L^2}{6\rho})\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 \\
&\quad + (\vartheta_t - \frac{3\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2 + \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2\Big] \\
&\overset{(i)}{\leq} \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{4\rho}\|\bar{u}_t - \nabla F(\bar{x}_t)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 + \frac{9\gamma\eta L^2}{2\rho}\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 + \lambda_t\|\bar{u}_t - \overline{\nabla f(x_t)}\|^2 \\
&\quad + \chi_t \frac{1}{m}\sum_{i=1}^m \|u_t^i - \nabla f^i(x_t^i)\|^2 + \alpha_t(1 - \frac{b}{2n})\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\quad + \frac{2\alpha_t n}{b}L^2\eta^2 \frac{1}{m}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 + (\theta_t - \frac{29\gamma\eta L^2}{6\rho})\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 \\
&\quad + (\vartheta_t - \frac{\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m \|w_t^i - \bar{w}_t\|^2\Big] \\
&\overset{(ii)}{\leq} \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m \|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m \|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m \|x_t^i - \bar{x}_t\|^2 \\
&\quad + \lambda_t(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \frac{2\lambda_t L^2\eta^2}{bm}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 + \frac{2\lambda_t\beta_t^2}{bm}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\quad + \chi_t(1-\beta_t)\frac{1}{m}\sum_{i=1}^m \|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 + \frac{2\chi_t L^2\eta^2}{bm}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 \\
&\quad + \frac{2\chi_t\beta_t^2}{b}\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 + \alpha_t(1 - \frac{b}{2n})\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\quad + \frac{2\alpha_t n}{b}L^2\eta^2 \frac{1}{m}\sum_{i=1}^m \|\tilde{x}_t^i - x_{t-1}^i\|^2 + \theta_t(1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m \|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \theta_t \frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m \|g_{t-1}^i\|^2 \\
&\quad + \vartheta_t\nu \frac{1}{m}\sum_{i=1}^m \|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \frac{3\vartheta_t\nu^2}{1-\nu}\frac{1}{m}\sum_{i=1}^m \Big(L^2\eta^2\|\tilde{x}_t^i - x_{t-1}^i\|^2 + \beta_t^2\|\nabla f^i(x_{t-1}^i) - u_{t-1}^i\|^2 \\
&\quad + \frac{\beta_t^2}{bn}\sum_{k=1}^n \|z_{k,t-1}^i - \nabla f^i(x_{t-1}^i)\|^2\Big)\Big], \qquad (99)
\end{aligned}
$$

where the inequality (i) is due to the above inequalities (89), (92) and (96); and the inequality (ii) holds by the above inequalities (90), (91), (93) and (97).

Then we have

$$\begin{aligned}
\Phi_{t+1} &\leq \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 \\
&\quad + \lambda_t(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \big(\chi_t - \chi_t\beta_t + \frac{3\vartheta_t\beta_t^2\nu^2}{1-\nu}\big)\frac{1}{m}\sum_{i=1}^m\|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\
&\quad + L^2\eta^2\big(\frac{2\lambda_t}{b} + \frac{2\chi_t}{b} + \frac{2\alpha_t n}{b} + \frac{3\vartheta_t\nu^2}{1-\nu}\big)\frac{1}{m}\sum_{i=1}^m\|\tilde{x}_t^i - x_{t-1}^i\|^2 \\
&\quad + \big(\alpha_t - \frac{\alpha_t b}{2n} + \frac{2\lambda_t\beta_t^2}{b} + \frac{2\chi_t\beta_t^2}{b} + \frac{3\beta_t^2\vartheta_t\nu^2}{b(1-\nu)}\big)\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\quad + \theta_t(1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \theta_t\frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2 + \vartheta_t\nu\frac{1}{m}\sum_{i=1}^m\|w_{t-1}^i - \bar{w}_{t-1}\|^2\Big] \\
&\overset{(i)}{\leq} \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 \\
&\quad + \lambda_t(1-\beta_t)\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \big(\chi_t - \chi_t\beta_t + \frac{3\vartheta_t\beta_t^2\nu^2}{1-\nu}\big)\frac{1}{m}\sum_{i=1}^m\|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\
&\quad + L^2\eta^2\big(\frac{2\lambda_t}{b} + \frac{2\chi_t}{b} + \frac{2\alpha_t n}{b} + \frac{3\vartheta_t\nu^2}{1-\nu}\big)\Big((3+\nu^2)\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \frac{2(1+\nu^2)}{1-\nu^2}\gamma^2\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2\Big) \\
&\quad + \big(\alpha_t - \frac{\alpha_t b}{2n} + \frac{2\lambda_t\beta_t^2}{b} + \frac{2\chi_t\beta_t^2}{b} + \frac{3\beta_t^2\vartheta_t\nu^2}{b(1-\nu)}\big)\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 \\
&\quad + \theta_t(1 - \frac{(1-\nu^2)\eta}{2})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i - \bar{x}_{t-1}\|^2 + \theta_t\frac{4\eta\gamma^2}{1-\nu^2}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2 + \vartheta_t\nu\frac{1}{m}\sum_{i=1}^m\|w_{t-1}^i - \bar{w}_{t-1}\|^2\Big],
\end{aligned} \tag{100}$$

where the inequality (i) is due to the above inequality (95).

Since $0 < \beta_t < 1$ for all $t \geq 1$, $\lambda_t = \frac{9\gamma\eta}{2\rho\beta_t} \geq \frac{9\gamma\eta}{2\rho}$ and $\lambda_t \leq \lambda_{t-1}$, then we have $\lambda_t(1-\beta_t) \leq \lambda_{t-1} - \frac{9\gamma\eta}{2\rho}$. Let $\chi_t = \frac{3\vartheta_t\nu^2}{1-\nu} \geq \frac{3\vartheta_t\beta_t\nu^2}{1-\nu}$ and $\chi_t \leq \chi_{t-1}$, we have $\chi_t - \chi_t\beta_t + \frac{3\vartheta_t\beta_t^2\nu^2}{1-\nu} \leq \chi_{t-1}$. Let $\alpha_t = \frac{2n\beta_t^2}{b^2}(2\lambda_t + 2\chi_t + \frac{3\vartheta_t\nu^2}{1-\nu}) = \frac{2n\beta_t^2}{b^2}(2\lambda_t + \frac{9\vartheta_t\nu^2}{1-\nu})$ and $\alpha_t \leq \alpha_{t-1}$, then we have $\alpha_t - \frac{\alpha_t b}{2n} + \frac{2\lambda_t\beta_t^2}{b} + \frac{2\chi_t\beta_t^2}{b} + \frac{3\beta_t^2\vartheta_t\nu^2}{b(1-\nu)} \leq \alpha_{t-1}$. Let $\vartheta_t = \frac{\gamma\eta}{\rho(1-\nu)}$ for all $t \geq 1$, since $0 < \nu < 1$, we have $\vartheta_t\nu = \vartheta_t - (1-\nu)\vartheta_t \leq \vartheta_{t-1} - \frac{3\gamma\eta}{4\rho}$. Meanwhile, let $\theta_t \leq \theta_{t-1}$ for all $t \geq 1$, $\gamma \leq \min\big(\frac{\rho(1-\nu^2)}{48\theta_t}, \frac{3\rho(1-\nu^2)\theta_t}{58L^2}\big)$, $\eta \leq \min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}\sqrt{H_t}}, \frac{\sqrt{\rho(1-\nu^2)\theta_t}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)$ with $H_t = \frac{9}{b\beta_t} + \frac{6\nu^2}{b(1-\nu)^2} + \frac{4n^2\beta_t^2}{b^3}\big(\frac{9}{\beta_t} + \frac{9\nu^2}{(1-\nu)^2}\big) + \frac{3\nu^2}{(1-\nu)^2}$ for all $t \geq 1$, we can obtain

$$\theta_t(1 - \frac{(1-\nu^2)\eta}{2}) + L^2\eta^2(3+\nu^2)\big(\frac{2\lambda_t}{b} + \frac{2\chi_t}{b} + \frac{2\alpha_t n}{b} + \frac{3\vartheta_t\nu^2}{1-\nu}\big) \leq \theta_{t-1} - \frac{29\gamma\eta L^2}{6\rho}, \tag{101}$$

$$\theta_t\frac{4\eta\gamma^2}{1-\nu^2} + \frac{2(1+\nu^2)}{1-\nu^2}L^2\eta^2\gamma^2\big(\frac{2\lambda_t}{b} + \frac{2\chi_t}{b} + \frac{2\alpha_t n}{b} + \frac{3\vartheta_t\nu^2}{1-\nu}\big) \leq \frac{\rho\gamma\eta}{6}. \tag{102}$$

Based on the choice of these parameters and the above inequality (100), we have

$$\begin{aligned}\Phi_{t+1} &\leq \mathbb{E}_{t+1}\Big[F(\bar{x}_t) - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 \\ &\quad + (\lambda_{t-1} - \frac{9\gamma\eta}{2\rho})\|\bar{u}_{t-1} - \overline{\nabla f(x_{t-1})}\|^2 + \chi_{t-1}\frac{1}{m}\sum_{i=1}^m\|u_{t-1}^i - \nabla f^i(x_{t-1}^i)\|^2 \\ &\quad + \alpha_{t-1}\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_{t-1}^i) - z_{k,t-1}^i\|^2 + (\theta_{t-1} - \frac{29\gamma\eta L^2}{6\rho})\frac{1}{m}\sum_{i=1}^m\|x_{t-1}^i - \bar{x}_{t-1}\|^2 \\ &\quad + (\vartheta_{t-1} - \frac{3\gamma\eta}{4\rho})\frac{1}{m}\sum_{i=1}^m\|w_{t-1}^i - \bar{w}_{t-1}\|^2 + \frac{\rho\gamma\eta}{6}\frac{1}{m}\sum_{i=1}^m\|g_{t-1}^i\|^2\Big] \\ &= \Phi_t - \frac{\gamma\eta}{12\rho}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 - \frac{\rho\gamma\eta}{12}\frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 - \frac{\gamma\eta L^2}{12\rho}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2. \end{aligned} \tag{103}$$

Then we can obtain

$$\frac{1}{\rho^2}\frac{1}{m}\sum_{i=1}^m\|w_t^i - \nabla F(x_t^i)\|^2 + \frac{L^2}{\rho^2}\frac{1}{m}\sum_{i=1}^m\|x_t^i - \bar{x}_t\|^2 + \frac{1}{m}\sum_{i=1}^m\|g_t^i\|^2 \leq \frac{12(\Phi_t - \Phi_{t+1})}{\gamma\rho\eta}. \tag{104}$$

Since $x_0^1 = \tilde{x}_0^1 = \cdots = x_0^m = \tilde{x}_0^m$, $z_{1,0}^i = z_{2,0}^i = \cdots = z_{n,0}^i = 0$ and $u_0^i = w_0^i = 0$ for any $i \in [m]$, we have

$$\begin{aligned}\Phi_1 &= \mathbb{E}\big[F(\bar{x}_1) + \alpha_0\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_0^i)\|^2 + (\lambda_0 - \frac{9\gamma\eta}{2\rho})\|\overline{\nabla f(x_0)}\|^2 + \chi_0\frac{1}{m}\sum_{i=1}^m\|\nabla f^i(x_0^i)\|^2\big] \\ &\leq F(\bar{x}_1) + \big(\alpha_0 + \chi_0 + \lambda_0 - \frac{9\gamma\eta}{2\rho}\big)\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_0^i)\|^2 \\ &= F(\bar{x}_1) + \big(\frac{18\gamma\eta\beta_0}{\rho} + \frac{18\beta_0^2\gamma\eta\nu^2}{\rho(1-\nu)^2} + \frac{3\gamma\eta\nu^2}{\rho(1-\nu)^2} + \frac{9\gamma\eta}{2\rho\beta_0} - \frac{9\gamma\eta}{2\rho}\big)\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_0^i)\|^2. \end{aligned} \tag{105}$$

Let $\mathcal{M}_t^i = \|g_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\|$. Then we have

$$\begin{aligned}\mathcal{M}_t^i &= \|g_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\ &\overset{(i)}{=} \|(A_t^i)^{-1}w_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\ &= \frac{1}{\|A_t^i\|}\|A_t^i\|\|(A_t^i)^{-1}w_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\ &\geq \frac{1}{\|A_t^i\|}\|w_t^i\| + \frac{1}{\rho}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\rho}\|\bar{x}_t - x_t^i\| \\ &\overset{(ii)}{\geq} \frac{1}{\|A_t^i\|}\|w_t^i\| + \frac{1}{\|A_t^i\|}\|\nabla F(x_t^i) - w_t^i\| + \frac{L}{\|A_t^i\|}\|\bar{x}_t - x_t^i\| \\ &\geq \frac{1}{\|A_t^i\|}\big(\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\|\big), \end{aligned} \tag{106}$$

where the equality (i) holds by $g_t^i = (A_t^i)^{-1}w_t^i$, and the inequality (ii) holds by $\|A_t^i\| \geq \rho$ for all $t \geq 1$ due to Assumption 3.4. Then we have

$$\|\nabla F(x_t^i)\| + L\|\bar{x}_t - x_t^i\| \leq \mathcal{M}_t^i\|A_t^i\|. \tag{107}$$

According to the above inequality (104), we have

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\mathcal{M}_t^i]^2 \leq \frac{1}{T}\sum_{t=1}^{T}\big[\frac{3}{\rho^2}\frac{1}{m}\sum_{i=1}^{m}\|w_t^i-\nabla F(x_t^i)\|^2+\frac{3L^2}{\rho^2}\frac{1}{m}\sum_{i=1}^{m}\|x_t^i-\bar{x}_t\|^2+\frac{3}{m}\sum_{i=1}^{m}\|g_t^i\|^2\big]$$
$$\leq \frac{1}{T}\sum_{t=1}^{T}\frac{36(\Phi_t-\Phi_{t+1})}{\gamma\rho\eta} \leq \frac{36(\Phi_1-F^*)}{T\gamma\rho\eta}. \tag{108}$$

By using Cauchy-Schwarz inequality, we have

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|] \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\big[\mathcal{M}_t^i\|A_t^i\|\big]$$
$$\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\mathcal{M}_t^i]^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{109}$$

By plugging the above inequalities (109) into (108), we can obtain

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|] \leq \frac{6\sqrt{\Phi_1-F^*}}{\sqrt{T\gamma\rho\eta}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{110}$$

Since $F(x)$ is L-smooth, we have

$$\|\nabla F(\bar{x}_t)\| = \|\nabla F(\bar{x}_t)-\nabla F(x_t^i)+\nabla F(x_t^i)\| \leq \|\nabla F(x_t^i)\|+L\|x_t^i-\bar{x}_t\|. \tag{111}$$

Meanwhile, let $G=\frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta}+\big(\frac{18\beta_0}{\rho^2}+\frac{18\beta_0^2\nu^2}{\rho^2(1-\nu)^2}+\frac{3\nu^2}{\rho^2(1-\nu)^2}+\frac{9}{2\rho^2\beta_0}-\frac{9}{2\rho^2}\big)\frac{1}{m}\sum_{i=1}^m\frac{1}{n}\sum_{k=1}^n\|\nabla f_k^i(x_0^i)\|^2$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla F(x_t^i)\|+L\|\bar{x}_t-x_t^i\|]$$
$$\leq \frac{6\sqrt{\Phi_1-F^*}}{\sqrt{T\gamma\rho\eta}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}$$
$$= \frac{6\sqrt{G}}{\sqrt{T}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|A_t^i\|^2}. \tag{112}$$

Let $\theta_t=\theta\geq\frac{29\gamma\eta L^2}{6\rho}$ for all $t\geq 1$. Let $b=\sqrt{n}$ and $\beta_t=\frac{b}{n}$ for all $t\geq 1$, we have

$$H_t=\frac{9}{b\beta_t}+\frac{6\nu^2}{b(1-\nu)^2}+\frac{4n^2\beta_t^2}{b^3}\big(\frac{9}{\beta_t}+\frac{9\nu^2}{(1-\nu)^2}\big)+\frac{3\nu^2}{(1-\nu)^2}\leq 45+\frac{45\nu^2}{(1-\nu)^2}. \tag{113}$$

Then we have $\frac{1}{\sqrt{H_t}}\geq\frac{1}{\sqrt{45+\frac{45\nu^2}{(1-\nu)^2}}}$, and

$$\min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}\sqrt{H_t}},\frac{\sqrt{\rho(1-\nu^2)\theta}}{2L\sqrt{\gamma(3+\nu^2)}\sqrt{H_t}}\big)\geq\min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}},\frac{\sqrt{\rho(1-\nu^2)\theta}}{2L\sqrt{\gamma(3+\nu^2)}}\big)\sqrt{45+\frac{45\nu^2}{(1-\nu)^2}}.$$

Thus we can let $\eta=\min\big(\frac{\rho\sqrt{1-\nu^2}}{2L\gamma\sqrt{6(1+\nu^2)}},\frac{\sqrt{\rho(1-\nu^2)\theta}}{2L\sqrt{\gamma(3+\nu^2)}}\big)\sqrt{45+\frac{45\nu^2}{(1-\nu)^2}}$ and $\gamma=\min\big(\frac{\rho(1-\nu^2)}{48\theta},\frac{3\rho(1-\nu^2)\theta}{58L^2}\big)$. Note that we set $\beta_t=\frac{b}{n}$ for all $t\geq 1$, while we can set $\beta_0\in(0,1)$, which is independent on T, b and n. Thus we have

$G = \frac{F(\bar{x}_1)-F^*}{\rho\gamma\eta} + \big(\frac{18\beta_0}{\rho^2} + \frac{18\beta_0^2\nu^2}{\rho^2(1-\nu)^2} + \frac{3\nu^2}{\rho^2(1-\nu)^2} + \frac{9}{2\rho^2\beta_0} - \frac{9}{2\rho^2}\big)\frac{1}{m}\sum_{i=1}^m \frac{1}{n}\sum_{k=1}^n \|\nabla f_k^i(x_0^i)\|^2$ is independent on T, b and n.

Let $\rho = O(1)$, $\eta = O(1)$ and $\gamma = O(1)$, then we have $G = O(1)$ is independent on T, b and n. Given $\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} = O(1)$, set

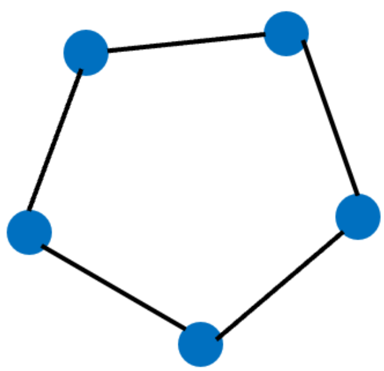
$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{6\sqrt{G}}{\sqrt{T}}\sqrt{\frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m \mathbb{E}\|A_t^i\|^2} \leq \epsilon, \tag{114}$$

we have $T = O(\epsilon^{-2})$. Since our Algorithm 2 requires b samples, it obtains a sample complexity of $Tb = O(\sqrt{n}\epsilon^{-2})$.

□

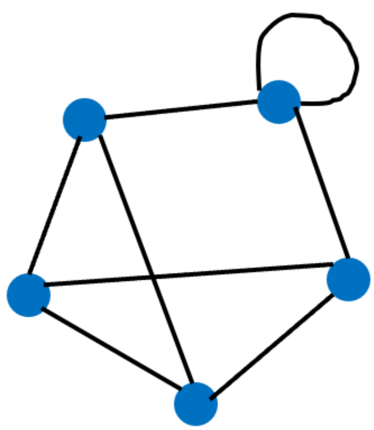
B. Additional Experiment Details

In the experiments, we consider two classical undirected networks that connect all clients, i.e., the *ring* and *3-regular* expander networks (Hoory et al., 2006), illustrated in Figures 6 and 7, respectively.



$$W = \begin{bmatrix} 0.4 & 0.3 & 0 & 0 & 0.3 \\ 0.3 & 0.4 & 0.3 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0 & 0.3 & 0.4 & 0.3 \\ 0.3 & 0 & 0 & 0.3 & 0.4 \end{bmatrix}$$

Figure 6: An illustration of the ring network with 5 nodes and its mixing matrix.



$$W = \begin{bmatrix} 0.5 & 0.25 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

Figure 7: An illustration of the 3-regular expander network with 5 nodes and its mixing matrix.