Decentralized Multi-Level Compositional Optimization Algorithms with Level-Independent Convergence Rate

Hongchang Gao
Temple University

Abstract

Stochastic multi-level compositional optimization problems cover many new machine learning paradigms, e.g., multi-step model-agnostic meta-learning, which require efficient optimization algorithms for large-scale data. This paper studies the decentralized stochastic multi-level optimization algorithm, which is challenging because the multi-level structure and decentralized communication scheme may make the number of levels significantly affect the order of the convergence rate. To this end, we develop two novel decentralized optimization algorithms to optimize the multi-level compositional optimization problem. Our theoretical results show that both algorithms can achieve the level-independent convergence rate for non-convex problems under much milder conditions compared with existing single-machine algorithms. To the best of our knowledge, this is the first work that achieves the level-independent convergence rate under the decentralized setting. Moreover, extensive experiments confirm the efficacy of our proposed algorithms.

1 Introduction

In recent years, some new learning paradigms, such as model-agnostic meta-learning (Finn et al., 2017), have been proposed to handle realistic machine learning applications, which are typically beyond the class of traditional stochastic optimization. Some examples include bilevel optimization, minimax optimization, compositional optimization, and so on. Of particular interest in this paper is the learning paradigm that can be formulated as the *stochastic multi-level compositional optimization problem*. More particularly, we are interested in the decentralized setting where data are distributed on different devices and the device performs peer-to-peer communication to exchange information with its neighboring devices. Mathematically, the loss function is defined as follows:

$$\min_{x\in\mathbb{R}^d} F(x) = \frac{1}{N}\sum_{n=1}^{N} F_n(x) \ , \tag{1}$$

where $F_n(x) = f_n^{(K)} \circ f_n^{(K-1)} \circ \cdots \circ f_n^{(2)} \circ f_n^{(1)}(x)$, $x \in \mathbb{R}^d$ is the model parameter of a machine learning model, N devices compose a communication network, and $F_n(x)$ is the loss function on the n-th device, for any $k \in \{1, 2, \cdots, K-1, K\}$, $f_n^{(k)}(\cdot) = \mathbb{E}_{\xi_n^{(k)}}[f_n^{(k)}(\cdot; \xi_n^{(k)})] : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ is the k-th level function on the n-th device, where $\xi_n^{(k)}$ denotes the data distribution for the k-th level function on the n-th device. It can observed that the input of $f_n^{(k)}(\cdot)$ is the output of $f_n^{(k-1)}(\cdot)$.

The stochastic multi-level compositional optimization (multi-level SCO) problem covers a wide range of machine learning models. For instance, the multi-step model-agnostic meta-learning (Finn et al., 2017) can be formulated as a multi-level SCO problem. The stochastic training of graph neural networks also belongs to the class of multi-level SCO problem (Yu et al., 2022; Cong et al., 2021). The neural network with batch-normalization is actually a multi-level SCO problem (Lian and Liu, 2018). The challenge of optimizing the multi-level SCO problem lies in that the stochastic gradient is not an unbiased estimator of the full gradient when the inner-level functions are nonlinear. To address this challenge, a couple of stochastic multi-level compositional gradient descent (multi-level SCGD) algorithms have been proposed recently. For instance, Yang et al. (2019) proposed the first stochastic multi-level compositional gradient descent algorithm. However, due to the nested structure of the loss function, the order of its convergence rate depends on the number of levels K exponentially [1], where a larger

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

[1] Following (Balasubramanian et al., 2022), throughout

K results in a slower convergence rate. As such, it cannot match the traditional stochastic gradient descent (SGD) algorithm's convergence rate. Later, some algorithms (Zhang and Xiao, 2021; Balasubramanian et al., 2022; Jiang et al., 2022) were proposed to achieve the level-independent convergence rate via leveraging the variance-reduced estimator. For instance, Zhang and Xiao (2021) exploited the SPIDER (Nguyen et al., 2017; Fang et al., 2018) estimator for both the stochastic function value $f_n^{(k)}(\cdot;\xi_n^{(k)})$ and the stochastic Jacobian matrix $\nabla f_n^{(k)}(\cdot;\xi_n^{(k)})$ of each level function to improve the convergence rate.

However, existing stochastic multi-level compositional optimization algorithms have some limitations. On the one hand, they only focus on the single-machine setting. As such, they cannot be used to solve the distributed multi-level SCO problem in Eq. (1). In particular, it is unclear if the *level-independent convergence rate* is still achievable under the decentralized setting. More particularly, it is unclear whether the consensus error caused by the decentralized communication scheme will make the level-independent convergence rate unachievable. On the other hand, under the single-machine setting, those algorithms with the variance-reduced estimator have some unrealistic operations, limiting their applications in real-world tasks. In particular, they apply the variance reduction technique to the stochastic Jacobian matrix of every level function $\nabla f_n^{(k)}(\cdot,\xi_n^{(k)})$ where $k \in \{1,2,\cdots,K\}$, which requires the clipping operation, e.g., Algorithm 3 in (Zhang and Xiao, 2021), or the projection operation, e.g., Eq. (3) in (Jiang et al., 2022), to upper bound the variance-reduced gradient. These operations either result in a very small learning rate or depend on unknown hyperparameters. These limitations motivate us to *1) develop decentralized optimization algorithms for Eq. (1) to enable multi-level SCO problems for distributed data, 2) propose the practical algorithm based on the variance-reduced stochastic gradient under mild conditions, and 3) establish the level-independent convergence rate for the proposed algorithm.*

To this end, we developed two novel decentralized multi-level stochastic compositional gradient descent algorithms, both of which can achieve the level-independent convergence rate. Specifically, they have the following contributions. 1) Our first algorithm demonstrates how to achieve the level-independent convergence rate with a novel combination of the inner-level function estimator and the momentum technique. Our second algorithm improves the convergence rate with a novel strategy of utilizing the variance-reduced estimator without impractical operations. In particular, unlike existing algorithms (Zhang and Xiao, 2021; Jiang et al., 2022), which apply the variance reduction technique to both *all stochastic inner-level function values* $\{f_n^{(k)}(\cdot;\xi_n^{(k)})\}_{k=1}^{K-1}$ and *all stochastic Jacobian matrices* $\{\nabla f_n^{(k)}(\cdot;\xi_n^{(k)})\}_{k=1}^{K}$, our algorithm leverages the variance-reduced estimator for stochastic inner-level function values $\{f_n^{(k)}(\cdot;\xi_n^{(k)})\}_{k=1}^{K-1}$ and *the gradient* $\nabla F_n(x;\xi_n)$. As such, our algorithm does not require the clipping operation for the learning rate or the projection operation for $\{\nabla f_n^{(k)}(\cdot;\xi_n^{(k)})\}_{k=1}^{K}$ as (Zhang and Xiao, 2021; Jiang et al., 2022). Thus, it is more friendly to implement. 2) Besides the novel algorithmic design, we established the level-independent convergence rate of our two algorithms for nonconvex problems under the decentralized setting. In particular, our first algorithm, which leverages the momentum technique for the gradient, enjoys the convergence rate of $O(\epsilon^{-4})$ to achieve the ϵ-stationary point. Our second algorithm, which exploits the variance reduction technique for the gradient, can achieve the convergence rate of $O(\epsilon^{-3})$ for nonconvex problems. As far as we know, this is the first decentralized optimization work for multi-level SCO with theoretical guarantees. 3) Extensive experiments on the multi-step model-agnostic meta-learning task confirm the effectiveness of our algorithms.

2 Related Work

2.1 Stochastic Two-Level Compositional Optimization

The stochastic two-level compositional optimization problem has been extensively studied in the past few years. In particular, to address the biased gradient estimator problem, Wang et al. (2017) developed the stochastic compositional gradient descent (SCGD) algorithm for the first time, where the moving-average technique was leveraged to the estimation of the inner-level function value to control the estimation error. However, its sample complexity is as large as $O(\epsilon^{-8})$ for nonconvex problems, which is worse than $O(\epsilon^{-4})$ of the standard SGD algorithm for non-compositional optimization problems. Then, Ghadimi et al. (2020) applied the momentum technique to stochastic compositional gradient so that it improved the sample complexity to $O(\epsilon^{-4})$. On the contrary, Chen et al. (2020) leveraged the variance-reduced estimator (Cutkosky and Orabona, 2019) for the inner-level function, which can also achieve the sample complexity of $O(\epsilon^{-4})$. To further improve the convergence rate, a couple

this paper, the level-dependent convergence rate means that the number of levels K affects the order of the convergence rate, e.g., $O(\epsilon^{-K})$, while the level-independent convergence rate indicates that K does not affect its order but may affect its coefficient, e.g., $O(K\epsilon^{-2})$.

of works exploited the variance-reduced technique to control the estimation error for both the inner-level function value and its Jacobian matrix. For instance, Yuan et al. (2019) leveraged the SPIDER variance reduction technique (Nguyen et al., 2017; Fang et al., 2018) to improve the sample complexity to $O(\epsilon^{-3})$ for stochastic nonconvex problems. However, this algorithm requires a large batch size. To address this problem, Yuan and Hu (2020) employed the STORM variance reduction technique (Cutkosky and Orabona, 2019), which can also achieve the sample complexity of $O(\epsilon^{-3})$, but with a small batch size. As for the nonconvex finite-sum compositional problem, a couple of works (Zhang and Xiao, 2019a,b; Yuan et al., 2019) also utilized the variance-reduction technique to improve the sample complexity to match the counterpart for non-compositional problems.

2.2 Stochastic Multi-Level Compositional Optimization

Even though the aforementioned algorithms can achieve desired sample complexity for the two-level compositional problem, it is non-trivial to extend them to the multi-level problem for achieving the same sample complexity. For instance, Yang et al. (2019) developed an accelerated stochastic compositional gradient descent algorithm for the stochastic multi-level compositional optimization problem, which can only achieve the sample complexity of $O(\epsilon^{-(7+K)/2})$ for nonconvex problems. Obviously, this sample complexity depends on the number of function levels K, which is far from satisfactory. Later, Balasubramanian et al. (2022) extended the momentum approach (Ghadimi et al., 2020) to the multi-level problem, obtaining the $O(\epsilon^{-6})$ sample complexity, which is worse than the counterpart (Ghadimi et al., 2020) for the two-level problem. Then, they added a correction term when using the moving-average technique to estimate each level function so that the sample complexity was improved to $O(\epsilon^{-4})$, which can match the standard momentum stochastic gradient descent algorithm. In (Chen et al., 2020), the STORM variance-reduction technique is leveraged to estimate each level function, which can also result in the sample complexity of $O(\epsilon^{-4})$. In (Zhang and Xiao, 2021), the SPIDER variance-reduction technique is exploited to estimate both each level function and its gradient so that it can achieve the sample complexity of $O(\epsilon^{-3})$. However, this algorithm requires a large batch size. Moreover, it requires a small learning rate to guarantee the Lipschitz continuousness of the variance-reduced gradient. Recently, Jiang et al. (2022) leveraged the STORM variance-reduction technique to estimate each level function value and its Jacobian matrix, resulting in the sample complexity of $O(\epsilon^{-3})$ with the mini-batch size of $O(1)$. However, this algorithm requires the projection operation for Jacobian matrices such that they are upper bounded. Thus, these algorithms with the sample complexity $O(\epsilon^{-3})$ are not practical for real-world applications. Moreover, it is unclear how to obtain the level-independent sample complexity under the decentralized setting.

2.3 Decentralized Compositional Optimization

Decentralized optimization has been extensively studied for the non-compositional optimization problem from both the computation (Lian et al., 2017; Sun et al., 2020; Xin et al., 2020) and communication (Koloskova et al., 2019a,b; Gao and Huang, 2020; Song et al., 2022; Hua et al., 2022; Ying et al., 2021) perspectives in recent years. Those algorithms are based on the stochastic gradient, which is an unbiased estimator of the full gradient. Thus, they cannot be directly extended to the stochastic compositional optimization problem because its stochastic gradient is a biased estimator of the full gradient. Recently, to address this problem, Gao and Huang (2021) developed the decentralized stochastic compositional gradient descent algorithm for the two-level stochastic compositional problem for the first time, which can achieve the sample complexity of $O(\epsilon^{-6})$. Zhao and Liu (2022) leveraged the STORM-like technique to estimate the inner-level function and improved the sample complexity to $O(\epsilon^{-4})$. Moreover, Gao (2023) developed the decentralized stochastic compositional gradient descent ascent algorithm for stochastic compositional minimax problems. On the other hand, a series of decentralized bilevel optimization algorithms have been proposed recently, e.g., (Gao et al., 2023; Zhang et al., 2023; Lu et al., 2022) and the related works therein. However, all those existing compositional and bilevel optimization algorithms only focus on the two-level problem. It is unclear how to apply them to the multi-level compositional optimization problem to achieve the level-independent sample complexity.

3 Decentralized Stochastic Multi-Level Compositional Optimization

In this section, we present the details of our proposed algorithms under the decentralized setting. Here, it is assumed the devices compose a communication graph and perform peer-to-peer communication. The adjacency matrix W of this graph satisfies the following assumption.

Assumption 1. *$W = [w_{ij}] \in \mathbb{R}^{N\times N}$ is a symmetric and doubly stochastic matrix. Its eigenvalues satisfy*

$|\lambda_N| \leq |\lambda_{N-1}| \leq \cdots \leq |\lambda_2| < |\lambda_1| = 1.$

Under this assumption, we can denote the spectral gap as $1-\lambda$ where $\lambda = |\lambda_2|$. Then, we propose two decentralized optimization algorithms for solving Eq. (1) in the following two subsections.

Algorithm 1 DSMCGDM

Input: $x_{n,0} = x_0$, $\alpha > 0$, $\beta > 0$, $\mu > 0$, $\eta > 0$.

1: **for** $t = 0, \cdots, T-1$ **do**
2: $u_{n,t}^{(0)} = x_{n,t}$,
3: **for** $k = 1, \cdots, K-1$ **do**
4: **if** $t == 0$ **then**
5: $u_{n,t}^{(k)} = f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$,
6: **else**
7: $u_{n,t}^{(k)} = (1-\beta\eta)(u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)}; \xi_{n,t}^{(k)})) + f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$,
8: **end if**
9: $v_{n,t}^{(k)} = \nabla f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$,
10: **end for**
11: $v_{n,t}^{(K)} = \nabla f_n^{(K)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)})$,
12: $g_{n,t} = v_{n,t}^{(1)} v_{n,t}^{(2)} \cdots v_{n,t}^{(K-1)} v_{n,t}^{(K)}$,
13: **if** $t == 0$ **then**
14: $m_{n,t} = g_{n,t}$, $y_{n,t} = m_{n,t}$
15: **else**
16: $m_{n,t} = (1-\mu\eta) m_{n,t-1} + \mu\eta g_{n,t}$,
17: $y_{n,t} = \sum_{n' \in \mathcal{N}_n} w_{nn'} y_{n',t-1} + m_{n,t} - m_{n,t-1}$,
18: **end if**
19: $x_{n,t+\frac{1}{2}} = \sum_{n' \in \mathcal{N}_n} w_{nn'} x_{n',t} - \alpha y_{n,t}$,
$x_{n,t+1} = x_{n,t} + \eta(x_{n,t+\frac{1}{2}} - x_{n,t})$,
20: **end for**

3.1 Decentralized Stochastic Multi-level Compositional Gradient Descent with Momentum

Challenges. The momentum technique is commonly used in optimization. However, facilitating it to multi-level SCGD is non-trivial. Under the single-machine setting, Balasubramanian et al. (2022) developed the first multi-level SCGD with momentum algorithm, which applies the moving-average technique to each inner-level function and the gradient. However, this straightforward extension can only achieve the $O(\epsilon^{-6})$ sample complexity, which is worse than $O(\epsilon^{-4})$ of the two-level algorithm. Then, Balasubramanian et al. (2022) introduced a correction term to the inner-level function estimator to address this problem. However, this correction term requires to compute the Jacobian matrix (See its Algorithm 2), which is too complicated and unclear if it works under the decentralized setting. Especially, *it is unclear whether the consensus error caused by the decentralized communication topology will worsen the convergence rate in the presence of multi-level inner functions.* Therefore, a natural question follows: **How to design an efficient decentralized multi-level SCGD with momentum algorithm to achieve the level-independent sample complexity $O(\epsilon^{-4})$?**

To answer this question, in Algorithm 1, we develop the Decentralized Stochastic Multi-level Compositional Gradient Descent with Momentum (DSMCGDM) algorithm. Specifically, to achieve the level-independent sample complexity, which can match the decentralized SGD with momentum algorithm for non-compositional problem, we leverage the STORM-like technique to estimate the k-th level function (where $k \in \{1, 2, \cdots, K-1\}$), which is shown below:

$$u_{n,t}^{(k)} = (1-\beta\eta)(u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)}; \xi_{n,t}^{(k)})) + f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)}) \ , \tag{2}$$

where $\beta > 0$, $\eta > 0$ are two hyperparameters satisfying $\beta\eta < 1$, $u_{n,t}^{(k)}$ is the estimation of the k-th level function $f_n^{(k)}(u_{n,t}^{(k-1)})$ on the n-th device.

It is worth noting that we do not apply this variance-reduction technique to the stochastic Jacobian matrix $v_{n,t}^{(k)} \triangleq \nabla f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$. After we obtain the stochastic Jacobian matrix $v_{n,t}^{(k)}$ of each level function, we combine them to get the stochastic compositional gradient $g_{n,t}$ of the objective function $F_n(x)$, which is shown in Line 12. Then, we compute the momentum of this stochastic compositional gradient in Line 16, where $\mu > 0$ is a hyperparameter satisfying $\mu\eta < 1$. After that, we leverage the gradient-tracking technique in Line 17 to communicate the momentum between different devices according to the communication topology, which is defined below:

$$y_{n,t} = \sum_{n' \in \mathcal{N}_n} w_{nn'} y_{n,t-1} + m_{n,t} - m_{n,t-1} \ , \tag{3}$$

where $\mathcal{N}_n = \{n' | w_{nn'} > 0\}$ denotes the neighbors of the n-th device and $w_{nn'}$ is the edge weight of the communication graph. Finally, we can leverage $y_{n,t}$ to update the model parameter on the corresponding device, which is shown in Line 19, where $\alpha > 0$ is a hyperparameter.

Note that Eq. (2) has been used for *non-momentum* algorithm under the single-machine setting in (Chen et al., 2020), rather then the decentralized setting. Therefore, it is still unclear how it affects the convergence for the *momentum* algorithm or the *decentralized* setting. In fact, this is the first time to apply Eq. (2) to the momentum algorithm. We believe this novel algorithmic design can also be applied to the single-machine setting to accelerate existing algorithms, e.g., (Chen et al., 2020). Moreover, to the

best of our knowledge, this is the first decentralized optimization algorithm for the stochastic multi-level compositional optimization problem. Meanwhile, this algorithmic design brings new challenges for convergence analysis due to the interaction between the estimator of each level function and momentum. We will address these challenges and show this algorithm can achieve the $O(\epsilon^{-4})$ sample complexity in Section 4.

3.2 Decentralized Stochastic Multi-Level Compositional Variance-Reduced Gradient Descent

To improve the convergence rate, in Algorithm 2, we propose our second algorithm: Decentralized Stochastic Multi-level Compositional Variance-Reduced Gradient descent algorithm (DSMCVRG).

Similar to Algorithm 1, we leverage the standard STORM technique [2] to estimate each level function, which is shown in Line 7, where $\beta > 0$ and $\beta\eta^2 < 1$. Different from Algorithm 1, we do not exploit the momentum to update model parameters. Instead, we leverage the variance-reduced gradient for local update, which is defined below:

$$m_{n,t} = (1-\mu\eta^2)(m_{n,t-1} - g_{n,t-1}^{\xi_t}) + g_{n,t}^{\xi_t} , \quad (4)$$

where $\mu > 0$ is a hyperparameter satisfying $\mu\eta^2 < 1$, the stochastic gradients $g_{n,t}^{\xi_t}$ and $g_{n,t-1}^{\xi_t}$ are defined as:

$$\begin{aligned} g_{n,t-1}^{\xi_t} &= \nabla f_n^{(1)}(u_{n,t-1}^{(0)}; \xi_{n,t}^{(1)}) \nabla f_n^{(2)}(u_{n,t-1}^{(1)}; \xi_{n,t}^{(2)}) \cdots \\ &\quad \times \nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)}; \xi_{n,t}^{(K-1)}) \nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}; \xi_{n,t}^{(K)}) , \\ g_{n,t}^{\xi_t} &= \nabla f_n^{(1)}(u_{n,t}^{(0)}; \xi_{n,t}^{(1)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}; \xi_{n,t}^{(2)}) \cdots \\ &\quad \times \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}; \xi_{n,t}^{(K-1)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)}) . \end{aligned} \quad (5)$$

Then, based on this variance-reduced gradient, we exploit the gradient-tracking technique to update the model parameter on each device, which is shown in Lines 17 and 19.

Novelty. Here, we would like to emphasize the novelty on the algorithmic design in Algorithm 2. Under the single-machine setting, existing variance-reduced multi-level compositional gradient descent algorithms (Zhang and Xiao, 2021; Jiang et al., 2022) apply the variance-reduction technique to each level function and its stochastic Jacobian matrix. For instance, Jiang et al. (2022) computes the variance-reduced Jacobian matrix for each level function as follows:

$$\begin{aligned} v_{n,t}^{(k)} &= (1-\beta\eta^2)(v_{n,t-1}^{(k)} - \nabla f_n^{(k)}(u_{n,t-1}^{(k-1)}; \xi_{n,t}^{(k)})) \\ &\quad + \nabla f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)}) . \end{aligned} \quad (6)$$

[2] Compared with Algorithm 1, η is replaced with η^2 when estimating each level function.

Algorithm 2 DSMCVRG

Input: $x_{n,0} = x_0$, $\alpha > 0$, $\beta > 0$, $\mu > 0$, $\eta > 0$.

1: **for** $t = 0, \cdots, T-1$ **do**
2: $u_{n,t}^{(0)} = x_{n,t}$,
3: **for** $k = 1, \cdots, K-1$ **do**
4: **if** $t == 0$ **then**
5: With batch size S, compute $u_{n,t}^{(k)} = f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$, $v_{n,t}^{(k)} = \nabla f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$,
6: **else**
7: $u_{n,t}^{(k)} = (1-\beta\eta^2)(u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)}; \xi_{n,t}^{(k)})) + f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$,
8: $v_{n,t}^{(k)} = \nabla f_n^{(k)}(u_{n,t}^{(k-1)}; \xi_{n,t}^{(k)})$,
9: **end if**
10: **end for**
11: **if** $t == 0$ **then**
12: $v_{n,t}^{(K)} = \nabla f_n^{(k)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)})$ with batch size S,
13: $m_{n,t} = g_{n,t}^{\xi_t}$, $\quad y_{n,t} = m_{n,t}$,
14: **else**
15: $v_{n,t}^{(K)} = \nabla f_n(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)})$,
16: $m_{n,t} = (1-\mu\eta^2)(m_{n,t-1} - g_{n,t-1}^{\xi_t}) + g_{n,t}^{\xi_t}$,
17: $y_{n,t} = \sum_{n' \in \mathcal{N}_n} w_{nn'} y_{n,t-1} + m_{n,t} - m_{n,t-1}$,
18: **end if**
19: $x_{n,t+\frac{1}{2}} = \sum_{n' \in \mathcal{N}_n} w_{nn'} x_{n',t} - \alpha y_{n,t}$, $x_{n,t+1} = x_{n,t} + \eta(x_{n,t+\frac{1}{2}} - x_{n,t})$,
20: **end for**

This kind of variance-reduced estimator for each level function suffers from some limitations. On the theoretical analysis side, when bounding the gradient estimation error for $\nabla F_n(\cdot)$, it requires $v_{n,t}^{(k)}$ to be upper bounded in all levels and iterations. To do that, Zhang and Xiao (2021) uses a clipping operation, which may result in a very tiny update (See γ_t in Algorithm 3 of (Zhang and Xiao, 2021)), while Jiang et al. (2022) employs a projection operation to guarantee $v_{n,t}^{(k)}$ is upper bounded by the Lipschitz constant of the deterministic Jacobian matrix (See Eq. (3) in Jiang et al. (2022)), which is an unknown hyperparameter so that it is not feasible in practice. On the implementation side, these algorithms are not friendly for practical applications. For instance, when applying them to the stochastic training of graph neural networks (GNN), computing the variance-reduced Jacobian matrix for each level function (i.e., each layer of GNN) requires to intervene the backpropagation in each layer, which is not easy to implement.

On the contrary, our Algorithm 2 just computes the standard stochastic Jacobian matrix $v_{n,t}^{(k)}$ for each level function. This can naturally avoid the aforementioned impractical operations since the standard stochastic Jacobian matrix is easy to bound under the commonly

used assumptions. Meanwhile, it is easy to compute. However, using standard stochastic Jacobian matrix of each level function may introduce a large estimation error. Then, a natural question follows: *Can Algorithm 2 achieve the $O(\epsilon^{-3})$ sample complexity as (Zhang and Xiao, 2021; Jiang et al., 2022) when not using the variance reduction technique for each level function's Jacobian?* In Section 4, we provide an affirmative answer: Our Algorithm 2 can still achieve the $O(\epsilon^{-3})$ sample complexity, even though we don't use the variance reduced Jacobian for each level function.

All in all, our algorithm is novel and we believe our idea can be leveraged to improve existing single-machine algorithms (Zhang and Xiao, 2021; Jiang et al., 2022).

4 Convergence Analysis

To establish the convergence rate of our algorithms, we introduce the following assumptions, which are commonly used in existing multi-level compositional optimization works (Yang et al., 2019; Zhang and Xiao, 2021; Jiang et al., 2022).

Assumption 2. *For any $k \in \{1, 2, \cdots, K\}$ and any $y_1, y_2 \in \mathbb{R}^{d_{k-1}}$, there exists $L_k > 0$ such that $\|\nabla f^{(k)}(y_1) - \nabla f^{(k)}(y_2)\| \leq L_k \|y_1 - y_2\|$ and $\mathbb{E}[\|\nabla f^{(k)}(y_1; \xi^{(k)}) - \nabla f^{(k)}(y_2; \xi^{(k)})\|] \leq L_k \|y_1 - y_2\|$. Additionally, $F_n(x)$ is L_F-smooth [3] where $L_F > 0$.*

Assumption 3. *For any $k \in \{1, 2, \cdots, K\}$ and any $y \in \mathbb{R}^{d_{k-1}}$, there exists $C_k > 0$ such that $\mathbb{E}[\|\nabla f^{(k)}(y; \xi)\|^2] \leq C_k^2$ and $\|\nabla f^{(k)}(y)\|^2 \leq C_k^2$.*

Assumption 4. *For any $k \in \{1, 2, \cdots, K\}$ and any $y_1, y_2 \in \mathbb{R}^{d_{k-1}}$, there exist $\sigma_k > 0$ and $\delta_k > 0$ such that $\mathbb{E}[\|\nabla f^{(k)}(y; \xi) - \nabla f^{(k)}(y)\|^2] \leq \sigma_k^2$ and $\mathbb{E}[\|f^{(k)}(y; \xi) - f^{(k)}(y)\|^2] \leq \delta_k^2$.*

Based on these assumptions, we denote $A_k = (\sum_{j=k}^{K-1} (\frac{L_{j+1} \prod_{i=1}^{K} C_i}{C_{j+1}} \prod_{i=k+1}^{j} C_i))^2$ and $B_k = \frac{\prod_{j=1}^{K} C_j^2}{C_k^2}$ for $k \in \{1, \cdots, K-1\}$, as well as $D_k = \frac{(\prod_{j=1}^{K} C_j^2) L_{k+1}^2}{C_{k+1}^2}$ for $k \in \{0, \cdots, K-1\}$. Moreover, we use $\bar{z}_t$ to denote the mean value across devices for any variables throughout this paper. Then, we established the convergence rate of our two algorithms.

Theorem 1. *Given Assumptions 1-4, by setting $\mu > 0$, $\beta > 0$, $\alpha \leq \min\{(1-\lambda)^2/\sqrt{\tilde{\alpha}_1}, 1/(4\sqrt{\tilde{\alpha}_2})\}$, $\eta \leq \min\{\tilde{\omega}_k/(8\beta \sum_{j=1}^{K-1} \tilde{\omega}_j C_j^2 \prod_{i=k+1}^{j}(2C_i^2)), 1/(2\alpha L_F), 1/\beta, 1/\mu, 1\}$ for any $k \in \{1, 2, \cdots, K-1\}$, Algorithm 1 has the following convergence rate:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \leq \frac{2(F(x_0)-F(x_*))}{\alpha\eta T} + O(\frac{\mu K}{T}) \\ + O(\frac{K}{\eta T}) + O(\frac{K}{\mu\eta T}) + O(\beta^2\mu^2\eta^3 K) + O(\mu^2\eta K) \\ + O(\mu^3\eta^2 K) + O(\beta^2\eta^2 K) + O(\mu\eta K) + O(\beta^2\eta K)\ ,\end{aligned} \tag{7}$$

where $\tilde{\omega}_k = \frac{2}{\beta}((12A_k + 8D_k)\mu + 2A_k + 2\beta\sum_{j=k+1}^{K-1}(20A_jC_j^2 + 8D_jC_j^2)\prod_{i=k+1}^{j}(2C_i^2))$ for $k \in \{1, 2, \cdots, K-1\}$, and $\tilde{\omega}_{K+1} = L_F^2 + 8(\frac{2L_F^2}{\mu^2} + 8L_F^2 + 4KD_0 + K\sum_{k=1}^{K-1}(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2)(\prod_{j=1}^{k-1}(2C_j^2)))$, $\tilde{\alpha}_1 = 4\tilde{\omega}_{K+1} + 8L_F^2/\mu^2 + 32L_F^2 + 16KD_0 + 4K\sum_{k=1}^{K-1}(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2)(\prod_{j=1}^{k-1}(2C_j^2))$, $\tilde{\alpha}_2 = 2L_F^2/\mu^2 + 8L_F^2 + 4KD_0 + K\sum_{k=1}^{K-1}(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2)(\prod_{j=1}^{k-1}(2C_j^2))$.

Corollary 1. *Given Assumptions 1-4, by setting $\mu = O(1)$, $\beta = O(1)$, $\alpha = O((1-\lambda)^2)$, $\eta = O(\epsilon^2)$, $T = O((1-\lambda)^{-2}\epsilon^{-4})$, Algorithm 1 can achieve the ϵ-stationary point, i.e., $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \leq \epsilon^2$.*

Remark 1. *Given $\mu = O(1)$ and $\beta = O(1)$, the hyperparameters $\tilde{\alpha}_i$ $(i = 1, 2)$ and $\tilde{\omega}_k$ $(k \in \{1, 2, \cdots, K-1, K+1\})$ are independent of the learning rate and spectral gap. Thus, they do not affect the order of the convergence rate.*

Remark 2. *From Corollary 1, we can know that the convergence rate of Algorithm 1 is $O((1-\lambda)^{-2}\epsilon^{-4})$, which is independent of the number of function levels. Meanwhile, it indicates that the dependence on the spectral gap is $O((1-\lambda)^{-2})$. When the communication graph is fully connected, the convergence rate becomes $O(\epsilon^{-4})$, which can match the single-machine momentum algorithm (Balasubramanian et al., 2022). All in all, the level-independent convergence rate is achievable under the dencetralized setting.*

Remark 3. *Since the mini-batch size is $O(1)$, the sample complexity is $O((1-\lambda)^{-2}\epsilon^{-4})$. Moreover, the communication complexity is $O((1-\lambda)^{-2}\epsilon^{-4})$.*

Theorem 2. *Given Assumptions 1-4, by setting $\mu > 0$, $\beta > 0$, $\alpha \leq \min\{(1-\lambda)^2/\sqrt{\tilde{\alpha}_1}, 1/(4\sqrt{\tilde{\alpha}_2})\}$, $\eta \leq \min\{0.5\sqrt{\tilde{\omega}_k/(2\beta\sum_{j=1}^{K-1}\tilde{\omega}_jC_j^2(\prod_{i=k+1}^{j}(2C_i^2)))}, 1/(2\alpha L_F), 1/\sqrt{\beta}, 1/\sqrt{\mu}, 1\}$ for any $k \in \{1, 2, \cdots, K-1\}$, Algorithm 2 has the following convergence rate:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \leq \frac{2(F(x_0)-F(x_*))}{\alpha\eta T} + O(\frac{K}{\eta^2 TS}) \\ + O(\frac{K}{\mu\eta^2 TS}) + O(\frac{\mu\eta K}{TS}) + O(\frac{K}{\eta T}) + O(\beta^2\eta^3 K) \\ + O(\mu^2\eta^3 K) + O(\beta^2\eta^2 K) + O(\frac{\beta^2\eta^2 K}{\mu}) \\ + O(\mu\eta^2 K) + O(\mu\beta^2\eta^5 K) + O(\mu^3\eta^5 K)\ ,\end{aligned} \tag{8}$$

[3] Based on the smoothness of each level function, it is easy to prove F_n is smooth (Yang et al., 2019; Zhang and Xiao, 2021; Jiang et al., 2022) so that we directly assume it is smooth.

where $\tilde{\omega}_k = \frac{16D_k}{\mu N} + 24D_k + \frac{4A_k}{\beta} + 16\sum_{j=1}^{K-1}((\frac{2}{\mu N} + 3)D_jC_j^2)(\prod_{i=k+1}^{j}(2C_i^2))$ *for* $k \in \{1, 2, \cdots, K-1\}$, $\tilde{\omega}_{K+2} = 16(\sum_{k=1}^{K-1}((\frac{8K}{\mu N} + 12K)D_kC_k^2 + 2\tilde{\omega}_kC_k^2)(\prod_{j=1}^{k-1}(2C_j^2)) + \frac{4KD_0}{\mu N} + 6KD_0) + 2L_F^2$, $\tilde{\alpha}_1 = 2\tilde{\omega}_{K+2} + 4K[\sum_{k=1}^{K-1}((\frac{8}{\mu N} + 12)D_kC_k^2 + 2\tilde{\omega}_kC_k^2)(\prod_{j=1}^{k-1}(2C_j^2)) + \frac{4D_0}{\mu N} + 6D_0]$, $\tilde{\alpha}_2 = K\sum_{k=1}^{K-1}((\frac{8}{\mu N} + 12)D_kC_k^2 + 2\tilde{\omega}_kC_k^2)(\prod_{j=1}^{k-1}(2C_j^2)) + \frac{4KD_0}{\mu N} + 6KD_0$.

Corollary 2. *Given Assumptions 1-4, by setting* $\mu = O(1)$, $\beta = O(1)$, $\alpha = O((1-\lambda)^2)$, $S = O(\epsilon^{-1})$, $\eta = O(\epsilon)$, $T = O((1-\lambda)^{-2}\epsilon^{-3})$, *Algorithm 2 can achieve* ϵ*-stationary point.*

Remark 4. *Given* $\mu = O(1)$ *and* $\beta = O(1)$, *the hyperparameters* $\tilde{\alpha}_i$ $(i = 1, 2)$ *and* $\tilde{\omega}_k$ $(k \in \{1, 2, \cdots, K-1, K+2\})$ *also do not affect the order of the convergence rate.*

Remark 5. *From Corollary 2, we can know that the convergence rate of Algorithm 2 is* $O((1-\lambda)^{-2}\epsilon^{-3}))$, *which is also independent of the number of function levels and has the dependence on the spectral gap with* $O((1-\lambda)^{-2})$. *Moreover, this convergence rate is better than Algorithm 1. Additionally, when the communication graph is fully connected, the convergence rate can match the single-machine algorithms (Zhang and Xiao, 2021; Jiang et al., 2022), but our Algorithm 2 requires much milder operations than (Zhang and Xiao, 2021; Jiang et al., 2022).*

Remark 6. *Since the mini-batch size is* $O(1)$ *except the first iteration, the sample complexity is* $O((1-\lambda)^{-2}\epsilon^{-3}))$. *Similarly, we can know that the communication complexity is* $O((1-\lambda)^{-2}\epsilon^{-3}))$.

Discussions. Due to the multi-level nested structure and the decentralized communication scheme, it is quite challenging to establish the convergence rate of our algorithms. Specifically, compared with the decentralized two-level compositional optimization problem, the multi-level nested structure makes the convergence analysis more difficult. For instance, when bounding $\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2]$ in Lemma 7, its upper bound depends on the update of the lower-level function estimator $\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2]$, which further has a quite complicated upper bound as below:

$$\begin{aligned}\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \leq& \Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(0)} - u_{n,t}^{(0)}\|^2] \\ &+ 2\beta^2\eta^2\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(j)} - f_n^{(j)}(u_{n,t-1}^{(j-1)})\|^2] \\ &+ 2\beta^2\eta^2\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\delta_j^2 \ . \end{aligned} \tag{9}$$

On the contrary, in the two-level compositional optimization problem, $\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2]$ becomes the update of model parameters, which is much easier to bound. On the other hand, compared with the single-machine multi-level compositional optimization problem, $\mathbb{E}[\|u_{n,t-1}^{(0)} - u_{n,t}^{(0)}\|^2]$ in Eq. (9) involves the decentralized communication operation, which makes it more difficult to bound.

Furthermore, the multi-level structure and the decentralized communication scheme bring more challenges to bound the consensus error, e.g., Lemma 11 and Lemma 24. Last but not least, our algorithm does not apply the variance-reduction technique to the stochastic Jacobian matrix of each level function. Thus, we need to carefully bound the gradient estimation error to guarantee the desired convergence rate. This has never been studied before so that we need to develop new strategies to bound the gradient estimation error, e.g., Lemma 22. All in all, the theoretical analysis is challenging.

To address those challenges, we developed novel potential functions to establish the convergence rate of our algorithms. In particular, to prove Theorem 1, we proposed the following potential function:

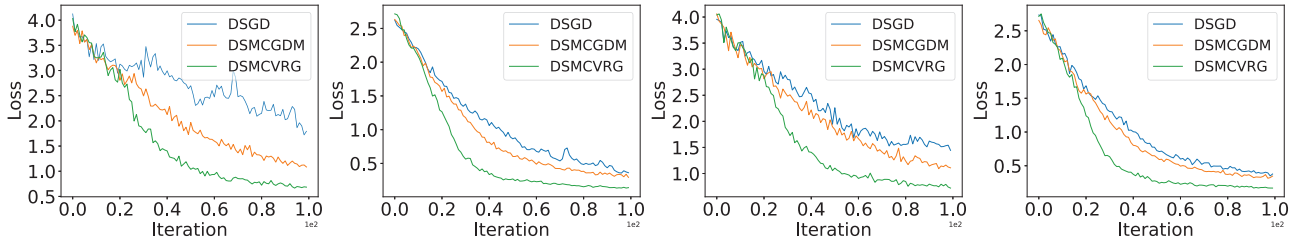
$$\begin{aligned}\mathcal{H}_t =& \mathbb{E}[F(\bar{x}_t)] + \omega_0\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla F_n(x_{n,t})\Big\|^2\Big] \\ &+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\ &+ \omega_K\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\ &+ \omega_{K+1}\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \omega_{K+2}\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \ , \end{aligned} \tag{10}$$

where $\omega_i > 0$ $(i \in \{0, 1, \cdots, K+2\})$ are determined in our proof, which actually is challenging due to the interaction between the multi-level structure and the decentralized communication scheme.

Moreover, since this potential function cannot be applied to Theorem 2, we proposed the following potential function to prove Theorem 2:

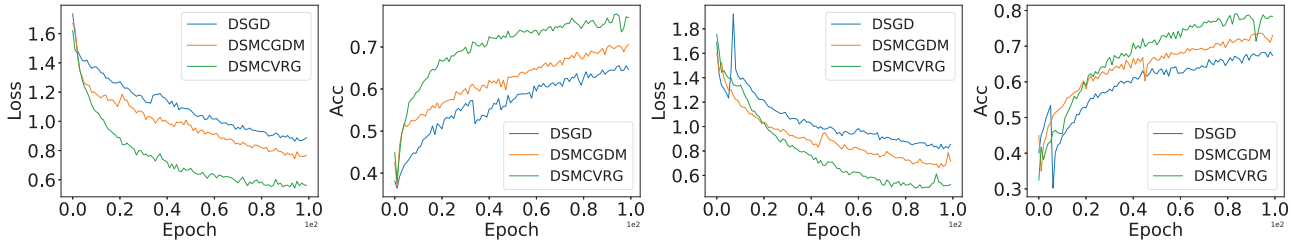
$$\begin{aligned}\mathcal{H}_t =& \mathbb{E}[F(\bar{x}_t)] + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\ &+ \omega_K\mathbb{E}[\|\bar{m}_t - \bar{h}_t\|^2] + \omega_{K+1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|m_{n,t} - h_{n,t}\|^2] \\ &+ \omega_{K+2}\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \omega_{K+3}\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \ , \end{aligned} \tag{11}$$

where $h_{n,t} = \nabla f_n^{(1)}(u_{n,t}^{(0)}) \cdots \nabla f_n^{(K)}(u_{n,t}^{(K-1)})$, and $\omega_i > 0$ $(i \in \{1, \cdots, K+3\})$ are determined in our proof.



(a) Ring Graph (Support) (b) Ring Graph (Query) (c) Random Graph (Support) (d) Random Graph (Query)

Figure 1: Regression: The loss function value on support and query sets versus the number of iterations for the ring and random graph.



(a) Ring Graph (Support) (b) Ring Graph (Query) (c) Random Graph (Query) (d) Random Graph (Support)

Figure 2: Classification: The loss function value on support set and test accuracy on query set versus the number of epochs for the ring and random graph.

Based on these two novel potential functions, the task boils down to studying how each term evolves across iterations and determining its coefficient. The detailed proof can be found in Appendix.

5 Experiment

In this section, we apply our proposed algorithms to the multi-step model-agnostic meta-learning task to verify the performance of our algorithms.

5.1 Multi-Step Model-Agnostic Meta-Learning

Model-agnostic meta-learning (MAML) (Finn et al., 2017) is to learn an initialization model that can be adapted to a new task via a couple of steps of stochastic gradient descent. Basically, the one-step MAML under the decentralized setting is defined as below:

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{i\sim\mathcal{P}_{n,\text{task}},\zeta_n\sim\mathcal{D}_{n,\text{query}_i}} \left[\mathcal{L}_{n,i}\left(y;\zeta_n\right)\right], \quad (12)$$

$$\text{where } y = x - \nu\mathbb{E}_{\xi_n\sim\mathcal{D}_{n,\text{support}_i}} \nabla\mathcal{L}_{n,i}\left(x;\xi_n\right) \ , \quad (13)$$

where Eq. (13) denotes one-step gradient descent, ν is the learning rate, $\mathcal{P}_{n,\text{task}}$ denotes the task distribution on the n-th device, $\mathcal{D}_{n,\text{query}_i}$ ($\mathcal{D}_{n,\text{support}_i}$) represents the query (support) set of the i-th task on the n-th device. This one-step update can be viewed as a two-level compositional optimization problem. If taking multiple gradient descent steps, this problem becomes a multi-level compositional optimization problem (Jiang et al., 2022; Chen et al., 2020). Therefore, we can apply our algorithms to the multi-step MAML problem. In our experiment, we will focus on two tasks: regression and classification tasks.

5.2 Experimental Settings and Results

Regression. For the regression problem, we follow (Finn et al., 2017) to generate a sinewave dataset. Specifically, when generating the sine wave, the amplitude is randomly picked from $[0.1, 5.0]$, the phase is from $[0, \pi]$, and the input is from $[-5, 5]$. The model used for this task is a fully-connected neural network with the dimensionality as $[1, 40, 40, 1]$. For the support set, the meta-batch size (tasks) on each device is set to 200 and the number of samples for each task is 10. For the query set, the meta-batch size is 500 and the number of samples in each task is also 10. Moreover, the number of gradient descent updates in Eq. (13) is 3 so that it is a four-level compositional optimization problem. The learning rate ν is 0.01.

Classification. In this experiment, we use Omniglot dataset, which has 1,623 characters (tasks) and each character has 20 images. 1,200 tasks are used as the support set and the left tasks are used as the query

set. Following (Finn et al., 2017), we employ the 5-way-1-shot setting. The model we used has four convolutional layers, where each layer has 64 3×3 filters, and one linear layer. The meta-batch size (tasks) on each device is set to 8. The number of gradient descent updates in Eq. (13) is set to 3 so that it is also a four-level compositional optimization problem. The learning rate ν is 0.01 too.

In our experiments, we select μ and β from $\{1, 3, 5, 7, 9\}$, and fix α to 1.0. Additionally, we set $\epsilon^2 = 0.1$. Then, we set the learning rate $\eta = \epsilon^2$ for Algorithm 1 in terms of Corollary 1, and $\eta = \epsilon$ for Algorithm 2 according to Corollary 2. Moreover, we use four devices in our experiments. The topology we used includes the ring graph and random graph. Here, the random graph is generated from an Erdos-Renyi random graph with the edge probability being 0.4. As for the baseline algorithm, we use the standard decentralized SGD (DSGD) (Lian et al., 2017) since there does not exist other decentralized multi-level compositional algorithms. In our experiments, the learning rate of DSGD is 0.1

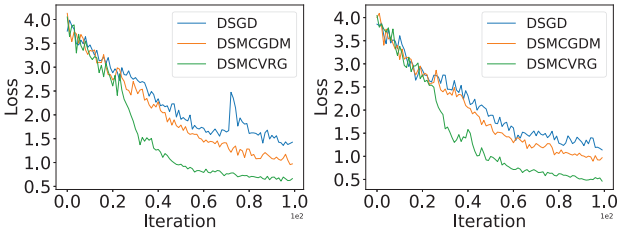
In Figure 1, we report the support and query loss function values versus the number of iterations for the regression task. It is easy to find that our two algorithms outperform the standard DSGD algorithm. The reason is that our algorithms leverage the variance-reduction technique to control the estimation error for each level function. Moreover, our second algorithm DSMCVRG converges faster than the first algorithm DSMCGDM, which confirms the correctness of our theoretical results.

In Figure 2, we show the loss function value on the support set and the accuracy on the query set for the classification task. It can also be found that our two algorithms outperform the baseline algorithm and DSMCVRG converges faster than DSMCGDM, which further confirms the correctness of our theoretical results.

5.3 More Experiments

To further demonstrate the performance of our algorithms, we set the number of inner steps of multi-step MAML to 4 and 5 so that we have the five-level and six-level compositional optimization problems. In Figure 3, we show the loss function values on the support set versus the number of iterations for sinewave dataset. From this figure, we can still find that our two algorithms outperform DSGD and our second algorithm DSMCVRG converges faster than DSMCGDM, which confirms the effectiveness and correctness of our proposed algorithms.

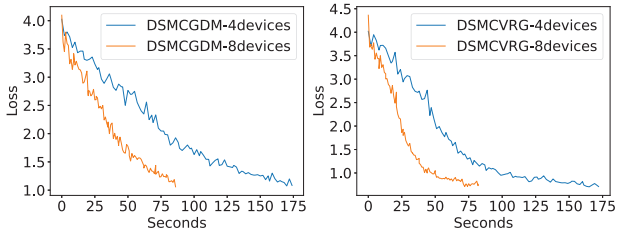
Moreover, we show the acceleration benefit of our decentralized optimization algorithms. In particular, we



(a) Five-level SCO problem (b) Six-level SCO problem

Figure 3: The loss function value on support test versus the number of iterations for the regression task. Ring graph is used.

compare the convergence performance when using four and eight devices. Here, the meta-batch size is set to 200 when using four devices and it is set to 100 when using eight devices. Other hyperparameters are the same as previous experiments. In Figure 4, we show the loss function value on the support set versus the consumed time for the regression task when using the ring graph. It is easy to find that using more devices can accelerate the convergence speed, which confirms the efficacy of our algorithms.



(a) DSMCGDM (b) DSMCVRG

Figure 4: The loss function value on support set versus the consumed time for the regression task and ring graph.

6 Conclusion

In this paper, we developed two novel decentralized stochastic multi-level compositional optimization algorithms. They both can achieve the level-independent convergence rate with practical operations. In particular, we developed a novel strategy for applying the variance reduction technique to estimate the gradient. Extensive experimental results confirm the effectiveness of our algorithms. We believe our novel algorithmic design and theoretical analysis strategies can benefit the development of multi-level compositional optimization problems for both single-machine and distributed settings.

References

K. Balasubramanian, S. Ghadimi, and A. Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

T. Chen, Y. Sun, and W. Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *arXiv preprint arXiv:2008.10847*, 2020.

W. Cong, M. Ramezani, and M. Mahdavi. On the importance of sampling in training gcns: Tighter analysis and variance reduction. *arXiv e-prints*, pages arXiv–2103, 2021.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 15236–15245, 2019.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

H. Gao. Achieving linear speedup in decentralized stochastic compositional minimax optimization. *arXiv preprint arXiv:2307.13430*, 2023.

H. Gao and H. Huang. Periodic stochastic gradient descent with momentum for decentralized training. *arXiv preprint arXiv:2008.10435*, 2020.

H. Gao and H. Huang. Fast training method for stochastic compositional optimization problems. *Advances in Neural Information Processing Systems*, 34, 2021.

H. Gao, B. Gu, and M. T. Thai. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *International Conference on Artificial Intelligence and Statistics*, pages 9238–9281. PMLR, 2023.

S. Ghadimi, A. Ruszczynski, and M. Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

Y. Hua, K. Miller, A. L. Bertozzi, C. Qian, and B. Wang. Efficient and reliable overlay networks for decentralized federated learning. *SIAM Journal on Applied Mathematics*, 82(4):1558–1586, 2022.

W. Jiang, B. Wang, Y. Wang, L. Zhang, and T. Yang. Optimal algorithms for stochastic multi-level compositional optimization. *arXiv preprint arXiv:2202.07530*, 2022.

A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019a.

A. Koloskova, S. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019b.

X. Lian and J. Liu. Revisit batch normalization: New understanding from an optimization view and a refinement via composition optimization. *arXiv preprint arXiv:1810.06177*, 2018.

X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *arXiv preprint arXiv:1705.09056*, 2017.

S. Lu, S. Zeng, X. Cui, M. Squillante, L. Horesh, B. Kingsbury, J. Liu, and M. Hong. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 35:30638–30650, 2022.

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

Z. Song, W. Li, K. Jin, L. Shi, M. Yan, W. Yin, and K. Yuan. Communication-efficient topologies for decentralized learning with $o(1)$ consensus rate. *arXiv preprint arXiv:2210.07881*, 2022.

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

M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

R. Xin, U. A. Khan, and S. Kar. A near-optimal stochastic gradient method for decentralized non-convex finite-sum optimization. *arXiv preprint arXiv:2008.07428*, 2020.

S. Yang, M. Wang, and E. X. Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1): 616–659, 2019.

B. Ying, K. Yuan, Y. Chen, H. Hu, P. Pan, and W. Yin. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34:13975–13987, 2021.

H. Yu, L. Wang, B. Wang, M. Liu, T. Yang, and S. Ji. Graphfm: Improving large-scale gnn training via feature momentum. In *International Conference on Machine Learning*, pages 25684–25701. PMLR, 2022.

H. Yuan and W. Hu. Stochastic recursive momentum method for non-convex compositional optimization. *arXiv preprint arXiv:2006.01688*, 2020.

H. Yuan, X. Lian, and J. Liu. Stochastic recursive variance reduction for efficient smooth non-convex compositional optimization. *arXiv preprint arXiv:1912.13515*, 2019.

J. Zhang and L. Xiao. A composite randomized incremental gradient method. In *International Conference on Machine Learning*, pages 7454–7462, 2019a.

J. Zhang and L. Xiao. A stochastic composite gradient method with incremental variance reduction. In *Advances in Neural Information Processing Systems*, pages 9078–9088, 2019b.

J. Zhang and L. Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.

Y. Zhang, M. T. Thai, J. Wu, and H. Gao. On the communication complexity of decentralized bilevel optimization. *arXiv preprint arXiv:2311.11342*, 2023.

S. Zhao and Y. Liu. Distributed stochastic compositional optimization problems over directed networks. *arXiv preprint arXiv:2203.11074*, 2022.

Supplementary Materials

A Appendix

A.1 Terminologies

Before presenting the detailed proof, we first introduce some terminologies as below. First, we denote the function up to the k-th level as below:

$$F_n^{(k)}(x) = f_n^{(1)}(x) f_n^{(2)}(F_n^{(1)}(x)) \cdots f_n^{(k-1)}(F_n^{(k-2)}(x)) f^{(k)}(F^{(k-1)}(x)) \ , \tag{14}$$

where $f_n^{(k)}(\cdot) = \mathbb{E}[f_n^{(k)}(\cdot; \xi^{(k)})]$ and $\xi^{(k)}$ denotes the random sample. It is easy to know $F_n(x) = F_n^{(K)}(x) = f_n^{(1)}(x) f_n^{(2)}(F_n^{(1)}(x)) \cdots f_n^{(K-1)}(F_n^{(K-2)}(x)) f_n^{(K)}(F_n^{(K-1)}(x))$. Then, the gradient of $\nabla F_n^{(k)}(x)$ can be represented as below:

$$\nabla F_n^{(k)}(x) = \nabla f_n^{(1)}(x) \nabla f_n^{(2)}(F_n^{(1)}(x)) \cdots \nabla f_n^{(k-1)}(F_n^{(k-2)}(x)) \nabla f_n^{(k)}(F_n^{(k-1)}(x)) \ , \tag{15}$$

where $\nabla f_n^{(k)}(\cdot) = \mathbb{E}[\nabla f_n^{(k)}(\cdot; \xi^{(k)})]$.

Throughout the proof, we assume $\prod_i^k a_i = 1$ when $i > k$. Additionally, we denote $X_t = [x_{1,t}, \cdots, x_{N,t}]$, $Y_t = [y_{1,t}, \cdots, y_{N,t}]$, $M_t = [m_{1,t}, \cdots, m_{N,t}]$, $G_t = [g_{1,t}, \cdots, g_{N,t}]$, $\bar{X}_t = [\frac{1}{N}\sum_{n=1}^N x_{n,t}, \cdots, \frac{1}{N}\sum_{n=1}^N x_{n,t}]$, $\bar{Y}_t = [\frac{1}{N}\sum_{n=1}^N y_{n,t}, \cdots, \frac{1}{N}\sum_{n=1}^N y_{n,t}]$, $\bar{M}_t = [\frac{1}{N}\sum_{n=1}^N m_{n,t}, \cdots, \frac{1}{N}\sum_{n=1}^N m_{n,t}]$.

A.2 Proof of Theorem 1

Lemma 1. *For $k \in \{1, \cdots, K-1\}$, given Assumptions 2-4, we can get*

$$\|u_{n,t}^{(k)} - F_n^{(k)}(x_{n,t})\| \leq \sum_{j=1}^k \Big(\prod_{i=j+1}^k C_i \Big) \|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\| \ . \tag{16}$$

Proof. When $k = 1$, we have $\|u_{n,t}^{(1)} - F_n^{(1)}(x_{n,t})\| = \|u_{n,t}^{(1)} - f_n^{(1)}(u_{n,t}^{(0)})\|$. Assume for $k > 1$, we have

$$\|u_{n,t}^{(k)} - F_n^{(k)}(x_{n,t})\| \leq \sum_{j=1}^k \Big(\prod_{i=j+1}^k C_i \Big) \|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\| \ . \tag{17}$$

Then, for $k+1$, we have

$$\begin{aligned}
&\|u_{n,t}^{(k+1)} - F_n^{(k+1)}(x_{n,t})\| \\
&= \|u_{n,t}^{(k+1)} - f_n^{(k+1)}(F_n^{(k)}(x_{n,t}))\| \\
&\leq \|u_{n,t}^{(k+1)} - f_n^{(k+1)}(u_{n,t}^{(k)})\| + \|f_n^{(k+1)}(u_{n,t}^{(k)}) - f_n^{(k+1)}(F_n^{(k)}(x_{n,t}))\| \\
&\leq \|u_{n,t}^{(k+1)} - f_n^{(k+1)}(u_{n,t}^{(k)})\| + C_{k+1}\|u_{n,t}^{(k)} - F_n^{(k)}(x_{n,t})\| \\
&\leq \|u_{n,t}^{(k+1)} - f_n^{(k+1)}(u_{n,t}^{(k)})\| + C_{k+1}\sum_{j=1}^k \Big(\prod_{i=j+1}^k C_i \Big) \|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\| \\
&= \sum_{j=1}^{k+1} \Big(\prod_{i=j+1}^{k+1} C_i \Big) \|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\| \ ,
\end{aligned} \tag{18}$$

which completes the proof.

□

Lemma 2. *For $k \in \{2, \cdots, K\}$, given Assumptions 2-4, we can get*

$$\begin{aligned}\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \leq \Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(0)} - u_{n,t}^{(0)}\|^2] + 2\beta^2\eta^2\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(j)} - f_n^{(j)}(u_{n,t-1}^{(j-1)})\|^2] \\ + 2\beta^2\eta^2\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\delta_j^2 \ ,\end{aligned} \tag{19}$$

and

$$\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \leq 2C_{k-1}^2\mathbb{E}[\|u_{n,t-1}^{(k-2)} - u_{n,t}^{(k-2)}\|^2 + 2\beta^2\eta^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - f_n^{(k-1)}(u_{n,t-1}^{(k-2)})\|^2] + 2\beta^2\eta^2\delta_{k-1}^2 \ . \tag{20}$$

Proof. For any $k > 1$, we can get

$$\begin{aligned}&\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \\ &= \mathbb{E}[\|(1-\beta\eta)(u_{n,t-1}^{(k-1)} - f_n^{(k-1)}(u_{n,t-1}^{(k-2)};\xi_{n,t}^{(k-1)})) + f_n^{(k-1)}(u_{n,t}^{(k-2)};\xi_{n,t}^{(k-1)}) - u_{n,t-1}^{(k-1)}\|^2] \\ &= \mathbb{E}[\| - \beta\eta(u_{n,t-1}^{(k-1)} - f_n^{(k-1)}(u_{n,t-1}^{(k-2)}) + f_n^{(k-1)}(u_{n,t-1}^{(k-2)}) - f_n^{(k-1)}(u_{n,t-1}^{(k-2)};\xi_{n,t}^{(k-1)})) \\ &\quad - f_n^{(k-1)}(u_{n,t-1}^{(k-2)};\xi_{n,t}^{(k-1)}) + f_n^{(k-1)}(u_{n,t}^{(k-2)};\xi_{n,t}^{(k-1)})\|^2 \\ &\leq 2\mathbb{E}[\| - \beta\eta(u_{n,t-1}^{(k-1)} - f_n^{(k-1)}(u_{n,t-1}^{(k-2)}) + f_n^{(k-1)}(u_{n,t-1}^{(k-2)}) - f_n^{(k-1)}(u_{n,t-1}^{(k-2)};\xi_{n,t}^{(k-1)}))\|^2] \\ &\quad + 2\mathbb{E}[\| - f_n^{(k-1)}(u_{n,t-1}^{(k-2)};\xi_{n,t}^{(k-1)}) + f_n^{(k-1)}(u_{n,t}^{(k-2)};\xi_{n,t}^{(k-1)})\|^2] \\ &\leq 2C_{k-1}^2\mathbb{E}[\|u_{n,t-1}^{(k-2)} - u_{n,t}^{(k-2)}\|^2 + 2\beta^2\eta^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - f_n^{(k-1)}(u_{n,t-1}^{(k-2)})\|^2] + 2\beta^2\eta^2\delta_{k-1}^2 \ ,\end{aligned} \tag{21}$$

where the last step holds due to Assumption 3 and Assumption 4, and $\mathbb{E}[f_n^{(k-1)}(u_{n,t-1}^{(k-2)};\xi_{n,t}^{(k-1)})] = f_n^{(k-1)}(u_{n,t-1}^{(k-2)})$. Then, by recursively expanding this inequality, we can get

$$\begin{aligned}\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \leq \Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(0)} - u_{n,t}^{(0)}\|^2] \\ + 2\beta^2\eta^2\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(j)} - f_n^{(j)}(u_{n,t-1}^{(j-1)})\|^2] + 2\beta^2\eta^2\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\delta_j^2 \ .\end{aligned} \tag{22}$$

□

Lemma 3. *Given Assumptions 2-4, we can get*

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N}\|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\ &- \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\|^2 \\ &\leq \frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}A_k\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2 \ ,\end{aligned} \tag{23}$$

where $A_k = \left(\sum_{j=k}^{K-1}\left(\frac{L_{j+1}\prod_{i=1}^{K}C_i}{C_{j+1}}\prod_{i=k+1}^{j}C_i\right)\right)^2$.

Proof. Because $u_{n,t}^{(0)} = x_{n,t}$, we can get

$$
\begin{aligned}
&\|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\| \\
&= \|\nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\nabla f_n^{(3)}(u_{n,t}^{(2)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad + \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\nabla f_n^{(3)}(u_{n,t}^{(2)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\nabla f_n^{(3)}(F_n^{(2)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad + \cdots \\
&\quad + \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\| \\
&\le \|\nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\nabla f_n^{(3)}(u_{n,t}^{(2)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\| \\
&\quad + \|\nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\nabla f_n^{(3)}(u_{n,t}^{(2)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\nabla f_n^{(3)}(F_n^{(2)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\| \\
&\quad + \cdots \\
&\quad + \|\nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\| \\
&\le \frac{L_2\prod_{k=1}^K C_k}{C_2}\|u_{n,t}^{(1)} - F_n^{(1)}(x_{n,t})\| + \frac{L_3\prod_{k=1}^K C_k}{C_3}\|u_{n,t}^{(2)} - F_n^{(2)}(x_{n,t})\| \\
&\quad + \cdots + \frac{L_K\prod_{k=1}^K C_k}{C_K}\|u_{n,t}^{(K-1)} - F_n^{(K-1)}(x_{n,t})\| \\
&= \sum_{k=1}^{K-1}\frac{L_{k+1}\prod_{j=1}^K C_j}{C_{k+1}}\|u_{n,t}^{(k)} - F_n^{(k)}(x_{n,t})\| \\
&\le \sum_{k=1}^{K-1}\frac{L_{k+1}\prod_{j=1}^K C_j}{C_{k+1}}\sum_{j=1}^{k}\Big(\prod_{i=j+1}^{k} C_i\Big)\|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\| \\
&\le \sum_{k=1}^{K-1}\left(\sum_{j=k}^{K-1}\Big(\frac{L_{j+1}\prod_{i=1}^K C_i}{C_{j+1}}\prod_{i=k+1}^{j} C_i\Big)\right)\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|\ ,
\end{aligned}
\tag{24}
$$

where the second to last step holds due to Lemma 1. Then, we complete the proof by taking the squared operation on both sides.

□

Lemma 4. *Given Assumptions 2-4 and* $D_k = \frac{(\prod_{j=1}^K C_j^2)L_{k+1}^2}{C_{k+1}^2}$ *where* $k \in \{0,\cdots,K-1\}$*, we can get*

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\le KD_0\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + 2K\sum_{k=1}^{K-1} D_k C_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] \\
&\quad + 2\beta^2\eta^2 K\sum_{k=1}^{K-1} D_k\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 2\beta^2\eta^2 K\sum_{k=1}^{K-1} D_k\delta_k^2\ .
\end{aligned}
\tag{25}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
=&\,\mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad+\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad\cdots \\
&\quad+\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
\le&\, K\mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&+K\mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&+\cdots \\
&+K\mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
\le&\, K\frac{(\prod_{j=1}^K C_j^2)L_1^2}{C_1^2}\mathbb{E}[\|u_{n,t}^{(0)}-u_{n,t-1}^{(0)}\|^2]+K\frac{(\prod_{j=1}^K C_j^2)L_2^2}{C_2^2}\mathbb{E}[\|u_{n,t}^{(1)}-u_{n,t-1}^{(1)}\|^2] \\
&+\cdots+K\frac{(\prod_{j=1}^K C_j^2)L_K^2}{C_K^2}\mathbb{E}[\|u_{n,t}^{(K-1)}-u_{n,t-1}^{(K-1)}\|^2] \\
\le&\, KD_0\mathbb{E}[\|x_{n,t}-x_{n,t-1}\|^2]+K\sum_{k=1}^{K-1}D_k\mathbb{E}[\|u_{n,t}^{(k)}-u_{n,t-1}^{(k)}\|^2] \\
\le&\, KD_0\mathbb{E}[\|x_{n,t}-x_{n,t-1}\|^2]+2K\sum_{k=1}^{K-1}D_kC_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)}-u_{n,t}^{(k-1)}\|^2] \\
&+2\beta^2\eta^2K\sum_{k=1}^{K-1}D_k\mathbb{E}[\|u_{n,t-1}^{(k)}-f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2]+2\beta^2\eta^2K\sum_{k=1}^{K-1}D_k\delta_k^2\ ,
\end{aligned}
\tag{26}
$$

where $D_k=\frac{(\prod_{j=1}^K C_j^2)L_{k+1}^2}{C_{k+1}^2}$, and the last step holds due to Lemma 2.

□

Lemma 5. *Given Assumptions 2-4, we can get*

$$
\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N(g_{n,t}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}))\Big\|^2\Big]\le K\sum_{k=1}^K B_k\frac{\sigma_k^2}{N}\ , \tag{27}
$$

where $B_k=\frac{\prod_{j=1}^K C_j^2}{C_k^2}$.

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(g_{n,t}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}))\Big\|^2\Big]\\
=&\,\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(v_{n,t}^{(1)}v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\\
&+\nabla f_n^{(1)}(u_{n,t}^{(0)})v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\\
&+\cdots\\
&+\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})v_{n,t}^{(K)}\\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big)\Big\|^2\Big]\\
\le&\,K\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(v_{n,t}^{(1)}v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\Big)\Big\|^2\Big] \qquad (28)\\
&+K\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(\nabla f_n^{(1)}(u_{n,t}^{(0)})v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\Big)\Big\|^2\Big]\\
&+\cdots\\
&+K\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})v_{n,t}^{(K)}\Big)\Big\|^2\Big]\\
&\quad-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big)\Big\|^2\Big]\\
\le&\,K\sum_{k=1}^{K}\frac{\prod_{j=1}^{K}C_j^2}{C_k^2}\frac{\sigma_k^2}{N}\ ,
\end{aligned}
$$

where the last step holds due to the fact that the sampling procedure on different workers and different levels are independent. For instance, for the first level, we can get

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(v_{n,t}^{(1)}v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\Big)\Big\|^2\Big]\\
=&\,\frac{1}{N^2}\mathbb{E}\Big[\sum_{n=1}^{N}\Big\|\Big(v_{n,t}^{(1)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\Big)v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\Big\|^2\Big]\\
&+\frac{1}{N^2}\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{n'=1,n'\neq n}^{N}\Big\langle\Big(v_{n,t}^{(1)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\Big)v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)},\Big(v_{n',t}^{(1)}-\nabla f_{n'}^{(1)}(u_{n',t}^{(0)})\Big)v_{n',t}^{(2)}\cdots v_{n',t}^{(K-1)}v_{n',t}^{(K)}\Big\rangle\Big] \qquad (29)\\
=&\,\frac{1}{N^2}\mathbb{E}\Big[\sum_{n=1}^{N}\Big\|\Big(v_{n,t}^{(1)}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\Big)v_{n,t}^{(2)}\cdots v_{n,t}^{(K-1)}v_{n,t}^{(K)}\Big\|^2\Big]\\
\le&\,\frac{\prod_{j=1}^{K}C_j^2}{C_1^2}\frac{\sigma_1^2}{N}\ ,
\end{aligned}
$$

where the third step follows from the fact that the sampling procedure on different workers and different levels are independent.

□

Similarly, we can prove the following lemma regarding the stochastic gradient on each device.

Lemma 6. *Given Assumptions 2-4, we can get*

$$
\mathbb{E}\Big[\Big\|g_{n,t}-\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big]\le K\sum_{k=1}^{K}B_k\sigma_k^2\ , \qquad (30)
$$

where $B_k = \frac{\prod_{j=1}^{K} C_j^2}{C_k^2}$.

Lemma 7. *For* $k \in \{1, \cdots, K-1\}$*, given Assumptions 2-4 and* $\eta \leq \frac{1}{\beta}$*, we can get*

$$\begin{aligned}\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] &\leq (1-\beta\eta)\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] \\ &+ 2C_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] + 2\beta^2\eta^2\delta_k^2 \ .\end{aligned} \tag{31}$$

Proof.

$$\begin{aligned}&\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\ &= \mathbb{E}[\|(1-\beta\eta)(u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)})) + f_n^{(k)}(u_{n,t}^{(k-1)};\xi_{n,t}^{(k)}) - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\ &= \mathbb{E}[\|(1-\beta\eta)(u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})) \\ &\quad + (f_n^{(k)}(u_{n,t-1}^{(k-1)}) - f_n^{(k)}(u_{n,t}^{(k-1)}) - f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) + f_n^{(k)}(u_{n,t}^{(k-1)};\xi_{n,t}^{(k)})) \\ &\quad + \beta\eta(f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) - f_n^{(k)}(u_{n,t-1}^{(k-1)}))\|^2] \\ &= \mathbb{E}[\|(1-\beta\eta)(u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)}))\|^2] \\ &\quad + \mathbb{E}[\|(f_n^{(k)}(u_{n,t-1}^{(k-1)}) - f_n^{(k)}(u_{n,t}^{(k-1)}) - f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) + f_n^{(k)}(u_{n,t}^{(k-1)};\xi_{n,t}^{(k)})) \\ &\quad + \beta\eta(f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) - f_n^{(k)}(u_{n,t-1}^{(k-1)}))\|^2] \\ &\leq (1-\beta\eta)^2\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] \\ &\quad + 2\mathbb{E}[\|f_n^{(k)}(u_{n,t-1}^{(k-1)}) - f_n^{(k)}(u_{n,t}^{(k-1)}) - f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) + f_n^{(k)}(u_{n,t}^{(k-1)};\xi_{n,t}^{(k)})\|^2] \\ &\quad + 2\beta^2\eta^2\mathbb{E}[\|f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] \\ &\leq (1-\beta\eta)^2\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] \\ &\quad + 2\mathbb{E}[\|f_n^{(k)}(u_{n,t-1}^{(k-1)};\xi_{n,t}^{(k)}) - f_n^{(k)}(u_{n,t}^{(k-1)};\xi_{n,t}^{(k)})\|^2] + 2\beta^2\eta^2\delta_k^2 \\ &\leq (1-\beta\eta)\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 2C_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] + 2\beta^2\eta^2\delta_k^2 \ ,\end{aligned} \tag{32}$$

where the second to last step holds due to Assumption 4, the last step holds due to Assumption 3. □

Lemma 8. *Given Assumptions 2-4, if* $\mu\eta \in (0,1)$*, we can get*

$$\begin{aligned}&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\ &\leq (1-\mu\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t-1})\Big\|^2\Big] \\ &\quad + 2\mu\eta K\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1} A_k\mathbb{E}\Big[\Big\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\Big\|^2\Big] \\ &\quad + 4\mu\eta K\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1} A_k C_k^2\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\Big\|^2\Big] \\ &\quad + \frac{2L_F^2}{\mu\eta}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t-1}\Big\|^2\Big] + 4\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \mu^2\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} \ .\end{aligned} \tag{33}$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big((1-\mu\eta)(m_{n,t-1} - \nabla F_n(x_{n,t-1})) + (1-\mu\eta)(\nabla F_n(x_{n,t-1}) - \nabla F_n(x_{n,t})) \\
&\quad + \mu\eta\Big(g_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad + \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\Big)\Big)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big((1-\mu\eta)(m_{n,t-1} - \nabla F_n(x_{n,t-1})) + (1-\mu\eta)(\nabla F_n(x_{n,t-1}) - \nabla F_n(x_{n,t})) \\
&\quad + \mu\eta\Big(\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\Big)\Big)\Big\|^2\Big] \\
&\quad + \mu^2\eta^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(g_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big)\Big\|^2\Big] \\
&\le (1-\mu\eta)^2(1+a)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t-1})\Big\|^2\Big] \\
&\quad + 2(1+a^{-1})(1-\mu\eta)^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(\nabla F_n(x_{n,t-1}) - \nabla F_n(x_{n,t}))\Big\|^2\Big] \\
&\quad + 2(1+a^{-1})\mu^2\eta^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\Big)\Big\|^2\Big] \\
&\quad + \mu^2\eta^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(g_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big)\Big\|^2\Big] \\
&\le (1-\mu\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t-1})\Big\|^2\Big] + \frac{2L_F^2}{\mu\eta}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t-1}\Big\|^2\Big] \\
&\quad + 2\mu\eta\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\qquad - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\Big)\Big\|^2\Big] \\
&\quad + \mu^2\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} \\
&\le (1-\mu\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t-1})\Big\|^2\Big] + \frac{2L_F^2}{\mu\eta}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t-1}\Big\|^2\Big] \\
&\quad + 2\mu\eta K\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1} A_k\mathbb{E}\Big[\Big\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\Big\|^2\Big] + \mu^2\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N}\ ,
\end{aligned}
\tag{34}
$$

where the second to last step holds due to $a = \frac{\mu\eta}{1-\mu\eta}$ and Lemma 5, the last step holds due to Lemma 3. Then, by combining it with Lemma 7, we complete the proof.

□

Lemma 9. *Given Assumptions 1-4, we can get*

$$\mathbb{E}[\|X_{t+1}-\bar{X}_{t+1}\|_F^2] \leq (1-\eta\frac{1-\lambda^2}{2})\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \frac{2\eta\alpha^2}{1-\lambda^2}\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] \ . \tag{35}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|X_{t+1}-\bar{X}_{t+1}\|_F^2] \\
&= \mathbb{E}[\|X_t+\eta(X_{t+\frac{1}{2}}-X_t)-\bar{X}_t-\eta(\bar{X}_{t+\frac{1}{2}}-\bar{X}_t)\|_F^2] \\
&\leq (1-\eta)^2(1+a)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta^2(1+a^{-1})\mathbb{E}[\|X_{t+\frac{1}{2}}-\bar{X}_{t+\frac{1}{2}}\|_F^2] \\
&\leq (1-\eta)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta\mathbb{E}[\|X_{t+\frac{1}{2}}-\bar{X}_{t+\frac{1}{2}}\|_F^2] \\
&\leq (1-\eta)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta\mathbb{E}[\|X_tW-\alpha Y_t-\bar{X}_t+\alpha\bar{Y}_t\|_F^2] \\
&\leq (1-\eta)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta(1+a)\mathbb{E}[\|X_tW-\bar{X}_t\|_F^2] + \eta\alpha^2(1+a^{-1})\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] \\
&\leq (1-\eta)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta\lambda^2(1+a)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta\alpha^2(1+a^{-1})\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] \\
&\leq (1-\eta)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \eta\frac{1+\lambda^2}{2}\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \frac{\eta\alpha^2(1+\lambda^2)}{1-\lambda^2}\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] \\
&\leq \Big(1-\eta\frac{1-\lambda^2}{2}\Big)\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + \frac{2\eta\alpha^2}{1-\lambda^2}\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] \ ,
\end{aligned} \tag{36}$$

where the third step holds due to $a=\frac{1-\lambda}{\lambda}$, the last step holds due to $a=\frac{1-\lambda^2}{2\lambda^2}$. □

Lemma 10. *Given Assumptions 1-4, we can get*

$$\mathbb{E}[\|X_{t+1}-X_t\|_F^2] \leq 8\eta^2\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + 4\alpha^2\eta^2\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] + 4\alpha^2\eta^2\mathbb{E}[\|\bar{M}_t\|_F^2] \ . \tag{37}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|X_{t+1}-X_t\|_F^2] \\
&= \mathbb{E}[\|X_t+\eta(X_{t+\frac{1}{2}}-X_t)-X_t\|_F^2] \\
&= \eta^2\mathbb{E}[\|X_{t+\frac{1}{2}}-X_t\|_F^2] \\
&= \eta^2\mathbb{E}[\|X_tW-\alpha Y_t-X_t\|_F^2] \\
&\leq 2\eta^2\mathbb{E}[\|X_tW-X_t\|_F^2] + 2\alpha^2\eta^2\mathbb{E}[\|Y_t\|_F^2] \\
&\leq 2\eta^2\mathbb{E}[\|(X_t-\bar{X}_t)(W-I)\|_F^2] + 2\alpha^2\eta^2\mathbb{E}[\|Y_t-\bar{Y}_t+\bar{Y}_t\|_F^2] \\
&\leq 8\eta^2\mathbb{E}[\|X_t-\bar{X}_t\|_F^2] + 4\alpha^2\eta^2\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] + 4\alpha^2\eta^2\mathbb{E}[\|\bar{M}_t\|_F^2] \ .
\end{aligned} \tag{38}$$

□

Lemma 11. *Given Assumptions 1-4, we can get*

$$\begin{aligned}
\mathbb{E}[\|Y_{t+1}-\bar{Y}_{t+1}\|_F^2] &\leq \lambda\mathbb{E}[\|Y_t-\bar{Y}_t\|_F^2] + \frac{4\mu^2\eta^2}{1-\lambda}\sum_{n=1}^N\mathbb{E}[\|m_{n,t}-\nabla F_n(x_{n,t})\|^2] \\
&+ \sum_{n=1}^N\sum_{k=1}^{K-1}\frac{\mu^2\eta^2}{1-\lambda}K(4A_k+8\beta^2\eta^2D_k)\mathbb{E}[\|u_{n,t}^{(k)}-f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&+ \frac{4\mu^2\eta^2}{1-\lambda}KD_0\sum_{n=1}^N\mathbb{E}[\|x_{n,t+1}-x_{n,t}\|^2] + \frac{8\mu^2\eta^2}{1-\lambda}K\sum_{n=1}^N\sum_{k=1}^{K-1}D_kC_k^2\mathbb{E}[\|u_{n,t}^{(k-1)}-u_{n,t+1}^{(k-1)}\|^2] \\
&+ \frac{8\beta^2\mu^2\eta^4}{1-\lambda}KN\sum_{k=1}^{K-1}D_k\delta_k^2 + \frac{4\mu^2\eta^2}{1-\lambda}KN\sum_{k=1}^K B_k\sigma_k^2 \ .
\end{aligned} \tag{39}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] = \mathbb{E}[\|Y_t W + M_{t+1} - M_t - \bar{Y}_t - \bar{M}_{t+1} + \bar{M}_t\|_F^2] \\
&\leq (1+a)\mathbb{E}[\|Y_t W - \bar{Y}_t\|_F^2] + (1+a^{-1})\mathbb{E}[\|M_{t+1} - M_t - \bar{M}_{t+1} + \bar{M}_t\|_F^2] \\
&\leq (1+a)\lambda^2\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + (1+a^{-1})\mathbb{E}[\|M_{t+1} - M_t\|_F^2] \\
&\leq \lambda\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{1}{1-\lambda}\mathbb{E}[\|(1-\mu\eta)M_t + \mu\eta G_{t+1} - M_t\|_F^2] \\
&= \lambda\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{\mu^2\eta^2}{1-\lambda}\mathbb{E}[\|M_t - G_{t+1}\|_F^2] \ ,
\end{aligned} \tag{40}$$

where the second inequality holds due to $a = \frac{1-\lambda}{\lambda}$. Furthermore, by combining it with the following inequality, we can complete the proof.

$$\begin{aligned}
&\mathbb{E}[\|M_t - G_{t+1}\|_F^2] \leq 4\sum_{n=1}^N \mathbb{E}[\|m_{n,t} - \nabla F_n(x_{n,t})\|^2] \\
&+ 4\sum_{n=1}^N \mathbb{E}[\|\nabla F_n(x_{n,t}) - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2] \\
&+ 4\sum_{n=1}^N \mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\
&\quad - \nabla f_n^{(1)}(u_{n,t+1}^{(0)})\nabla f_n^{(2)}(u_{n,t+1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t+1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t+1}^{(K-1)})\|^2] \\
&+ 4\sum_{n=1}^N \mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t+1}^{(0)})\nabla f_n^{(2)}(u_{n,t+1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t+1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t+1}^{(K-1)}) - g_{n,t+1}\|^2] \\
&\leq 4\sum_{n=1}^N \mathbb{E}[\|m_{n,t} - \nabla F_n(x_{n,t})\|^2] + 4K\sum_{n=1}^N\sum_{k=1}^{K-1} A_k\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&+ 4KD_0\sum_{n=1}^N \mathbb{E}[\|x_{n,t+1} - x_{n,t}\|^2] + 8K\sum_{n=1}^N\sum_{k=1}^{K-1} D_k C_k^2\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] \\
&+ 8\beta^2\eta^2 K\sum_{n=1}^N\sum_{k=1}^{K-1} D_k\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] + 8\beta^2\eta^2 KN\sum_{k=1}^{K-1} D_k\delta_k^2 + 4KN\sum_{k=1}^K B_k\sigma_k^2 \\
&\leq 4\sum_{n=1}^N \mathbb{E}[\|m_{n,t} - \nabla F_n(x_{n,t})\|^2] + K\sum_{n=1}^N\sum_{k=1}^{K-1}(4A_k + 8\beta^2\eta^2 D_k)\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] + 8\beta^2\eta^2 KN\sum_{k=1}^{K-1} D_k\delta_k^2 \\
&+ 4KD_0\sum_{n=1}^N \mathbb{E}[\|x_{n,t+1} - x_{n,t}\|^2] + 8K\sum_{n=1}^N\sum_{k=1}^{K-1} D_k C_k^2\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] + 4KN\sum_{k=1}^K B_k\sigma_k^2 \ ,
\end{aligned} \tag{41}$$

where the third step holds due to Lemma 3, Lemma 4, and Lemma 5.

□

By following the proof of Lemma 8, it is easy to prove the following lemma.

Lemma 12. *Given Assumptions 2-4, if $\mu\eta \in (0,1)$, we can get*

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^N \mathbb{E}\Big[\Big\|m_{n,t+1} - \nabla F_n(x_{n,t+1})\Big\|^2\Big] \leq (1-\mu\eta)\frac{1}{N}\sum_{n=1}^N \mathbb{E}\Big[\Big\|m_{n,t} - \nabla F_n(x_{n,t})\Big\|^2\Big] \\
&+ 2\mu\eta K\frac{1}{N}\sum_{n=1}^N\sum_{k=1}^{K-1} A_k\mathbb{E}\Big[\Big\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\Big\|^2\Big] + 4\mu\eta K\frac{1}{N}\sum_{n=1}^N\sum_{k=1}^{K-1} A_k C_k^2\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\Big\|^2\Big] \\
&+ \frac{2L_F^2}{\mu\eta}\frac{1}{N}\sum_{n=1}^N \mathbb{E}\Big[\Big\|x_{n,t+1} - x_{n,t}\Big\|^2\Big] + 4\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \mu^2\eta^2 K\sum_{k=1}^K B_k\sigma_k^2 \ .
\end{aligned} \tag{42}$$

Based on these lemmas, we prove Theorem 1 below.

Proof.

$$
\begin{aligned}
&F\left(\bar{x}_{t+1}\right) \leq F(\bar{x}_t) + \langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L_F}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&= F(\bar{x}_t) - \alpha\eta \langle \nabla F(\bar{x}_t), \bar{m}_t \rangle + \frac{\alpha^2\eta^2 L_F}{2}\|\bar{m}_t\|^2 \\
&= F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - (\frac{\alpha\eta}{2} - \frac{\alpha^2\eta^2 L_F}{2})\|\bar{m}_t\|^2 + \frac{\alpha\eta}{2}\|\bar{m}_t - \nabla F(\bar{x}_t)\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \frac{\alpha\eta}{2}\|\bar{m}_t - \nabla F(\bar{x}_t)\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \alpha\eta\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\|^2 + \alpha\eta\|\frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t}) - \nabla F(\bar{x}_t)\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \alpha\eta\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\|^2 + \alpha\eta L_F^2\frac{1}{N}\sum_{n=1}^{N}\|x_{n,t} - \bar{x}_t\|^2 \ ,
\end{aligned}
\tag{43}
$$

where the fourth step holds due to $\eta \leq \frac{1}{2\alpha L_F}$.

To prove Theorem 1, we introduce the following potential function:

$$
\begin{aligned}
&\mathcal{H}_{t+1} = \mathbb{E}[F(\bar{x}_{t+1})] + \omega_0\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t+1} - \nabla F_n(x_{n,t+1})\Big\|^2\Big] + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\mathbb{E}[\|u_{n,t+1}^{(k)} - f_n^{(k)}(u_{n,t+1}^{(k-1)})\|^2] \\
&+ \omega_K\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,t+1} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t+1})\Big\|^2\Big] + \omega_{K+1}\frac{1}{N}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] + \omega_{K+2}\frac{1}{N}\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \ .
\end{aligned}
\tag{44}
$$

Then, based on Lemmas 7, 8, 12, 9, 11, we can get

$$
\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\alpha\eta}{4}\mathbb{E}[\|\bar{m}_t\|^2] + 4\omega_K\mu\beta^2\eta^3 K\sum_{k=1}^{K-1}A_k\delta_k^2 + \omega_K\mu^2\eta^2 K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N} + 2\beta^2\eta^2\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&+ \omega_{K+2}\frac{8\beta^2\mu^2\eta^4}{1-\lambda}K\sum_{k=1}^{K-1}D_k\delta_k^2 + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}K\sum_{k=1}^{K}B_k\sigma_k^2 + 4\omega_0\mu\beta^2\eta^3 K\sum_{k=1}^{K-1}A_k\delta_k^2 + \omega_0\mu^2\eta^2 K\sum_{k=1}^{K}B_k\sigma_k^2 \\
&+ \Big(\omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda} - \mu\eta\omega_0\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla F_n(x_{n,t})\Big\|^2\Big] \\
&+ \Big(\alpha\eta - \mu\eta\omega_K\Big)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(\frac{\mu^2\eta^2}{1-\lambda}K(4A_k + 8\beta^2\eta^2 D_k)\omega_{K+2} + 2\omega_K\mu\eta K A_k + 2\omega_0\mu\eta K A_k - \omega_k\beta\eta\Big)\mathbb{E}\Big[\Big\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\Big\|^2\Big] \\
&+ \Big(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0\Big)\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&+ \Big(\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+1}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \Big(\frac{2\eta\alpha^2}{1-\lambda^2}\omega_{K+1} - (1-\lambda)\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_kC_k^2\Big)\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] \ .
\end{aligned}
\tag{45}
$$

Then, according to Lemma 2, we can get

$$
\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\alpha\eta}{4}\mathbb{E}[\|\bar{m}_t\|^2] + 4\omega_K\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \omega_K\mu^2\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 2\beta^2\eta^2\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&\quad + \omega_{K+2}\frac{8\beta^2\mu^2\eta^4}{1-\lambda}K\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}K\sum_{k=1}^{K} B_k\sigma_k^2 + 4\omega_0\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \omega_0\mu^2\eta^2 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + \Big(\omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda} - \mu\eta\omega_0\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla F_n(x_{n,t})\Big\|^2\Big] + \Big(\alpha\eta - \mu\eta\omega_K\Big)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(\frac{\mu^2\eta^2}{1-\lambda}K(4A_k + 8\beta^2\eta^2 D_k)\omega_{K+2} + 2\omega_K\mu\eta K A_k + 2\omega_0\mu\eta K A_k - \omega_k\beta\eta\Big)\mathbb{E}\Big[\Big\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + \Big(\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+1}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \Big(\frac{2\eta\alpha^2}{1-\lambda^2}\omega_{K+1} - (1-\lambda)\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&\quad + \Big(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0\Big)\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_k C_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\mathbb{E}[\|u_{n,t}^{(0)} - u_{n,t+1}^{(0)}\|^2] \\
&\quad + 2\beta^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 \\
&\qquad + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_k C_k^2\Big)\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\mathbb{E}[\|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\|^2] \\
&\quad + 2\beta^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_k C_k^2\Big)\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\delta_j^2 \ . \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\alpha\eta}{4}\mathbb{E}[\|\bar{m}_t\|^2] + 4\omega_K\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \omega_K\mu^2\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 2\beta^2\eta^2\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&\quad + \omega_{K+2}\frac{8\beta^2\mu^2\eta^4}{1-\lambda}K\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}K\sum_{k=1}^{K} B_k\sigma_k^2 + 4\omega_0\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \omega_0\mu^2\eta^2 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + \Big(\omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda} - \mu\eta\omega_0\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla F_n(x_{n,t})\Big\|^2\Big] + \Big(\alpha\eta - \mu\eta\omega_K\Big)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(\frac{\mu^2\eta^2}{1-\lambda}K(4A_k + 8\beta^2\eta^2 D_k)\omega_{K+2} + 2\omega_K\mu\eta K A_k + 2\omega_0\mu\eta K A_k - \omega_k\beta\eta\Big)\mathbb{E}\Big[\Big\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + \Big(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0 \\
&\qquad + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_k C_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Big)\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&\quad + \Big(\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+1}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \Big(\frac{2\eta\alpha^2}{1-\lambda^2}\omega_{K+1} - (1-\lambda)\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&\quad + 2\beta^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Big(4\omega_K\mu\eta K A_j C_j^2 + 4\omega_0\mu\eta K A_j C_j^2 + 2\omega_j C_j^2 \\
&\qquad + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_j C_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&\quad + 2\beta^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Big(4\omega_K\mu\eta K A_j C_j^2 + 4\omega_0\mu\eta K A_j C_j^2 + 2\omega_j C_j^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_j C_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2 \ .
\end{aligned}
\tag{46}
$$

Based on Lemma 10, we can get

$$
\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + 4\omega_K\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \omega_K\mu^2\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 2\beta^2\eta^2\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&\quad + \omega_{K+2}\frac{8\beta^2\mu^2\eta^4}{1-\lambda}K\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}K\sum_{k=1}^{K} B_k\sigma_k^2 + 4\omega_0\mu\beta^2\eta^3 K\sum_{k=1}^{K-1} A_k\delta_k^2 + \omega_0\mu^2\eta^2 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + 2\beta^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\left[\sum_{j=k+1}^{K-1}\Big(4\omega_K\mu\eta K A_j C_j^2 + 4\omega_0\mu\eta K A_j C_j^2 + 2\omega_j C_j^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}K D_j C_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]\delta_k^2 \\
&\quad + \Big(\omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda} - \mu\eta\omega_0\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla F_n(x_{n,t})\Big\|^2\Big] \\
&\quad + \Big(\alpha\eta - \mu\eta\omega_K\Big)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\Big\|^2\Big] \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\left[\frac{\mu^2\eta^2}{1-\lambda}K(4A_k + 8\beta^2\eta^2 D_k)\omega_{K+2} + 2\omega_K\mu\eta K A_k + 2\omega_0\mu\eta K A_k - \omega_k\beta\eta + 2\beta^2\eta^2\sum_{j=k+1}^{K-1}\Big(2\omega_j C_j^2\right. \\
&\qquad \left. + 4\omega_K\mu\eta K A_j C_j^2 + 4\omega_0\mu\eta K A_j C_j^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}K D_j C_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]\mathbb{E}\Big[\Big\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + \left[\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+1} + 8\eta^2\left(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0\right.\right. \\
&\qquad \left.\left. + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}K D_k C_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\right)\right]\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] \\
&\quad + \left[\frac{2\eta\alpha^2}{1-\lambda^2}\omega_{K+1} - (1-\lambda)\omega_{K+2} + 4\alpha^2\eta^2\left(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0\right.\right. \\
&\qquad \left.\left. + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}K D_k C_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\right)\right]\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&\quad + \left[4\alpha^2\eta^2\left(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0\right.\right. \\
&\qquad \left.\left. + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta K A_k C_k^2 + 4\omega_0\mu\eta K A_k C_k^2 + 2\omega_k C_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}K D_k C_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\right) - \frac{\alpha\eta}{4}\right]\mathbb{E}[\|\bar{m}_t\|^2]\ .
\end{aligned}
\tag{47}
$$

In the following, we enforce the coefficient of the last six terms to be non-positive. Specifically, by setting $\omega_K = \frac{\alpha}{\mu}$, we can get $\alpha\eta - \mu\eta\omega_K \leq 0$. Moreover, we set $\omega_{K+2} = \alpha(1-\lambda)$ and $\omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda} - \mu\eta\omega_0 = 0$ so that we can get $\omega_0 = \omega_{K+2}\frac{4\mu\eta}{1-\lambda} = 4\alpha\mu\eta$.

Then, we enforce

$$
\begin{aligned}
&\frac{\mu^2\eta^2}{1-\lambda}K(4A_k + 8\beta^2\eta^2 D_k)\omega_{K+2} + 2\omega_K\mu\eta K A_k + 2\omega_0\mu\eta K A_k - \omega_k\beta\eta \\
&\quad + 2\beta^2\eta^2\sum_{j=k+1}^{K-1}\Big(4\omega_K\mu\eta K A_j C_j^2 + 4\omega_0\mu\eta K A_j C_j^2 + 2\omega_j C_j^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}K D_j C_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big) \leq 0\ .
\end{aligned}
\tag{48}
$$

This is equivalent to enforce

$$\begin{aligned}
&\frac{\mu^2\eta^2}{1-\lambda}K(4A_k+8\beta^2\eta^2D_k)\omega_{K+2}+2\frac{\alpha}{\mu}\mu\eta KA_k+2\omega_{K+2}\frac{4\mu\eta}{1-\lambda}\mu\eta KA_k-\omega_k\beta\eta\\
&+2\beta^2\eta^2\sum_{j=k+1}^{K-1}\Big(4\frac{\alpha}{\mu}\mu\eta KA_jC_j^2+4\omega_{K+2}\frac{4\mu\eta}{1-\lambda}\mu\eta KA_jC_j^2+2\omega_jC_j^2+\omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\\
&\le\frac{\mu^2\eta^2}{1-\lambda}K(4A_k+8\beta^2\eta^2D_k)\alpha(1-\lambda)+2\frac{\alpha}{\mu}\mu\eta KA_k+2\alpha(1-\lambda)\frac{4\mu\eta}{1-\lambda}\mu\eta KA_k-\omega_k\beta\eta\\
&+2\beta^2\eta^2\sum_{j=k+1}^{K-1}\Big(4\frac{\alpha}{\mu}\mu\eta KA_jC_j^2+4\alpha(1-\lambda)\frac{4\mu\eta}{1-\lambda}\mu\eta KA_jC_j^2+2\omega_jC_j^2+\alpha(1-\lambda)\frac{8\mu^2\eta^2}{1-\lambda}KD_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\\
&\le K(4A_k+8\beta^2\eta^2D_k)\alpha\mu^2\eta^2+2\alpha\eta KA_k+8\alpha\mu^2\eta^2KA_k-\omega_k\beta\eta\\
&+2\beta^2\eta^2\sum_{j=k+1}^{K-1}\Big(4\alpha\eta KA_jC_j^2+16\alpha\mu^2\eta^2KA_jC_j^2+2\omega_jC_j^2+8\alpha\mu^2\eta^2KD_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\le 0\ .
\end{aligned}\tag{49}$$

It can be done by enforcing

$$\begin{aligned}
&2\beta^2\eta^2\sum_{j=k+1}^{K-1}\Big(2\omega_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)-\omega_k\beta\eta\le-\frac{1}{2}\omega_k\beta\eta\ ,\\
&K(4A_k+8\beta^2\eta^2D_k)\alpha\mu^2\eta+2\alpha KA_k+8\alpha\mu^2\eta KA_k\\
&\quad+2\beta^2\eta\sum_{j=k+1}^{K-1}\Big(4\alpha\eta KA_jC_j^2+16\alpha\mu^2\eta^2KA_jC_j^2+8\alpha\mu^2\eta^2KD_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\le\frac{1}{2}\omega_k\beta\ .
\end{aligned}\tag{50}$$

As for the first inequality, we can get

$$\eta\le\frac{\omega_k}{4\beta\sum_{j=1}^{K-1}\Big(2\omega_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)}\ .\tag{51}$$

As for the second inequality, we can get

$$\begin{aligned}
\frac{1}{2}\omega_k\beta&\ge K(4A_k+8\beta^2\eta^2D_k)\alpha\mu^2\eta+2\alpha KA_k+8\alpha\mu^2\eta KA_k\\
&+2\beta^2\eta\sum_{j=k+1}^{K-1}\Big(4\alpha\eta KA_jC_j^2+16\alpha\mu^2\eta^2KA_jC_j^2+8\alpha\mu^2\eta^2KD_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\ .
\end{aligned}\tag{52}$$

Then, due to $\beta\eta<1$, $\mu\eta<1$, we can set

$$\omega_k=\frac{2\alpha K}{\beta}\Bigg((12A_k+8D_k)\mu+2A_k+2\beta\sum_{j=k+1}^{K-1}\Big(20A_jC_j^2+8D_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg)\ .\tag{53}$$

Here, we represent $\omega_k\triangleq\alpha K\tilde{\omega}_k$, where $\tilde{\omega}_k=\frac{2}{\beta}\Bigg((12A_k+8D_k)\mu+2A_k+2\beta\sum_{j=k+1}^{K-1}\Big(20A_jC_j^2+8D_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg)$. Then, we can simplify the upper bound of η as follows:

$$\eta\le\frac{\tilde{\omega}_k}{8\beta\sum_{j=1}^{K-1}\tilde{\omega}_jC_j^2\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)}\ .\tag{54}$$

Based on the value of ω_k where $k \in \{0, 1, \cdots, K\}$ and ω_{K+2} , due to $\eta < 1$, we can get

$$\begin{aligned}
&\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0 \\
&\quad + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta KA_kC_k^2 + 4\omega_0\mu\eta KA_kC_k^2 + 2\omega_kC_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) \\
&= \frac{2\alpha L_F^2}{\mu^2\eta} + 8\alpha L_F^2 + 4\alpha KD_0 + \sum_{k=1}^{K-1}\Big(4\alpha\eta KA_kC_k^2 + 16\alpha KA_kC_k^2 + 2\alpha K\tilde{\omega}_kC_k^2 + 8\alpha KD_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) \\
&\leq \frac{2\alpha L_F^2}{\mu^2\eta} + 8\alpha L_F^2 + 4\alpha KD_0 + \alpha K\sum_{k=1}^{K-1}\Big(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) .
\end{aligned} \tag{55}$$

Furthermore, we enforce

$$\begin{aligned}
&\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+1} + 8\eta^2\Bigg(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0 \\
&\quad + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta KA_kC_k^2 + 4\omega_0\mu\eta KA_kC_k^2 + 2\omega_kC_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg) \\
&\leq \alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+1} + 8\eta^2\Bigg(\frac{2\alpha L_F^2}{\mu^2\eta} + 8\alpha L_F^2 + 4\alpha KD_0 + \alpha K\sum_{k=1}^{K-1}\Big(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg) \\
&\leq 0 .
\end{aligned} \tag{56}$$

Similarly, due to $\eta < 1$, we can set

$$\omega_{K+1} = \frac{2\alpha}{(1-\lambda^2)}\Bigg[L_F^2 + 8\Bigg(\frac{2L_F^2}{\mu^2} + 8L_F^2 + 4KD_0 + K\sum_{k=1}^{K-1}\Big(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg)\Bigg] . \tag{57}$$

Here, we represent $\omega_{K+1} \triangleq \frac{2\alpha}{(1-\lambda^2)}\tilde{\omega}_{K+1}$, where $\tilde{\omega}_{K+1} = \Bigg[L_F^2 + 8\Bigg(\frac{2L_F^2}{\mu^2} + 8L_F^2 + 4KD_0 + K\sum_{k=1}^{K-1}\Big(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg)\Bigg]$.

In addition, we enforce

$$\begin{aligned}
&\frac{2\eta\alpha^2}{1-\lambda^2}\omega_{K+1} - (1-\lambda)\omega_{K+2} + 4\alpha^2\eta^2\Bigg(\frac{2L_F^2}{\mu\eta}\omega_K + \omega_0\frac{2L_F^2}{\mu\eta} + \omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0 \\
&\quad + \sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta KA_kC_k^2 + 4\omega_0\mu\eta KA_kC_k^2 + 2\omega_kC_k^2 + \omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg) \\
&\leq \frac{2\eta\alpha^2}{1-\lambda^2}\frac{2\alpha}{(1-\lambda^2)}\tilde{\omega}_{K+1} - \alpha(1-\lambda)^2 \\
&\quad + 4\alpha^2\eta^2\Bigg(\frac{2\alpha L_F^2}{\mu^2\eta} + 8\alpha L_F^2 + 4\alpha KD_0 + \alpha K\sum_{k=1}^{K-1}\Big(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg) \leq 0 .
\end{aligned} \tag{58}$$

Due to $\eta < 1$ and $1 + \lambda > 1$, we can get

$$\alpha \leq \frac{(1-\lambda)^2}{\sqrt{4\tilde{\omega}_{K+1} + 8L_F^2/\mu^2 + 32L_F^2 + 16KD_0 + 4K\sum_{k=1}^{K-1}\Big(20A_kC_k^2 + 2\tilde{\omega}_kC_k^2 + 8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)}} . \tag{59}$$

And we enforce

$$\begin{aligned}
&4\alpha^2\eta^2\Bigg(\frac{2L_F^2}{\mu\eta}\omega_K+\omega_0\frac{2L_F^2}{\mu\eta}+\omega_{K+2}\frac{4\mu^2\eta^2}{1-\lambda}KD_0\\
&\quad+\sum_{k=1}^{K-1}\Big(4\omega_K\mu\eta KA_kC_k^2+4\omega_0\mu\eta KA_kC_k^2+2\omega_kC_k^2+\omega_{K+2}\frac{8\mu^2\eta^2}{1-\lambda}KD_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg)-\frac{\alpha\eta}{4}\\
&\leq 4\alpha^2\eta^2\Bigg(\frac{2\alpha L_F^2}{\mu^2\eta}+8\alpha L_F^2+4\alpha KD_0+\alpha K\sum_{k=1}^{K-1}\Big(20A_kC_k^2+2\tilde{\omega}_kC_k^2+8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg)-\frac{\alpha\eta}{4}\leq 0\ .
\end{aligned}\tag{60}$$

Similarly, due to $\eta<1$, we can get

$$\alpha\leq\frac{1}{4\sqrt{2L_F^2/\mu^2+8L_F^2+4KD_0+K\sum_{k=1}^{K-1}\Big(20A_kC_k^2+2\tilde{\omega}_kC_k^2+8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)}}\ .\tag{61}$$

In summary, by setting

$$\begin{aligned}
&\omega_0=4\alpha\mu\eta\ ,\\
&\omega_k=\alpha K\tilde{\omega}_k\ ,\forall k\in\{1,2,\cdots,K-1\}\ ,\\
&\omega_K=\frac{\alpha}{\mu}\ ,\\
&\omega_{K+1}=\frac{2\alpha}{1-\lambda^2}\tilde{\omega}_{K+1}\ ,\\
&\omega_{K+2}=\alpha(1-\lambda)\ ,\\
&\eta\leq\frac{\tilde{\omega}_k}{8\beta\sum_{j=1}^{K-1}\tilde{\omega}_jC_j^2\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)}\ ,\\
&\alpha\leq\frac{(1-\lambda)^2}{\sqrt{4\tilde{\omega}_{K+1}+8L_F^2/\mu^2+32L_F^2+16KD_0+4K\sum_{k=1}^{K-1}\Big(20A_kC_k^2+2\tilde{\omega}_kC_k^2+8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)}}\ ,\\
&\alpha\leq\frac{1}{4\sqrt{2L_F^2/\mu^2+8L_F^2+4KD_0+K\sum_{k=1}^{K-1}\Big(20A_kC_k^2+2\tilde{\omega}_kC_k^2+8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)}}\ ,
\end{aligned}\tag{62}$$

where $\tilde{\omega}_k=\frac{2}{\beta}\Bigg((12A_k+8D_k)\mu+2A_k+2\beta\sum_{j=k+1}^{K-1}\Big(20A_jC_j^2+8D_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg)$ and $\tilde{\omega}_{K+1}=L_F^2+8\Bigg(\frac{2L_F^2}{\mu^2}+8L_F^2+4KD_0+K\sum_{k=1}^{K-1}\Big(20A_kC_k^2+2\tilde{\omega}_kC_k^2+8D_kC_k^2\Big)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\Bigg)$, we can get

$$\begin{aligned}
&\mathcal{H}_{t+1}-\mathcal{H}_t\\
&\leq-\frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2+8\alpha\beta^2\mu^2\eta^4K\sum_{k=1}^{K-1}D_k\delta_k^2+4\alpha\mu^2\eta^2K\sum_{k=1}^{K}B_k\sigma_k^2+16\alpha\mu^2\beta^2\eta^4K\sum_{k=1}^{K-1}A_k\delta_k^2+4\alpha\mu^3\eta^3K\sum_{k=1}^{K}B_k\sigma_k^2\\
&\quad+4\alpha\beta^2\eta^3K\sum_{k=1}^{K-1}A_k\delta_k^2+\alpha\mu\eta^2K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+2\beta^2\eta^2\alpha K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2\\
&\quad+2\alpha\beta^2\eta^2K\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Big(20A_jC_j^2+2\tilde{\omega}_jC_j^2+8D_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2\ .
\end{aligned}\tag{63}$$

Then, it is easy to get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \\
&\leq \frac{2(\mathcal{H}_0-\mathcal{H}_T)}{\alpha\eta T}+16\beta^2\mu^2\eta^3K\sum_{k=1}^{K-1}D_k\delta_k^2+8\eta\mu^2K\sum_{k=1}^{K}B_k\sigma_k^2+32\mu^2\beta^2\eta^3K\sum_{k=1}^{K-1}A_k\delta_k^2+8\mu^3\eta^2K\sum_{k=1}^{K}B_k\sigma_k^2 \\
&\quad+8\beta^2\eta^2K\sum_{k=1}^{K-1}A_k\delta_k^2+2\mu\eta K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+4\eta\beta^2K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2 \\
&\quad+4\eta\beta^2K\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Big(20A_jC_j^2+2\tilde{\omega}_jC_j^2+8D_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2\ .
\end{aligned} \tag{64}$$

According to the initial value, we can get

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}[\|Y_0-\bar{Y}_0\|_F^2] \\
&=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|v_{n,0}^{(1)}v_{n,0}^{(2)}\cdots v_{n,0}^{(K-1)}v_{n,0}^{(K)}-\frac{1}{N}\sum_{n'=1}^{N}v_{n',0}^{(1)}v_{n',0}^{(2)}\cdots v_{n',0}^{(K-1)}v_{n',0}^{(K)}\|^2] \\
&=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|v_{n,0}^{(1)}v_{n,0}^{(2)}\cdots v_{n,0}^{(K-1)}v_{n,0}^{(K)}-\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)}) \\
&\quad+\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)}) \\
&\quad-\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\cdots\nabla f_{n'}^{(K-1)}(u_{n',0}^{(K-2)})\nabla f_{n'}^{(K)}(u_{n',0}^{(K-1)}) \\
&\quad+\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\cdots\nabla f_{n'}^{(K-1)}(u_{n',0}^{(K-2)})\nabla f_{n'}^{(K)}(u_{n',0}^{(K-1)}) \\
&\quad-\frac{1}{N}\sum_{n'=1}^{N}v_{n',0}^{(1)}v_{n',0}^{(2)}\cdots v_{n',0}^{(K-1)}v_{n',0}^{(K)}\|^2] \\
&\leq 3\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|v_{n,0}^{(1)}v_{n,0}^{(2)}\cdots v_{n,0}^{(K-1)}v_{n,0}^{(K)}-\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\|^2] \\
&\quad+3\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)}) \\
&\qquad-\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\cdots\nabla f_{n'}^{(K-1)}(u_{n',0}^{(K-2)})\nabla f_{n'}^{(K)}(u_{n',0}^{(K-1)})\|^2] \\
&\quad+3\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\cdots\nabla f_{n'}^{(K-1)}(u_{n',0}^{(K-2)})\nabla f_{n'}^{(K)}(u_{n',0}^{(K-1)}) \\
&\qquad-\frac{1}{N}\sum_{n'=1}^{N}v_{n',0}^{(1)}v_{n',0}^{(2)}\cdots v_{n',0}^{(K-1)}v_{n',0}^{(K)}\|^2] \\
&\leq 6K\sum_{k=1}^{K}B_k\sigma_k^2+12K\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K}C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1}8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\ ,
\end{aligned} \tag{65}$$

where the last step holds due to the following inequality:

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)}) \\
&\quad-\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\cdots\nabla f_{n'}^{(K-1)}(u_{n',0}^{(K-2)})\nabla f_{n'}^{(K)}(u_{n',0}^{(K-1)})\|^2] \\
&\le K\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\Big(\nabla f_n^{(1)}(u_{n,0}^{(0)})-\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\Big)\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)}) \\
&\quad+K\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\Big(\nabla f_n^{(2)}(u_{n,0}^{(1)})-\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\Big)\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\|^2] \\
&\quad+\cdots \\
&\quad+K\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\frac{1}{N}\sum_{n'=1}^{N}\nabla f_{n'}^{(1)}(u_{n',0}^{(0)})\nabla f_{n'}^{(2)}(u_{n',0}^{(1)})\cdots\nabla f_{n'}^{(K-1)}(u_{n',0}^{(K-2)})\Big(\nabla f_n^{(K)}(u_{n,0}^{(K-1)})-\nabla f_{n'}^{(K)}(u_{n',0}^{(K-1)})\Big)\|^2] \\
&\le 0+4K\sum_{k=2}^{K}\frac{1}{N}\sum_{n=1}^{N}\frac{(\prod_{j=1}^{K}C_j^2)L_k^2}{C_k^2}\mathbb{E}[\|u_{n,0}^{(k-1)}-\bar{u}_0^{(k-1)}\|^2] \\
&\le 4K\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K}C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1}8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\ ,
\end{aligned}
\tag{66}
$$

where the last step holds due to the following inequality:

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|u_{n,0}^{(k-1)}-\bar{u}_0^{(k-1)}\|^2] \\
&=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|f_n^{(k-1)}(u_{n,0}^{(k-2)};\xi_{n,t}^{(k-1)})-\frac{1}{N}\sum_{n'=1}^{N}f_{n'}^{(k-1)}(u_{n',0}^{(k-2)};\xi_{n',t}^{(k-1)})\|^2] \\
&=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|f_n^{(k-1)}(u_{n,0}^{(k-2)};\xi_{n,t}^{(k-1)})-f_n^{(k-1)}(u_{n,0}^{(k-2)})+f_n^{(k-1)}(u_{n,0}^{(k-2)})-f^{(k-1)}(\bar{u}_0^{(k-2)}) \\
&\quad+f^{(k-1)}(\bar{u}_0^{(k-2)})-\frac{1}{N}\sum_{n'=1}^{N}f_{n'}^{(k-1)}(u_{n',0}^{(k-2)})+\frac{1}{N}\sum_{n'=1}^{N}f_{n'}^{(k-1)}(u_{n',0}^{(k-2)})-\frac{1}{N}\sum_{n'=1}^{N}f_{n'}^{(k-1)}(u_{n',0}^{(k-2)};\xi_{n',t}^{(k-1)})\|^2] \\
&\le 8C_{k-1}^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|u_{n,0}^{(k-2)}-\bar{u}_0^{(k-2)}\|^2]+8\delta_{k-1}^2 \\
&\le\sum_{i=1}^{k-1}8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\ .
\end{aligned}
\tag{67}
$$

Moreover, we can get

$$
\mathbb{E}[\|u_{n,0}^{(k)}-f_n^{(k)}(u_{n,0}^{(k-1)})\|^2]=\mathbb{E}[\|f_n^{(k)}(u_{n,0}^{(k-1)};\xi_{n,0}^{(k)})-f_n^{(k)}(u_{n,0}^{(k-1)})\|^2]\le\delta_k^2\ , \tag{68}
$$

and

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,0}-\frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,0})\Big\|^2\Big]\\
&\leq 2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}v_{n,0}^{(1)}v_{n,0}^{(2)}\cdots v_{n,0}^{(K-1)}v_{n,0}^{(K)}-\frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\Big\|^2\Big]\\
&\quad+2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\\
&\qquad-\frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t}))\cdots\nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\Big\|^2\Big]\\
&\leq 2K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}A_k\mathbb{E}[\|u_{n,0}^{(k)}-f_n^{(k)}(u_{n,0}^{(k-1)})\|^2]\\
&\leq 2K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+2K\sum_{k=1}^{K-1}A_k\delta_k^2\ ,
\end{aligned}
\tag{69}
$$

as well as $\mathbb{E}\Big[\Big\|m_{n,0}-\nabla F_n(x_{n,0})\Big\|^2\Big]\leq 2K\sum_{k=1}^{K}B_k\sigma_k^2+2K\sum_{k=1}^{K-1}A_k\delta_k^2$. Then, we can get

$$
\begin{aligned}
\mathcal{H}_0&=F(x_0)+\omega_0\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,0}-\nabla F_n(x_{n,0})\Big\|^2\Big]+\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\mathbb{E}[\|u_{n,0}^{(k)}-f_n^{(k)}(u_{n,0}^{(k-1)})\|^2]\\
&\quad+\omega_K\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,0}-\frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,0})\Big\|^2\Big]+\omega_{K+1}\frac{1}{N}\mathbb{E}[\|X_0-\bar{X}_0\|_F^2]+\omega_{K+2}\frac{1}{N}\mathbb{E}[\|Y_0-\bar{Y}_0\|_F^2]\\
&\leq F(x_0)+8\alpha\mu\eta K(\sum_{k=1}^{K}B_k\sigma_k^2+\sum_{k=1}^{K-1}A_k\delta_k^2)+\alpha K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2\\
&\quad+\frac{2\alpha K}{\mu}(\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+\sum_{k=1}^{K-1}A_k\delta_k^2)+6\alpha K\Big(\sum_{k=1}^{K}B_k\sigma_k^2+2\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K}C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1}8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\Big)\ .
\end{aligned}
\tag{70}
$$

Finally, we can get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2]\\
&\leq\frac{2(F(x_0)-F(x_*))}{\alpha\eta T}+\frac{16\mu K}{T}(\sum_{k=1}^{K}B_k\sigma_k^2+\sum_{k=1}^{K-1}A_k\delta_k^2)+\frac{2K}{\eta T}\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2\\
&\quad+\frac{4K}{\mu\eta T}(\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+\sum_{k=1}^{K-1}A_k\delta_k^2)+\frac{12K}{\eta T}\Big(\sum_{k=1}^{K}B_k\sigma_k^2+2\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K}C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1}8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\Big)\\
&\quad+16\beta^2\mu^2\eta^3K\sum_{k=1}^{K-1}D_k\delta_k^2+8\eta\mu^2K\sum_{k=1}^{K}B_k\sigma_k^2+32\mu^2\beta^2\eta^3K\sum_{k=1}^{K-1}A_k\delta_k^2+8\mu^3\eta^2K\sum_{k=1}^{K}B_k\sigma_k^2\\
&\quad+8\beta^2\eta^2K\sum_{k=1}^{K-1}A_k\delta_k^2+2\mu\eta K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}+4\eta\beta^2K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2\\
&\quad+4\eta\beta^2K\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Big(20A_jC_j^2+2\tilde{\omega}_jC_j^2+8D_jC_j^2\Big)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2\ .
\end{aligned}
\tag{71}
$$

□

A.3 Proof of Theorem 2

Lemma 13. *Given Assumptions 2-4, we can get*

$$\begin{aligned}\mathbb{E}[\|g_{n,t}^{\xi_t} - g_{n,t-1}^{\xi_t}\|^2] \leq KD_0\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + 2K\sum_{k=1}^{K-1} D_k C_k^2 \mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] \\ + 2\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k \mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 2\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 \ ,\end{aligned} \tag{72}$$

where $D_k = \frac{(\prod_{j=1}^K C_i^2)L_{k+1}^2}{C_{k+1}^2}$.

Proof.

$$\begin{aligned}
&\mathbb{E}[\|g_{n,t}^{\xi_t} - g_{n,t-1}^{\xi_t}\|^2] \\
&= \mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)}) \\
&\quad - \nabla f_n^{(1)}(u_{n,t-1}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}; \xi_{n,t}^{(K)})\|^2] \\
&\leq K\mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)}) \\
&\quad - \nabla f_n^{(1)}(u_{n,t-1}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)})\|^2] \\
&\quad + K\mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t-1}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}; \xi_{n,t}^{(K)}) \\
&\quad - \nabla f_n^{(1)}(u_{n,t-1}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}; \xi_{n,t}^{(K)})\|^2] \\
&\quad \cdots \\
&\quad + K\mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}; \xi_{n,t}^{(K)}) \\
&\quad - \nabla f_n^{(1)}(u_{n,t-1}^{(0)}; \xi_{n,t}^{(1)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)}; \xi_{n,t}^{(2)}) \cdots \nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)}; \xi_{n,t}^{(K-1)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}; \xi_{n,t}^{(K)})\|^2] \\
&\leq K\frac{(\prod_{j=1}^K C_i^2)L_1^2}{C_1^2}\mathbb{E}[\|u_{n,t}^{(0)} - u_{n,t-1}^{(0)}\|^2] + K\frac{(\prod_{j=1}^K C_i^2)L_2^2}{C_2^2}\mathbb{E}[\|u_{n,t}^{(1)} - u_{n,t-1}^{(1)}\|^2] \\
&\quad + \cdots + K\frac{(\prod_{j=1}^K C_i^2)L_K^2}{C_K^2}\mathbb{E}[\|u_{n,t}^{(K-1)} - u_{n,t-1}^{(K-1)}\|^2] \\
&= KD_0\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + K\sum_{k=1}^{K-1} D_k\mathbb{E}[\|u_{n,t}^{(k)} - u_{n,t-1}^{(k)}\|^2] \\
&\leq KD_0\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + 2K\sum_{k=1}^{K-1} D_k C_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] \\
&\quad + 2\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 2\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 \ ,
\end{aligned} \tag{73}$$

where $D_k = \frac{(\prod_{j=1}^K C_j^2)L_{k+1}^2}{C_{k+1}^2}$, the third to last step holds due to Assumption 3 and Assumption 2, the last step holds due to Lemma 15.

□

Lemma 14. *For $k \in \{1, \cdots, K-1\}$, given Assumptions 2-4, we can get*

$$\|u_{n,t}^{(k)} - F_n^{(k)}(x_{n,t})\| \leq \sum_{j=1}^{k}\Big(\prod_{i=j+1}^{k} C_i\Big)\|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\| \ , \tag{74}$$

This lemma is the same as Lemma 1.

Lemma 15. *For $k \in \{2, \cdots, K\}$, given Assumptions 2-4, we can get*

$$\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \leq 2C_{k-1}^2 \mathbb{E}[\|u_{n,t-1}^{(k-2)} - u_{n,t}^{(k-2)}\|^2] + 2\beta^2\eta^4 \mathbb{E}[\|u_{n,t-1}^{(k-1)} - f_n^{(k-1)}(u_{n,t-1}^{(k-2)})\|^2] + 2\beta^2\eta^4\delta_{k-1}^2 \ , \tag{75}$$

and

$$\begin{aligned}\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t-1}^{(k-1)}\|^2] \leq & \Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(0)} - u_{n,t}^{(0)}\|^2] + 2\beta^2\eta^4 \sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\mathbb{E}[\|u_{n,t-1}^{(j)} - f_n^{(j)}(u_{n,t-1}^{(j-1)})\|^2] \\ & + 2\beta^2\eta^4 \sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\delta_j^2 \ .\end{aligned} \tag{76}$$

This lemma can be proved by following Lemma 2 through replacing η with η^2.

Lemma 16. *Given Assumptions 2-4, we can get*

$$\begin{aligned}& \frac{1}{N}\sum_{n=1}^{N} \|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) \\ & - \nabla f_n^{(1)}(x_{n,t})\nabla f_n^{(2)}(F_n^{(1)}(x_{n,t})) \cdots \nabla f_n^{(K-1)}(F_n^{(K-2)}(x_{n,t}))\nabla f_n^{(K)}(F_n^{(K-1)}(x_{n,t}))\|^2 \\ \leq & \frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1} A_k \|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2 \ ,\end{aligned} \tag{77}$$

where $A_k = \left(\sum_{j=k}^{K-1}\left(\frac{L_{j+1}\prod_{i=1}^{K}C_i}{C_{j+1}}\prod_{i=k+1}^{j}C_i\right)\right)^2$.

This lemma is the same as Lemma 3

Lemma 17. *Given Assumptions 2-4, we can get*

$$\begin{aligned}& \mathbb{E}\Big[\Big\|\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\ & \quad - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\ \leq & KD_0\|x_{n,t} - x_{n,t-1}\|^2 + 2K\sum_{k=1}^{K-1} D_k C_k^2 \|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2 \\ & + 2\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k \|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2 + 2\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 \ ,\end{aligned} \tag{78}$$

where $D_k = \frac{(\prod_{j=1}^{K}C_j^2)L_{k+1}^2}{C_{k+1}^2}$.

This lemma can be proved by following Lemma 4 through replacing η with η^2.

Lemma 18. *Given Assumptions 2-4, we can get*

$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(g_{n,t}^{\xi_t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}))\Big\|^2\Big] \leq K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} \ , \tag{79}$$

where $B_k = \frac{\prod_{j=1}^{K}C_j^2}{C_k^2}$.

This lemma is the same as Lemma 5.

Lemma 19. *Given Assumptions 2-4, we can get*

$$\mathbb{E}\Big[\Big\|g_{n,t}^{\xi_t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \leq K\sum_{k=1}^{K} B_k\sigma_k^2 \ , \tag{80}$$

where $B_k = \frac{\prod_{j=1}^{K} C_j^2}{C_k^2}$.

This lemma is the same as Lemma 6.

Lemma 20. *For* $k \in \{1, \cdots, K-1\}$*, given Assumptions 2-4, we can get*

$$\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \leq (1-\beta\eta^2)\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 2C_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] + 2\beta^2\eta^4\delta_k^2 \ . \quad (81)$$

This lemma can be proved by following Lemma 7 through replacing η with η^2.

Lemma 21. *Given Assumptions 1-4, we can get*

$$\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] \leq (1 - \eta\frac{1-\lambda^2}{2})\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{2\eta\alpha^2}{1-\lambda^2}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \ . \quad (82)$$

This lemma is the same as Lemma 9.

Lemma 22. *Given Assumptions 2-4 and* $\mu\eta^2 \in (0,1)$*, we can get*

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\leq (1-\mu\eta^2)\mathbb{E}\Big[\Big\|\bar{m}_{t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&\quad + 2KD_0\frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + 4K\frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{K-1}D_kC_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] \\
&\quad + 4\beta^2\eta^4K\frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{K-1}D_k\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 4\beta^2\eta^4K\sum_{k=1}^{K-1}D_k\frac{\delta_k^2}{N} + 2\mu^2\eta^4K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N} \ .
\end{aligned} \quad (83)$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big((1-\mu\eta^2)(m_{n,t-1} - \nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})) \\
&\quad + (\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad\quad - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) - g_{n,t-1} + g_{n,t}) \\
&\quad + \mu\eta^2(g_{n,t-1} - \nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}))\Big)\Big\|^2\Big] \\
&= (1-\mu\eta^2)^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&\quad + 2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)}) \\
&\quad\quad - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) - g_{n,t-1} + g_{n,t}\Big)\Big\|^2\Big] \\
&\quad + 2\mu^2\eta^4\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(g_{n,t-1} - \nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big)\Big\|^2\Big] \\
&\leq (1-\mu\eta^2)^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}m_{n,t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&\quad + 2\frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_{n,t} - g_{n,t-1}\Big\|^2\Big] \\
&\quad + 2\mu^2\eta^4\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\Big(g_{n,t-1} - \nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big)\Big\|^2\Big] \\
&\leq (1-\mu\eta^2)\mathbb{E}\Big[\Big\|\bar{m}_{t-1} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&\quad + 2KD_0\frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + 4K\frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{K-1}D_kC_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] \\
&\quad + 4\beta^2\eta^4K\frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{K-1}D_k\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 4\beta^2\eta^4K\sum_{k=1}^{K-1}D_k\frac{\delta_k^2}{N} + 2\mu^2\eta^4K\sum_{k=1}^{K}B_k\frac{\sigma_k^2}{N}\ .
\end{aligned}
\tag{84}
$$

□

Lemma 23. *Given Assumptions 2-4 and $\mu\eta^2 \in (0,1)$, we can get*

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\leq (1-\mu\eta^2)\mathbb{E}\Big[\Big\|m_{n,t-1} - \nabla f_n^{(1)}(u_{n,t-1}^{(0)})\nabla f_n^{(2)}(u_{n,t-1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t-1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t-1}^{(K-1)})\Big\|^2\Big] \\
&\quad + 2KD_0\mathbb{E}[\|x_{n,t} - x_{n,t-1}\|^2] + 4K\sum_{k=1}^{K-1}D_kC_k^2\mathbb{E}[\|u_{n,t-1}^{(k-1)} - u_{n,t}^{(k-1)}\|^2] \\
&\quad + 4\beta^2\eta^4K\sum_{k=1}^{K-1}D_k\mathbb{E}[\|u_{n,t-1}^{(k)} - f_n^{(k)}(u_{n,t-1}^{(k-1)})\|^2] + 4\beta^2\eta^4K\sum_{k=1}^{K-1}D_k\delta_k^2 + 2\mu^2\eta^4K\sum_{k=1}^{K}B_k\sigma_k^2\ .
\end{aligned}
\tag{85}
$$

This lemma is easy to prove by following Lemma 22.

Lemma 24. *Given Assumptions 1-4, we can get*

$$\begin{aligned}
&\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \\
&\leq \lambda \mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{2\mu^2\eta^4}{1-\lambda} \sum_{n=1}^{N} \mathbb{E}[\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2] \\
&\quad + \frac{2KD_0}{1-\lambda} \sum_{n=1}^{N} \mathbb{E}[\|x_{n,t+1} - x_{n,t}\|^2] + \frac{4K}{1-\lambda} \sum_{n=1}^{N} \sum_{k=1}^{K-1} D_k C_k^2 \mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] \\
&\quad + \frac{4\beta^2\eta^4 K}{1-\lambda} \sum_{n=1}^{N} \sum_{k=1}^{K-1} D_k \mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] + \frac{4\beta^2\eta^4 KN}{1-\lambda} \sum_{k=1}^{K-1} D_k \delta_k^2 + \frac{\mu^2\eta^4 KN}{1-\lambda} \sum_{k=1}^{K} B_k \sigma_k^2 \ .
\end{aligned} \tag{86}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|M_{t+1} - M_t\|_F^2] = \sum_{n=1}^{N} \mathbb{E}[\|m_{n,t+1} - m_{n,t}\|^2] = \sum_{n=1}^{N} \mathbb{E}[\|(1-\mu\eta^2)(m_{n,t} - g_{n,t}^{\xi_{t+1}}) + g_{n,t+1}^{\xi_{t+1}} - m_{n,t}\|^2] \\
&= \sum_{n=1}^{N} \mathbb{E}[\|g_{n,t+1}^{\xi_{t+1}} - g_{n,t}^{\xi_{t+1}} - \mu\eta^2(m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)})) \\
&\quad - \mu\eta^2(\nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)}) - g_{n,t}^{\xi_{t+1}})\|^2] \\
&= \sum_{n=1}^{N} \mathbb{E}[\|g_{n,t+1}^{\xi_{t+1}} - g_{n,t}^{\xi_{t+1}} - \mu\eta^2(m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)}))\|^2] \\
&\quad + \mu^2\eta^4 \sum_{n=1}^{N} \mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)}) - g_{n,t}^{\xi_{t+1}}\|^2] \\
&\leq 2 \sum_{n=1}^{N} \mathbb{E}[\|g_{n,t+1}^{\xi_{t+1}} - g_{n,t}^{\xi_{t+1}}\|^2] + 2\mu^2\eta^4 \sum_{n=1}^{N} \mathbb{E}[\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2] \\
&\quad + \mu^2\eta^4 \sum_{n=1}^{N} \mathbb{E}[\|\nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)}) - g_{n,t}^{\xi_{t+1}}\|^2] \\
&\leq 2\mu^2\eta^4 \sum_{n=1}^{N} \mathbb{E}[\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2] \\
&\quad + 2KD_0 \sum_{n=1}^{N} \mathbb{E}[\|x_{n,t+1} - x_{n,t}\|^2] + 4K \sum_{n=1}^{N} \sum_{k=1}^{K-1} D_k C_k^2 \mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] \\
&\quad + 4\beta^2\eta^4 K \sum_{n=1}^{N} \sum_{k=1}^{K-1} D_k \mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] + 4\beta^2\eta^4 KN \sum_{k=1}^{K-1} D_k \delta_k^2 + \mu^2\eta^4 KN \sum_{k=1}^{K} B_k \sigma_k^2 \ .
\end{aligned} \tag{87}$$

Then, we can get

$$\begin{aligned}
&\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \leq \lambda \mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{1}{1-\lambda} \mathbb{E}[\|M_{t+1} - M_t\|_F^2] \\
&\leq \lambda \mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{2\mu^2\eta^4}{1-\lambda} \sum_{n=1}^{N} \mathbb{E}[\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)}) \nabla f_n^{(2)}(u_{n,t}^{(1)}) \cdots \nabla f_n^{(K-1)}(u_{n,t}^{(K-2)}) \nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2] \\
&\quad + \frac{2KD_0}{1-\lambda} \sum_{n=1}^{N} \mathbb{E}[\|x_{n,t+1} - x_{n,t}\|^2] + \frac{4K}{1-\lambda} \sum_{n=1}^{N} \sum_{k=1}^{K-1} D_k C_k^2 \mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] \\
&\quad + \frac{4\beta^2\eta^4 K}{1-\lambda} \sum_{n=1}^{N} \sum_{k=1}^{K-1} D_k \mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] + \frac{4\beta^2\eta^4 KN}{1-\lambda} \sum_{k=1}^{K-1} D_k \delta_k^2 + \frac{\mu^2\eta^4 KN}{1-\lambda} \sum_{k=1}^{K} B_k \sigma_k^2 \ .
\end{aligned} \tag{88}$$

□

Based on these lemmas, we begin to prove Theorem 2.

Proof. At first, we can get

$$
\begin{aligned}
&F\left(\bar{x}_{t+1}\right) \leq F(\bar{x}_t) + \langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L_F}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&= F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - (\frac{\alpha\eta}{2} - \frac{\alpha^2\eta^2 L_F}{2})\|\bar{m}_t\|^2 + \frac{\alpha\eta}{2}\|\bar{m}_t - \nabla F(\bar{x}_t)\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \frac{\alpha\eta}{2}\|\bar{m}_t - \nabla F(\bar{x}_t)\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \alpha\eta\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\|^2 + \alpha\eta\|\frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t}) - \nabla F(\bar{x}_t)\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \alpha\eta\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla F_n(x_{n,t})\|^2 + \alpha\eta L_F^2\frac{1}{N}\sum_{n=1}^{N}\|x_{n,t} - \bar{x}_t\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \alpha\eta L_F^2\frac{1}{N}\sum_{n=1}^{N}\|x_{n,t} - \bar{x}_t\|^2 \\
&\quad + 2\alpha\eta\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2 \\
&\quad + 2\alpha\eta\frac{1}{N}\sum_{n=1}^{N}\|\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)}) - \nabla F_n(x_{n,t})\|^2 \\
&\leq F(\bar{x}_t) - \frac{\alpha\eta}{2}\|\nabla F(\bar{x}_t)\|^2 - \frac{\alpha\eta}{4}\|\bar{m}_t\|^2 + \alpha\eta L_F^2\frac{1}{N}\sum_{n=1}^{N}\|x_{n,t} - \bar{x}_t\|^2 \\
&\quad + 2\alpha\eta\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\|^2 \\
&\quad + 2\alpha\eta\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}A_k\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2 \ ,
\end{aligned}
\tag{89}
$$

where the fourth step follows from $\eta \leq \frac{1}{2\alpha L_F}$, the last step follows from Lemma 16. Similarly, we define a novel potential function below:

$$
\begin{aligned}
&\mathcal{H}_{t+1} = \mathbb{E}[F(\bar{x}_{t+1})] + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\mathbb{E}[\|u_{n,t+1}^{(k)} - f_n^{(k)}(u_{n,t+1}^{(k-1)})\|^2] \\
&+ \omega_K\mathbb{E}\Big[\Big\|\bar{m}_{t+1} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t+1}^{(0)})\nabla f_n^{(2)}(u_{n,t+1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t+1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t+1}^{(K-1)})\Big\|^2\Big] \\
&+ \omega_{K+1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|m_{n,t+1} - \nabla f_n^{(1)}(u_{n,t+1}^{(0)})\nabla f_n^{(2)}(u_{n,t+1}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t+1}^{(K-2)})\nabla f_n^{(K)}(u_{n,t+1}^{(K-1)})\|^2] \\
&+ \omega_{K+2}\frac{1}{N}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] + \omega_{K+3}\frac{1}{N}\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \ .
\end{aligned}
\tag{90}
$$

Then, we can get

$$
\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
\leq& -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\alpha\eta}{4}\mathbb{E}[\|\bar{m}_t\|^2] + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+3}\frac{\mu^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K} B_k\sigma_k^2 + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&+ 4\omega_K\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + 2\omega_K\mu^2\eta^4 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 4\omega_{K+1}\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 2\omega_{K+1}\mu^2\eta^4 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&+ \Big(2\alpha\eta - \mu\eta^2\omega_K\Big)\mathbb{E}\Big[\Big\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\beta^2\eta^4 K D_k\frac{1}{N} + 4\omega_{K+1}\beta^2\eta^4 K D_k + 2\alpha\eta K A_k + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}D_k - \beta\eta^2\omega_k\Big)\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&+ \Big(+\omega_{K+3}\frac{2\mu^2\eta^4}{1-\lambda} - \mu\eta^2\omega_{K+1}\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&+ \Big(\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \Big(\omega_{K+2}\frac{2\eta\alpha^2}{1-\lambda^2} - (1-\lambda)\omega_{K+3}\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&+ \Big(2\omega_K K D_0\frac{1}{N} + 2\omega_{K+1}K D_0 + \omega_{K+3}\frac{2K D_0}{1-\lambda}\Big)\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\left(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\right)\mathbb{E}[\|u_{n,t}^{(k-1)} - u_{n,t+1}^{(k-1)}\|^2] \\
\leq& -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\alpha\eta}{4}\mathbb{E}[\|\bar{m}_t\|^2] + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+3}\frac{\mu^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K} B_k\sigma_k^2 + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&+ 4\omega_K\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + 2\omega_K\mu^2\eta^4 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 4\omega_{K+1}\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 2\omega_{K+1}\mu^2\eta^4 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&+ \Big(2\alpha\eta - \mu\eta^2\omega_K\Big)\mathbb{E}\Big[\Big\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\beta^2\eta^4 K D_k\frac{1}{N} + 4\omega_{K+1}\beta^2\eta^4 K D_k + 2\alpha\eta K A_k + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}D_k - \beta\eta^2\omega_k\Big)\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&+ \Big(+\omega_{K+3}\frac{2\mu^2\eta^4}{1-\lambda} - \mu\eta^2\omega_{K+1}\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&+ \Big(\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \Big(\omega_{K+2}\frac{2\eta\alpha^2}{1-\lambda^2} - (1-\lambda)\omega_{K+3}\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&+ \Big(2\omega_K K D_0\frac{1}{N} + 2\omega_{K+1}K D_0 + \omega_{K+3}\frac{2K D_0}{1-\lambda}\Big)\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&+ \sum_{k=1}^{K-1}\left(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\left(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\right)2\beta^2\eta^4\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\mathbb{E}[\|u_{n,t}^{(j)} - f_n^{(j)}(u_{n,t}^{(j-1)})\|^2] \\
&+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\left(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\right)2\beta^2\eta^4\sum_{j=1}^{k-1}\Big(\prod_{i=j+1}^{k-1}(2C_i^2)\Big)\delta_j^2\ ,
\end{aligned}
\tag{91}
$$

where the last step holds due to Lemma 15. It ca be reformulated as below:

$$
\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\alpha\eta}{4}\mathbb{E}[\|\bar{m}_t\|^2] + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+3}\frac{\mu^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K} B_k\sigma_k^2 + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&\quad + 4\omega_K\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + 2\omega_K\mu^2\eta^4 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 4\omega_{K+1}\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 2\omega_{K+1}\mu^2\eta^4 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + \Big(2\alpha\eta - \mu\eta^2\omega_K\Big)\mathbb{E}\Big[\Big\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\beta^2\eta^4 K D_k\frac{1}{N} + 4\omega_{K+1}\beta^2\eta^4 K D_k + 2\alpha\eta K A_k + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}D_k - \beta\eta^2\omega_k\Big)\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&\quad + \Big(+\omega_{K+3}\frac{2\mu^2\eta^4}{1-\lambda} - \mu\eta^2\omega_{K+1}\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\quad + \Big(\alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \Big(\omega_{K+2}\frac{2\eta\alpha^2}{1-\lambda^2} - (1-\lambda)\omega_{K+3}\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&\quad + \Bigg[\sum_{k=1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) \\
&\qquad + 2\omega_K K D_0\frac{1}{N} + 2\omega_{K+1}K D_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]\frac{1}{N}\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
&\quad + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_j C_j^2 \\
&\qquad\qquad + 2\omega_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&\quad + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_j C_j^2 + 2\omega_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2 \,.
\end{aligned}
\tag{92}
$$

Then, according to Lemma 10, we can get

$$
\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K-1} D_k\delta_k^2 + \omega_{K+3}\frac{\mu^2\eta^4 K}{1-\lambda}\sum_{k=1}^{K} B_k\sigma_k^2 + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\delta_k^2 \\
&\quad + 4\omega_K\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + 2\omega_K\mu^2\eta^4 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 4\omega_{K+1}\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 2\omega_{K+1}\mu^2\eta^4 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + \Big(2\alpha\eta - \mu\eta^2\omega_K\Big)\mathbb{E}\Big[\Big\|\bar{m}_t - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\quad + \Big(\omega_{K+3}\frac{2\mu^2\eta^4}{1-\lambda} - \mu\eta^2\omega_{K+1}\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|m_{n,t} - \nabla f_n^{(1)}(u_{n,t}^{(0)})\nabla f_n^{(2)}(u_{n,t}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,t}^{(K-2)})\nabla f_n^{(K)}(u_{n,t}^{(K-1)})\Big\|^2\Big] \\
&\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(4\omega_K\beta^2\eta^4 K D_k\frac{1}{N} + 4\omega_{K+1}\beta^2\eta^4 K D_k + 2\alpha\eta K A_k + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}D_k - \beta\eta^2\omega_k \\
&\qquad + 2\beta^2\eta^4\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_j C_j^2 + 2\omega_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\Big)\mathbb{E}[\|u_{n,t}^{(k)} - f_n^{(k)}(u_{n,t}^{(k-1)})\|^2] \\
&\quad + \Big(\Bigg[\sum_{k=1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) \\
&\qquad + 2\omega_K K D_0\frac{1}{N} + 2\omega_{K+1}K D_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]8\eta^2 + \alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+2}\Big)\frac{1}{N}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] \\
&\quad + \Big(\omega_{K+2}\frac{2\eta\alpha^2}{1-\lambda^2} - (1-\lambda)\omega_{K+3} + \Bigg[\sum_{k=1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) \\
&\qquad + 2\omega_K K D_0\frac{1}{N} + 2\omega_{K+1}K D_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]4\alpha^2\eta^2\Big)\frac{1}{N}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&\quad + \Bigg(\Bigg[\sum_{k=1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_k C_k^2 + 2\omega_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) \\
&\qquad + 2\omega_K K D_0\frac{1}{N} + 2\omega_{K+1}K D_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]4\alpha^2\eta^2 - \frac{\alpha\eta}{4}\Bigg)\mathbb{E}[\|\bar{m}_t\|^2] \\
&\quad + 2\beta^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_j C_j^2 + 2\omega_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2 \ .
\end{aligned}
\tag{93}
$$

At first, we set $\omega_K = \frac{2\alpha}{\mu\eta}$ such that $2\alpha\eta - \mu\eta^2\omega_K = 0$. Then, we set $\omega_{K+3} = \alpha(1-\lambda)$. By enforcing $\omega_{K+3}\frac{2\mu^2\eta^4}{1-\lambda} - \mu\eta^2\omega_{K+1} = 0$, we can get $\omega_{K+1} = \omega_{K+3}\frac{2\mu\eta^2}{1-\lambda} = 2\alpha\mu\eta^2$.

Then, we enforce

$$
\begin{aligned}
&4\omega_K\beta^2\eta^4 K D_k\frac{1}{N} + 4\omega_{K+1}\beta^2\eta^4 K D_k + 2\alpha\eta K A_k + \omega_{K+3}\frac{4\beta^2\eta^4 K}{1-\lambda}D_k - \beta\eta^2\omega_k \\
&\quad + 2\beta^2\eta^4\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_j C_j^2 + 2\omega_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg] \leq 0 \ ,
\end{aligned}
\tag{94}
$$

It is easy to get

$$
\begin{aligned}
&\frac{8\alpha\beta^2\eta^3}{\mu}\frac{KD_k}{N} + 4\omega_{K+3}K D_k\frac{2\mu\beta^2\eta^6}{1-\lambda} + 2\alpha\eta K A_k + \omega_{K+3}K D_k\frac{4\beta^2\eta^4}{1-\lambda} - \beta\eta^2\omega_k \\
&\quad + 2\beta^2\eta^4\Bigg[\sum_{j=1}^{K-1}\Bigg(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+3}\frac{2\mu\eta^2}{1-\lambda}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_j C_j^2 + 2\omega_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg] \leq 0 \ .
\end{aligned}
\tag{95}
$$

Then, we enforce

$$2\beta^2\eta^4\left[\sum_{j=1}^{K-1}2\omega_j C_j^2\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]-\beta\eta^2\omega_k\leq-\frac{1}{2}\beta\eta^2\omega_k\ , \tag{96}$$

and

$$\begin{aligned}&\frac{8\alpha\beta^2\eta^3}{\mu}\frac{KD_k}{N}+4\omega_{K+3}KD_k\frac{2\mu\beta^2\eta^6}{1-\lambda}+2\alpha\eta KA_k+\omega_{K+3}KD_k\frac{4\beta^2\eta^4}{1-\lambda}\\&\quad+2\beta^2\eta^4\left[\sum_{j=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N}+4\omega_{K+3}\frac{2\mu\eta^2}{1-\lambda}K+\omega_{K+3}\frac{4K}{1-\lambda}\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]-\frac{1}{2}\beta\eta^2\omega_k\leq 0\ .\end{aligned} \tag{97}$$

From the first inequality, we can get

$$\eta\leq\frac{1}{2}\sqrt{\frac{\omega_k}{\beta\sum_{j=1}^{K-1}2\omega_jC_j^2\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)}}\ . \tag{98}$$

As for the second inequality, due to $\mu\eta^2<1$, we have

$$\begin{aligned}&\frac{8\alpha\beta^2\eta^3}{\mu}\frac{KD_k}{N}+4\omega_{K+3}KD_k\frac{2\mu\beta^2\eta^6}{1-\lambda}+2\alpha\eta KA_k+\omega_{K+3}KD_k\frac{4\beta^2\eta^4}{1-\lambda}\\&\quad+2\beta^2\eta^4\left[\sum_{j=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N}+4\omega_{K+3}\frac{2\mu\eta^2}{1-\lambda}K+\omega_{K+3}\frac{4K}{1-\lambda}\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]-\frac{1}{2}\beta\eta^2\omega_k\\&\leq\frac{8\alpha\beta^2\eta^3}{\mu}\frac{KD_k}{N}+\omega_{K+3}KD_k\frac{8\beta^2\eta^4}{1-\lambda}+2\alpha\eta KA_k+\omega_{K+3}KD_k\frac{4\beta^2\eta^4}{1-\lambda}\\&\quad+2\beta^2\eta^4\left[\sum_{j=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N}+\omega_{K+3}\frac{8}{1-\lambda}K+\omega_{K+3}\frac{4K}{1-\lambda}\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]-\frac{1}{2}\beta\eta^2\omega_k\\&\leq\frac{8\alpha\beta^2\eta^3}{\mu}\frac{KD_k}{N}+\omega_{K+3}KD_k\frac{12\beta^2\eta^4}{1-\lambda}+2\alpha\eta KA_k\\&\quad+2\beta^2\eta^4\left[\sum_{j=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N}+\omega_{K+3}\frac{12K}{1-\lambda}\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]-\frac{1}{2}\beta\eta^2\omega_k\\&\leq\frac{8\alpha\beta^2\eta^3}{\mu}\frac{KD_k}{N}+12\beta^2\eta^4\alpha KD_k+2\alpha\eta KA_k\\&\quad+2\alpha\beta^2\eta^4\left[\sum_{j=1}^{K-1}\left(\Big(\frac{8}{\mu\eta}\frac{K}{N}+12K\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]-\frac{1}{2}\beta\eta^2\omega_k\\&\leq\beta\eta^2\left[\frac{8\alpha\beta\eta}{\mu}\frac{KD_k}{N}+12\beta\eta^2\alpha KD_k+2\frac{1}{\beta\eta}\alpha KA_k\right.\\&\quad\left.+2\alpha\beta\eta^2\sum_{j=1}^{K-1}\left(\Big(\frac{8}{\mu\eta}\frac{K}{N}+12K\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)-\frac{1}{2}\omega_k\right]\ .\end{aligned} \tag{99}$$

We enforce this upper bound to be non-positive, i.e.,

$$\begin{aligned}&\frac{8\alpha\beta\eta}{\mu}\frac{KD_k}{N}+12\beta\eta^2\alpha KD_k+2\frac{1}{\beta\eta}\alpha KA_k+2\alpha\beta\eta^2\sum_{j=1}^{K-1}\left(\Big(\frac{8}{\mu\eta}\frac{K}{N}+12K\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)-\frac{1}{2}\omega_k\leq 0\ ,\\&\omega_k\geq\frac{16\alpha\beta\eta}{\mu}\frac{KD_k}{N}+24\beta\eta^2\alpha KD_k+4\frac{1}{\beta\eta}\alpha KA_k+4\alpha\beta\eta^2\sum_{j=1}^{K-1}\left(\Big(\frac{8}{\mu\eta}\frac{K}{N}+12K\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\\&=\frac{\alpha}{\eta}\left[\frac{16\beta\eta^2}{\mu}\frac{KD_k}{N}+24\beta\eta^3KD_k+4\frac{1}{\beta}KA_k+4\beta\eta^3\sum_{j=1}^{K-1}\left(\Big(\frac{8}{\mu\eta}\frac{K}{N}+12K\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]\ .\end{aligned} \tag{100}$$

Due to $\beta\eta^2 \leq 1, \eta \leq 1$, we can set

$$\omega_k \triangleq \frac{\alpha K}{\eta}\tilde{\omega}_k = \frac{\alpha}{\eta}\left[\frac{16KD_k}{\mu N} + 24KD_k + \frac{4KA_k}{\beta} + 16K\sum_{j=1}^{K-1}\left(\Big(\frac{2}{\mu N}+3\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right], \tag{101}$$

where $\tilde{\omega}_k = \left[\frac{16D_k}{\mu N} + 24D_k + \frac{4A_k}{\beta} + 16\sum_{j=1}^{K-1}\left(\Big(\frac{2}{\mu N}+3\Big)D_jC_j^2\right)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\right]$. Then, we can simplified the upper bound of η as below:

$$\eta \leq \frac{1}{2}\sqrt{\frac{\tilde{\omega}_k}{2\beta\sum_{j=1}^{K-1}\tilde{\omega}_jC_j^2\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)}}\,. \tag{102}$$

Based on these values, we have

$$\begin{aligned}
&\Bigg[\sum_{k=1}^{K-1}\left(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_kC_k^2 + 2\omega_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\\
&\quad + 2\omega_K KD_0\frac{1}{N} + 2\omega_{K+1}KD_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]\\
&\leq \Bigg[\sum_{k=1}^{K-1}\left(\Big(4\frac{2\alpha}{\mu\eta}K\frac{1}{N} + 4\omega_{K+3}\frac{2\mu\eta^2}{1-\lambda}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_kC_k^2 + 2\omega_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\\
&\quad + 2\frac{2\alpha}{\mu\eta}KD_0\frac{1}{N} + 2\omega_{K+3}\frac{2\mu\eta^2}{1-\lambda}KD_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]\\
&\leq \Bigg[\sum_{k=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N} + \omega_{K+3}\frac{12K}{1-\lambda}\Big)D_kC_k^2 + 2\omega_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4\alpha}{\mu\eta}\frac{KD_0}{N} + \omega_{K+3}\frac{6KD_0}{1-\lambda}\Bigg]\\
&\leq \Bigg[\sum_{k=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N} + 12\alpha K\Big)D_kC_k^2 + \frac{2\alpha K}{\eta}\tilde{\omega}_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4\alpha}{\mu\eta}\frac{KD_0}{N} + 6\alpha KD_0\Bigg]\,.
\end{aligned} \tag{103}$$

Then, we enforce

$$\begin{aligned}
&\Bigg[\sum_{k=1}^{K-1}\left(\Big(4\omega_K K\frac{1}{N} + 4\omega_{K+1}K + \omega_{K+3}\frac{4K}{1-\lambda}\Big)D_kC_k^2 + 2\omega_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big)\\
&\quad + 2\omega_K KD_0\frac{1}{N} + 2\omega_{K+1}KD_0 + \omega_{K+3}\frac{2KD_0}{1-\lambda}\Bigg]8\eta^2 + \alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+2}\\
&\leq \Bigg[\sum_{k=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu\eta}\frac{K}{N} + 12\alpha K\Big)D_kC_k^2 + \frac{2\alpha K}{\eta}\tilde{\omega}_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4\alpha}{\mu\eta}\frac{KD_0}{N} + 6\alpha KD_0\Bigg]8\eta^2 + \alpha\eta L_F^2 - \eta\frac{1-\lambda^2}{2}\omega_{K+2}\\
&\leq 0\,.
\end{aligned} \tag{104}$$

It is easy to know

$$\omega_{K+2} \geq \frac{2}{1-\lambda^2}\left[\left(\sum_{k=1}^{K-1}\left(\Big(\frac{8\alpha}{\mu}\frac{K}{N} + 12\alpha\eta K\Big)D_kC_k^2 + 2\alpha\tilde{\omega}_kKC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4\alpha}{\mu}\frac{KD_0}{N} + 6\alpha\eta KD_0\right)8 + \alpha L_F^2\right]. \tag{105}$$

Since $\eta < 1$, we can set

$$\omega_{K+2} \triangleq \frac{\alpha}{1-\lambda^2}\tilde{\omega}_{K+2} = \frac{2\alpha}{1-\lambda^2}\left[\left(\sum_{k=1}^{K-1}\left(\Big(\frac{8}{\mu N}+12\Big)D_kC_k^2 + 2\tilde{\omega}_kC_k^2\right)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4D_0}{\mu N} + 6D_0\right)8K + L_F^2\right], \tag{106}$$

where $\tilde{\omega}_{K+2} = 2\left[\left(\sum_{k=1}^{K-1}\left(\left(\frac{8}{\mu N}+12\right)D_kC_k^2+2\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4D_0}{\mu N}+6D_0\right)8K+L_F^2\right]$.

Moreover, we enforce

$$\begin{aligned}
&\omega_{K+2}\frac{2\eta\alpha^2}{1-\lambda^2}-(1-\lambda)\omega_{K+3}+\left[\sum_{k=1}^{K-1}\left(\left(4\omega_KK\frac{1}{N}+4\omega_{K+1}K+\omega_{K+3}\frac{4K}{1-\lambda}\right)D_kC_k^2+2\omega_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)\right.\\
&\quad\left.+2\omega_KKD_0\frac{1}{N}+2\omega_{K+1}KD_0+\omega_{K+3}\frac{2KD_0}{1-\lambda}\right]4\alpha^2\eta^2\\
&\leq\tilde{\omega}_{K+2}\frac{2\eta\alpha^3}{(1-\lambda^2)^2}-\alpha(1-\lambda)^2\\
&\quad+\left[\sum_{k=1}^{K-1}\left(\left(\frac{8\alpha}{\mu\eta}\frac{K}{N}+12\alpha K\right)D_kC_k^2+\frac{2\alpha K}{\eta}\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4\alpha}{\mu\eta}\frac{KD_0}{N}+6\alpha KD_0\right]4\alpha^2\eta^2\leq 0\ .
\end{aligned}\tag{107}$$

Then, due to $\eta<1$ and $1+\lambda>1$, we can get

$$\alpha\leq\frac{(1-\lambda)^2}{\sqrt{2\tilde{\omega}_{K+2}+4K\left[\sum_{k=1}^{K-1}\left(\left(\frac{8}{\mu N}+12\right)D_kC_k^2+2\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4D_0}{\mu N}+6D_0\right]}}.\tag{108}$$

Moreover, with $\eta<1$, we enforce

$$\begin{aligned}
&\left[\sum_{k=1}^{K-1}\left(\left(4\omega_KK\frac{1}{N}+4\omega_{K+1}K+\omega_{K+3}\frac{4K}{1-\lambda}\right)D_kC_k^2+2\omega_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)\right.\\
&\quad\left.+2\omega_KKD_0\frac{1}{N}+2\omega_{K+1}KD_0+\omega_{K+3}\frac{2KD_0}{1-\lambda}\right]4\alpha^2\eta^2-\frac{\alpha\eta}{4}\\
&\leq\left[\sum_{k=1}^{K-1}\left(\left(\frac{8\alpha}{\mu\eta}\frac{K}{N}+12\alpha K\right)D_kC_k^2+\frac{2\alpha K}{\eta}\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4\alpha}{\mu\eta}\frac{KD_0}{N}+6\alpha KD_0\right]4\alpha^2\eta^2-\frac{\alpha\eta}{4}\\
&\leq\left[\sum_{k=1}^{K-1}\left(\left(\frac{8\alpha}{\mu}\frac{K}{N}+12\eta\alpha K\right)D_kC_k^2+2\alpha K\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4\alpha}{\mu}\frac{KD_0}{N}+6\alpha\eta KD_0\right]4\alpha^2\eta-\frac{\alpha\eta}{4}\\
&\leq\left[\sum_{k=1}^{K-1}\left(\left(\frac{8}{\mu}\frac{K}{N}+12K\right)D_kC_k^2+2K\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4}{\mu}\frac{KD_0}{N}+6KD_0\right]4\alpha^3\eta-\frac{\alpha\eta}{4}\leq 0\ ,
\end{aligned}\tag{109}$$

so that we can get

$$\alpha\leq\frac{1}{4}\Bigg/\sqrt{K\sum_{k=1}^{K-1}\left(\left(\frac{8}{\mu N}+12\right)D_kC_k^2+2\tilde{\omega}_kC_k^2\right)\left(\prod_{j=1}^{k-1}(2C_j^2)\right)+\frac{4KD_0}{\mu N}+6KD_0}\ .\tag{110}$$

In summary, by setting

$$\begin{aligned}
&\omega_k = \frac{\alpha K}{\eta}\tilde{\omega}_k,\ k \in \{1, 2, \cdots, K-1\}\ , \\
&\omega_K = \frac{2\alpha}{\mu\eta}\ , \omega_{K+1} = 2\alpha\mu\eta^2\ , \omega_{K+2} = \frac{\alpha}{1-\lambda^2}\tilde{\omega}_{K+2}\ , \omega_{K+3} = \alpha(1-\lambda)\ , \\
&\alpha \leq (1-\lambda)^2 \Bigg/ \sqrt{2\tilde{\omega}_{K+2} + 4K\Bigg[\sum_{k=1}^{K-1}\Bigg(\Big(\frac{8}{\mu N}+12\Big)D_k C_k^2 + 2\tilde{\omega}_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4D_0}{\mu N} + 6D_0\Bigg]}\ , \\
&\alpha \leq \frac{1}{4} \Bigg/ \sqrt{K\sum_{k=1}^{K-1}\Bigg(\Big(\frac{8}{\mu N}+12\Big)D_k C_k^2 + 2\tilde{\omega}_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4KD_0}{\mu N} + 6KD_0}\ , \\
&\eta \leq \frac{1}{2}\sqrt{\frac{\tilde{\omega}_k}{2\beta\sum_{j=1}^{K-1}\tilde{\omega}_j C_j^2\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)}}\ ,
\end{aligned} \tag{111}$$

where $\tilde{\omega}_k = \Bigg[\frac{16D_k}{\mu N} + 24D_k + \frac{4A_k}{\beta} + 16\sum_{j=1}^{K-1}\Bigg(\Big(\frac{2}{\mu N}+3\Big)D_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]$, $\tilde{\omega}_{K+2} = 2\Bigg[\Bigg(\sum_{k=1}^{K-1}\Bigg(\Big(\frac{8K}{\mu N} + 12K\Big)D_k C_k^2 + 2\tilde{\omega}_k C_k^2\Bigg)\Big(\prod_{j=1}^{k-1}(2C_j^2)\Big) + \frac{4KD_0}{\mu N} + 6KD_0\Bigg)8 + L_F^2\Bigg]$, we can get

$$\begin{aligned}
&\mathcal{H}_{t+1} - \mathcal{H}_t \\
&\leq -\frac{\alpha\eta}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + 4\alpha\beta^2\eta^4 K\sum_{k=1}^{K-1} D_k\delta_k^2 + \alpha\mu^2\eta^4 K\sum_{k=1}^{K} B_k\sigma_k^2 + 2\alpha\beta^2\eta^3 K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2 \\
&\quad + \frac{8\alpha}{\mu}\beta^2\eta^3 K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + \frac{4\alpha\mu^2\eta^3}{\mu}K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 8\alpha\mu\beta^2\eta^6 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 4\alpha\mu^3\eta^6 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + 2\alpha\beta^2\eta^3 K\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(\frac{8}{\mu N}+8\mu\eta^3+4\eta\Big)D_j C_j^2 + 2\tilde{\omega}_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2\ .
\end{aligned} \tag{112}$$

By summing over t from 0 to $T-1$, we can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \\
&\leq \frac{2(\mathcal{H}_0 - \mathcal{H}_T)}{\alpha\eta T} + 8\beta^2\eta^3 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 2\mu^2\eta^3 K\sum_{k=1}^{K} B_k\sigma_k^2 + 4\beta^2\eta^2 K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2 \\
&\quad + \frac{16\beta^2\eta^2}{\mu}K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + 8\mu\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 16\mu\beta^2\eta^5 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 8\mu^3\eta^5 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + 4\beta^2\eta^2 K\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(\frac{8}{\mu N}+8\mu\eta^3+4\eta\Big)D_j C_j^2 + 2\tilde{\omega}_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2\ .
\end{aligned} \tag{113}$$

In the following, we bound $\mathcal{H}_0$. Specifically, we have

$$\mathbb{E}[\|u_{n,0}^{(k)} - f_n^{(k)}(u_{n,0}^{(k-1)})\|^2] = \mathbb{E}[\|f_n^{(k)}(u_{n,0}^{(k-1)}; \xi_{n,0}^{(k)}) - f_n^{(k)}(u_{n,0}^{(k-1)})\|^2] \leq \frac{\delta_k^2}{S}\ , \tag{114}$$

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} m_{n,0} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} v_{n,0}^{(1)}v_{n,0}^{(2)}\cdots v_{n,0}^{(K-1)}v_{n,0}^{(K)} - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\Big\|^2\Big] \\
&\leq K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{SN} ,
\end{aligned}
\tag{115}
$$

as well as $\mathbb{E}\Big[\Big\|m_{n,0} - \nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\Big\|^2\Big] \leq K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{S}$. Similar to Theorem 1, we can get

$$
\frac{1}{N}\mathbb{E}[\|Y_0 - \bar{Y}_0\|_F^2] \leq 6K\sum_{k=1}^{K} B_k\sigma_k^2 + 12K\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K} C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1} 8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2) . \tag{116}
$$

As a result, we can get

$$
\begin{aligned}
H_0 &= F(x_0) + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\omega_k\mathbb{E}[\|u_{n,0}^{(k)} - f_n^{(k)}(u_{n,0}^{(k-1)})\|^2] \\
&+ \omega_K\mathbb{E}\Big[\Big\|\bar{m}_0 - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\Big\|^2\Big] \\
&+ \omega_{K+1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|m_{n,0} - \nabla f_n^{(1)}(u_{n,0}^{(0)})\nabla f_n^{(2)}(u_{n,0}^{(1)})\cdots\nabla f_n^{(K-1)}(u_{n,0}^{(K-2)})\nabla f_n^{(K)}(u_{n,0}^{(K-1)})\|^2] \\
&+ \omega_{K+2}\frac{1}{N}\mathbb{E}[\|X_0 - \bar{X}_0\|_F^2] + \omega_{K+3}\frac{1}{N}\mathbb{E}[\|Y_0 - \bar{Y}_0\|_F^2] \\
&\leq F(x_0) + \frac{\alpha K}{\eta}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\tilde{\omega}_k\frac{\delta_k^2}{S} + \frac{2\alpha}{\mu\eta}K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{SN} + 2\alpha\mu\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{S} \\
&\quad + \alpha\Big(6K\sum_{k=1}^{K} B_k\sigma_k^2 + 12K\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K} C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1} 8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\Big) .
\end{aligned}
\tag{117}
$$

Finally, we can get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] \\
&\leq \frac{2(F(x_0) - F(x_*))}{\alpha\eta T} + \frac{2K}{\eta^2 T}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\tilde{\omega}_k\frac{\delta_k^2}{S} + \frac{4K}{\mu\eta^2 T}\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{SN} + \frac{4\mu\eta K}{T}\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{S} \\
&\quad + \frac{12K}{\eta T}\Big(\sum_{k=1}^{K} B_k\sigma_k^2 + 2\sum_{k=2}^{K}\frac{(\prod_{j=1}^{K} C_j^2)L_k^2}{C_k^2}\sum_{i=1}^{k-1} 8\delta_i^2\prod_{j=i+1}^{k-1}(8C_j^2)\Big) \\
&\quad + 8\beta^2\eta^3 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 2\mu^2\eta^3 K\sum_{k=1}^{K} B_k\sigma_k^2 + 4\beta^2\eta^2 K\sum_{k=1}^{K-1}\tilde{\omega}_k\delta_k^2 \\
&\quad + \frac{16\beta^2\eta^2}{\mu}K\sum_{k=1}^{K-1} D_k\frac{\delta_k^2}{N} + 8\mu\eta^2 K\sum_{k=1}^{K} B_k\frac{\sigma_k^2}{N} + 16\mu\beta^2\eta^5 K\sum_{k=1}^{K-1} D_k\delta_k^2 + 8\mu^3\eta^5 K\sum_{k=1}^{K} B_k\sigma_k^2 \\
&\quad + 4\beta^2\eta^2 K\sum_{k=1}^{K-1}\Bigg[\sum_{j=k+1}^{K-1}\Bigg(\Big(\frac{8}{\mu N} + 8\mu\eta^3 + 4\eta\Big)D_j C_j^2 + 2\tilde{\omega}_j C_j^2\Bigg)\Big(\prod_{i=k+1}^{j}(2C_i^2)\Big)\Bigg]\delta_k^2 .
\end{aligned}
\tag{118}
$$

□