Stability and Generalization for Stochastic Recursive Momentum-based Algorithms for (Strongly-)Convex One to K-Level Stochastic Optimizations

Xiaokang Pan [1] Xingyu Li [2] Jin Liu [1] Tao Sun [3] Kai Sun [4] Lixing Chen [5] Zhe Qu [1]

Abstract

STOchastic Recursive Momentum (STORM)-based algorithms have been widely developed to solve one to K-level ($K \geq 3$) stochastic optimization problems. Specifically, they use estimators to mitigate the biased gradient issue and achieve near-optimal convergence results. However, there is relatively little work on understanding their generalization performance, particularly evident during the transition from one to K-level optimization contexts. This paper provides a comprehensive generalization analysis of three representative STORM-based algorithms: STORM, COVER, and SVMR, for one, two, and K-level stochastic optimizations under both convex and strongly convex settings based on algorithmic stability. Firstly, we define stability for K-level optimizations and link it to generalization. Then, we detail the stability results for three prominent STORM-based algorithms. Finally, we derive their excess risk bounds by balancing stability results with optimization errors. Our theoretical results provide strong evidence to complete STORM-based algorithms: (1) Each estimator may decrease their stability due to variance with its estimation target. (2) Every additional level might escalate the generalization error, influenced by the stability and the variance between its cumulative stochastic gradient and the true gradient. (3) Increasing the batch size for the initial computation of estimators presents a favorable trade-off, enhancing the generalization performance.

1. Introduction

In stochastic optimization problems, variance reduction techniques (Fang et al., 2018; Zhou et al., 2020; Wen et al., 2018; Qi et al., 2021; Liu et al., 2019; 2024) can significantly mitigate the negative impact of inherent variance due to the stochastic gradients. In particular, Stochastic Recursive Momentum (STORM) (Cutkosky & Orabona, 2019) stands out for its simple implementation and near-optimal convergence results. STORM carefully designs momentum-based estimators for model updating, which can dynamically adapt to the optimization challenge without a large batch or checkpoint gradient computations. Due to these advantages, STORM has been extensively used in various practical applications: reinforcement learning (Hu et al., 2019; Mao et al., 2022), model-agnostic meta-learning (Ji et al., 2022; Qu et al., 2023a), risk-averse portfolio optimization (Tran Dinh et al., 2020; Jiang et al., 2022), and deep AUC maximization (Yuan et al., 2021; Liu et al., 2024).

Subsequently, various STORM-based algorithms (Hu et al., 2019; Yuan et al., 2021; Chen et al., 2021; Jiang et al., 2022; Li et al., 2023a) have extended this methodology to address stochastic two-level and K-level (where $K \geq 3$) optimization problems. In their definitions, two-level stochastic optimizations are equivalent to stochastic compositional optimizations and similar to K-level stochastic optimizations (Wang et al., 2017; Ghadimi et al., 2020; Chen et al., 2021), which pose a challenge in obtaining a biased estimate of the objective function and gradients (Dann et al., 2014; Wang et al., 2017). By leveraging the high-precision estimations, STORM-based algorithms have successfully addressed the corresponding challenge.

In particular, in two-level optimizations, one of the most popular STORM-based algorithms COVER (Qi et al., 2021) employs estimators for both the value of the inner function and the value of the gradient. When increasing to K-level optimizations, inherent variances can be magnified, leading to significant gradient deviations and potential explosions. To mitigate this, the near-optimal algorithm SVMR (Jiang et al., 2022) employs estimators for all function values and gradients, except the outer function value, and applies gradient projection techniques to the function gradient estimator.

[1] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China [2] Department of Computer Science, Tulane University, New Orleans, USA [3] College of Computer, National University of Defense Technology, Changsha, China [4] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China [5] Institute of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: X. Li <xli82@tulane.edu>, Z. Qu <zhe_qu@csu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Although STORM-based algorithms have achieved many breakthroughs in algorithmic convergence, their effect on generalization performance is less understood (Hardt et al., 2016; Yang et al., 2023), i.e., how the model trained by the training samples would generalize to test samples, especially for optimizations with higher levels. To clearly understand the generalization of these algorithms, we consider the following two key questions.

(1) Compared to SGD-based Algorithms, do these estimators prove weaker or stronger generalization performance in STORM-based algorithms?
(2) As escalating to the K-level optimization, how does this increased complexity impact the generalization performance of the estimators?

Specifically, as the success of STORM lies in leveraging estimators to tackle biased gradient issues, exploring the influence of these estimators on generalization performance enriches the study (Yuan et al., 2019; Hu et al., 2019; Ghadimi et al., 2020; Balasubramanian et al., 2022; Qu et al., 2023b). Additionally, in K-level optimization, the gradient estimator at each level is influenced by the function value estimator at the preceding level, which, in turn, indirectly affects the function value estimator at the subsequent level (Chen et al., 2021; Jiang et al., 2022). Therefore, addressing the second question can offer guidance for designing corresponding estimators in more complex and general scenarios.

To answer the above two questions, this paper leverages the algorithmic stability to systematically explore the generalization of STORM-based algorithms from one to K-level stochastic optimizations. We believe that this exploration is important to gain insights into STORM's scalability and effectiveness across different tiers of stochastic optimization. In particular, our contributions are summarized as follows.

1. To achieve our goal, we first introduce a novel definition of uniform stability, specifically for K-level optimizations. Leveraging this definition, we establish a quantitative relationship between generalization error and stability in the context of K-level optimization. Then, we analyze the stability and optimization errors for three distinct algorithms: STORM, COVER, and SVMR, corresponding to one, two, and K-level stochastic optimizations in both convex and strongly convex settings. Finally, by analyzing the interplay between stability and optimization errors, we ascertain their excess risks in these settings.

2. Our theoretical results indicate that fewer iterations and proper step sizes will improve algorithm stability of stability in the convex setting. For the excess risk, our results demonstrate that we need about $T \asymp \max(n_k^{5/2})$, $\forall k \in [1, K]$, iterations to achieve the ideal excess risk rate. In the strongly convex setting, a proper step size will not necessarily make the algorithm stable enough, which must be combined with expanding the batch size to ensure stability. Moreover, $T \asymp \max(n_k^{7/6})$ iterations should be used, which is fewer than the convex setting.

3. Based on our analysis, we can successfully address the above questions. Firstly, we find that the stability of the algorithm can be compromised by each estimator, due to the variance between the estimator and its estimated target, which degrades the generalization performance. Moreover, as the number of levels increases, two main factors impact the algorithm's generalization error: the first is the influence of the new level on the algorithm's stability, and the second is the variance between the combined stochastic gradient and the true gradient across all levels. There is one more observation in our analysis: employing more samples for the initial computation of estimators may enhance performance without significantly increasing computational costs. This strategy presents a viable approach to improve the efficiency of STORM-based algorithms.

2. Related Work

Algorithmic stability and Generalization. In learning theory, the stability of an algorithm shows that small changes in the training data result in only minimal differences in the predictions made by the model (Kearns & Ron, 1997; Vapnik & Chapelle, 2000; Cucker & Smale, 2002). The landmark work (Bousquet & Elisseeff, 2002) introduces the notion of uniform stability and establishes the generalization of ERM based on stability, and it has a deep connection with (Cesa-Bianchi et al., 2004; Rakhlin et al., 2005; Kutin & Niyogi, 2012). Furthermore, (Bartlett & Mendelson, 2002; Poggio et al., 2004; Shalev-Shwartz et al., 2010) discuss the relationship between algorithmic stability and complexity measures, and use it on general conditions for predictivity. (Hardt et al., 2016) contribute significantly to the understanding of algorithmic stability in optimization algorithms, particularly gradient descent. More recently, (Li et al., 2023b) presents in-context learning, showing its effectiveness and stability in different data scenarios. (Sakaue & Oki, 2023) demonstrates that coordinate estimation leads to tighter generalization bounds.

Stochastic Compositional Optimization. Extensive studies have mitigated the issue of bias in gradient estimation due to combination functions. (Wang et al., 2017) uses stochastic gradients for internal function value computation. Variance reduction techniques can accelerate the efficiency of Stochastic Compositional Gradient Descent (SCGD). Algorithms such as SAGA (Zhang & Xiao, 2019), SPIDER (Fang et al., 2018), and STORM (Cutkosky & Orabona, 2019) have been integrated into SCGD. Later, some studies (Yuan et al., 2019; Zhang & Xiao, 2021; Tarzanagh et al., 2022) have successfully linked stochastic two-level or

K-level optimization challenges. In K-level optimization, (Yang et al., 2019) leads to the creation of an accelerated technique (A-TSCGD). Subsequently, (Balasubramanian et al., 2022) introduces the NLASG method, which expands the scope of the NASA (Ghadimi et al., 2020) algorithm to broader applications. In a similar vein, (Chen et al., 2021; Jiang et al., 2022) extend STORM for estimating function values to K levels. However, all the above works only focus on convergence analysis.

3. Preliminaries and Warm Up

In this section, we begin by introducing three optimization problems that we address, accompanied by three popular STORM-based algorithms designed for these specific problems. Then, we will present the concept of stability as applied in statistical learning theory (James et al., 2013). To this end, we present the first theorem in this paper.

3.1. One to K-level Stochastic Optimziations

In this paper, we extend algorithmic stability analysis to the most popular STORM-based algorithms: STORM (Cutkosky & Orabona, 2019), COVER (Qi et al., 2021), and SVMR (Jiang et al., 2022) for stochastic optimization problems with levels 1, 2, and $K \geq 3$, respectively. Detailed update rules for these algorithms are presented in Appendix A, Algorithms 1-3. Their optimization formulations are introduced subsequently.

One-level optimization. Typically, the one-level stochastic optimization problem (Hardt et al., 2016; Cutkosky & Orabona, 2019; Bousquet et al., 2020; Levy et al., 2021) can be formulated as follows

$$\min_{x\in\mathcal{X}}\Big\{F(x)=\mathbb{E}_{\nu}[f_{\nu}(x)]\Big\}, \tag{1}$$

where $f:\mathbb{R}^d\rightarrow\mathbb{R}^{d_1}$ on a convex domain $\mathcal{X}\in\mathbb{R}^d$, ν is an independent random data sample, and F is the empirical risk $\min_{x\in\mathcal{X}}\{F_S(x):=f_S(x)=\frac{1}{n}\sum_{i=1}^n f_{\nu_i}(x)\}$. Let $S=\{\nu_1,\cdots,\nu_n\}$ be a dataset from which the samples are drawn independently and identically (i.i.d.). To facilitate the expansion below, we give more symbol definitions: S' is the i.i.d copy of S, where $S'=\{\nu_1',\cdots,\nu_n'\}$, and S^i is the i.i.d. copy of S where only i-th data point ν_i in S in change to ν_i'. Compared with SGD which directly uses stochastic gradients for updates, the main part of STORM (Cutkosky & Orabona, 2019) is to leverage the corrected momentum variance reduction estimator for updates.

Two-level optimization. We consider the two-level stochastic optimization problem (Yuan et al., 2019; Yang et al., 2019; Balasubramanian et al., 2022) as follows

$$\min_{x\in\mathcal{X}}\Big\{F(x)=f\circ g(x)=\mathbb{E}_{\nu}[f_{\nu}(\mathbb{E}_{\omega}[g_{\omega}(x)])]\Big\}, \tag{2}$$

where $f:\mathbb{R}^{d_1}\rightarrow\mathbb{R}^{d_2}$ and $g:\mathbb{R}^d\rightarrow\mathbb{R}^{d_1}$ on a convex domain $\mathcal{X}\in\mathbb{R}^d$, ν and ω are independent random variables. Let $S=S_\nu\cup S_\omega$, where $S_\nu=\{\nu_1,\cdots,\nu_n\}$ and $S_\omega=\{\omega_1,\cdots,\omega_m\}$, and the empirical risk is defined as $\min_{x\in\mathcal{X}}\{F_S(x):=f_S(g_S(x))=\frac{1}{n}\sum_{i=1}^n f_{\nu_i}(\frac{1}{m}\sum_{j=1}^m g_{\omega_j}(x))\}$. In this scenario, altering a single data point can affect either S_ν or S_ω. For $\forall i\in[1,n]$ and $\forall j\in[1,m]$, $S^{i,\nu}$ denotes the version of S where only the i-th point in S_ν is replaced by ν_i', with S_ω remaining unchanged. $S^{j,\omega}$ is defined similarly. The i.i.d. copied dataset S' is represented as $S'=S_\nu'\cup S_\omega'$, where $S_\nu'=\{\nu_1',\ldots,\nu_n'\}$ and $S_\omega'=\{\omega_1',\ldots,\omega_m'\}$. Note that the two-level optimization problem in (2) can also be considered as the compositional optimization (Yuan et al., 2019; Yang et al., 2019; Balasubramanian et al., 2022; Hu et al., 2023). Among the STORM-based algorithms for two-level stochastic optimization, we will analyze the stability and generalization of the most popular algorithms, COVER (Qi et al., 2021). Specifically, COVER utilizes two estimators for both the function and gradient values of the inner function, namely u_t and v_t.

K-level optimization. The K-level stochastic optimization problem (Chen et al., 2021; Jiang et al., 2022) can be formulated as follows

$$\begin{aligned}\min_{x\in\mathcal{X}}\Big\{F(x)&=f_K\circ f_{K-1}\circ\cdots\circ f_1(x)\\&=\mathbb{E}_{\nu^{(K)}}[f_K^{\nu^{(K)}}(\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(x)])]\Big\},\end{aligned} \tag{3}$$

where $f_k:\mathbb{R}^{d_{k-1}}\rightarrow\mathbb{R}^{d_k}$ on a convex domain $\mathcal{X}\in\mathbb{R}^d$, $k\in[1,k]$ and $d_0=d$. $\nu^{(k)}$ are independent random variables, where $k\in[1,K]$. Similarly, let $S=\cup_{k=1}^K S_k$, where $S_k=\{\nu_1^{(k)},\cdots,\nu_{n_k}^{(k)}\}$, the empirical risk is defined as $\min_{x\in\mathcal{X}}\{F_S(x):=f_{K,S}\circ f_{K-1,S}\cdots f_{1,S}=\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\cdots(\frac{1}{n_1}\sum_{i_1=1}^{n_1}f_1^{\nu_{i_1}^{(1)}}(A(S))))\}$. In the K-level optimization, where changing one sample data can occur in any layer of the function, we define: $S^{l,k}$ be the i.i.d. copy of S where only the l-th data point ν_l^k in S_k is replaced with $\nu_l^{k'}$, where $k\in[1,k]$ and $l\in[1,n_k]$. Moreover, we denote $S'=\cup_{i=1}^K S^{(i)}$, where $S^{(i)}=\{\nu_1^{(i)'},\cdots,\nu_{n_i}^{(i)'}\}$. In this scenario, we consider SVMR (Jiang et al., 2022) with multiple estimators, which obtains the best convergence result. In particular, $u^{(k)}$ represents the estimate of the k-th layer function value and $v^{(k)}$ represents the estimate of the k-th layer function's gradient value.

3.2. Concept of Excess Risk

As we all know, excess risk is an evaluation for the generalization performance (Bousquet & Elisseeff, 2002; James et al., 2013; Charles & Papailiopoulos, 2018), which is used to analyze the three tackled STORM-based algorithms in this paper. For a randomized algorithm A, denote by $A(S)$

its output model based on the training data S. By denoting $F(x_*) = \inf_{x\in\mathcal{X}} F(x)$ and $F(x_*^S) = \inf_{x\in\mathcal{X}} F_S(x)$, then the excess risk is $\mathbb{E}_{S,A}[F(A(S) - F(x_*)]$. According to the decomposition in (Bousquet & Elisseeff, 2002) and $F_S(x_*^S) \leq F_S(x_*)$ by the definition of x_*^S, we can obtain the excess risk as follows

$$\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] \leq \mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] + \mathbb{E}_{S,A}[F_S(A(S)) - F_S(x_*^S)]. \tag{4}$$

We refer to the term $\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))]$ as the generalization error, as it quantifies the generalization shift from training to testing behavior. Similarly, $\mathbb{E}_{S,A}[F_S(A(S)) - F_S(x_*^S)]$ is termed the optimization error, measuring how effectively the algorithm minimizes empirical risk. The generalization error in this paper is informed by analyses from prior studies (Cutkosky & Orabona, 2019; Qi et al., 2021; Jiang et al., 2022). Unlike these works, which primarily focus on convergence analysis, our main objective is to estimate the generalization error through the algorithmic stability approach (Bousquet & Elisseeff, 2002). Next, we provide the definitions of stability.

Definition 1 (Uniform Stability)**.** The uniform stability of the three stochastic optimizations is defined as follows

(i) In the one-level optimization, an algorithm A is uniformly stable for (1) if $\forall i \in [1, n]$, there holds $\mathbb{E}_A[\|A(S) - A(S^i)\|] \leq \epsilon$.

(ii) In the two-level optimization, an algorithm A is uniformly stable for (2), if $\forall i \in [1, n]$ and $\forall j \in [1, m]$, there holds $\mathbb{E}_A[\|A(S) - A(S^{i,\nu})\|] \leq \epsilon_\nu$ and $\mathbb{E}_A[\|A(S) - A(S^{j,\omega})\|] \leq \epsilon_\omega$.

(iii) In the K-level optimization, an algorithm A is uniformly stable for (3), if $\forall k \in [1, K]$ and $\forall l \in [1, n_k]$, there holds $\mathbb{E}_A[\|A(S) - A(S^{l,k})\|] \leq \epsilon_k$.

The expectation $\mathbb{E}_A[\cdot]$ is taken w.r.t. the internal randomness of A not the data points for the above definition.

We aim to elucidate the connection between uniform stability (as outlined in Definition 1) and the generalization error, a relation applicable across all randomized algorithms. To achieve this, we state the following assumption.

Assumption 1 (Lipschitz Continuity)**.** The Lipschitz continuity of our focused problems is proposed as follows

(i) In the one-level optimization problem, there exists a constant L_f, such that f_ν is Lipschitz continuous with parameters L_f, i.e., $\sup_\nu \|f_\nu(x) - f_\nu(\hat{x})\| \leq L_f\|x - \hat{x}\|$, for all $x, \hat{x} \in \mathbb{R}^d$.

(ii) In the two-level optimization problem, there exist two constants L_f and L_g, such that f_ν and g_ω are Lipschitz continuous with parameters L_f and L_g, respectively, i.e., $\sup_\nu \|f_\nu(y) - f_\nu(\hat{y})\| \leq L_f\|y - \hat{y}\|$ for all $y, \hat{y} \in \mathbb{R}^{d_1}$, and $\sup_\omega \|g_\omega(x) - g_\omega(\hat{x})\| \leq L_g\|x - \hat{x}\|$ for all $x, \hat{x} \in \mathbb{R}^d$.

(iii) In the K-level optimization problem, there exists a constant L_f, such that $\forall k \in [1, K]$, $f_k^{\nu^{(k)}}$ are Lipschitz continuous with parameter L_f, respectively, i.e., $\sup_{\nu^{(k)}} \|f_k^{\nu^{(k)}}(y) - f_k^{\nu^{(k)}}(\hat{y})\| \leq L_f\|y - \hat{y}\|$, $\forall\, y, \hat{y} \in \mathbb{R}^{d_{k-1}}$.

3.3. Generalization of the K-level Optimization

Although existing studies have established relationships between the generalization error and the stability under one-level (Hardt et al., 2016) and two-level (Yang et al., 2023) stochastic optimizations, the more complex and general K-level stochastic optimization remains unexplored. Therefore, by integrating the stability concept, we specifically define the following theorem for the K-level optimization, which aims to show the quantitative relationship between the generalization error and the stability.

Theorem 1. *If Assumption 1 (iii) holds true and the randomized algorithm A is uniformly stable, then for $K \geq 3$, $\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))]$ is bounded by*

$$L_f^K \epsilon_K + \sum_{k=1}^{K-1} \Big(4L_f^K \epsilon_k + L_f \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]}{n_k}} \Big),$$

where $\mathrm{Var}_k(A(S)) = \mathbb{E}_{\nu^{(k)}}[\|f_k \circ f_{k-1} \circ \cdots \circ f_1(A(S) - f_k^{\nu^{(k)}} \circ f_{k-1} \circ \cdots \circ f_1(A(S)\|^2]$.

Remark 1. Theorem 1 establishes the quantitative relationship between the generalization and the uniform stability for any randomized algorithm applied to K-level stochastic optimizations. In particular, when $K = 1$, i.e., the one-level stochastic optimization, where $F(x) = \mathbb{E}_\nu[f_\nu(x)]$ and $F_S(x) = \frac{1}{n}\sum_{i=1}^n f_{\nu_i}(x)$, we can see the absence of randomness with respect to ϵ_k, $\forall k \in [2, K]$. Consequently, we derive $\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] \leq L_f\epsilon$, consistent with the findings in (Hardt et al., 2016). For the two-level scenario, i.e., $k = 2$, we obtain $L_f^2\epsilon_2 + 4L_f^2\epsilon_1 + L_f\sqrt{\mathbb{E}_{S,A}[\mathrm{Var}_1(A(S))]/n_1}$. Here, the variance term $\mathbb{E}_{S,A}[\mathrm{Var}_1(A(S))]$ arises from the estimator used for the inner function values. We only need to alter the notations in Assumption 1 (iii) to obtain results consistent with (Yang et al., 2023).

Remark 2. In Theorem 1, we can find the generalization error depends not only on stability but also on the variance term, i.e., $\sqrt{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]/n_k}$ due to the estimators. An interesting observation is that the variance term is not only determined by the current layer function but also by the combined function of the total number of layers, i.e., for $\mathrm{Var}_k(A(S))$, which is determined by $f_k \circ f_{k-1} \circ \cdots \circ f_1$, instead of f_k. This implies that with an increasing number of levels, we should enlarge the sample size in order to achieve a better generalization error.

After establishing the quantitative relationship between the generalization error and the stability bound, the next goal is to establish stability bounds for these corresponding algorithms, i.e., STORM, COVER, and SVMR. In next section, we will introduce how to approach this in detail.

4. Stability and Generalization

In this section, we present the main results for various optimization problems, which include stability bounds and optimization errors, and ultimately derive the excess risks. Different results for the convex and strongly convex settings will be shown in separate subsections. Before giving the theoretical results, we state the following assumptions to facilitate our proofs.

Assumption 2 (Empirical Variance)**.** With probability 1 w.r.t. S, there exist constants to bound the following:

(i) In the one-level optimization problem, there exist two constants σ_f and σ_J, such that $\sup_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n}[\|f_{\nu_i}(x) - f_S(x)\|^2] \leq \sigma_f^2$ and $\sup_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n}[\|\nabla f_{\nu_i}(x) - \nabla f_S(x)\|^2] \leq \sigma_J^2$.

(ii) In the two-level optimization problem, there exist three constants σ_g, σ_g' and σ_f, such that $\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^{m}[\|g_{\omega_j}(x) - g_S(x)\|^2] \leq \sigma_g^2$, $\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^{m}[\|\nabla g_{\omega_j}(x) - \nabla g_S(x)\|^2] \leq \sigma_{g'}^2$ and $\sup_{y\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}[\|\nabla f_{\nu_i}(y) - \nabla f_S(y)\|^2] \leq \sigma_f^2$.

(iii) In the K-level optimization problem, there exist two constants σ_f and σ_J, such that for $1 \leq k \leq K$, there holds $\sup_{y\in\mathbb{R}_{d_{k-1}}} \frac{1}{n_k}\sum_{j=1}^{n_k}[\|f_k^{\nu^{(j)}}(y) - f_{k,S}(y)\|^2] \leq \sigma_f^2$ and $\sup_{y\in\mathbb{R}_{d_{k-1}}} \frac{1}{n_k}\sum_{j=1}^{n_k}[\|\nabla f_k^{\nu^{(j)}}(y) - \nabla f_{k,S}(y)\|^2] \leq \sigma_J^2$.

Assumption 3 (Smoothness and Lipschitz continuous gradient)**.** With probability 1 w.r.t. S, there exist constants to make following conditions hold true.

(i) In the one-level optimization, the problem $f_S(\cdot)$ is L-smooth, i.e., $\|\nabla f_\nu(x) - \nabla f_\nu(x')\| \leq L\|x - x'\|$, $\forall x, x' \in \mathcal{X}$.

(ii) In the two-level optimization, the problem $f_S(g_S(\cdot))$ is L-smooth, i.e., $\|\nabla g_S(x)\nabla f_S(g_S(x)) - \nabla g_S(x')\nabla f_S(g_S(x'))\| \leq L\|x - x'\|$, $\forall x, x' \in \mathcal{X}$. Also, $f_S(\cdot)$ has Lipschitz continuous gradients, i.e.,$\|\nabla f_S(y) - \nabla f_S(\bar{y})\| \leq C_f\|y - \bar{y}\|$ for all $y, \bar{y} \in \mathbb{R}^d$.

(iii) In the K-level optimization, the problem $F_S(\cdot)$ is L-smooth, i.e., $\|\Pi_{i=1}^{K}\nabla F_{i,S}(x) - \Pi_{i=1}^{K}\nabla F_{i,S}(x')\| \leq L\|x - x'\|$, $\forall x, x' \in \mathcal{X}$, where $\nabla F_{k,S}(x) = \nabla f_{k,S}(f_{k-1,S}(\cdots(f_{1,S}(x)))))$ and $\nabla F_{1,S}(x) = \nabla f_{1,S}(x)$. Additionally, $\forall k \in [1, K]$, the k-level function has Lipschitz continuous gradients, i.e., $\|\nabla f_{k,S}(y) - \nabla f_{k,S}(\bar{y})\| \leq L_f\|y - \bar{y}\|$ for all $y, \bar{y} \in \mathbb{R}^{d_{k-1}}$.

Assumptions 2-3 are widely used in convergence and generalization analysis (Charles & Papailiopoulos, 2018; Cutkosky & Orabona, 2019; Zhang et al., 2021; Qi et al., 2021; Jiang et al., 2022; Yang et al., 2023), which ensure the convergence and stability. It is important to note that Assumption 2 in generalization analysis shows the difference between the stochastic gradient and the empirical risk gradient $\nabla f_S(x)$. We also present the following definition for our focused settings, i.e., convex and strongly convex.

Definition 2. A function F is μ-strongly convex if for all $x, x' \in \mathcal{X}$, we have $F(x) \geq F(x') + \langle\nabla F(x'), x - x'\rangle + \frac{\mu}{2}\|x - x'\|^2$, and if $\mu = 0$, we say that F is convex.

4.1. Convex setting

Stability Results. The following theorems establish the uniform stability for the three optimizations under the convex setting, i.e., convex F_S. All the theoretical results in this subsection are under Assumptions 1-3.

Theorem 2 (One-level, Stability, Convex)**.** *Consider STORM in Algorithm 1 with* $\eta_t = \eta \leq \frac{2}{3L}$ *and* $\beta_t = \beta \in (0, 1)$, $\forall t \in [0, T-1]$. *Then, the outputs* $A(S) = x_T$ *at iteration* T *are uniformly stable with*

$$\epsilon = O\Big(\sup_S \eta \sum_{j=0}^{T-1} \mathrm{Var}(v_j) + \frac{L_f\eta T}{n}\Big),$$

where $\mathrm{Var}(v_j) = (\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}$.

Remark 3. We can find that in (Hardt et al., 2016), the uniform stability for SGD with the same setting is of the order $O(\frac{L_f\eta T}{n})$. However, using STORM adds another term $\sup_S \eta\sum_{j=0}^{T-1}\mathrm{Var}(v_j)$ caused by the estimator. This new term is determined by the difference between the estimate v_j and the gradient of the empirical risk $\nabla f_S(x_j)$. In other words, STORM may not be as stable as SGD.

Theorem 3 (Two-level, Stability, Convex)**.** *Consider COVER in Algorithm 2 with* $\eta_t = \eta \leq \frac{1}{4L}$ *and* $\beta_t = \beta \in (0, 1)$, $\forall t \in [0, T-1]$. *Then, the outputs* $A(S) = x_T$ *at iteration* T *are uniformly stable with*

$$\begin{aligned}\epsilon_\nu + \epsilon_\omega =& O\Big(L_g C_f \sup_S \eta \sum_{j=0}^{T-1}(\mathrm{Var}(u_j) + \mathrm{Var}(v_j)) \\ &+ L_f\sigma_f\eta\sqrt{T} + \frac{L_g L_f \eta T}{m} + \frac{L_g L_f \eta T}{n}\Big).\end{aligned}$$

where $\mathrm{Var}(u_j) = (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2}$ *and* $\mathrm{Var}(v_j) = (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2}$

Remark 4. When comparing the stability of COVER in Theorem 3 with STORM, particularly under the condition where $n = m$, COVER in the two-level stochastic optimization is characterized by two additional terms: $L_f \sigma_f \eta \sqrt{T}$ and $L_g C_f \sup_S \eta \sum_{j=0}^{T-1} \mathrm{Var}(v_j)$. The first term emerges due to the empirical error of the outer function. The second term is generated by the provided estimator from COVER for the inner function values, which accounts for the difference between the inner function estimator and the empirical risk of the inner function value.

Theorem 4 (K-level, Stability, Convex)**.** *Consider SVMR in Algorithm 3 with $\eta_t = \frac{2}{LK(K+2)}$ and $\beta_t = \beta \in (0,1)$, $\forall t \in [0, T-1]$. Then, the outputs $A(S) = x_T$ at iteration T are uniformly stable with*

$$\sum_{k=1}^{K} \epsilon_k = O\Big(\sup_S \eta \sum_{s=1}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{K-j+\frac{(i-1)i}{2}} \mathrm{Var}_{j,s}(u) + \sup_S \eta \sum_{s=1}^{T-1} \sum_{i=1}^{K} L_f^{K+\frac{(i-3)i}{2}} \mathrm{Var}_{i,s}(v) + \sum_{k=1}^{K} \frac{\eta L_f^K T}{n_k} \Big),$$

where $\mathrm{Var}_{j,s}(u) = (\mathbb{E}_A \|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}$ *and* $\mathrm{Var}_{i,s}(v) = (\mathbb{E}_A \|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}$.

Remark 5. Compared to the stability of COVER, especially when n_k is equal $\forall k \in [1, K]$, SVMR introduces additional terms due to its estimators. Let us discuss the term introduced by the function gradient estimator $\sup_S \eta \sum_{s=1}^{T-1} \sum_{i=1}^{K} L_f^{K+\frac{(i-3)i}{2}} \mathrm{Var}_{i,s}(v)$, accumulating an extra factor of K due to the need for K estimators to estimate the function gradient at each level. As for the term from the function value estimator $\sup_S \eta \sum_{s=1}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{K-j+\frac{(i-1)i}{2}} \mathrm{Var}_{j,s}(u)$, it becomes more complex in K-level optimization, involving three cumulative summations. This complexity arises from interactions between multiple levels, where estimators at different levels have influence instead of them at the same level. The derivatives of the function at the each level are affected by the function value estimator at the previous level and, in turn, impact the function value estimator at the next level, indicating their increased importance. The omitted term relates to the use of the gradient value estimator for the outer function and is equal to $L_f \sigma_f \eta \sqrt{T}$ in Theorem 4. This omission transforms the empirical variance of the outer function into a discrepancy between the gradient estimator and the empirical gradient value of the outer function.

Remark 6. Regardless of any algorithm, i.e., SGD or STORM-based, or any number of levels, the choice of step size η will affect the stability bound, which indicates proper selection of η. In addition, we can find that using fewer iterations can make the algorithms more stable, which may be a potential approach to enhance the generalization of STORM-based algorithms.

Combining Theorems 1 and 5, we have established generalization results for the three algorithms. To get excess risk bounds, we also need the optimization error results, i.e., $\mathbb{E}[F_S(A(S) - F_S(x_*^S)]$.

Generalization results. Before giving the theorems, we give some clarification. We use the assumption that the $\mathcal{X}$ domain is bounded in $\mathbb{R}^d$ to give the upper bound, i.e., $\mathbb{E}_A[\|x_t - x_*^S\|^2] \le D_x, \forall t \in [0, T-1]$. Let c be an arbitrary constant, the following three theorems hold.

Theorem 5 (Optimization, Convex)**.** *Let $A(S) = \frac{1}{T}\sum_{t=1}^{T} x_t$ be the solution produced by STORM, COVER, and SVMR in Algorithms 1-3, respectively. The following results bound the optimization error $\mathbb{E}[F_S(A(S)) - F_S(x_*^S)]$.*

(One-level). *For the problem in* (1)*, by selecting $\eta_t = \eta$ and $\beta_t = \beta$, then it holds*

$$O\Big(\frac{D_x}{\eta T} + (D_x + \sigma_J^2)\beta^{\frac{1}{2}} + L_f^2 \eta + V(T\beta)^{-c}\beta^{-\frac{1}{2}} + \frac{L_f^2 \eta^2}{\beta^{3/2}}\Big),$$

where $\mathbb{E}_A\|v_0 - \nabla f_S(x_0)\|^2 \le V$.

(Two-level). *For the problem in* (2)*, by selecting $\eta_t = \eta$ and $\beta_t = \beta$, then it holds*

$$O\Big(\frac{D_x}{\eta T} + \Phi_1 \beta^{\frac{1}{2}} + \Phi_2 \eta + \Phi_3 (T\beta)^{-c}\beta^{-\frac{1}{2}} + \frac{\Phi_4 \eta^2}{\beta^{3/2}}\Big),$$

where $\Phi_1 = L_g C_f \sigma_g^2 + L_f \sigma_{g'}^2 + (L_f + L_g C_f) D_x$, $\Phi_2 = L_g^2 L_f^2$, $\Phi_3 = L_g C_f U + L_f V$, $\Phi_4 = L_g^5 L_f^2 C_f + L_g^4 L_f^3$, $\mathbb{E}_A\|u_0 - g_S(x_0)\|^2 \le U$, and $\mathbb{E}_A\|v_0 - \nabla g_S(x_0)\|^2 \le V$.

(K-level). *For the problem in* (3)*, by selecting $\eta_t = \eta$ and $\beta_t = \beta < \max\left(\frac{1}{8K\sum_{i=1}^{K}(2L_f^2)^i}, 1\right)$, then it holds*

$$O\Big(\frac{D_x}{\eta T} + \Phi_5 \beta^{\frac{1}{2}} + L_f^K \eta + \Phi_6 (T\beta)^{-c}\beta^{-\frac{1}{2}} + \frac{\Phi_7 \eta^2}{\beta^{3/2}}\Big).$$

where $\Phi_5 = L_f^m(\sigma_f^2 + \sigma_J^2 + \sigma_f^2 \sum_{i=1}^{K} L_f^{2i} + D_x) + D_x$, $\Phi_6 = L_f^m(\sum_{i=1}^{K} U_i + V_i)$, $\Phi_7 = L_f^m \sum_{i=1}^{K} L_f^{2i}$, $L_f^m = \max(L_f^{K-j+\frac{(i-1)i}{2}}, L_f^{K+\frac{(i-3)i}{2}})$ for any $i, j \in [1, K]$, $\mathbb{E}_A\|u_1^{(i)} - f_{i,S}(u_0^{(i-1)})\|^2 \le U_i$, and $\mathbb{E}_A\|v_1^{(i)} - \nabla f_{i,S}(u_0^{(i-1)})\|^2 \le V_i$, $\forall i \in [1, K]$.

Remark 7. In Theorem 5, we can see that various factors affect optimization errors. Note that selecting β_t and η_t should be tailored to the specific requirements of different problems. In particular, when adjusting η_t to minimize the optimization error in one-level optimizations, η_t impacts $\frac{D_x}{\eta T}$, $L_f^K \eta$, and $\frac{L_f^2 \eta^2}{\beta^{3/2}}$. Unfortunately, the unknown value of L_f during training complicates determining the optimal η. In addition, each theorem features a term influenced by the first estimation error, i.e., $V(T\beta)^{-c}\beta^{-\frac{1}{2}}$, $\Phi_3(T\beta)^{-c}\beta^{-\frac{1}{2}}$, and $\Phi_6(T\beta)^{-c}\beta^{-\frac{1}{2}}$, where V, Φ_3, and Φ_6 all include the discrepancy between the estimators and the empirical risk at the first iteration. This suggests that employing a larger batch size to compute the estimators in the first iteration could effectively reduce the optimization error of the algorithm without significantly increasing computational costs.

By combining Theorems 1-4, we obtain the generalization error. Further, integrating this with the optimization error outlined in Theorem 5 allows us to derive the following excess risk bounds.

Theorem 6 (Excess Risk Bound, Convex). *Let $A(S) = \frac{1}{T}\sum_{t=1}^{T} x_t$ be the solution produced by STORM, COVER, and SVMR in Algorithms 1-3, respectively.*

(One-level). *For the problem in (1), by selecting $T \asymp n^{\frac{5}{2}}$, $\eta = T^{-\frac{4}{5}}$, and $\beta = T^{-\frac{4}{5}}$, we can obtain that $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\frac{1}{\sqrt{n}}\Big)$.*

(Two-level). *For the problem in (2), by selecting $T \asymp \max(n^{5/2}, m^{5/2})$, $\eta = T^{-\frac{4}{5}}$, and $\beta = T^{-\frac{4}{5}}$, we can obtain that $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)$.*

(K-level). *For the problem in (3), by selecting $T \asymp \max(n_k^{5/2})$, $\forall k \in [1, K]$, $\eta = T^{-\frac{4}{5}}$, and $\beta = T^{-\frac{4}{5}}$, we can obtain that $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\sum_{k=1}^{K} \frac{1}{\sqrt{n_k}}\Big)$.*

Remark 8. Theorem 6 demonstrates that STORM, by choosing $T \asymp n^{\frac{5}{2}}$ and appropriately selecting iteration number T and parameters η, β, achieves a generalization error rate of $O(\frac{1}{\sqrt{n}})$ in a convex setting. This is in contrast to SGD, which requires fewer iterations ($T \asymp n$) to reach the same bound (Hardt et al., 2016). This difference may be caused by the estimator in STORM, potentially leading to increased generalization error and excess risk due to reduced algorithm stability. This contrast is further highlighted when comparing with Theorem 6, where each additional level, denoted as $K+1$, requires reassessing iterations and selecting the maximum sample size $T \asymp \max(n_k^{5/2})$, $\forall k \in [1, K+1]$, which results in an incremental excess risk increase of $O(\frac{1}{\sqrt{n_{K+1}}})$ with each level while maintaining constant settings for η and β relative to T.

Remark 9. It should be noted that in Theorems 2-4, we discuss the stability of the final iterate $A(S) = x_T$. Conversely, in Theorem 5, we address the generalization bound of $A(S) = \frac{1}{T}\sum_{t=1}^{T} x_t$, representing the average of the intermediate iterates $x_1, \ldots, x_T$. This distinction arises from the understanding that generalization encompasses both stability and optimization. In the convex setting, the primary focus of optimization is often on the average of intermediate iterates, as exemplified in sources such as (Wang et al., 2017; Yang et al., 2023).

4.2. The Strongly Convex Setting

Note that we follow a similar process in the convex setting to analyze the generalization performance in the strongly convex setting.

Stability Results. The following theorem establishes the uniform Stability in the strongly convex setting. Before proceeding, we assume that Assumptions 1-3 and Definition 2 apply to F_S, which is strongly convex at the corresponding level, as outlined in Section 4.2.

Theorem 7 (One-level, Stability, Strongly Convex). *Consider STORM in Algorithm 1 with $\eta_t = \eta \leq \frac{2}{3(L+\mu)}$ and $\beta_t = \beta \in (0,1)$, $\forall t \in [0, T-1]$. Then, the outputs $A(S) = x_T$ at iteration T are uniform stable with*

$$\epsilon = O\Big(\eta \sum_{j=0}^{T-1} (1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1} \operatorname{Var}(v_j) + \frac{L_f(L+\mu)}{L\mu n}\Big).$$

Theorem 8 (Two-level, Stability, Strongly Convex). *Consider COVER in Algorithm 2 with $\eta_t = \eta \leq \frac{1}{4L+4\mu}$ and $\beta_t = \beta \in (0,1)$, $\forall t \in [0, T-1]$ and the output $A(S) = x_T$. Then, the outputs $A(S) = x_T$ at iteration T are uniform stable with*

$$\begin{aligned}\epsilon_\nu + \epsilon_\omega = O\Big(& L_g C_f \eta \sup_S \sum_{j=0}^{T-1} (1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1} \operatorname{Var}(u_j) \\ &+ L_f \eta \sup_S \sum_{j=0}^{T-1} (1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1} \operatorname{Var}(v_j) \\ &+ \frac{(L+\mu)L_g L_f}{L\mu m} + \frac{(L+\mu)L_g L_f}{L\mu n} + L_g \sigma_f \sqrt{\frac{L+\mu}{L\mu}} \sqrt{\eta}\Big).\end{aligned}$$

Theorem 9 (K-level, Stability, Strongly Convex). *Consider SVMR in Algorithm 3 with $\eta_t = \eta \leq \frac{2}{(L+\mu)K(K+2)}$ and $\beta_t = \beta \in (0,1)$, $\forall t \in [0, T-1]$ and the output $A(S) = x_T$. Then, the outputs $A(S) = x_T$ at iteration T are uniform stable with*

$$\begin{aligned}&\sum_{k=1}^{K} \epsilon_k \\ &= O\Big(\sum_{s=1}^{T-1} \sum_{i=1}^{K} \Big(1 - \frac{2\eta L\mu}{L+\mu}\Big)^{T-s} \eta L_f^{K + \frac{(i-3)i}{2}} \operatorname{Var}_{i,s}(v) \\ &+ \sum_{s=1}^{T-1} \sum_{i=1}^{K} \sum_{j=1}^{i-1} \Big(1 - \frac{2\eta L\mu}{L+\mu}\Big)^{T-s} \eta L_f^{K-j+\frac{(i-1)i}{2}} \operatorname{Var}_{j,s}(u) \\ &+ \sum_{k=1}^{K} \frac{L_f^K (L+\mu)}{L\mu n_k}\Big).\end{aligned}$$

Remark 10. Many conclusions from the strongly convex setting align with the convex setting, and we analyze them individually. First, in the one-level stochastic optimization, the stability of SGD is of the order $O(\frac{1}{\mu n})$ in (Hardt et al., 2016). Compared to SGD, our results include an additional term, $\eta \sum_{j=0}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}\operatorname{Var}(v_j)$, which is the same as in the convex setting. This implies that STORM may also be less stable under the strongly convex setting than SGD. Second, in the two-level scenario, considering $m = n$, COVER introduces two additional terms. The reasons for these terms are the same as under the convex setting, stemming from the additional estimator used and the empirical variance of the outer function. Lastly, in K-level optimization, SVMR includes only one additional coefficient, $(1-\frac{2\eta L\mu}{L+\mu})^{T-s}$, due to the strongly convex property.

Remark 11. Note that there are some significant differences in the strongly convex setting compared to the convex setting. Under the strongly-convex setting, each situation includes an item, such as $\sum_{k=1}^{K}\frac{L_f^K(L+\mu)}{L\mu n_k}$, that is independent of the step size but depends on the sample size used by each layer function. Therefore, in strongly convex settings, achieving satisfactory stability may require more than just selecting an appropriate step size; it becomes imperative to increase the sample size simultaneously to improve stability.

Generalization results. Let c be an arbitrary constant, the following theorems hold, which aim to show the optimization errors in the strongly convex setting.

Theorem 10 (Optimization, Strongly Convex)**.** *Let* $A(S) = (\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}x_t)/(\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t})$ *be the solution produced by STORM, COVER, and SVMR in Algorithms 1-3, respectively. The following results bound the optimization error* $\mathbb{E}[F_S(A(S)) - F_S(x_*^S)]$.

(One-level). *For the problem in* (1)*, by selecting* $\eta_t = \eta \leq \frac{2}{3(L+\mu)}$ *and* $\beta_t = \beta \in (0,1)$*, then it holds*

$$O\Big(\frac{D_x + U\eta}{(\eta T)^c} + L_f^2 L\eta + \frac{V}{(\beta T)^c} + \sigma_J^2\beta + \frac{L_f^2\eta^2}{\beta}\Big).$$

(Two-level). *For the problem in* (2)*, by selecting* $\eta_t = \eta$*, and* $\beta_t = \beta < \min\left(\frac{1}{8C_f^2}, 1\right)$*, then it holds*

$$O\Big(\frac{D_x + \Psi_1\eta}{(\eta T)^c} + LL_g^2L_f^2\eta + \frac{\Psi_2}{(\beta T)^c} + \Psi_3\beta + \frac{\Psi_3\eta^2}{\beta}\Big),$$

where $\Psi_1 = L_g^2C_f^2U + L_f^2V$, $\Psi_2 = L_g^2C_f^2\sigma_g^2 + L_f^2\sigma_{g'}^2$*, and* $\Psi_3 = L_g^6C_f^2L_f^2 + L_f^4L_g^4$.

(K-level). *For the problem in* (3)*, by selecting* $\eta_t = \eta$ *and* $\beta_t = \beta < \max\left(\frac{1}{8K\sum_{i=1}^{K}(2L_f^2)^i}, 1\right)$*, then it holds*

$$O\Big(\frac{D_x + \Psi_4\eta}{(\eta T)^c} + \eta L_f^K + \frac{\Psi_5}{(\beta T)^c} + \Psi_6\beta + \frac{\Psi_7\eta^2}{\beta}\Big).$$

where $\Psi_4 = L_f^m\sum_{j=1}^{K-1}(U_i + V_i)$, $\Psi_5 = L_f^m\sum_{i=1}^{K}(U_i + V_i)$, $\Psi_6 = L_f^m(\sigma_f^2 + \sigma_J^2 + \sigma_f^2(\sum_{i=1}^{K}(L_f^2)^i))$*, and* $\Psi_7 = L_f^m\sum_{i=1}^{K}(L_f^2)^i$.

Now, we come to derive the following excess risk bounds for the strongly convex setting.

Theorem 11 (Excess Risk Bound, Strongly Convex)**.** *Let* $A(S) = (\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}x_t)/(\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t})$ *be the solution produced by STORM, COVER, and SVMR in Algorithms 1-3, respectively.*

(One-level). *For the problem in* (1)*, by selecting* $T \asymp n^{\frac{7}{6}}$, $\eta = T^{-\frac{6}{7}}$*, and* $\beta = T^{-\frac{6}{7}}$*, we can obtain that* $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\frac{1}{\sqrt{n}}\Big)$.

(Two-level). *For the problem in* (2)*, by selecting* $T \asymp \max(n^{7/6}, m^{7/6})$ *and* $\eta = \beta = T^{-\frac{7}{6}}$*, we can obtain that* $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)$.

(K-level). *For the problem in* (3)*, by selecting* $T \asymp \max(n_k^{7/6})$, $\forall k \in [1, K]$ *and* $\eta = \beta = T^{-\frac{7}{6}}$*, we can obtain that* $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\sum_{k=1}^{K}\frac{1}{\sqrt{n_k}}\Big)$.

Remark 12. Theorem 11 demonstrates that, in the case of strong convexity, the generalization error for STORM can attain a rate of $O(\frac{1}{\sqrt{n}})$ by carefully choosing the iteration number T, along with constant step sizes η and β. We can find that under the strongly convex setting, we only need iteration $T \asymp n^{\frac{7}{6}}$, however, under the convex setting, we need more iteration $T \asymp n^{\frac{5}{2}}$. Summarizing these three theorems, we can easily discern the relationship between the excess risk bound and the number of levels. This conclusion is very similar to that in the convex setting. Specifically, for each additional level, denoted as $K+1$, it is necessary to reassess iterations and select the maximum sample size $T \asymp \max(n_k^{7/6})$, $\forall k \in [1, K+1]$. This results in an incremental excess risk increase of $O(\frac{1}{\sqrt{n_{K+1}}})$ with each level, while η and β remain constant relative to T.

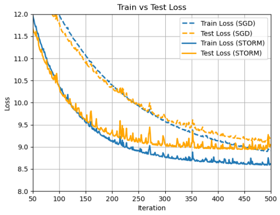
To make our paper easy to understand, Table 1 lists all of our theoretical results in Appendix A.

5. Experiments

In this section, we carried out a series of experiments using simulated data to validate our theoretical findings, consisting of four separate tests.

First, we examined the performance of STORM versus SGD in fitting a univariate quintic polynomial. We generated 2000 data points based on this polynomial and introduced Gaussian noise with a mean of 0 and variance of 3. The data was divided into a training and testing split of 60/40. Throughout 500 iterations, using a step size of 0.001 and a batch size of 128, we monitored both training and testing losses using the mean squared error metric. Although STORM demonstrated poorer generalization, indicated by a larger discrepancy between training and testing losses, it outperformed SGD in overall loss metrics.

Second, we investigated how varying the number of levels, k, affects generalization error within a two-level optimization framework. We represented our target function as $F(x) = f(g(\cdot))$, creating two sets of data points, S_1 and S_2, each contaminated with Gaussian noise (mean 0, variance 3). The dataset was split into a 60/40 train-test ratio. The goal was to optimize $g(\cdot)$ to fit S_1 and $f(\cdot)$ to fit S_2 using SVMR as the optimizer, with a step size of 0.01, a projection operation L_f set at 50, and a batch size of 128 over 500 iterations. We recorded the average generalization error during the last 10 iterations while incrementally increasing the level count from 1 to 50. Our results showed a steady rise in generalization error as the number of levels increased, particularly intensifying beyond 35 levels.



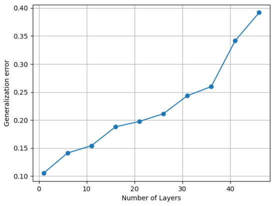
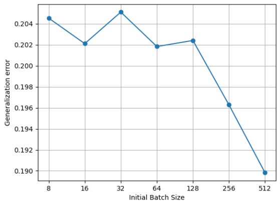


Figure 1. SGD VS STORM.

Figure 2. Effect of Level.

Third, we explored the impact of the initial iteration batch size on generalization. In this experiment, we maintained a fixed number of levels $k = 10$, with other parameters consistent with above, and varied only the batch size during the first five iterations before stabilizing it at 128. We observed that when the initial batch size is smaller than the standard value of 128, the generalization error is higher than at 128. Conversely, setting the initial batch size to 256 and 512 significantly improved the generalization error. This finding supports our observation that under the same initial conditions, increasing the batch size in the initial few iterations can enhance the generalization performance of SVMR.

Fourth, we investigated the impact of noise on generalization. In this experiment, while keeping the settings consistent with Experiment 2, we set $k = 10$ and maintained the batch size at 128. However, we varied the variance of Gaussian distribution noise. Specifically, we incrementally increased the Gaussian noise variance from 0.1 to 3 in steps of 0.1 to observe its effects on generalization. Noise can improve generalization by 1) aiding the model in escaping local minima to find lower values, and 2) preventing the model from overfitting the training data. The drawback of noise in terms of generalization is that it challenges an algorithm's stability; excessive noise can compromise this stability, thereby diminishing generalization performance. Our results indicated that when the noise variance does not exceed 1.5, it positively impacts generalization. However, beyond a variance of 1.5, the detrimental effects on algorithm stability outweigh the benefits, leading to poorer generalization outcomes.



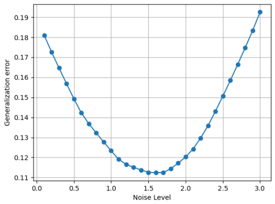


Figure 3. Effect of batch size.

Figure 4. Effect of noise.

6. Conclusion

This paper conducts a thorough generalization analysis of STORM-based algorithms: STORM, COVER, and SVMR, for one, two, and K-level stochastic optimizations. Firstly, for the K-level optimization, we introduce a tailored stability notion, paving the way for deeply understanding the relationship between generalization error, stability, and the number of levels. We further investigate their stability and excess risk bounds in both convex and strongly convex settings. Based on our analysis, we have found three observations for STORM-based algorithms: (1) Individual estimators can compromise algorithm stability due to target variances, harming generalization performance. (2) Increasing the number of levels also affects the algorithm's generalization error through stability and gradient variances. (3) Using more initial samples for estimation can boost performance without significantly raising computational costs.

Impact Statement

This paper contributes to the advancement of the Machine Learning field. While recognizing the potential societal consequences of our work, we believe it is unnecessary to specifically highlight any particular implications here.

Acknowledgements

This work at Central South University was supported by the National Science Foundation of China (NSFC) under Grants 62302525 and 62172444, the Natural Science Foundation of Hunan Province under Grant 2022JJ30753, the Hunan Provincial Education Department's Scientific Research Fund under Grant 23A0020, the Central South University Innovation-Driven Research Programme under Grant

2023CXQD018, and the High Performance Computing Center of Central South University.

References

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

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Charles, Z. and Papailiopoulos, D. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pp. 745–754. PMLR, 2018.

Chen, T., Sun, Y., and Yin, W. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

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

Dann, C., Neumann, G., Peters, J., et al. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

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

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Hu, Q., Zhu, D., and Yang, T. Non-smooth weakly-convex finite-sum coupled compositional optimization. *arXiv preprint arXiv:2310.03234*, 2023.

Hu, W., Li, C. J., Lian, X., Liu, J., and Yuan, H. Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Ji, K., Yang, J., and Liang, Y. Theoretical convergence of multi-step model-agnostic meta-learning. *The Journal of Machine Learning Research*, 23(1):1317–1357, 2022.

Jiang, W., Wang, B., Wang, Y., Zhang, L., and Yang, T. Optimal algorithms for stochastic multi-level compositional optimization. In *International Conference on Machine Learning*, pp. 10195–10216. PMLR, 2022.

Kearns, M. and Ron, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory*, pp. 152–162, 1997.

Kutin, S. and Niyogi, P. Almost-everywhere algorithmic stability and generalization error. *arXiv preprint arXiv:1301.0579*, 2012.

Levy, K. Y., Kavis, A., and Cevher, V. Storm+: Fully adaptive sgd with momentum for nonconvex optimization. *arXiv preprint arXiv:2111.01040*, 2021.

Li, X., Qu, Z., Tang, B., and Lu, Z. Fedlga: Toward system-heterogeneity of federated learning via local gradient approximation. *IEEE Transactions on Cybernetics*, 2023a.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023b.

Liu, J., Pan, X., Duan, J., Li, H.-D., Li, Y., and Qu, Z. Faster stochastic variance reduction methods for compositional minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13927–13935, 2024.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Mao, W., Yang, L., Zhang, K., and Basar, T. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 15007–15049. PMLR, 2022.

Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. General conditions for predictivity in learning theory. *Nature*, 428 (6981):419–422, 2004.

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in neural information processing systems*, 34:1752–1765, 2021.

Qu, Z., Li, X., Han, X., Duan, R., Shen, C., and Chen, L. How to prevent the poor performance clients for personalized federated learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12167–12176, 2023a.

Qu, Z., Li, X., Xu, J., Tang, B., Lu, Z., and Liu, Y. On the convergence of multi-server federated learning with overlapping area. *IEEE Transactions on Mobile Computing*, 22(11):6647–6662, 2023b.

Rakhlin, A., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

Sakaue, S. and Oki, T. Improved generalization bound and learning of sparsity patterns for data-driven low-rank approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1–10. PMLR, 2023.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pp. 21146–21179. PMLR, 2022.

Tran Dinh, Q., Liu, D., and Nguyen, L. Hybrid variance-reduced sgd algorithms for minimax problems with nonconvex-linear function. *Advances in Neural Information Processing Systems*, 33:11096–11107, 2020.

Vapnik, V. and Chapelle, O. Bounds on error expectation for support vector machines. *Neural computation*, 12(9): 2013–2036, 2000.

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

Wen, Y., Vicol, P., Ba, J., Tran, D., and Grosse, R. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

Yang, M., Wei, X., Yang, T., and Ying, Y. Stability and generalization of stochastic compositional gradient descent algorithms. *arXiv preprint arXiv:2307.03357*, 2023.

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

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

Yuan, Z., Guo, Z., Chawla, N., and Yang, T. Compositional training for end-to-end deep auc maximization. In *International Conference on Learning Representations*, 2021.

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

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

Zhang, J., Hong, M., Wang, M., and Zhang, S. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 568–576. PMLR, 2021.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. *The Journal of Machine Learning Research*, 21(1):4130–4192, 2020.

A. Results Summary and Corresponding Algorithms

A.1. Summary of Results

Table 1. Summary of our results.

Setting	Bound	Level	Reference	Result
	Generation	1	(Hardt et al., 2016)	$L_f \epsilon$
		2	(Yang et al., 2023)	$L_f^2 \epsilon_2 + 4L_f^2 \epsilon_1 + L_f \sqrt{\mathbb{E}_{S,A}[\mathrm{Var}_1(A(S))]/n_1}$
		K	Theorem 1	$L_f^K \epsilon_K + \sum_{k=1}^{K-1} \Big(4L_f^K \epsilon_k + L_f \sqrt{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]/n_k}\Big)$
C	Stability	1	Theorem 2	$O\Big(\eta \sum_{j=0}^{T-1} \mathrm{Var}(v_j) + \frac{L_f \eta T}{n}\Big)$
		2	Theorem 3	$O\Big(\eta \sum_{j=0}^{T-1} (\mathrm{Var}(u_j) + \mathrm{Var}(v_j)) + \eta\sqrt{T} + \frac{\eta T}{m} + \frac{\eta T}{n}\Big)$
		K	Theorem 4	$O\Big(\tilde{L}_f^{K,i} \sum_{j=1}^{i-1} L_f^{i-j} \mathrm{Var}^T(u,v) + \sum_{k=1}^{K} \frac{\eta L_f^K T}{n_k}\Big)$
	Excess Risk	1	Theorem 6	$O(\frac{1}{\sqrt{n}}),\ T \asymp n^{5/2}$
		2	Theorem 6	$O\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big),\ T \asymp \max(n^{5/2}, m^{5/2})$
		K	Theorem 6	$O\Big(\sum_{k=1}^{K} \frac{1}{\sqrt{n_k}}\Big),\ T \asymp \max(n_k^{5/2}),\ \forall k \in [1,K]$
SC	Stability	1	Theorem 7	$O\Big(\eta \sum_{j=0}^{T-1} \tilde{L}^{T-j-1} \mathrm{Var}(v_j) + \frac{L_f(L+\mu)}{L\mu n}\Big)$
		2	Theorem 8	$O\Big(\eta \sum_{j=0}^{T-1} \tilde{L}^{T-j-1} (\mathrm{Var}(u_j) + \mathrm{Var}(v_j)) + \frac{(L+\mu)L_g L_f}{L\mu m} + \frac{(L+\mu)L_g L_f}{L\mu n}\Big)$
		K	Theorem 9	$O\Big(\eta \sum_{s=1}^{T-1} \tilde{L}^{T-s} \tilde{L}_f^{K,i} \sum_{j=1}^{i-1} L_f^{i-j} \mathrm{Var}^T(u,v) + \sum_{k=1}^{K} \frac{L_f^K (L+\mu)}{L\mu n_k}\Big)$
	Excess Risk	1	Theorem 11	$O\Big(\frac{1}{\sqrt{n}}\Big),\ T \asymp n^{7/6}$
		2	Theorem 11	$O\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}),\ T \asymp \max(n^{7/6}, m^{7/6})$
		K	Theorem 11	$O\Big(\sum_{k=1}^{K} \frac{1}{\sqrt{n_k}}\Big),\ T \asymp \max(n_k^{5/2}),\ \forall k \in [1,K]$

We use the following parameters to simplify the notations: $\mathrm{Var}^T(u,v) = \sum_{s=1}^{T-1} (\mathrm{Var}_{j,s}(u) + \mathrm{Var}_{i,s}(v))$, $\tilde{L} = (1 - \frac{2\eta L\mu}{L+\mu})$, and $\tilde{L}_f^{K,i} = \sum_{i=1}^{K} L_f^{K+\frac{(i-3)i}{2}}$

A.2. Description of Algorithms

Algorithm 1 STORM.

```
Inputs: Training data $S = \{\nu_i : i = 1, \cdots, n\}$; Number of iterations $T$; Parameter $\eta_t$, $\beta_t$
Initialize $x_0 \in \mathcal{X}$, $v_0 \in \mathbb{R}^d$
Draw a sample $j_0 \in [1,n]$, obtain $\nabla f_{\nu_{j_0}}(x_0)$.
for $t = 0$ to $T-1$ do
    $x_{t+1} = x_t - \eta_t v_t$
    Draw a sample $j_{t+1} \in [1,n]$, obtain $\nabla f_{\nu_{j_{t+1}}}(x_t)$
    Compute estimators $v_{t+1} = \Pi_{L_f}[\nabla f_{\nu_{j_{t+1}}}(x_{t+1}) + (1-\beta_{t+1})(v_t - \nabla f_{\nu_{j_{t+1}}}(x_t))]$
end for
Outputs: $A(S) = x_T$ or $x_\tau \sim \mathtt{Unif}(\{x_t\}_{t=1}^T)$
```

Algorithm 2 COVER.

```
Inputs: Training data $S_\nu = \{\nu_i : i = 1, \cdots, n\}$, $S_\omega = \{\omega_j : j = 1, \cdots, m\}$; Number of iterations $T$, Parameter $\eta_t$, $\beta_t$
Initialize $x_0 \in \mathcal{X}$, $u_0, v_0 \in \mathbb{R}^d$
Draw a sample $j_0 \in [1, n]$ and , $i_0 \in [1, m]$, obtain $\nabla g_{\omega_{j_0}}(x_0)$ and $\nabla f_{\nu_{i_0}}(u_0)$.
for $t = 0$ to $T - 1$ do
    $x_{t+1} = x_t - \eta_t v_t \nabla f_{\nu_{i_t}}(u_t)$
    Draw a sample $j_{t+1} \in [1, m]$, obtain $g_{\omega_{j_{t+1}}}(x_{t+1})$ and $g_{\omega_{j_{t+1}}}(x_t)$
    Compute estimators $u_{t+1} = g_{\omega_{j_{t+1}}}(x_{t+1}) + (1 - \beta_{t+1})(u_t - g_{\omega_{j_{t+1}}}(x_t))$
    Draw a sample $j_{t+1} \in [1, m]$ , obtain $\nabla g_{\omega_{j_{t+1}}}(x_{t+1})$ and $\nabla g_{\omega_{j_{t+1}}}(x_t)$
    Compute estimators $v_{t+1} = \Pi_{L_f}[\nabla g_{\omega_{j_{t+1}}}(x_{t+1}) + (1 - \beta_{t+1})(v_t - \nabla g_{\omega_{j_{t+1}}}(x_t)]$
    Draw samples $i_{t+1} \in [1, n]$, obtain $\nabla f_{\nu_{i_{t+1}}}(u_{t+1})$
end for
Outputs: $A(S) = x_T$ or $x_\tau \sim \texttt{Unif}(\{x_t\}_{t=1}^T)$
```

Algorithm 3 SVMR.

```
Inputs: Training data $S = \{\nu_1^{(1)}, \cdots, \nu_{n_1}^{(1)}, \cdots, \nu_1^{(K)}, \cdots, \nu_{n_K}^{(K)}\}$.; Number of iterations $T$; Parameter $\eta_t$, $\beta_t$
Initialize $x_0 \in \mathcal{X}$, $u_0^{(i)}, v_0^{(i)} \in \mathbb{R}^d$ for all $i \in [0, K]$
Draw a sample $j_0 \in [1, n]$ and , $i_0 \in [1, m]$, obtain $\nabla g_{\omega_{j_0}}(x_0)$ and $\nabla f_{\nu_{i_0}}(u_0)$.
for $t = 0$ to $T - 1$ do
    $x_{t+1} = x_t - \eta_t \prod_{i=1}^K v_t^{(i)}$ and set $u_t^{(0)} = x_t$
    for level $i = 1$ to $K$ do
        Draw a sample $\nu_{t+1}^{(i)} \in [1, n_i]$, obtain $f_{\nu_{t+1}^{(i)}}(u_{t+1}^{(i-1)})$, $f_{\nu_{t+1}^{(i)}}(u_t^{(i-1)})$, $\nabla f_{\nu_{t+1}^{(i)}}(u_{t+1}^{(i-1)})$ and $\nabla f_{\nu_{t+1}^{(i)}}(u_t^{(i-1)})$
        Compute estimators $u_{t+1}^{(i)} = f_{\nu_{t+1}^{(i)}}(u_{t+1}^{(i-1)}) + (1 - \beta_{t+1})(u_t^{(i)} - f_{\nu_{t+1}^{(i)}}(u_t^{(i-1)}))$
        Compute estimators $v_{t+1}^{(i)} = \Pi_{L_f}[\nabla f_{\nu_{t+1}^{(i)}}(u_{t+1}^{(i-1)}) + (1 - \beta_{t+1})(u_t^{(i)} - \nabla f_{\nu_{t+1}^{(i)}}(u_t^{(i-1)}))]$
    end for
end for
Outputs: $A(S) = x_T$ or $x_\tau \sim \texttt{Unif}(\{x_t\}_{t=1}^T)$
```

B. Useful Lemmas

Before giving the detailed proof, we first give some useful lemmas.

Lemma 12 (Lemma 4 in (Yang et al., 2023)). *Let $\{a_i\}_{i=1}^T, \{b_i\}_{i=1}^T$ be two sequences of positive real numbers such that $a_i \leq a_{i+1}$ and $b_i \geq b_{i+1}$ for all i. Then we have $\frac{\sum_{i=1}^T a_i b_i}{\sum_{i=1}^T a_i} \leq \frac{\sum_{i=1}^T b_i}{T}$.*

Lemma 13. *Consider a sequence $\{\beta_t\}_{t\geq 0} \in (0,1]$ and define $\Upsilon_t = \prod_{i=1}^t (1-\beta_i)$, then we can get for any $q_t \leq (1-\beta_t)q_{t-1} + p_t$, $q_t \leq \Upsilon_t (q_0 + \sum_{i=1}^t \frac{p_i}{\Upsilon_i})$.*

Proof. We divide both side of $q_t \leq (1-\beta_t)q_{t-1} + p_t$ by Υ_t, then we have $\frac{q_t}{\Upsilon_t} \leq \frac{q_{t-1}}{\Upsilon_{t-1}} + \frac{p_t}{\Upsilon_t}$, $t \geq 1$. Summing up the above inequalities, we have $q_t \leq \Upsilon_k (q_0 + \sum_{i=1}^t \frac{p_i}{\Upsilon_i})$. □

Lemma 14 (Lemma 2 in (Yang et al., 2023)). *Assume that the non-negative sequence $u_t : t \in \mathbb{N}$ satisfies the following recursive inequality for all $t \in \mathbb{N}$,*

$$u_t^2 \leq S_t + \sum_{\tau=1}^{t-1} \alpha_\tau u_\tau.$$

where $\{S_\tau : \tau \in \mathbb{N}\}$ is an increasing sequence, $S_0 \geq u_0^2$ and α_τ for any $\tau \in \mathbb{N}$. Then, the following inequality holds true:

$$u_t \leq \sqrt{S_t} + \sum_{\tau=1}^{t-1} \alpha_\tau.$$

C. One-level Stochastic Optimizations

Lemma 15 (Theorem 3.7 in (Hardt et al., 2016)). *If Assumption 1(i), 2 (i) and 3 (i) holds true and the randomized algorithm A is ϵ-uniformly stable then*

$$\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] \leq L_f \epsilon.$$

Lemma 16 (Lemma 2 in (Cutkosky & Orabona, 2019)). *Let Assumption 1(i), 2 (i) and 3 (i) holds hold for the empirical risk F_S , and x_t, v_t is generated by Algorithm 1, then we have*

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2 | \mathcal{F}_t] \leq (1-\beta_t)\|v_{t-1} - \nabla f_S(x_{t-1})\|^2 | + 2\beta_t^2 \sigma_J^2 + 2L_f^2 \|x_t - x_{t-1}\|^2.$$

Lemma 17. *Let Assumption 1(i), 2 (i) and 3 (i) holds hold for the empirical risk F_S , and x_t, v_t is generated by Algorithm 1, then for any $c > 0$, we have*

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq (\frac{c}{e})^c (t\beta)^{-c} \mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2 \eta^2}{\beta}.$$

proof of lemma 17. According to Lemma 16, and note that $\mathbb{E}_A[\|x_t - x_{t-1}\|^2] \leq L_f^2 \eta_{t-1}^2$ we have

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq (1-\beta_t)\mathbb{E}_A[\|v_{t-1} - \nabla f_S(x_{t-1})\|^2] + 2\beta_t^2 \sigma_J^2 + L_f^2 \eta_{t-1}^2.$$

Telescoping the above inequality from 1 to t, according to Lemma 13, we have

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq \prod_{j=1}^t (1-\beta_j)\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + \prod_{j=1}^t (1-\beta_j)(\sum_{j=1}^t \frac{2\beta_j \sigma_J^2}{\prod_{i=1}^j (1-\beta_i)})$$
$$+ \prod_{j=1}^t (1-\beta_j)(\sum_{j=1}^t \frac{L_f^2 \eta_{j-1}^2}{\prod_{i=1}^j (1-\beta_i)}).$$

Setting $\beta_t = \beta$ and $\eta_t = \eta$, we have

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq \prod_{j=1}^t (1-\beta_j)\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + \sum_{j=1}^t (1-\beta)^{t-j}(2\beta^2 \sigma_J^2 + L_f^2 \eta^2).$$

Note that for all $K \leq N$ and $\beta_i > 0$, we have

$$\prod_{i=K}^{N}(1-\beta_i) \leq \exp(-\sum_{i=K}^{N}\beta_i), \tag{5}$$

then we have

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq \exp(-\beta t)\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + \sum_{j=1}^{t}(1-\beta)^{t-j}(2\beta^2\sigma_J^2 + L_f^2\eta^2).$$

According to the fact that for any $c > 0$, we have

$$e^{-x} \leq (\frac{c}{e})^c x^{-c}, \tag{6}$$

then we can get for any $c > 0$

$$\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq (\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + \sum_{j=1}^{t}(1-\beta)^{t-j}(2\beta^2\sigma_J^2 + L_f^2\eta^2).$$

Moreover, according to the fact that

$$\sum_{j=1}^{t}(1-\beta)^{t-j} \leq \frac{1}{\beta}, \tag{7}$$

we have $\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2] \leq (\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta}$. □

We first give some notations used in the one-level optimization to simplify our proof.

For any $k \in [1, n]$, let $S^k = \{\nu_1, \ldots, \nu_{k-1}, \nu_k', \nu_{k+1}, \ldots, \nu_n\}$ be formed from S by replacing the k-th element.

Let $\{x_{t+1}\}$, and $\{v_{t+1}\}$ be generated by Algorithm 1 based on S. Similarly, $\{x_{t+1}^k\}$ and $\{v_{t+1}^k\}$ be generated by Algorithm 1 based on S^k. Set $x_0 = x_0^k$ as starting points in $\mathcal{X}$.

Next, we give the detailed proof of Theorem 2.

proof of Theorem 2 . We will consider two cases, i.e., $i_t \neq k$ and $i_t = k$.

Case 1 ($i_t \neq k$)**.** We have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^k\|^2 &= \|x_t - \eta_t v_t - x_t^k + \eta_t v_t^k\|^2 \\ &\leq \|x_t - x_t^k\|^2 - 2\eta_t\langle v_t - v_t^k, x_t - x_t^k\rangle + \eta_t^2\|v_t - v_t^k\|^2.\end{aligned} \tag{8}$$

For the second term on the RHS of (8), we have

$$\begin{aligned}&-2\eta_t\langle v_t - v_t^k, x_t - x_t^k\rangle \\ &= -2\eta_t\langle v_t - \nabla f_S(x_t), x_t - x_t^k\rangle - 2\eta_t\langle \nabla f_S(x_t) - \nabla f_S(x_t^k), x_t - x_t^k\rangle - 2\eta_t\langle \nabla f_S(x_t^k) - v_t^k, x_t - x_t^k\rangle.\end{aligned}$$

Smoothness generally suggests that the gradient update of F is constrained from being excessively large. Additionally, the convexity and L-smoothness of F indicate co-coercivity in the gradients, leading to the following conclusion

$$\langle \nabla F(x) - \nabla F(x'), x - x'\rangle \geq \frac{1}{L}\|\nabla F(x) - \nabla F(x')\|^2.$$

Then using Assumption 3 (i), i.e., the smoothness of $f_S(\cdot)$, we can get

$$\begin{aligned}&-2\eta_t\langle v_t - v_t^k, x_t - x_t^k\rangle \\ &\leq 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| - \frac{2\eta_t}{L}\|\nabla f_S(x_t) - \nabla f_S(x_t^k)\|^2 + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\|.\end{aligned}$$

For the third term on the RHS of (8), we have

$$\eta_t^2\|v_t - v_t^k\|^2 \le 3\eta_t^2\|v_t - \nabla f_S(x_t)\|^2 + 3\eta_t^2\|\nabla f_S(x_t) - \nabla f_S(x_t^k)\|^2 + 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2.$$

Putting above two inequalities into (8), we have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^k\|^2 \le \|x_t - x_t^k\|^2 &+ 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\| \\ &+ (3\eta_t^2 - \frac{2\eta_t}{L})\|\nabla f_S(x_t) - \nabla f_S(x_t^k)\|^2 + 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2.\end{aligned}$$

By setting $\eta_t \le \frac{2}{3L}$, we have

$$\begin{aligned}&\|x_{t+1} - x_{t+1}^k\|^2 \\ &\le \|x_t - x_t^k\|^2 + 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\| + 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2.\end{aligned}$$

Case 2 ($i_t = k$). We have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^k\| &= \|x_t - \eta_t v_t - x_t^k + \eta_t v_t^k\| \\ &\le \|x_t - x_t^k\| + \eta_t\|v_t - v_t^k\| \le \|x_t - x_t^k\| + \eta_t L_f.\end{aligned}$$

Then we can get

$$\|x_{t+1} - x_{t+1}^k\|^2 \le \|x_t - x_t^k\|^2 + 2\eta_t L_f\|x_t - x_t^k\| + \eta_t^2 L_f^2.$$

Combining **Case 1** and **Case 2** we have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^k\|^2 \le \|x_t - x_t^k\|^2 &+ 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\| \\ &+ 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2 + 2\eta_t L_f\|x_t - x_t^k\|\mathbf{1}_{i_t=k} + \eta_t^2 L_f^2\mathbf{1}_{i_t=k}.\end{aligned}$$

Note that

$$\mathbb{E}_A[\|x_t - x_t^k\|\mathbf{1}_{[i_t=k]}] = \mathbb{E}_A[\|x_t - x_t^k\|\mathbf{1}_{[i_t=k]}] = \frac{1}{n}\mathbb{E}_A[\|x_t - x_t^k\|] \le \frac{1}{n}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2}. \tag{9}$$

Then using Cauchy-Schwarz inequality, we can get

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_{t+1}^k\|^2] \le \mathbb{E}_A[\|x_t - x_t^k\|^2] &+ 2\eta_t(\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2])^{1/2}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2} \\ &+ 2\eta_t(\mathbb{E}_A[\|v_t^k - \nabla f_S(x_t^k)\|^2])^{1/2}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2} + 3\eta_t^2\mathbb{E}_A[\|v_t^k - f_S(x_t^k)\|^2] \\ &+ \frac{2L_f\eta_t}{n}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2} + \frac{\eta_t^2 L_f^2}{n}.\end{aligned}$$

Telescoping the above inequality from 0 to t, and combining with $x_0 = x_0^k$, we have

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_{t+1}^k\|^2] \le\ & 2\sum_{j=1}^{t}\eta_j(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}(\mathbb{E}_A[\|x_j - x_j^k\|^2])^{1/2} \\ &+ 2\sum_{j=1}^{t}\eta_j(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2}(\mathbb{E}_A[\|x_j - x_j^k\|^2])^{1/2} + 3\sum_{j=1}^{t}\eta_j^2\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2] \\ &+ \sum_{j=0}^{t}\frac{2L_f\eta_j}{n}(\mathbb{E}_A[\|x_j - x_j^k\|^2])^{1/2} + \sum_{j=0}^{t}\frac{\eta_j^2 L_f^2}{n}.\end{aligned}$$

Denote $u_t = (\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2}$, then we can get

$$\begin{aligned}u_t^2 \le\ & 2\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}u_j + 2\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2}u_j \\ &+ 3\sum_{j=1}^{t-1}\eta_j^2\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2] + \sum_{j=0}^{t-1}\frac{2L_f\eta_j}{n}u_j + \sum_{j=0}^{t-1}\frac{\eta_j^2 L_f^2}{n}.\end{aligned}$$

Define

$$S_t \leq 3\sum_{j=1}^{t-1}\eta_j^2\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2] + \sum_{j=0}^{t-1}\frac{\eta_j^2 L_f^2}{n},$$

and

$$\alpha_j = 2\eta_j(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} + 2\eta_j(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} + \frac{2L_f\eta_j}{n}.$$

using Lemma 13 we can get

$$\begin{aligned}
u_t &\leq \sqrt{S_t} + \sum_{j=1}^{t-1}\alpha_j \\
&\leq 2(\sum_{j=1}^{t-1}\eta_j^2\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} + (\sum_{j=0}^{t-1}\frac{\eta_j^2 L_f^2}{n})^{1/2} + 2\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} \\
&\quad + 2\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} + \sum_{j=1}^{t-1}\frac{2L_f\eta_j}{n}.
\end{aligned}$$

Furthermore, setting $\eta_t = \eta$, we can get $\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} \leq \sup_S \eta\sum_{j=1}^{t-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}$ and $\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} \leq \sup_S \eta\sum_{j=1}^{t-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}$. Consequently, with T iterations, we obtain that

$$u_T \leq 6\sup_S \eta\sum_{j=0}^{T-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} + \frac{\eta L_f\sqrt{T}}{\sqrt{n}} + \frac{2L_f\eta T}{n}. \tag{10}$$

Because often we have $T \geq n$, and $\mathbb{E}_A[\|x_T - x_T^k\|] \leq u_T = (\mathbb{E}_A[\|x_T - x_T^k\|^2])^{1/2}$, then we can get

$$\mathbb{E}_A[\|x_T - x_T^k\|] \leq O(\sup_S \eta\sum_{j=0}^{T-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{\frac{1}{2}} + \frac{L_f\eta T}{n}).$$

This completes the proof. □

Corollary 1. *Consider STORM in Algorithm 1 with $\eta_t = \eta \leq \frac{2}{3L}$, and $\beta_t = \beta \in (0,1)$, for any $t \in [0, T-1]$. With the output $A(S) = x_T$, ϵ satisfies*

$$O\bigg(\eta T\big((\beta T)^{-\frac{c}{2}} + \beta^{1/2} + \eta\beta^{-1/2}\big) + \eta T\frac{1}{n}\bigg).$$

Next, we give the proof of Corollary 1.

proof of Corollary 1. Combining (10) and Lemma 17, we can get

$$\begin{aligned}
\epsilon &\leq 6\sup_S \eta\sum_{j=0}^{T}((\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta})^{\frac{1}{2}} + \frac{\eta L_f\sqrt{T}}{\sqrt{n}} + \frac{2L_f\eta T}{n} \\
&\leq 6\sup_S \eta\Big((\frac{c}{e})^c\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2]\beta^{-\frac{c}{2}}\sum_{j=0}^{T}t^{-\frac{c}{2}} + 2\sigma_J\sqrt{\beta}T + L_f\eta\beta^{-\frac{1}{2}}T\Big) + \frac{3L_f\eta T}{n}.
\end{aligned}$$

Then according to

$$\sum_{t=1}^{T}t^{-z} = O(T^{1-z}), \forall z \in (-1,0)\cup(-\infty,-1), \quad \sum_{t=1}^{T}t^{-1} = O(\log T), \tag{11}$$

we have

$$\epsilon = O(\eta(\beta T)^{-\frac{c}{2}}T + \eta\beta^{1/2}T + \eta^2\beta^{-1/2}T + \eta T n^{-1})$$

□

Before giving the proof of Theorem 5, we first introduce a useful lemma.

Lemma 18. *Suppose Assumption 1(i), 2 (i) and 3 (i) holds for the empirical risk F_S. By running Algorithm 1, we have for any $\gamma_t > 0$*

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \\
&\leq (1+\eta_t\gamma_t)\mathbb{E}_A[\|x_t - x_*^S\|^2|\mathcal{F}_t] - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \eta_t^2 L_f^2 + \frac{\eta_t}{\gamma_t}\mathbb{E}_A[\|\nabla f_S(x_t) - v_t\|^2|\mathcal{F}_t],
\end{aligned}$$

where $\mathcal{F}_t$ is the σ-field generated by $\{v_{i_0}, \cdots, v_{i_{t-1}}\}$.

proof of Lemma 18. According to the update rule of Algorithm 1, we have

$$\begin{aligned}
\|x_{t+1} - x_*^S\|^2 &= \|x_t - \eta_t v_t - x_*^S\|^2 \\
&= \|x_t - x_*^S\|^2 - 2\eta_t\langle v_t, x_t - x_*^S\rangle + \eta_t^2\|v_t\|^2 \\
&= \|x_t - x_*^S\|^2 - 2\eta_t\langle \nabla f_S(x_t), x_t - x_*^S\rangle + \eta_t^2\|v_t\|^2 + 2\eta_t\langle \nabla f_S(x_t) - v_t, x_t - x_*^S\rangle.
\end{aligned}$$

Let $\mathcal{F}_t$ be the σ-field generated by $\{v_{i_0}, \cdots, v_{i_{t-1}}\}$, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \\
&= \mathbb{E}_A[\|x_t - x_*^S\|^2|\mathcal{F}_t] - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \eta_t^2 L_f^2 + \mathbb{E}_A[2\eta_t\langle \nabla f_S(x_t) - v_t, x_t - x_*^S\rangle|\mathcal{F}_t] \\
&\leq \mathbb{E}_A[\|x_t - x_*^S\|^2|\mathcal{F}_t] - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \eta_t^2 L_f^2 + 2\eta_t\mathbb{E}_A[\frac{1}{2\gamma_t}\|\nabla f_S(x_t) - v_t\|^2 + \frac{\gamma_t}{2}\|x_t - x_*^S\|^2|\mathcal{F}_t] \\
&= (1+\eta_t\gamma_t)\mathbb{E}_A[\|x_t - x_*^S\|^2|\mathcal{F}_t] - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \eta_t^2 L_f^2 + \frac{\eta_t}{\gamma_t}\mathbb{E}_A[\|\nabla f_S(x_t) - v_t\|^2|\mathcal{F}_t].
\end{aligned}$$

This complete the proof. □

Then we give the proof of Theorem 5.

proof of Theorem 5. Setting $\eta_t = \eta$, $\beta_t = \beta$ and $\gamma_t = \sqrt{\beta}$, putting Lemma 17 into 18 we have

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2] \leq \mathbb{E}_A[\|x_t - x_*^S\|^2] + \eta\sqrt{\beta}\mathbb{E}_A[\|x_t - x_*^S\|^2] - 2\eta\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \eta^2 L_f^2 \\
+ \frac{\eta}{\sqrt{\beta}}((\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta}).
\end{aligned}$$

Re-arranging above inequality and telescoping from 1 to t we have

$$\begin{aligned}
&2\eta\sum_{t=1}^{T}\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\
&\qquad\leq D_x + D_x\eta\beta^{1/2}T + L_f^2\eta^2 T + (\frac{c}{e})^c V\beta^{-\frac{1}{2}-c}\eta\sum_{t=1}^{T}t^{-c} + 2\sigma_J^2\eta\beta^{1/2}T + L_f^2\eta^3\beta^{-3/2}T.
\end{aligned} \tag{12}$$

Then From the choice of $A(S)$, according to (11), as long as $c > 2$, we have

$$\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)] = O(D_x(\eta T)^{-1} + D_x\beta^{1/2} + L_f^2\eta + VT^{-c}\beta^{-1/2-c} + \sigma_J^2\beta^{1/2} + L_f^2\eta^2\beta^{-3/2}).$$

This complete the proof. □

Next we give the proof of Theorem 6.

proof of Theorem 6. Combining Lemma 17 and (10), we have

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^k\|] &\le 6\eta \sum_{j=0}^{t-1}((\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta})^{1/2} + \frac{\eta L_f\sqrt{t}}{\sqrt{n}} + \frac{2L_f\eta t}{n}\\ &\le 6(\frac{c}{e})^{c/2}V\eta\beta^{-c/2}\sum_{j=0}^{t-1} t^{-c/2} + 12\sigma_J\eta\beta^{1/2}t + 6L_f\eta^2\beta^{-1/2}t + \frac{\eta L_f\sqrt{t}}{\sqrt{n}} + \frac{2L_f\eta t}{n}.\end{aligned}$$

Then according to Theorem 15, we have

$$\mathbb{E}_{S,A}[F(x_t) - F_S(x_t)] \le L_f(6(\frac{c}{e})^c V\eta\beta^{-c/2}\sum_{j=0}^{t-1} t^{-c/2} + 12\sigma_J\eta\beta^{1/2}t + 6L_f\eta^2\beta^{-1/2}t + \frac{\eta L_f\sqrt{t}}{\sqrt{n}} + \frac{2L_f\eta t}{n}).$$

Combining above inequality with (12), and according to $F_S(x_*^S) \le F_S(x_*)$ we have

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_{S,A}[F(x_t) - F(x_*)]\\ &\le (D_x + D_x\eta\beta^{1/2}T + L_f^2\eta^2T + (\frac{c}{e})^cV\beta^{-\frac{1}{2}-c}\eta\sum_{t=1}^{T} t^{-c} + 2\sigma_J^2\eta\beta^{1/2}T + L_f^2\eta^3\beta^{-3/2}T)/2\eta\\ &\quad + L_f\sum_{t=1}^{T}(6(\frac{c}{e})^cV\eta\beta^{-c/2}\sum_{j=0}^{t-1} t^{-c/2} + 12\sigma_J\eta\beta^{1/2}\sum_{t=1}^{T} t + 6L_f\eta^2\beta^{-1/2}\sum_{t=1}^{T} t + \sum_{t=1}^{T}\frac{3L_f\eta t}{n}).\end{aligned}$$

According to (11), we have

$$\sum_{t=1}^{T}\sum_{j=1}^{T} j^{-\frac{c}{2}} = O(\sum_{t=1}^{T} t^{1-\frac{c}{2}}(\log t)^{\mathbf{1}_{c=2}}) = O(T^{2-\frac{c}{2}}(\log T)^{\mathbf{1}_{c=2}}). \tag{13}$$

Combining above two inequalities, we have

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_{S,A}[F(x_t) - F(x_*)]\\ &= O\Big(\eta^{-1} + \beta^{1/2}T + \eta T + (\beta T)^{-c}\beta^{-1/2}T + \beta^{1/2}T + \eta^2\beta^{-3/2}T + \eta\beta^{-c/2}T^{2-\frac{c}{2}}(\log T)^{\mathbf{1}_{c=2}}\\ &\qquad + \eta\beta^{1/2}T^2 + \eta^2\beta^{-1/2}T^2 + \eta T^2 n^{-1}\Big).\end{aligned}$$

Setting $\eta = T^{-a}$ and $\beta = T^{-b}$, dividing both sides of above inequality with T, then from the choice of $A(S)$ we get

$$\begin{aligned}&\mathbb{E}_{S,A}[F(A(S)) - F(x_*)]\\ &\le O\Big(T^{a-1} + T^{-b/2} + T^{-a} + T^{1/b-c(1-b)+T^{-b/2-1}} + T^{-b/2} + T^{3b/2-2a} + T^{1-a+c/2(b-1)}(\log T)^{\mathbf{1}_{c=2}}\\ &\qquad + T^{1-a-b/2} + T^{1-2a+b/2} + T^{1-a}n^{-1}\Big).\end{aligned}$$

As long as $c > 4$, the dominating terms are $O(T^{1-a-\frac{b}{2}})$, $O(T^{1+\frac{b}{2}-2a})$, $O(n^{-1}T^{1-a})$, $O(T^{a-1})$, and $O(T^{\frac{3}{2}b-2a})$. Setting $a = b = 4/5$, then we have

$$\mathbb{E}[F(A(S)) - F(x_*)] = O(T^{-\frac{1}{5}} + \frac{T^{\frac{1}{5}}}{n}).$$

Choosing $T = O(n^{2.5})$, we have the following bound

$$\mathbb{E}[F(A(S)) - F(x_*)] = O(\frac{1}{\sqrt{n}}).$$

This completes the proof. □

C.1. Strongly-convex-setting

proof of Theorem 7. Similar to the proof for convex setting, we use the same notations.

We will consider two cases, i.e., $i_t \neq k$ and $i_t = k$.

Case 1 ($i_t \neq k$)**.** We have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^k\|^2 &= \|x_t - \eta_t v_t - x_t^k + \eta_t v_t^k\|^2 \\ &\leq \|x_t - x_t^k\|^2 - 2\eta_t\langle v_t - v_t^k, x_t - x_t^k\rangle + \eta_t^2\|v_t - v_t^k\|^2.\end{aligned} \tag{14}$$

For the second term on the RHS of (14), we have

$$\begin{aligned}&- 2\eta_t\langle v_t - v_t^k, x_t - x_t^k\rangle \\ &= -2\eta_t\langle v_t - \nabla f_S(x_t), x_t - x_t^k\rangle - 2\eta_t\langle \nabla f_S(x_t) - \nabla f_S(x_t^k), x_t - x_t^k\rangle - 2\eta_t\langle \nabla f_S(x_t^k) - v_t^k, x_t - x_t^k\rangle.\end{aligned}$$

Note that if F is μ strongly convex, then $\varphi(x) = F(x) - \frac{\sigma}{2}\|x\|^2$ is convex with $(L-\mu)$-smooth. Then, applying above to φ yields the following inequality

$$\langle \nabla F(x) - \nabla F(x'), x - x'\rangle \geq \frac{L\mu}{L+\mu}\|x - x'\|^2 + \frac{1}{L+\mu}\|\nabla F(x) - \nabla F(x')\|^2.$$

Then using Assumption 3 (i), i.e., the smoothness, and combining with the strong convexity of $f_S(\cdot)$ we can get

$$\begin{aligned}&- 2\eta_t\langle v_t - v_t^k, x_t - x_t^k\rangle \\ &\leq 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| - 2\eta_t(\frac{1}{L+\mu}\|\nabla f_S(x_t) - \nabla f_S(x_t^k)\|^2 + \frac{L\mu}{L+\mu}\|x_t - x_t^k\|^2) \\ &\quad + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\|.\end{aligned}$$

For the third term on the RHS of (14), we have

$$\eta_t^2\|v_t - v_t^k\|^2 \leq 3\eta_t^2\|v_t - \nabla f_S(x_t)\|^2 + 3\eta_t^2\|\nabla f_S(x_t) - \nabla f_S(x_t^k)\|^2 + 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2.$$

Putting above two inequalities into (14), we have

$$\begin{aligned}&\|x_{t+1} - x_{t+1}^k\|^2 \\ &\leq (1 - \frac{2\eta_t L\mu}{L+\mu})\|x_t - x_t^k\|^2 + 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\| \\ &\quad + (3\eta_t^2 - \frac{2\eta_t}{L+\mu})\|\nabla f_S(x_t) - \nabla f_S(x_t^k)\|^2 + 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2.\end{aligned}$$

By setting $\eta_t \leq \frac{2}{3(L+\mu)}$, we have

$$\begin{aligned}(1 - \frac{2\eta_t L\mu}{L+\mu})\|x_{t+1} - x_{t+1}^k\|^2 \leq (1 - \frac{2\eta_t L\mu}{L+\mu})&\|x_t - x_t^k\|^2 + 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| \\ &+ 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\| + 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2.\end{aligned}$$

Case 2 ($i_t = k$)**.** We have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^k\| &= \|x_t - \eta_t v_t - x_t^k + \eta_t v_t^k\| \\ &\leq \|x_t - x_t^k\| + \eta_t\|v_t - v_t^k\| \leq \|x_t - x_t^k\| + \eta_t L_f.\end{aligned}$$

Then we can get

$$\|x_{t+1} - x_{t+1}^k\|^2 \leq \|x_t - x_t^k\|^2 + 2\eta_t L_f\|x_t - x_t^k\| + \eta_t^2 L_f^2.$$

Combining **Case 1** and **Case 2** we have

$$\|x_{t+1} - x_{t+1}^k\|^2 \leq (1 - \frac{2\eta_t L\mu}{L+\mu})\|x_t - x_t^k\|^2 + 2\eta_t\|v_t - \nabla f_S(x_t)\| \cdot \|x_t - x_t^k\| + 2\eta_t\|v_t^k - f_S(x_t^k)\| \cdot \|x_t - x_t^k\|$$
$$+ 3\eta_t^2\|v_t^k - f_S(x_t^k)\|^2 + 2\eta_t L_f\|x_t - x_t^k\|\mathbf{1}_{i_t=k} + \eta_t^2 L_f^2 \mathbf{1}_{i_t=k}.$$

According to (9), we have

$$\mathbb{E}_A[\|x_{t+1} - x_{t+1}^k\|^2] \leq (1 - \frac{2\eta_t L\mu}{L+\mu})\mathbb{E}_A[\|x_t - x_t^k\|^2] + 2\eta_t(\mathbb{E}_A[\|v_t - \nabla f_S(x_t)\|^2])^{1/2}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2}$$
$$+ 2\eta_t(\mathbb{E}_A[\|v_t^k - \nabla f_S(x_t^k)\|^2])^{1/2}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2} + 3\eta_t^2\mathbb{E}_A[\|v_t^k - f_S(x_t^k)\|^2]$$
$$+ \frac{2L_f\eta_t}{n}(\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2} + \frac{\eta_t^2 L_f^2}{n}.$$

According to Lemma 13, setting $\eta_t = \eta$ and $\beta_t = \beta$, we can get

$$\mathbb{E}_A[\|x_{t+1} - x_{t+1}^k\|^2] \leq 2\eta\sum_{j=1}^{t}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}(\mathbb{E}_A[\|x_j - x_j^k\|^2])^{1/2}$$
$$+ 2\eta\sum_{j=1}^{t}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2}(\mathbb{E}_A[\|x_j - x_j^k\|^2])^{1/2}$$
$$+ \sum_{j=1}^{t}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j}\frac{2L_f\eta}{n}(\mathbb{E}_A[\|x_j - x_j^k\|^2])^{1/2} + \frac{\eta^2 L_f^2}{n}\sum_{j=0}^{t}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j}$$
$$+ 3\eta^2\sum_{j=0}^{t}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2].$$

Denote $u_t = (\mathbb{E}_A[\|x_t - x_t^k\|^2])^{1/2}$, then we can get

$$u_t^2 \leq 2\eta\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}u_j$$
$$+ 2\eta\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2}u_j + \frac{2L_f\eta}{n}\sum_{j=0}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}u_j$$
$$+ 3\eta^2\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2] + \frac{\eta^2 L_f^2}{n}\sum_{j=0}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}.$$

Define $S_t = 3\eta^2\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2] + \frac{\eta^2 L_f^2}{n}\sum_{j=0}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}$ and $\alpha_j = 2\eta(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} + 2\eta(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} + \frac{2L_f\eta}{n}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}$. using Lemma 14, we can get

$$u_t \leq \sqrt{S_t} + \sum_{t=1}^{t-1}\alpha_j$$
$$\leq 2\eta(\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2])^{1/2} + \sqrt{\frac{(L+\mu)\eta L_f^2}{2L\mu n}}$$
$$+ 2\eta\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2}$$
$$+ 2\eta\sum_{j=1}^{t-1}(1 - \frac{2\eta L\mu}{L+\mu})^{t-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} + \frac{L_f(L+\mu)}{L\mu n},$$

where the last inequality holds by (7). Consequently, with T iterations, because of the inequality $\mathbb{E}_A[\|x_T - x_T^k\|] \leq u_T$, we have

$$\begin{aligned}
\mathbb{E}_A[\|x_T - x_T^k\|] \leq{}& 2\eta(\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2])^{1/2} + \sqrt{\frac{(L+\mu)\eta L_f^2}{2L\mu n}} \\
&+ 2\eta\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} \\
&+ 2\eta\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} + \frac{L_f(L+\mu)}{L\mu n},
\end{aligned}$$

Then we analyze which one of $(\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2])^{1/2}$ and $\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2}$ is the dominant term.

For the first term, according to Lemma 17 we have

$$\begin{aligned}
&(\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2])^{1/2} \\
&\leq (\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}((\frac{c}{e})^c(j\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta}))^{1/2} \\
&\leq (\frac{c}{e})^{\frac{c}{2}}\beta^{-\frac{c}{2}}\sqrt{V}(\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}j^{-c})^{\frac{1}{2}} + \sigma_J\sqrt{\frac{(L+\mu)\beta}{\eta L\mu}} + L_f\sqrt{\frac{\eta(L+\mu)}{2\beta L\mu}},
\end{aligned}$$

where the last inequality holds by (7), as for $\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}j^{-c}$, according to Lemma 12, we have

$$\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}j^{-c} \leq \frac{\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}\sum_{j=1}^{T-1}j^{-\frac{c}{2}}}{T} \leq \frac{(L+\mu)\sum_{j=1}^{T-1}j^{-c}}{2T\eta L\mu}, \tag{15}$$

then according to (11) we can get

$$\begin{aligned}
&(\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}\mathbb{E}_A[\|v_j^k - f_S(x_j^k)\|^2])^{1/2} \\
&\leq (\frac{c}{e})^{\frac{c}{2}}\sqrt{\frac{V(L+\mu)}{2\eta L\mu}}(T\beta)^{-\frac{c}{2}} + \sigma_J\sqrt{\frac{(L+\mu)\beta}{\eta L\mu}} + L_f\sqrt{\frac{\eta(L+\mu)}{2\beta L\mu}}.
\end{aligned}$$

For the second term, according to Lemma 17 we have

$$\begin{aligned}
&\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} \\
&\leq \sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}((\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta})^{\frac{1}{2}}.
\end{aligned}$$

Similar to the first term, we can get

$$\sum_{j=1}^{T-1}(1 - \frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j^k - \nabla f_S(x_j^k)\|^2])^{1/2} \leq (\frac{c}{e})^{\frac{c}{2}}\sqrt{V}\frac{L+\mu}{\eta L\mu}(T\beta)^{-\frac{c}{2}} + \sigma_J\sqrt{\beta}\frac{L+\mu}{\eta L\mu} + L_f\frac{L+\mu}{2\sqrt{\beta}L\mu}. \tag{16}$$

It's easily to get the dominating term is the second term $\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j^k-\nabla f_S(x_j^k)\|^2])^{1/2}$. Therefore

$$\begin{aligned}
\mathbb{E}_A[\|x_T-x_T^k\|] \leq & 2\eta(\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}\mathbb{E}_A[\|v_j^k-f_S(x_j^k)\|^2])^{1/2}+\sqrt{\frac{(L+\mu)\eta L_f^2}{2L\mu n}} \\
& +2\eta\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j-\nabla f_S(x_j)\|^2])^{1/2} \\
& +2\eta\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j^k-\nabla f_S(x_j^k)\|^2])^{1/2}+\frac{L_f(L+\mu)}{L\mu n} \\
\leq & 6\eta\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j-\nabla f_S(x_j)\|^2])^{1/2}+\sqrt{\frac{(L+\mu)\eta L_f^2}{2L\mu n}}+\frac{L_f(L+\mu)}{L\mu n} \\
\leq & 6\eta\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j-\nabla f_S(x_j)\|^2])^{1/2}+\frac{2L_f(L+\mu)}{L\mu n},
\end{aligned} \tag{17}$$

where the last inequality holds since often we have $\eta \leq \frac{1}{n}$. Then we get the final result

$$\epsilon \leq O(\eta\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j-\nabla f_S(x_j)\|^2])^{1/2}+\frac{L_f(L+\mu)}{L\mu n}).$$

This completes the proof. □

Corollary 2 (One-level Optimization)**.** *Consider STORM in Algorithm 1 with $\eta_t=\eta\leq\frac{2}{3(L+\mu)}$, and $\beta_t=\beta\in(0,1)$ for any $t\in[0,T-1]$ and the output $A(S)=x_T$. Then, we have the following results*

$$\epsilon \leq O((T\beta)^{-\frac{c}{2}}+\beta^{\frac{1}{2}}+\eta\beta^{-\frac{1}{2}}+n^{-1}).$$

Next, we give the proof of Corollary 2.

proof of Corollary 2. Combining Theorem 10 and (16), we have

$$\epsilon \leq O((T\beta)^{-\frac{c}{2}}+\beta^{\frac{1}{2}}+\eta\beta^{-\frac{1}{2}}+n^{-1}).$$

This complete the proof. □

Before give the detailed proof of Theorem 10, we first give a useful lemma.

Lemma 19. *Let Assumption 1(i), 2 (i) and 3 (i) holds, as F_S is μ-strongly convex. By running Algorithm 1, we have*

$$\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq F_S(x_t)-\frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2+\frac{L\eta_t^2L_f^2}{2}+2\eta_t\|v_t-\nabla F_S(x_t)\|^2.$$

where $\mathcal{F}_t$ is the σ-field generated by $\{v_{i_0},\cdots,v_{i_{t-1}}\}$.

proof of Lemma 19. According to the smoothness of $F_S(\cdot)$, then we have

$$\begin{aligned}
F_S(x_{t+1}) &\leq F_S(x_t)+\langle\nabla F_S(x_t),x_{t+1}-x_t\rangle+\frac{L}{2}\|x_{t+1}-x_t\|^2 \\
&\leq F_S(x_t)-\eta_t\|\nabla F_S(x_t)\|^2+\frac{L\eta_t^2L_f^2}{2}-\eta_t\langle\nabla F_S(x_t),v_t-\nabla F_S(x_t)\rangle \\
&\leq F_S(x_t)-\eta_t\|\nabla F_S(x_t)\|^2+\frac{L\eta_t^2L_f^2}{2}+\frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2+2\eta_t\|v_t-\nabla F_S(x_t)\|^2,
\end{aligned}$$

where the last inequality holds by Cauchy-Schwartz. Then we can get

$$\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq F_S(x_t) - \frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2 + \frac{L\eta_t^2 L_f^2}{2} + 2\eta_t\|v_t - \nabla F_S(x_t)\|^2.$$

This complete the proof. □

Now we move on the proof of Theorem 10.

proof of Theorem 10. Satisfying strong convexity also satisfies Polyak-Łojasiewicz (PL) inequality, then we can get for all x

$$\frac{1}{2}\|\nabla F_S(x)\|^2 \geq \mu(F_S(x) - F_S(x_*^S)).$$

According to Lemma 19, we have

$$\mathbb{E}_A[F_S(x_{t+1}) - F_S(x_*^S)] \leq (1-\mu\eta_t)\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \frac{L\eta_t^2 L_f^2}{2} + 2\eta_t\|v_t - \nabla F_S(x_t)\|^2.$$

Setting $\eta_t = \eta$ and $\beta_t = \beta$, using Lemma 17, we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1}) - F_S(x_*^S)] \leq (1-\mu\eta)\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \frac{L\eta^2 L_f^2}{2} \\ + 2\eta((\frac{c}{e})^c(t\beta)^{-c}\mathbb{E}_A[\|v_0 - \nabla f_S(x_0)\|^2] + 2\beta\sigma_J^2 + \frac{L_f^2\eta^2}{\beta}).\end{aligned}$$

Telescoping the above inequality from 1 to T, according to Lemma 13, we can get

$$\begin{aligned}&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)^{T-1}\mathbb{E}_A[F_S(x_1) - F_S(x_*^S)] + \frac{L\eta^2 L_f^2}{2}\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1} \\ &\quad + 2\eta V(\frac{c}{e})^c\beta^{-c}\sum_{t=1}^{T-1} t^{-c}(1-\mu\eta)^{T-t-1} + 2\sigma_J^2\eta\beta\sum_{t=1}^{T-t-1}(1-\mu\eta)^{T-t-1} + \frac{2L_f^2\eta^3}{\beta}\sum_{t=1}^{T-t-1}(1-\mu\eta)^{T-t-1}.\end{aligned}$$

For $t = 0$, we have

$$\mathbb{E}_A[F_S(x_1) - F_S(x_*^S)] \leq (1-\mu\eta)\mathbb{E}_A[F_S(x_0) - F_S(x_*^S)] + \frac{L\eta^2 L_f^2}{2} + 2\eta V.$$

Combining the above two inequalities, we have

$$\begin{aligned}&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)^T\mathbb{E}_A[F_S(x_0) - F_S(x_*^S)] + \frac{L\eta^2 L_f^2}{2}\sum_{t=1}^{T}(1-\mu\eta)^{T-t} + 2\eta V(1-\mu\eta)^{T-1} \\ &\quad + 2\eta V(\frac{c}{e})^c\beta^{-c}\sum_{t=1}^{T-1} t^{-c}(1-\mu\eta)^{T-t-1} + 2\sigma_J^2\eta\beta\sum_{t=1}^{T-t-1}(1-\mu\eta)^{T-t-1} + \frac{2L_f^2\eta^3}{\beta}\sum_{t=1}^{T-t-1}(1-\mu\eta)^{T-t-1}.\end{aligned}$$

According to (5), (6) and (7), we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \leq (\frac{c}{e\mu})^c(\eta T)^{-c}D_x + \frac{L\eta L_f^2}{2\mu} + 2\eta(\frac{c}{e\mu})^c(\eta T)^{-c}V \\ + 2\eta V(\frac{c}{e})^c\beta^{-c}\sum_{t=1}^{T-1} t^{-c}(1-\mu\eta)^{T-t-1} + \frac{2\sigma_J^2\beta}{\mu} + \frac{2L_f^2\eta^2}{\beta\mu}.\end{aligned}$$

Then according to (15) we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \le{}& (\frac{c}{e\mu})^c(\eta T)^{-c}D_x + \frac{L\eta L_f^2}{2\mu} + 2\eta(\frac{c}{e\mu})^c(\eta T)^{-c}V \\ &+ \frac{2V(c/e)^c(\beta T)^{-c}}{\mu} + \frac{2\sigma_J^2\beta}{\mu} + \frac{2L_f^2\eta^2}{\beta\mu}.\end{aligned}$$

Then we can get

$$\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] = O(D_x(\eta T)^{-c} + L_f^2 L\eta + V\eta(\eta T)^{-c} + V(\beta T)^{-c} + \sigma_J^2\beta + L_f^2\eta^2\beta^{-1}).$$

This completes the proof. □

Next, we move on to the proof of Theorem 11

proof of Theorem 11. Combining (17) and Theorem 15, we have

$$\mathbb{E}_{S,A}[F(x_T) - F_S(x_T)] \le L_f(6\eta\sum_{j=1}^{T-1}(1-\frac{2\eta L\mu}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j - \nabla f_S(x_j)\|^2])^{1/2} + \frac{2L_f(L+\mu)}{L\mu n}).$$

Then according to (16), we have

$$\mathbb{E}_{S,A}[F(x_T) - F_S(x_T)] \le 6L_f\eta((\frac{c}{e})^{\frac{c}{2}}\sqrt{V}\frac{L+\mu}{\eta L\mu}(T\beta)^{-\frac{c}{2}} + \sigma_J\sqrt{\beta}\frac{L+\mu}{\eta L\mu} + L_f\frac{L+\mu}{2\beta L\mu}) + \frac{2L_f^2(L+\mu)}{L\mu n}.$$

Combining with Theorem 10, and using the fact that $F_S(x_*^S) \le F_S(x_*)$ we have

$$\begin{aligned}\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] \le{}& 6L_f\eta((\frac{c}{e})^{\frac{c}{2}}\sqrt{V}\frac{L+\mu}{\eta L\mu}(T\beta)^{-\frac{c}{2}} + \sigma_J\sqrt{\beta}\frac{L+\mu}{\eta L\mu} + L_f\frac{L+\mu}{2\sqrt{\beta}L\mu}) + \frac{2L_f^2(L+\mu)}{L\mu n} \\ &+ (\frac{c}{e\mu})^c(\eta T)^{-c}V + \frac{L\eta L_f^2}{2\mu} + 2\eta(\frac{c}{e\mu})^c(\eta T)^{-c}V \\ &+ \frac{2V(c/e)^c(\beta T)^{-c}}{\mu} + \frac{2\sigma_J^2\beta}{\mu} + \frac{2L_f^2\eta^2}{\beta\mu}.\end{aligned}$$

Setting $\eta = T^{-a}$ and $\beta = T^{-b}$ with $a, b \in (0, 1]$, we have

$$\begin{aligned}&\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] \\ &= O(T^{\frac{c}{2}(b-1)} + T^{-\frac{b}{2}} + T^{\frac{b}{2}-a} + T^{-c(1-a)} + T^{-a} + T^{-c(1-a)-a} + T^{-c(1-b)} + T^{-b} + T^{b-2a}).\end{aligned}$$

Setting $c = 3$, the dominating terms are $O(T^{\frac{b}{2}-a}), \quad O(T^{-\frac{b}{2}}), \quad O(T^{\frac{3}{2}(b-1)}), \quad O(T^{-\frac{a}{2}}), \quad O(T^{3(a-1)})$.
Setting $a = b = \frac{6}{7}$, we have

$$\mathbb{E}_{S,A}\left[F(A(S)) - F\left(x_*\right)\right] = O\left(T^{-\frac{3}{7}}\right).$$

Setting $T = O(n^{\frac{7}{6}})$, we have the following

$$\mathbb{E}_{S,A}\left[F(A(S)) - F\left(x_*\right)\right] = O\left(\frac{1}{\sqrt{n}}\right).$$

The proof is completed.

□

D. Two-level Stochastic Optimizations

Lemma 20 (Theorem 1 in (Yang et al., 2023)). *If Assumption 1 (ii) holds true and the randomized algorithm A is ϵ-uniformly stable then*

$$\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] \leq L_f L_g \epsilon_\nu + 4L_f L_g \epsilon_\omega + L_f \sqrt{m^{-1}\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(A(S)))]},$$

where the variance term $\mathrm{Var}_\omega(g_\omega(A(S))) = \mathbb{E}_\omega[\|g_\omega(A(S)) - g(A(S))\|^2]$.

Lemma 21 (Lemma 7 in (Qi et al., 2021)). *Let Assumption 1(ii), 2 (ii) and 3 (ii) hold for the empirical risk, and x_t, u_t are generated by Algorithm 2 2, then we have*

$$\mathbb{E}[\|u_t - g_S(x_t)\|^2] \leq (1-\beta_t)\mathbb{E}[\|u_{t-1} - g_S(x_{t-1})\|^2] + 2\beta_t^2\sigma_g^2 + 2L_g^2\|x_t - x_{t-1}\|^2.$$

Lemma 22 (Lemma 7 in (Qi et al., 2021)). *Let Assumption 1(ii), 2 (ii) and 3 (ii) hold for the empirical risk, and x_t, v_t are generated by Algorithm 2, then we have*

$$\mathbb{E}[\|v_t - \nabla g_S(x_t)\|^2] \leq (1-\beta_t)\mathbb{E}[\|v_{t-1} - \nabla g_S(x_{t-1})\|^2] + 2\beta_t^2\sigma_{g'}^2, + 2L_g^2\|x_t - x_{t-1}\|^2.$$

Lemma 23. *Let Assumption 1(ii), 2 (ii) and 3 (ii) hold and x_t, u_t are generated by Algorithm 2, let $0 < \eta_t = \eta < 1$ and let $0 < \beta_t = \beta < 1$, for any $c > 0$, we have*

$$\mathbb{E}[\|u_t - g_S(x_t)\|^2] \leq (\frac{c}{e})^c (t\beta)^{-c}\mathbb{E}[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}.$$

Proof. According to the rule of update we have $\mathbb{E}[\|x_t - x_{t-1}\|^2] \leq L_f^2L_g^2\eta_{t-1}^2$, then using Lemma 21 and 13, we have

$$\mathbb{E}[\|u_t - g_S(x_t)\|^2] \leq \prod_{i=1}^{t}(1-\beta_i)\mathbb{E}[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2\Upsilon_t\sum_{i=1}^{t}\frac{\beta_i^2}{\Upsilon_i} + 2L_g^4L_f^2\Upsilon_t\sum_{i=1}^{t}\frac{\eta_{i-1}^2}{\Upsilon_i}.$$

For the term $\Upsilon_t\sum_{i=1}^t \beta_i^2/\Upsilon_i$, according to the setting that $\beta_t = \beta$, we have

$$\Upsilon_t\sum_{i=1}^{t}\frac{\beta_i^2}{\Upsilon_i} = \beta(\Upsilon_t\frac{\beta_1}{\Upsilon_1} + \Upsilon_t\sum_{i=2}^{t}\frac{\beta_i}{\Upsilon_i}) = \beta(\Upsilon_t\frac{\beta_1}{\Upsilon_1} + \Upsilon_t\sum_{i=2}^{t}(\frac{1}{\Upsilon_i} - \frac{1}{\Upsilon_{i-1}})) = \beta(1-\Upsilon_t) = \beta.$$

Then according to the setting that $\eta_t = \eta$, we have

$$\mathbb{E}[\|u_t - g_S(x_t)\|^2] \leq \prod_{i=1}^{t}(1-\beta_i)\mathbb{E}[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}.$$

Then using (5) and (11), we can get

$$\mathbb{E}[\|u_t - g_S(x_t)\|^2] \leq (\frac{c}{e})^c (t\beta)^{-c}\mathbb{E}[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}.$$

□

And the proof of $\mathbb{E}[\|v_t - \nabla g_S(x_t)\|^2]$ is similarly to Lemma 23, we won't repeat it.

Lemma 24. *Let Assumption 1(ii), 2 (ii) and 3 (ii) hold and x_t, v_t are generated by Algorithm 2, let $0 < \eta_t \leq \eta < 1$ and let $0 < \beta_t \leq \beta < 1$, for any $c > 0$, we have*

$$\mathbb{E}[\|v_t - \nabla g_S(x_t)\|^2] \leq (\frac{c}{e})^c (t\beta)^{-c}\mathbb{E}[\|v_0 - \nabla g_S(x_0)\|^2] + 2\sigma_{g'}^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}.$$

We first give some notations used in the two-level optimization to simplify our proof.

For any $k \in [n]$, let $S^{k,\nu} = \{\nu_1, \ldots, \nu_{k-1}, \nu_k', \nu_{k+1}, \ldots, \nu_n, \omega_1, \ldots, \omega_m\}$ be formed from S_ν by replacing the k-th element. Similarly, for any $l \in [m]$, define $S^{l,\omega} = \{\nu_1, \ldots, \nu_n, \omega_1, \ldots, \omega_{l-1}, \omega_l', \omega_{l+1}, \ldots, \omega_m\}$ as formed from S_ω by replacing the l-th element. Let $\{x_{t+1}\}$, $\{u_{t+1}\}$ and $\{v_{t+1}\}$ be generated by COVER based on S, $\{x_{t+1}^{k,\nu}\}$, $\{u_{t+1}^{k,\nu}\}$ and $\{v_{t+1}^{k,\nu}\}$ be generated by COVER based on $S^{k,\nu}$, $\{x_{t+1}^{l,\omega}\}$, $\{u_{t+1}^{l,\omega}\}$ and $\{v_{t+1}^{l,\omega}\}$ be generated by COVER based on $S^{l,\omega}$. Set $x_0 = x_0^{k,\nu}$ and $x_0 = x_0^{l,\omega}$ as starting points in $\mathcal{X}$.

D.1. Convex-setting

Proof of Theorem 3. Since a change in one sample data can occur in either S_ν or S_ω, we estimate $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{k,\nu}\|]$ and $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|]$ as follows.

Estimation of $\mathbb{E}_A\big[\|x_{t+1} - x_{t+1}^{k,\nu}\|\big]$

We first give the estimation of $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{k,\nu}\|]$. For this purpose, we will consider two cases, i.e., $i_t \neq k$ and $i_t = k$.

Case 1 ($i_t \neq k$). We have

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{k,\nu}\|^2 \\
&\le \|x_t - \eta_t v_t \nabla f_{\nu_{i_t}}(u_t) - x_t^{k,\nu} + \eta_t v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu})\|^2 \\
&\le \|x_t - x_t^{k,\nu}\|^2 - 2\eta_t\langle v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t\nabla f_{\nu_{i_t}}(u_t), x_t^{k,\nu} - x_t\rangle + \eta_t^2\|v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t\nabla f_{\nu_{i_t}}(u_t)\|^2.
\end{aligned} \tag{18}$$

We begin to estimate the second term in (18).

$$\begin{aligned}
&-2\eta_t\langle v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t\nabla f_{\nu_{i_t}}(u_t), x_t^{k,\nu} - x_t\rangle \\
&= -2\eta_t\langle v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t^{k,\nu}\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t\langle v_t^{k,\nu}\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})) - v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t\langle v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t\langle \nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t)\nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t\langle \nabla g_S(x_t)\nabla f_S(g_S(x_t)) - v_t\nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle - 2\eta_t\langle v_t\nabla f_S(g_S(x_t)) - v_t\nabla f_S(u_t), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t\langle v_t\nabla f_S(u_t) - v_t\nabla f_{\nu_{i_t}}(u_t), x_t^{k,\nu} - x_t\rangle.
\end{aligned} \tag{19}$$

Now we estimate the terms on the right hand side of (19) one by one.

For the first term of the RHS, we have

$$\begin{aligned}
&-2\eta_t\langle v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t^{k,\nu}\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\le 2\eta_t\|v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t^{k,\nu}\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\| \cdot \|x_t^{k,\nu} - x_t\| \le 2L_g C_f\eta_t\|u_t^{k,\nu} - g_S(x_t^{k,\nu})\| \cdot \|x_t^{k,\nu} - x_t\|.
\end{aligned} \tag{20}$$

For the second term of the RHS, according to $\mathbb{E}_{j_t}[v_t^{k,\nu}\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))] = v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu}))$, we have

$$-2\eta_t\mathbb{E}_{j_t}[\langle v_t^{k,\nu}\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})) - v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle] = 0. \tag{21}$$

For the third term of the RHS, we have

$$\begin{aligned}
&-2\eta_t\langle v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\le 2\eta_t\|\nabla f_S(g_S(x_t^{k,\nu}))(v_t^{k,\nu} - \nabla g_S(x_t^{k,\nu}))\| \cdot \|x_t^{k,\nu} - x_t\| \le 2\eta_t L_f\|v_t^{k,\nu} - \nabla g_S(x_t^{k,\nu})\| \cdot \|x_t^{k,\nu} - x_t\|
\end{aligned} \tag{22}$$

Then according to Assumption 3, for the fourth term of the RHS, we have

$$\begin{aligned}
&2\eta_t\langle \nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t)\nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle \\
&\ge \frac{2\eta_t}{L}\|\nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t)\nabla f_S(g_S(x_t))\|^2.
\end{aligned} \tag{23}$$

Analogous to the above four terms, we can easily get

$$-2\eta_t\langle \nabla g_S(x_t)\nabla f_S(g_S(x_t)) - v_t\nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle \le 2L_f\eta_t\|v_t - \nabla g_S(x_t)\| \cdot \|x_t^{k,\nu} - x_t\|, \tag{24}$$

$$-2\eta_t\langle v_t\nabla f_S(g_S(x_t)) - v_t\nabla f_S(u_t), x_t^{k,\nu} - x_t\rangle \le 2L_g C_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{k,\nu} - x_t\|, \tag{25}$$

$$-2\eta_t\mathbb{E}_{j_t}[\langle v_t\nabla f_{\nu_{i_t}}(g_S(x_t)) - v_t\nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle] = 0. \tag{26}$$

Putting (20) - (26) into (19) we have

$$\begin{aligned}
&-2\eta_t\mathbb{E}_{j_t}[\langle v_t^{k,\nu}\nabla f(u_t^{k,\nu})-v_t\nabla f(u_t),x_t^{k,\nu}-x_t\rangle]\\
&\le 2L_gC_f\eta_t\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|+2L_f\eta_t\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|\\
&\quad-\frac{2\eta_t}{L}\|\nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu}))-\nabla g_S(x_t)\nabla f_S(g_S(x_t))\|^2\\
&\quad+2L_gC_f\eta_t\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|+2L_f\eta_t\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|.
\end{aligned}\tag{27}$$

Now we begin to bound the third term of the RHS in (18).

$$\begin{aligned}
&\|v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu})-v_t\nabla f_{\nu_{i_t}}(u_t)\|\\
&\le\|v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu})-v_t^{k,\nu}\nabla f_S(u_t^{k,\nu})\|+\|v_t^{k,\nu}\nabla f_S(u_t^{k,\nu})-v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu}))\|\\
&\quad+\|v_t^{k,\nu}\nabla f_S(g_S(x_t^{k,\nu}))-\nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu}))\|+\|\nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu}))-\nabla g_S(x_t)\nabla f_S(g_S(x_t))\|\\
&\quad+\|\nabla g_S(x_t)\nabla f_S(g_S(x_t))-v_t\nabla f_S(g_S(x_t))\|+\|v_t\nabla f_S(g_S(x_t))-v_t\nabla f_S(u_t)\|+\|v_t\nabla f_S(u_t)-v_t\nabla f_{\nu_{i_t}}(u_t)\|.
\end{aligned}$$

Now because of the fact that $(\sum_{i=1}^k a_i)^2\le k\sum_{i=1}^k a_i^2$, we have

$$\begin{aligned}
&\eta_t^2\|v_t^{k,\nu}\nabla f(u_t^{k,\nu})-v_t\nabla f(u_t)\|^2\\
&\le 7\eta_t^2L_g^2\|\nabla f_{\nu_{i_t}}(u_t^{k,\nu})-\nabla f_S(u_t^{k,\nu})\|^2+7\eta_t^2L_g^2C_f^2\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|^2+7\eta_t^2L_f^2\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|^2\\
&\quad+7\eta_t^2\|\nabla g_S(x_t^{k,\nu})\nabla f_S(g_S(x_t^{k,\nu}))-\nabla g_S(x_t)\nabla f_S(g_S(x_t))\|^2\\
&\quad+7\eta_t^2L_f^2\|\nabla g_S(x_t)-v_t\|^2+7\eta_t^2L_g^2C_f^2\|u_t-g_S(x_t)\|^2+7\eta_t^2L_g^2\|\nabla f_{\nu_{i_t}}(u_t)-\nabla f_S(u_t)\|^2.
\end{aligned}\tag{28}$$

Putting (27) and (28) into (18), we have

$$\begin{aligned}
&\mathbb{E}_{j_t}[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2]\\
&\le\|x_t-x_t^{k,\nu}\|^2+2L_gC_f\eta_t\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|+2L_f\eta_t\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|\\
&\quad+2L_gC_f\eta_t\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|+2L_f\eta_t\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|\\
&\quad+7\eta_t^2L_g^2\|\nabla f_{\nu_{i_t}}(u_t^{k,\nu})-\nabla f_S(u_t^{k,\nu})\|^2+7\eta_t^2L_g^2C_f^2\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|^2+7\eta_t^2L_f^2\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|^2\\
&\quad+7\eta_t^2L_f^2\|\nabla g_S(x_t)-v_t\|^2+7\eta_t^2L_g^2C_f^2\|u_t-g_S(x_t)\|^2+7\eta_t^2L_g^2\|\nabla f_{\nu_{i_t}}(u_t)-\nabla f_S(u_t)\|^2.
\end{aligned}$$

where we use $\eta_t\le\frac{2}{7L}$ in the inequality.

Case 2 ($i_t=k$). We have

$$\begin{aligned}
\|x_{t+1}-x_{t+1}^{k,\nu}\|&=\|x_t-\eta_tv_t\nabla f_{\nu_{i_t}}(u_t)-x_t^{k,\nu}+\eta_tv_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu})\|\\
&\le\|x_t-x_t^{k,\nu}\|+\eta_t\|v_t^{k,\nu}\nabla f_{\nu_{i_t}}(u_t^{k,\nu})-v_t\nabla f_{\nu_{i_t}}(u_t)\|\le\|x_t-x_t^{k,\nu}\|+2\eta_tL_gL_f,
\end{aligned}\tag{29}$$

where the first inequality holds by Assumption 1, then we have

$$\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\le\|x_t-x_t^{k,\nu}\|^2+4\eta_tL_gL_f\|x_t-x_t^{k,\nu}\|+4\eta_t^2L_g^2L_f^2.$$

Combining above **Case 1** and **Case 2**, we can get

$$\begin{aligned}
&\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\\
&\le\|x_t-x_t^{k,\nu}\|^2+2L_gC_f\eta_t\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|+2L_f\eta_t\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|\\
&\quad+2L_gC_f\eta_t\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|+2L_f\eta_t\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|\\
&\quad+7\eta_t^2L_g^2\|\nabla f_{\nu_{i_t}}(u_t^{k,\nu})-\nabla f_S(u_t^{k,\nu})\|^2+7\eta_t^2L_g^2C_f^2\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|^2+7\eta_t^2L_f^2\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|^2\\
&\quad+7\eta_t^2L_f^2\|\nabla g_S(x_t)-v_t\|^2+7\eta_t^2L_g^2C_f^2\|u_t-g_S(x_t)\|^2+7\eta_t^2L_g^2\|\nabla f_{\nu_{i_t}}(u_t)-\nabla f_S(u_t)\|^2\\
&\quad+4\eta_tL_gL_f\|x_t^{k,\nu}-x_t\|\cdot\mathbf{1}_{i_t=k}+4\eta_t^2L_g^2L_f^2\cdot\mathbf{1}_{i_t=k}.
\end{aligned}$$

According to Cauchy-Schwarz inequality, we can get

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2] \le{}& \mathbb{E}_A[\|x_t - x_t^{k,\nu}\|^2] + 2L_gC_f\eta_t(\mathbb{E}_A[\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|^2])^{1/2}\cdot(\mathbb{E}_A\|x_t^{k,\nu}-x_t\|^2])^{1/2}\\ &+2L_f\eta_t(\mathbb{E}_A[\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|^2])^{1/2}\cdot(\mathbb{E}_A\|x_t^{k,\nu}-x_t\|^2])^{1/2}\\ &+2L_gC_f\eta_t(\mathbb{E}_A[\|u_t-g_S(x_t)\|^2])^{1/2}\cdot(\mathbb{E}[\|x_t^{k,\nu}-x_t\|^2])^{1/2}\\ &+2L_f\eta_t(\mathbb{E}_A[\|v_t-\nabla g_S(x_t)\|^2])^{1/2}\cdot(\mathbb{E}_A[\|x_t^{k,\nu}-x_t\|^2])^{1/2}\\ &+7\eta_t^2L_g^2C_f^2\mathbb{E}_A[\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|^2]+7\eta_t^2L_f^2\mathbb{E}_A[\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|^2]\\ &+7\eta_t^2L_f^2\mathbb{E}_A[\|\nabla g_S(x_t)-v_t\|^2]+7\eta_t^2L_g^2C_f^2\mathbb{E}_A[\|u_t-g_S(x_t)\|^2]+14\eta_t^2L_f^2\sigma_f^2\\ &+4\eta_tL_gL_f\mathbb{E}_A[\|x_t^{k,\nu}-x_t\|\cdot\mathbf{1}_{i_t=k}]+4\eta_t^2L_g^2L_f^2\cdot\mathbb{E}_A[\mathbf{1}_{i_t=k}].\end{aligned}$$

Besides, according to

$$\mathbb{E}_A[\|x_t^{k,\nu}-x_t\|\mathbf{1}_{[i_t=k]}]=\mathbb{E}_A[\|x_t^{k,\nu}-x_t\|\mathbb{E}_{i_t}[\mathbf{1}_{[i_t=k]}]]=\frac{1}{n}\mathbb{E}_A[\|x_t^{k,\nu}-x_t\|]\le\frac{1}{n}(\mathbb{E}_A[\|x_t^{k,\nu}-x_t\|^2])^{1/2},\tag{30}$$

we can get

$$\begin{aligned}&\mathbb{E}_A[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2]\\ &\le 2L_gC_f\sum_{j=0}^{t}\eta_j((\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2])^{1/2}+(\mathbb{E}_A[\|u_j-g_S(x_j)\|^2])^{1/2})\cdot(\mathbb{E}_A\|x_j-x_j^{k,\nu}\|^2])^{1/2}\\ &\quad+2L_f\sum_{j=0}^{t}\eta_j((\mathbb{E}_A[\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2])^{1/2}+(\mathbb{E}_A[\|v_j-\nabla g_S(x_j)\|^2])^{1/2})\cdot(\mathbb{E}_A\|x_j-x_j^{k,\nu}\|^2])^{1/2}\\ &\quad+7\sum_{j=0}^{t}\eta_j^2L_g^2C_f^2\mathbb{E}_A[\|u_j-g_S(x_j)\|^2]+7\sum_{j=0}^{t}\eta_j^2L_g^2C_f^2\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2]\\ &\quad+7\sum_{j=0}^{t}\eta_j^2L_f^2\mathbb{E}_A[\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2]+7\sum_{j=0}^{t}\eta_j^2L_f^2\mathbb{E}_A[\|v_j-\nabla g_S(x_j)\|^2]+14L_f^2\sigma_f^2\sum_{j=0}^{t}\eta_j^2\\ &\quad+\frac{4L_gL_f}{n}\sum_{j=0}^{t}\eta_j(\mathbb{E}_A[\|x_j-x_j^{k,\nu}\|^2])^{1/2}+\frac{4L_g^2L_f^2}{n}\sum_{j=0}^{t}\eta_j^2.\end{aligned}$$

For notational convenience, we denote by $u_t=(\mathbb{E}_A[\|x_t-x_t^{k,\nu}\|^2])^{1/2}$, define

$$\begin{aligned}\alpha_j={}&2L_gC_f\eta_j(\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2])^{1/2}+2L_gC_f\eta_j(\mathbb{E}_A[\|u_j-g_S(x_j)\|^2])^{1/2}\\ &+2L_f\eta_j(\mathbb{E}_A[\|v_j-\nabla g_S(x_j)\|^2])^{1/2}+2L_f\eta_j(\mathbb{E}_A[\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2])^{1/2}+\frac{4L_gL_f}{n}\eta_j,\end{aligned}$$

and

$$\begin{aligned}S_t={}&7\sum_{j=0}^{t-1}\eta_j^2L_g^2C_f^2\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2]+7\sum_{j=0}^{t-1}\eta_j^2L_g^2C_f^2\mathbb{E}_A[\|u_j-g_S(x_j)\|^2]+14L_f^2\sigma_f^2\sum_{j=0}^{t-1}\eta_j^2\\ &+7\sum_{j=0}^{t-1}\eta_j^2L_f^2\mathbb{E}_A[\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2]+7\sum_{j=0}^{t-1}\eta_j^2L_f^2\mathbb{E}_A[\|\nabla g_S(x_j)-v_j\|^2]+\frac{4L_g^2L_f^2}{n}\sum_{j=0}^{t-1}\eta_j^2.\end{aligned}$$

Using Lemma 14, we can get

$$
\begin{aligned}
u_t &\le \sqrt{S_t} + \sum_{j=1}^{t-1} \alpha_j \\
&\le (7L_g^2 C_f^2 \sum_{j=0}^{t-1} \eta_j^2 \mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{1/2} + (7L_g^2 C_f^2 \sum_{j=0}^{t-1} \eta_j^2 \mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} + (14L_f^2\sigma_f^2 \sum_{j=0}^{t-1} \eta_j^2)^{\frac{1}{2}} \\
&\quad + (7L_f^2 \sum_{j=0}^{t-1} \eta_j^2 \mathbb{E}_A[\|v_j^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2])^{1/2} + (7L_f^2 \sum_{j=0}^{t-1} \eta_j^2 \mathbb{E}_A[\|\nabla g_S(x_j) - v_j\|^2])^{1/2} + (\frac{4L_g^2 L_f^2}{n} \sum_{j=0}^{t-1} \eta_j^2)^{1/2} \\
&\quad + 2L_g C_f \sum_{j=1}^{t-1} \eta_j (\mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{1/2} + 2L_g C_f \sum_{j=1}^{t-1} \eta_j (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\
&\quad + 2L_f \sum_{j=1}^{t-1} \eta_j (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} + 2L_f \sum_{j=1}^{t-1} \eta_j (\mathbb{E}_A[\|v_j^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2])^{1/2} + \frac{4L_g L_f}{n} \sum_{j=1}^{t-1} \eta_j
\end{aligned}
$$

Then according to the inequality that $(\sum_{i=1}^k a_i)^{1/2} \le \sum_{i=1}^k (a_i)^{1/2}$ we can get

$$
\begin{aligned}
u_t \le\ & 5L_g C_f \sum_{j=0}^{t} \eta_j (\mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{1/2} + 5L_g C_f \sum_{j=0}^{t-1} \eta_j (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\
& + 5L_f \sum_{j=0}^{t} \eta_j (\mathbb{E}_A[\|v_j^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2])^{1/2} + 5L_f \sum_{j=0}^{t-1} \eta_j (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\
& + (14L_f^2\sigma_f^2 \sum_{j=0}^{t-1} \eta_j^2)^{\frac{1}{2}} + (\frac{4L_g^2 L_f^2}{n} \sum_{j=0}^{t-1} \eta_j^2)^{1/2} + \frac{4L_g L_f}{n} \sum_{j=1}^{t-1} \eta_j.
\end{aligned}
$$

By setting $\eta_t = \eta$, with T iterations, we have

$$
\begin{aligned}
u_T &\le 10L_g C_f \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} + 10L_f \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\
&\quad + 4L_f\sigma_f \eta\sqrt{T} + \frac{2L_g L_f \eta \sqrt{T}}{\sqrt{n}} + \frac{4L_g L_f \eta T}{n} \\
&\le 10L_g C_f \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} + 10L_f \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\
&\quad + 4L_f\sigma_f \eta\sqrt{T} + \frac{6L_g L_f \eta T}{n},
\end{aligned}
$$

where the first inequality holds by $\sum_{j=1}^t \eta_j (\mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{1/2} \le \sup_S \eta \sum_{j=1}^t (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2}$ and $\sum_{j=1}^t \eta_j (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \le \sup_S \eta (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2}$. The other terms to the RHS are treated similarly. And the second inequality follows by the fact that we often have $n \le T$, therefore $\sqrt{\frac{T}{n}} \le \frac{T}{n}$. We further get

$$
\begin{aligned}
\mathbb{E}_A[\|x_T - x_T^{k,\nu}\|] &\le u_T \\
&\le 10L_g C_f \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} + 10L_f \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\
&\quad + 4L_f\sigma_f \eta\sqrt{T} + \frac{6L_g L_f \eta T}{n}.
\end{aligned} \tag{31}
$$

Then we can get the following result

$$\begin{aligned}\mathbb{E}_A[\|x_T - x_T^{k,\nu}\|] = O\Big(\frac{L_g L_f \eta T}{n} + L_f\sigma_f\eta\sqrt{T} + L_g C_f \sup_S \eta \sum_{j=0}^{T-1}((\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ + (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2}\Big).\end{aligned}$$

Estimation of $\mathbb{E}_A\left[\|x_{t+1} - x_{t+1}^{l,\omega}\|\right]$

Next we give the estimation of $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}]$. Similarly, we consider two cases, $j_t \neq l$ and $j_t = l$.

Case 1 ($j_t \neq l$**).** We have

$$\begin{aligned}&\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \\ &\leq \|x_t - \eta_t v_t \nabla f_{\nu_{i_t}}(u_t) - x_t^{l,\omega} + \eta_t v_t^{l,\omega}\nabla f_{\nu_{i_t}}(u_t^{l,\omega})\|^2 \\ &\leq \|x_t - x_t^{l,\omega}\|^2 - 2\eta_t\langle v_t^{l,\omega}\nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - v_t\nabla f_{\nu_{i_t}}(u_t), x_t^{l,\omega} - x_t\rangle + \eta_t^2\|v_t^{l,\omega}\nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - v_t\nabla f_{\nu_{i_t}}(u_t)\|^2.\end{aligned} \tag{32}$$

Similarly to the process of (18), we have

$$\begin{aligned}&\mathbb{E}_{j_t}[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2] \\ &\leq \|x_t - x_t^{l,\omega}\|^2 + 2L_g C_f \eta_t\|u_t^{l,\omega} - g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| \\ &\quad + 2L_g C_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| \\ &\quad + 7\eta_t^2 L_g^2\|\nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - \nabla f_S(u_t^{l,\omega})\|^2 + 7\eta_t^2 L_g^2 C_f^2\|u_t^{l,\omega} - g_S(x_t^{l,\omega})\|^2 + 7\eta_t^2 L_f^2\|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\|^2 \\ &\quad + 7\eta_t^2 L_f^2\|\nabla g_S(x_t) - v_t\|^2 + 7\eta_t^2 L_g^2 C_f^2\|u_t - g_S(x_t)\|^2 + 7\eta_t^2 L_g^2\|\nabla f_{\nu_{i_t}}(u_t) - \nabla f_S(u_t)\|^2.\end{aligned}$$

where we use $\eta_t \leq \frac{2}{7L}$ in the inequality.

Case 2 ($j_t = l$**).** We have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^{l,\omega}\| &= \|x_t - \eta_t v_t\nabla f_{\nu_{i_t}}(u_t) - x_t^{l,\omega} + \eta_t v_t^{l,\omega}\nabla f_{\nu_{i_t}}(u_t^{l,\omega})\| \\ &\leq \|x_t - x_t^{l,\omega}\| + \eta_t\|v_t^{l,\omega}\nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - v_t\nabla f_{\nu_{i_t}}(u_t)\| \leq \|x_t - x_t^{l,\omega}\| + 2\eta_t L_g L_f,\end{aligned} \tag{33}$$

where the first inequality holds by Assumption 1, then we have

$$\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - x_t^{l,\omega}\|^2 + 4\eta_t L_g L_f\|x_t - x_t^{l,\omega}\| + 4\eta_t^2 L_g^2 L_f^2.$$

Combining **Case 1** and **Case 2** we have

$$\begin{aligned}&\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \\ &\leq \|x_t - x_t^{l,\omega}\|^2 + 2L_g C_f\eta_t\|u_t^{l,\omega} - g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| \\ &\quad + 2L_g C_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| \\ &\quad + 7\eta_t^2 L_g^2 C_f^2\|u_t^{l,\omega} - g_S(x_t^{l,\omega})\|^2 + 7\eta_t^2 L_f^2\|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\|^2 + 7\eta_t^2 L_f^2\|\nabla g_S(x_t) - v_t\|^2 \\ &\quad + 7\eta_t^2 L_g^2 C_f^2\|u_t - g_S(x_t)\|^2 + 14\eta_t^2 L_g^2\sigma_f^2 + 4\eta_t L_g L_f\|x_t^{l,\omega} - x_t\| \cdot \mathbf{1}_{j_t = l} + 4\eta_t^2 L_g^2 L_f^2 \cdot \mathbf{1}_{j_t = l}.\end{aligned}$$

Besides, according to the fact that

$$\mathbb{E}_A[\|x_t^{l,\omega} - x_t\|\mathbf{1}_{[j_t = l]}] = \mathbb{E}_A[\|x_t^{l,\omega} - x_t\|\mathbb{E}_{j_t}[\mathbf{1}_{[j_t = l]}]] = \frac{1}{m}\mathbb{E}_A[\|x_t^{l,\omega} - x_t\|] \leq \frac{1}{m}(\mathbb{E}_A[\|x_t^{l,\omega} - x_t\|^2])^{1/2}.$$

Then similarly to the estimation of $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{k,\nu}\|]$, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_T - x_T^{l,\omega}\|] \leq u_T \\
&\leq 10L_gC_f \sup_S \eta \sum_{j=0}^{T-1}(\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} + 10L_f \sup_S \eta \sum_{j=0}^{T-1}(\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\
&\quad + 4L_f\sigma_f\eta\sqrt{T} + \frac{6L_gL_f\eta T}{m}.
\end{aligned} \tag{34}$$

Then we can get the following result

$$\begin{aligned}
\mathbb{E}_A[\|x_T - x_T^{l,\omega}\|] = O\Big(&\frac{L_gL_f\eta T}{n} + L_f\sigma_f\eta\sqrt{T} + L_gC_f \sup_S \eta \sum_{j=0}^{T-1}((\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\
&+ (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2}\Big).
\end{aligned}$$

Now we combine the above two estimations, we can conclude that

$$\begin{aligned}
\epsilon_\nu + \epsilon_\omega =& O\big(L_f\sigma_f\eta\sqrt{T} + L_gC_f \sup_S \eta \sum_{j=0}^{T-1}((\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} + (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\
&+ \frac{L_gL_f\eta T}{m} + \frac{L_gL_f\eta T}{n}\big).
\end{aligned}$$

This completes the proof. □

Corollary 3 (Two-level Optimization)**.** *Consider Algorithm 2 with* $\eta_t = \eta \leq \frac{1}{4L}$ *and* $\beta_t = \beta < \min\{1/8C_f^2, 1\}$*, for any* $t \in [0, T-1]$*. With the output* $A(S) = x_T$*, then* $\epsilon_\nu + \epsilon_\omega$ *satisfies*

$$O\Big(\eta T\big((\beta T)^{-\frac{c}{2}} + \beta^{1/2} + \eta\beta^{-1/2}\big) + \eta T(\frac{1}{m} + \frac{1}{n})\Big).$$

Proof of Corollary 3. Considering the upadte rule of Algorithm 2, according to Lemma 23 and 24 we have

$$\begin{aligned}
\epsilon_\nu + \epsilon_\omega &= O\big(\eta Tm^{-1} + \eta Tn^{-1} + \eta\sqrt{T} + \eta\sum_{j=0}^{T-1}((j\beta)^{-c} + \sqrt{\beta} + \sqrt{\frac{\eta^2}{\beta}}) \\
&= O\big(\eta Tm^{-1} + \eta Tn^{-1} + \eta\sqrt{T} + \eta T^{-c/2+1}\beta^{-c/2} + \eta T\beta^{1/2} + \eta^2\beta^{-1/2}T\big).
\end{aligned}$$

This complete the proof. □

Before giving the detailed proof of Theorem 5, we first give a useful lemma.

Lemma 25. *Let Assumption 1(ii), 2 (ii) and 3 (ii) hold for the empirical risk* F_S*, for Algorithm 2 and any* $\gamma_t > 0$*, we have*

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \leq& (1 + \frac{\eta_t(L_f + L_gC_f)}{\gamma_t})\|x_t - x_*^S\|^2 + L_g^2L_f^2\eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S)) \\
&+ \eta_t\gamma_tL_f\mathbb{E}_A[\|v_t - \nabla g_S(x_t)\|^2\mathcal{F}_t] + \eta_t\gamma_tL_gC_f\mathbb{E}_A[\|u_t - g_S(x_t)\|^2\mathcal{F}_t].
\end{aligned}$$

where $\mathcal{F}_t$ *is the* σ*-field generated by* $\{\omega_{j_0}, \cdots, \omega_{j_{t-1}}, v_{i_0}, \cdots, v_{i_{t-1}}\}$.

Proof. According to the update rule of Algorithm 2, we have

$$\begin{aligned}
\|x_{t+1} - x_*^S\| &\leq \|x_t - \eta_tv_t\nabla f_{\nu_{i_t}}(u_t) - x_*^S\|^2 \\
&= \|x_t - x_*^S\|^2 + \eta_t^2\|v_t\nabla f_{\nu_{i_t}}(u_t)\|^2 - 2\eta_t\langle x_t - x_*^S, v_t\nabla f_{\nu_{i_t}}(u_t)\rangle \\
&= \|x_t - x_*^S\|^2 + \eta_t^2\|v_t\nabla f_{\nu_{i_t}}(u_t)\|^2 - 2\eta_t\langle x_t - x_*^S, \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\rangle + \theta_t,
\end{aligned}$$

where

$$\theta_t = 2\eta_t \langle x_t - x_*^S, \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - v_t \nabla f_{\nu_{i_t}}(u_t)\rangle.$$

Let $\mathcal{F}_t$ be the σ-field generated by $\{\omega_{j_0}, \cdots, \omega_{j_{t-1}}, v_{i_0}, \cdots, v_{i_{t-1}}\}$. Taking expectation to the above inequality and using Assumption 1, we have

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] &\le \|x_t - x_*^S\|^2 + L_g^2 L_f^2 \eta_t^2 - 2\eta_t \mathbb{E}_A[\langle x_t - x_*^S, \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\rangle \mathcal{F}_t] + \mathbb{E}_A[\theta_t|\mathcal{F}_t] \\
&\le \|x_t - x_*^S\|^2 + L_g^2 L_f^2 \eta_t^2 - 2\eta_t \langle x_t - x_*^S, \nabla F_S(x_t)\rangle + \mathbb{E}_A[\theta_t|\mathcal{F}_t] \\
&\le \|x_t - x_*^S\|^2 + L_g^2 L_f^2 \eta_t^2 - 2\eta_t (F_S(x_t) - F_S(x_*^S)) + \mathbb{E}_A[\theta_t|\mathcal{F}_t],
\end{aligned}$$

where the last inequality holds by the convexity of F_S. As for the term $\mathbb{E}_A[\theta_t|\mathcal{F}_t]$, we have

$$\begin{aligned}
\theta_t &= 2\eta_t \langle x_t - x_*^S, \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - v_t \nabla f_{\nu_{i_t}}(u_t)\rangle \\
&= 2\eta_t \langle x_t - x_*^S, \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - v_t \nabla f_{\nu_{i_t}}(g_S(x_t))\rangle + 2\eta_t \langle x_t - x_*^S, v_t \nabla f_{\nu_{i_t}}(g_S(x_t)) - v_t \nabla f_{\nu_{i_t}}(u_t)\rangle \\
&\le 2\eta_t L_f \|x_t - x_*^S\| \cdot \|v_t - \nabla g_S(x_t)\| + 2\eta_t L_g C_f \|x_t - x_*^S\| \cdot \|u_t - g_S(x_t)\| \\
&\le \frac{\eta_t(L_f + L_g C_f)}{\gamma_t} \|x_t - x_*^S\|^2 + \eta_t \gamma_t L_f \|v_t - \nabla g_S(x_t)\|^2 + \eta_t \gamma_t L_g C_f \|u_t - g_S(x_t)\|^2,
\end{aligned}$$

where the last inequality holds by Cauchy-Schwartz inequality.

Combining above two inequalities, we have

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \le (1 + \frac{\eta_t(L_f + L_g C_f)}{\gamma_t})\|x_t - x_*^S\|^2 + L_g^2 L_f^2 \eta_t^2 - 2\eta_t (F_S(x_t) - F_S(x_*^S)) \\
+ \eta_t \gamma_t L_f \mathbb{E}_A[\|v_t - \nabla g_S(x_t)\|^2 \mathcal{F}_t] + \eta_t \gamma_t L_g C_f \mathbb{E}_A[\|u_t - g_S(x_t)\|^2 \mathcal{F}_t].
\end{aligned}$$

This complete the proof. □

Proof of Theorem 5. Now we begin to proof the Theorem 5. According to Lemma 25, setting $\eta_t = \eta$, $\beta_t = \beta$ and let $\gamma_t = \frac{1}{\sqrt{\beta}}$, by rearranging and adding up, we get

$$\begin{aligned}
2\eta \sum_{t=1}^T \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \le \mathbb{E}_A[\|x_1 - x_*^S\|^2] + \eta\sqrt{\beta} \sum_{t=1}^T (L_f + L_g C_f)\|x_t - x_*^S\|^2 + L_g^2 L_f^2 \eta^2 T \\
+ \frac{\eta}{\sqrt{\beta}} \sum_{t=1}^T (L_f \|v_t - \nabla g_S(x_t)\|^2 + L_g C_f \sum_{t=1}^T \|u_t - g_S(x_t)\|^2).
\end{aligned}$$

Then according to the definition that $\mathbb{E}[\|x_t - x_*^S\|^2]$ is bounded by D_x, and Lemma 23 and Lemma 24, we have

$$\begin{aligned}
&2\eta \sum_{t=1}^T \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\
&\le D_x + \eta\sqrt{\beta}(L_f + L_g C_f) D_x T + L_g^2 L_f^2 \eta^2 T \\
&\quad + \frac{\eta}{\sqrt{\beta}} L_g C_f (\frac{c}{e})^c \sum_{t=1}^T (t\beta)^{-c} \mathbb{E}_A[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2 \eta \beta^{1/2} L_g C_f T + 2L_g^5 L_f^2 C_f \frac{\eta^3}{\beta^{3/2}} T \\
&\quad + \frac{\eta}{\sqrt{\beta}} L_f (\frac{c}{e})^c \sum_{t=1}^T (t\beta)^{-c} \mathbb{E}_A[\|v_0 - \nabla g_S(x_0)\|^2] + 2\sigma_{g'}^2 \eta \beta^{1/2} L_f T + 2L_g^4 L_f^3 \frac{\eta^3}{\beta^{3/2}} T.
\end{aligned} \tag{35}$$

According to (11), without losing generality, let $c \ne 1$, we have

$$\begin{aligned}
&\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)] \\
&\le O\Big(D_x(\eta T)^{-1} + (L_g C_f U + L_f V)(\beta T)^{-c}\beta^{-\frac{1}{2}} + (L_g C_f \sigma_g^2 + L_f \sigma_{g'}^2 + (L_f + L_g C_f) D_x)\beta^{1/2} \\
&\quad + L_g^2 L_f^2 \eta + (L_g^5 L_f^2 C_f + L_g^4 L_f^3)\eta^2 \beta^{-\frac{3}{2}}\Big),
\end{aligned}$$

where $\mathbb{E}[\|u_0 - g_S(x_0)\|^2] \leq U$ and $\mathbb{E}[\|v_0 - \nabla g_S(x_0)\|^2] \leq V$.

This complete the proof. □

proof of Theorem 6. Putting Lemma 23 and 24 into (31) and (34), for any $c > 0$, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] \\
&\leq 10L_gC_f \sup_S \eta\sqrt{(\frac{c}{e})^c U}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 10L_gC_f\eta\sqrt{2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}}t \\
&\quad + 10L_f \sup_S \eta\sqrt{(\frac{c}{e})^c V}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 10L_f\eta\sqrt{2\beta\sigma_{g'}^2 + \frac{2L_g^4L_f^2\eta^2}{\beta}}t + \frac{6L_gL_f\eta t}{n} + 4L_f\sigma_f\eta\sqrt{t}.
\end{aligned}$$

Similarly, we can get

$$\begin{aligned}
&\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \\
&\leq 10L_gC_f \sup_S \eta\sqrt{(\frac{c}{e})^c U}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 10L_gC_f\eta\sqrt{2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}}t \\
&\quad + 10L_f \sup_S \eta\sqrt{(\frac{c}{e})^c V}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 10L_f\eta\sqrt{2\beta\sigma_{g'}^2 + \frac{2L_g^4L_f^2\eta^2}{\beta}}t + \frac{6L_gL_f\eta t}{m} + 4L_f\sigma_f\eta\sqrt{t}.
\end{aligned}$$

Combining above two inequalities, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] + 4\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \\
&\leq \frac{24L_gL_f\eta t}{m} + \frac{6L_gL_f\eta t}{n} + 50L_gC_f \sup_S \eta\sqrt{(\frac{c}{e})^c U}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 50L_gC_f\eta\sqrt{2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}}t \\
&\quad + 50L_f \sup_S \eta\sqrt{(\frac{c}{e})^c V}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 50L_f\eta\sqrt{2\beta\sigma_{g'}^2 + \frac{2L_g^4L_f^2\eta^2}{\beta}}t + 20L_f\sigma_f\eta\sqrt{t}.
\end{aligned}$$

Putting above inequality into Lemma 20, we have

$$\begin{aligned}
\mathbb{E}_{S,A}[F(x_t) - F_S(x_t)] \leq{}& 50L_g^2L_fC_f \sup_S \eta\sqrt{(\frac{c}{e})^c U}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 50L_g^2L_fC_f\eta\sqrt{2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}}t \\
&+ 50L_gL_f^2 \sup_S \eta\sqrt{(\frac{c}{e})^c V}\beta^{-\frac{c}{2}} \sum_{j=0}^{t-1} \sqrt{j^{-c}} + 50L_gL_f^2\eta\sqrt{2\beta\sigma_{g'}^2 + \frac{2L_g^4L_f^2\eta^2}{\beta}}t \qquad (36) \\
&+ \frac{24L_g^2L_f^2\eta t}{m} + \frac{6L_g^2L_f^2\eta t}{n} + L_f\sqrt{m^{-1}\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(A(S)))]}.
\end{aligned}$$

Due to

$$\sum_{t=1}^{T} \mathbb{E}_{S,A}[F(x_t) - F(x_*)] \leq \sum_{t=1}^{T} \mathbb{E}_{S,A}[F(x_t) - F_S(x_*)] = \sum_{t=1}^{T} \mathbb{E}_{S,A}[F(x_t) - F_S(x_t) + F_S(x_t) - F_S(x_*^S)],$$

then combining (36) and (35), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{S,A}[F(x_t)-F(x_*)] \\
&\leq \frac{1}{2\sqrt{\beta}}L_gC_f(\frac{c}{e})^c\sum_{t=1}^{T}(t\beta)^{-c}\mathbb{E}[\|u_0-g_S(x_0)\|^2]+\sigma_g^2\beta^{1/2}L_gC_fT+L_g^5L_f^2C_f\frac{\eta^2}{\beta^{3/2}}T \\
&\quad+\frac{1}{2\sqrt{\beta}}L_gL_f(\frac{c}{e})^c\sum_{t=1}^{T}(t\beta)^{-c}\mathbb{E}[\|v_0-\nabla g_S(x_0)\|^2]+\sigma_{g'}^2\beta^{1/2}L_gL_fT+L_g^5L_f^2C_f\frac{\eta^2}{\beta^{3/2}}T \\
&\quad+\frac{D_x}{2\eta}+\sqrt{\beta}(L_gL_f+L_gC_f+L_g)D_xT/2+L_g^2L_f^2\eta T/2+\frac{24L_g^2L_f^2\eta}{m}\sum_{t=1}^{T}t+\frac{6L_g^2L_f^2\eta}{n}\sum_{t=1}^{T}t \\
&\quad+50L_g^2L_fC_f\sup_S\eta\sqrt{(\frac{c}{e})^cU}\beta^{-\frac{c}{2}}\sum_{t=1}^{T}\sum_{j=0}^{t-1}\sqrt{j^{-c}}+50L_g^2L_fC_f\eta\sqrt{2\sigma_g^2\beta+\frac{2L_g^4L_f^2\eta^2}{\beta}}\sum_{t=1}^{T}t \\
&\quad+50L_gL_f^2\sup_S\eta\sqrt{(\frac{c}{e})^cU_{\dot{v}}}\beta^{-\frac{c}{2}}\sum_{t=1}^{T}\sum_{j=0}^{t-1}\sqrt{j^{-c}}+50L_gL_f^2\eta\sqrt{2\beta\sigma_{g'}^2+\frac{2L_g^4L_f^2\eta^2}{\beta}}\sum_{t=1}^{T}t \\
&\quad+L_fT\sqrt{m^{-1}\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(A(S)))]}.
\end{aligned}$$

Using (13) we have, for any $c>0$,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{S,A}[F(x_t)-F(x_*)] \leq O\Big(&\beta^{-\frac{1}{2}-c}T^{1-c}(\log T)^{\mathbf{1}_{c=1}}+\beta^{\frac{1}{2}}T+\frac{\eta^2T}{\beta^{\frac{3}{2}}}+\frac{1}{\eta}+\eta T+T^2\eta(\frac{1}{n}+\frac{1}{m}) \\
&+\eta\beta^{-\frac{c}{2}}T^{2-\frac{c}{2}}(\log T)^{\mathbf{1}_{c=2}}+T^2\eta\sqrt{\beta}+\frac{\eta^2T^2}{\sqrt{\beta}}+\frac{T}{\sqrt{m}}\Big).
\end{aligned}$$

Dividing both sides of the above inequality, setting $\eta=T^{-a}$ and $\beta=T^{-b}$, and from the choice of $A(S)$, we have

$$\begin{aligned}
&\mathbb{E}_{S,A}[F(A(S))-F(x_*)] \\
&\leq O\Big(T^{-(1-b)c+\frac{b}{2}}(\log T)^{\mathbf{1}_{c=1}}+T^{-\frac{b}{2}}+T^{\frac{3}{2}b-2a}+T^{a-1}+T^{-a}+T^{1-a}(\frac{1}{n}+\frac{1}{m}) \\
&\quad+T^{1-a-\frac{c}{2}(b-1)}(\log T)^{\mathbf{1}_{c=2}}+T^{1-a-\frac{b}{2}}+T^{1+\frac{b}{2}-2a}+\frac{1}{\sqrt{m}}\Big).
\end{aligned}$$

Since $a,b\in(0,1]$, as long as we have $c>4$, the dominating terms are the following $O(T^{1-a-\frac{b}{2}})$, $O(T^{1+\frac{b}{2}-2a})$, $O(n^{-1}T^{1-a})$, $O(m^{-1}T^{1-a})$, $O(T^{a-1})$, and $O(T^{\frac{3}{2}b-2a})$. Setting $a=b=4/5$, then we have

$$\mathbb{E}_{S,A}[F(A(S))-F(x_*)]=O(T^{-\frac{1}{5}}+\frac{T^{\frac{1}{5}}}{n}+\frac{T^{\frac{1}{5}}}{m}+\frac{1}{\sqrt{m}}).$$

Choosing $T=O(\max\{n^{2.5},m^{2.5}\})$, we have the following bound

$$\mathbb{E}_{S,A}[F(A(S))-F(x_*)]=O(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}).$$

This complete the proof. □

D.2. Strongly-convex-setting

proof of Theorem 8. Similar to the proof for convex setting, we use the same notations. Since changing one sample data can happen in either S_ν or S_ω, we estimate $\mathbb{E}[\|x_{t+1}-x_{t+1}^{k,\nu}\|]$ and $\mathbb{E}[\|x_{t+1}-x_{t+1}^{l,\omega}\|]$.

Estimation of $\mathbb{E}_A\left[\|x_{t+1} - x_{t+1}^{k,\nu}\|\right]$

we will consider two cases: $i_t \neq k$ and $i_t = k$.

Case 1 ($i_t \neq k$)**.** We have

$$
\begin{aligned}
&\|x_{t+1} - x_{t+1}^{k,\nu}\|^2 \\
&\leq \|x_t - \eta_t v_t \nabla f_{\nu_{i_t}}(u_t) - x_t^{k,\nu} + \eta_t v_t^{k,\nu} \nabla f_{\nu_{i_t}}(u_t^{k,\nu})\|^2 \\
&\leq \|x_t - x_t^{k,\nu}\|^2 - 2\eta_t \langle v_t^{k,\nu} \nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t \nabla f_{\nu_{i_t}}(u_t), x_t^{k,\nu} - x_t\rangle + \eta_t^2 \|v_t^{k,\nu} \nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t \nabla f_{\nu_{i_t}}(u_t)\|^2.
\end{aligned}
\tag{37}
$$

We begin to estimate the second term in (37).

$$
\begin{aligned}
&- 2\eta_t \langle v_t^{k,\nu} \nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t \nabla f_{\nu_{i_t}}(u_t), x_t^{k,\nu} - x_t\rangle \\
&= -2\eta_t \langle v_t^{k,\nu} \nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - v_t^{k,\nu} \nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t \langle v_t^{k,\nu} \nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})) - v_t^{k,\nu} \nabla f_S(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t \langle v_t^{k,\nu} \nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t^{k,\nu}) \nabla f_S(g_S(x_t^{k,\nu})), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t \langle \nabla g_S(x_t^{k,\nu}) \nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t) \nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t \langle \nabla g_S(x_t) \nabla f_S(g_S(x_t)) - v_t \nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle - 2\eta_t \langle v_t \nabla f_S(g_S(x_t)) - v_t \nabla f_S(u_t), x_t^{k,\nu} - x_t\rangle \\
&\quad - 2\eta_t \langle v_t \nabla f_S(u_t) - v_t \nabla f_{\nu_{i_t}}(u_t), x_t^{k,\nu} - x_t\rangle.
\end{aligned}
\tag{38}
$$

Changing the setting from convex to strongly convex will only affect the fourth item on the RHS of (19), and the other items will remain the same as before. Now we estimate the fourth term on the RHS of (38).

$$
\begin{aligned}
&\langle \nabla g_S(x_t^{k,\nu}) \nabla f_S(g_S(x_t^{k,\nu}) - \nabla g_S(x_t) \nabla f_S(g_S(x_t)), x_t^{k,\nu} - x_t\rangle \\
&\geq \frac{L\mu}{L+\mu} \|x_t^{k,\nu} - x_t\|^2 + \frac{1}{L+\mu} \|\nabla g_S(x_t^{k,\nu}) \nabla f_S(g_S(x_t^{k,\nu})) - \nabla g_S(x_t) \nabla f_S(g_S(x_t))\|^2.
\end{aligned}
$$

Then substituting above inequality into (37) we have

$$
\begin{aligned}
&\|x_{t+1} - x_{t+1}^{k,\nu}\| \\
&\leq (1 - \frac{2L\mu\eta_t}{L+\mu}) \|x_t - x_t^{k,\nu}\|^2 + 2L_g C_f \eta_t \|u_t^{k,\nu} - g_S(x_t^{k,\nu})\| \cdot \|x_t^{k,\nu} - x_t\| + 2L_f \eta_t \|v_t^{k,\nu} - \nabla g_S(x_t^{k,\nu})\| \cdot \|x_t^{k,\nu} - x_t\| \\
&\quad + 2L_g C_f \eta_t \|u_t - g_S(x_t)\| \cdot \|x_t^{k,\nu} - x_t\| + 2L_f \eta_t \|u_t - g_S(x_t)\| \cdot \|x_t^{k,\nu} - x_t\| \\
&\quad + 7\eta_t^2 L_g^2 \|\nabla f_{\nu_{i_t}}(u_t^{k,\nu}) - \nabla f_S(u_t^{k,\nu})\|^2 + 7\eta_t^2 L_g^2 C_f^2 \|u_t^{k,\nu} - g_S(x_t^{k,\nu})\|^2 + 7\eta_t^2 L_f^2 \|v_t^{k,\nu} - \nabla g_S(x_t^{k,\nu})\|^2 \\
&\quad + 7\eta_t^2 L_f^2 \|\nabla g_S(x_t) - v_t\|^2 + 7\eta_t^2 L_g^2 C_f^2 \|u_t - g_S(x_t)\|^2 + 7\eta_t^2 L_g^2 \|\nabla f_{\nu_{i_t}}(u_t) - \nabla f_S(u_t)\|^2.
\end{aligned}
$$

where the inequality holds by $\eta_t \leq \frac{2}{7(L+\mu)}$.

Case 2 ($i_t = k$)**.** We have

$$
\begin{aligned}
\|x_{t+1} - x_{t+1}^{l,\omega}\| &= \|x_t - \eta_t v_t \nabla f_{\nu_{i_t}}(u_t) - x_t^{l,\omega} + \eta_t v_t^{l,\omega} \nabla f_{\nu_{i_t}}(u_t^{l,\omega})\| \\
&\leq \|x_t - x_t^{l,\omega}\| + \eta_t \|v_t^{l,\omega} \nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - v_t \nabla f_{\nu_{i_t}}(u_t)\| \leq \|x_t - x_t^{l,\omega}\| + 2\eta_t L_g L_f,
\end{aligned}
\tag{39}
$$

where the first inequality holds by Assumption 1, then we have

$$
\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - x_t^{l,\omega}\|^2 + 4\eta_t L_g L_f \|x_t - x_t^{l,\omega}\| + 4\eta_t^2 L_g^2 L_f^2.
$$

Combining above two cases, we have

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2] \le{}& (1-\frac{2L\mu\eta_t}{L+\mu})\mathbb{E}_A[\|x_t-x_t^{k,\nu}\|^2] + 2L_gC_f\eta_t\mathbb{E}_A[\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|] \\
&+ 2L_f\eta_t\mathbb{E}_A[\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|\cdot\|x_t^{k,\nu}-x_t\|] \\
&+ 2L_gC_f\eta_t\mathbb{E}_A[\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|] + 2L_f\eta_t\mathbb{E}_A[\|u_t-g_S(x_t)\|\cdot\|x_t^{k,\nu}-x_t\|] \\
&+ 7\eta_t^2L_g^2C_f^2\mathbb{E}_A[\|u_t^{k,\nu}-g_S(x_t^{k,\nu})\|^2] + 7\eta_t^2L_f^2\mathbb{E}_A[\|v_t^{k,\nu}-\nabla g_S(x_t^{k,\nu})\|^2] \\
&+ 7\eta_t^2L_f^2\mathbb{E}_A[\|\nabla g_S(x_t)-v_t\|^2] + 7\eta_t^2L_g^2C_f^2\mathbb{E}_A[\|u_t-g_S(x_t)\|^2] + 14\eta_t^2L_g^2\sigma_f^2 \\
&+ 4\eta_tL_gL_f\mathbb{E}_A[\|x_t-x_t^{k,\nu}\|\mathbf{1}_{[i_t=k]}] + 4\eta_t^2L_f^2L_g^2\mathbb{E}_A[\mathbf{1}_{[i_t=k]}].
\end{aligned}$$

By setting $\eta_t=\eta$, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2] \\
&\le 2L_gC_f\eta\sum_{j=0}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}((\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2])^{1/2}+(\mathbb{E}_A[\|u_j-g_S(x_j)\|^2])^{1/2})(\mathbb{E}_A[\|x_j-x_j^{k,\nu}\|^2])^{1/2} \\
&\quad + 2L_f\eta\sum_{j=0}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}((\mathbb{E}_A\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2)^{1/2}+(\mathbb{E}_A\|v_j-\nabla g_S(x_j)\|^2)^{1/2})(\mathbb{E}_A[\|x_j-x_j^{k,\nu}\|^2])^{1/2} \\
&\quad + 7\eta^2L_g^2C_f^2\sum_{j=1}^{t}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2] + 7\eta^2L_g^2C_f^2\sum_{j=1}^{t}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|u_j-g_S(x_j)\|^2] \\
&\quad + 7\eta^2L_f^2\sum_{j=1}^{t}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2] + 7\eta^2L_f^2\sum_{j=1}^{t}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|\nabla g_S(x_j)-v_j\|^2] \\
&\quad + \frac{4\eta L_gL_f}{n}\sum_{j=1}^{t}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|x_j-x_j^{k,\nu}\|] + \frac{4\eta^2L_f^2L_g^2}{n}\sum_{j=1}^{t}(1-\frac{2L\mu\eta}{L+\mu})^{t-j} + 14\eta^2L_g^2\sigma_f^2\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j},
\end{aligned}$$

where the inequality holds by Lemma 13, Cauchy-Schwarz inequality, and the fact that $x_0=x_0^{k,\nu}$. Define $u_t=(\mathbb{E}_A[\|x_t-x_t^{k,\nu}\|^2])^{1/2}$, we have

$$\begin{aligned}
u_t^2 \le{}& 2L_gC_f\eta\sum_{j=0}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}((\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2])^{1/2}+(\mathbb{E}_A[\|u_j-g_S(x_j)\|^2])^{1/2})u_j \\
&+ 2L_f\eta\sum_{j=0}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}((\mathbb{E}_A\|v_t^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2)^{1/2}+(\mathbb{E}_A\|v_j-\nabla g_S(x_j)\|^2)^{1/2})u_j \\
&+ 7\eta^2L_g^2C_f^2\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|u_j^{k,\nu}-g_S(x_j^{k,\nu})\|^2] + 7\eta^2L_g^2C_f^2\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|u_j-g_S(x_j)\|^2] \\
&+ 7\eta^2L_f^2\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|v_j^{k,\nu}-\nabla g_S(x_j^{k,\nu})\|^2] + 7\eta^2L_f^2\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|\nabla g_S(x_j)-v_j\|^2] \\
&+ \frac{4\eta L_gL_f}{n}\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j}\mathbb{E}_A[\|x_j-x_j^{k,\nu}\|] + \frac{4\eta^2L_f^2L_g^2}{n}\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j} + 14\eta^2L_g^2\sigma_f^2\sum_{j=1}^{t-1}(1-\frac{2L\mu\eta}{L+\mu})^{t-j},
\end{aligned}$$

Furthermore, define

$$\begin{aligned}\alpha_j \leq\ & 2L_g C_f \eta \sum_{j=0}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} ((\mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{1/2} + (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2}) \\ & + 2L_f \eta \sum_{j=0}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} ((\mathbb{E}_A\|v_t^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2)^{1/2} + (\mathbb{E}_A\|v_j - \nabla g_S(x_j)\|^2)^{1/2}) \\ & + \frac{4\eta L_g L_f}{n} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j-1},\end{aligned}$$

and

$$\begin{aligned}S_t \leq\ & 7\eta^2 L_g^2 C_f^2 \sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2] + 7\eta^2 L_g^2 C_f^2 \sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|u_j - g_S(x_j)\|^2] \\ & + 7\eta^2 L_f^2 \sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|v_j^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2] + 7\eta^2 L_f^2 \sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|\nabla g_S(x_j) - v_j\|^2] \\ & + \frac{4\eta^2 L_f^2 L_g^2}{n} \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j-1} + 14\eta^2 L_g^2 \sigma_f^2 \sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j},\end{aligned}$$

using Lemma 14, we have

$$\begin{aligned}u_t \leq\ & 3\eta L_g C_f (\sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{\frac{1}{2}} + 3\eta L_g C_f (\sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{\frac{1}{2}} \\ & + 3\eta L_f (\sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|v_j^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2])^{\frac{1}{2}} + 3\eta L_f (\sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|\nabla g_S(x_j) - v_j\|^2])^{\frac{1}{2}} \\ & + 2L_g C_f \eta \sum_{j=0}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} ((\mathbb{E}_A[\|u_j^{k,\nu} - g_S(x_j^{k,\nu})\|^2])^{1/2} + (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2}) \\ & + 2L_f \eta \sum_{j=0}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} ((\mathbb{E}_A\|v_j^{k,\nu} - \nabla g_S(x_j^{k,\nu})\|^2)^{1/2} + (\mathbb{E}_A\|v_j - \nabla g_S(x_j)\|^2)^{1/2}) \\ & + \frac{2L_g L_f (L+\mu)}{n} + \sqrt{\frac{2L_f^2 L_g^2 \eta (L+\mu)}{L\mu n}} + \sqrt{\frac{14 L_g^2 \sigma_f^2 \eta (L+\mu)}{L\mu}},\end{aligned}$$

where we use the inequality that $(\sum_{i=1}^{k} a_i)^{1/2} \leq \sum_{i=1}^{k} (a_i)^{1/2}$ and (7). Then, with T iterations, we have

$$\begin{aligned}\mathbb{E}_A[\|x_T - x_T^{k,\nu}\|] \leq\ & 6L_g C_f \eta \sup_S (\sum_{j=0}^{T-1} (1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1} \mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & + 6L_f \eta \sup_S (\sum_{j=0}^{T-1} (1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1} \mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\ & + 4L_g C_f \eta \sup_S \sum_{j=0}^{T-1} (1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1} (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & + 4L_f \eta \sup_S \sum_{j=0}^{T-1} (1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1} (\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\ & + \frac{2L_g L_f (L+\mu)}{n} + \sqrt{\frac{2L_f^2 L_g^2 \eta (L+\mu)}{L\mu n}} + \sqrt{\frac{14 L_g^2 \sigma_f^2 \eta (L+\mu)}{L\mu}}.\end{aligned}$$

Estimation of $\mathbb{E}_A\left[\|x_{t+1} - x_{t+1}^{l,\omega}\|\right]$

Similarly, we will consider two cases: $j_t \neq l$ and $j_t = l$.

Case 1 ($j_t \neq l$).

Similarly, we have

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,\omega}\| \\
&\leq (1 - \frac{2L\mu\eta_t}{L+\mu})\|x_t - x_t^{l,\omega}\|^2 + 2L_g C_f \eta_t \|u_t^{l,\omega} - g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| \\
&\quad + 2L_g C_f \eta_t \|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| \\
&\quad + 7\eta_t^2 L_g^2 \|\nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - \nabla f_S(u_t^{l,\omega})\|^2 + 7\eta_t^2 L_g^2 C_f^2 \|u_t^{l,\omega} - g_S(x_t^{l,\omega})\|^2 + 7\eta_t^2 L_f^2 \|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\|^2 \\
&\quad + 7\eta_t^2 L_f^2 \|\nabla g_S(x_t) - v_t\|^2 + 7\eta_t^2 L_g^2 C_f^2 \|u_t - g_S(x_t)\|^2 + 7\eta_t^2 L_g^2 \|\nabla f_{\nu_{i_t}}(u_t) - \nabla f_S(u_t)\|^2.
\end{aligned}$$

where the inequality holds by $\eta_t \leq \frac{2}{7(L+\mu)}$.

Case 2 ($j_t = l$). We have

$$\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - x_t^{l,\omega}\|^2 + 4\eta_t L_g L_f \|x_t - x_t^{l,\omega}\| + 4\eta_t^2 L_g^2 L_f^2.$$

Combining the above two cases, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2] \\
&\leq (1 - \frac{2L\mu\eta_t}{L+\mu})\|x_t - x_t^{l,\omega}\|^2 + 2L_g C_f \eta_t \|u_t^{l,\omega} - g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\| \cdot \|x_t^{l,\omega} - x_t\| \\
&\quad + 2L_g C_f \eta_t \|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| + 2L_f\eta_t\|u_t - g_S(x_t)\| \cdot \|x_t^{l,\omega} - x_t\| \\
&\quad + 7\eta_t^2 L_g^2 \|\nabla f_{\nu_{i_t}}(u_t^{l,\omega}) - \nabla f_S(u_t^{l,\omega})\|^2 + 7\eta_t^2 L_g^2 C_f^2 \|u_t^{l,\omega} - g_S(x_t^{l,\omega})\|^2 + 7\eta_t^2 L_f^2 \|v_t^{l,\omega} - \nabla g_S(x_t^{l,\omega})\|^2 \\
&\quad + 7\eta_t^2 L_f^2 \|\nabla g_S(x_t) - v_t\|^2 + 7\eta_t^2 L_g^2 C_f^2 \|u_t - g_S(x_t)\|^2 + 7\eta_t^2 L_g^2 \|\nabla f_{\nu_{i_t}}(u_t) - \nabla f_S(u_t)\|^2 \\
&\quad + 4\eta_t L_g L_f \mathbb{E}_A[\|x_t - x_t^{l,\omega}\| \mathbf{1}_{[j_t = l]}] + 4\eta_t^2 L_f^2 L_g^2 \mathbb{E}_A[\mathbf{1}_{[j_t = l]}].
\end{aligned}$$

Setting $\eta_t = \eta$, telescoping above inequality from 1 to t we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2] \\
&\leq 2L_g C_f \eta \sum_{j=0}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} ((\mathbb{E}_A[\|u_j^{l,\omega} - g_S(x_j^{l,\omega})\|^2])^{1/2} + (\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2})(\mathbb{E}_A[\|x_j - x_j^{l,\omega}\|^2])^{1/2} \\
&\quad + 2L_f \eta \sum_{j=0}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} ((\mathbb{E}_A\|v_j^{l,\omega} - \nabla g_S(x_j^{l,\omega})\|^2)^{1/2} + (\mathbb{E}_A\|v_j - \nabla g_S(x_j)\|^2)^{1/2})(\mathbb{E}_A[\|x_j - x_j^{l,\omega}\|^2])^{1/2} \\
&\quad + 7\eta^2 L_g^2 C_f^2 \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|u_j^{l,\omega} - g_S(x_j^{l,\omega})\|^2] + 7\eta^2 L_g^2 C_f^2 \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|u_j - g_S(x_j)\|^2] \\
&\quad + 7\eta^2 L_f^2 \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|v_j^{l,\omega} - \nabla g_S(x_j^{l,\omega})\|^2] + 7\eta^2 L_f^2 \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|\nabla g_S(x_j) - v_j\|^2] \\
&\quad + \frac{4\eta L_g L_f}{m} \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} \mathbb{E}_A[\|x_j - x_j^{l,\omega}\|] + \frac{4\eta^2 L_f^2 L_g^2}{m} \sum_{j=1}^{t} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j} + 14\eta^2 L_g^2 \sigma_f^2 \sum_{j=1}^{t-1} (1 - \frac{2L\mu\eta}{L+\mu})^{t-j},
\end{aligned}$$

Then with T iterations, we have

$$\begin{aligned}\mathbb{E}_A[\|x_T - x_T^{l,\omega}\|] \leq {} & 6L_g C_f \eta \sup_S (\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & + 6L_f \eta \sup_S (\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\ & + 4L_g C_f \eta \sup_S \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & + 4L_f \eta \sup_S \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\ & + \frac{2L_g L_f (L+\mu)}{m} + \sqrt{\frac{2L_f^2 L_g^2 \eta (L+\mu)}{L\mu m}} + \sqrt{\frac{14L_g^2 \sigma_f^2 \eta (L+\mu)}{L\mu}}.\end{aligned} \tag{40}$$

Now we combine the above results for estimating $\mathbb{E}_A[\|x_T - x_T^{k,\nu}\|]$ and $\mathbb{E}_A[\|x_T - x_T^{l,\omega}\|]$, we have

$$\begin{aligned}\epsilon_\nu + \epsilon_\omega \leq {} & 20L_g C_f \eta \sup_S \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & + 20L_f \eta \sup_S \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} + \sqrt{\frac{2L_f^2 L_g^2 \eta (L+\mu)}{L\mu m}} \\ & + \frac{2(L+\mu)L_g L_f}{L\mu m} + \sqrt{\frac{2L_f^2 L_g^2 \eta (L+\mu)}{L\mu n}} + \frac{2(L+\mu)L_g L_f}{L\mu n} + 8\sqrt{\frac{L_g^2 \sigma_f^2 \eta (L+\mu)}{L\mu}}.\end{aligned} \tag{41}$$

Now we will illustrate why the second inequality of above holds true. According to Lemma 23, we have

$$\begin{aligned} & (\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & \leq (\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}((\frac{c}{e})^c (j\beta)^{-c})\mathbb{E}[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2\beta + \frac{2L_g^4 L_f^2 \eta^2}{\beta})^{1/2} \\ & \leq (\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(2\sigma_g^2\beta + \frac{2L_g^4 L_f^2 \eta^2}{\beta}))^{\frac{1}{2}} + ((\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\frac{c}{e})^c (j\beta)^{-c})\mathbb{E}[\|u_0 - g_S(x_0)\|^2])^{\frac{1}{2}} \\ & \leq \frac{2\sigma_g\sqrt{\beta(L+\mu)}}{\sqrt{L\mu\eta}} + \frac{2L_g^2 L_f \eta \sqrt{L+\mu}}{\sqrt{L\mu\eta\beta}} + (\frac{c}{e})^{\frac{c}{2}} \frac{\sqrt{U_u}(L+\mu)}{L\mu} T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}.\end{aligned}$$

Likewise,

$$\begin{aligned} & \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ & \leq \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}((\frac{c}{e})^c (j\beta)^{-c})\mathbb{E}[\|u_0 - g_S(x_0)\|^2] + 2\sigma_g^2\beta + \frac{2L_g^4 L_f^2 \eta^2}{\beta})^{\frac{1}{2}} \\ & \leq \sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(\sigma_g\sqrt{2\beta} + \frac{L_g^2 L_f \eta \sqrt{2}}{\sqrt{\beta}}) + (\frac{c}{e})^{\frac{c}{2}}\sqrt{U_u}\sum_{j=0}^{T-1}(1 - \frac{2L\mu\eta}{L+\mu})^{T-j-1}(j\beta)^{-\frac{c}{2}} \\ & \leq \frac{(L+\mu)\sigma_g\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2 L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}} \frac{\sqrt{U_u}(L+\mu)}{L\mu\eta} T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}.\end{aligned} \tag{42}$$

According to the above two inequalities, we can get the dominating term is $\sum_{j=0}^{T-1}(1-\frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|u_j-g_S(x_j)\|^2])^{1/2}$, then the inequality (41) holds true. The treatment of the other items is similar, so we won't go into details. Since often we have $\eta \leq \min(\frac{1}{n}, \frac{1}{m})$, then we have

$$\begin{aligned}\epsilon_\nu + \epsilon_\omega \leq O\Big(& L_g C_f \eta \sup_S \sum_{j=0}^{T-1}(1-\frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|u_j - g_S(x_j)\|^2])^{1/2} \\ &+ L_f \eta \sup_S \sum_{j=0}^{T-1}(1-\frac{2L\mu\eta}{L+\mu})^{T-j-1}(\mathbb{E}_A[\|v_j - \nabla g_S(x_j)\|^2])^{1/2} \\ &+ \frac{(L+\mu)L_g L_f}{L\mu m} + \frac{(L+\mu)L_g L_f}{L\mu n} + L_g \sigma_f \sqrt{\frac{L+\mu}{L\mu}}\sqrt{\eta}\Big).\end{aligned} \tag{43}$$

This completes the proof. □

Corollary 4 (Two-level Optimization)**.** *Consider Algorithm 2 with $\eta_t = \eta \leq 1/(4L+4\mu)$, and $\beta_t = \beta < \min\{1/8C_f^2, 1\}$ for any $t \in [0, T-1]$ and the output $A(S) = x_T$. Then, we have the following results*

$$\epsilon_\nu + \epsilon_\omega \leq O((T\beta)^{-\frac{c}{2}} + \beta^{\frac{1}{2}} + \eta^{\frac{1}{2}} + \eta\beta^{-\frac{1}{2}} + n^{-1} + m^{-1}).$$

proof of Corollary 4. Next, we move on to the Corollary 4. Combining (42) and (43) we have

$$\epsilon_\nu + \epsilon_\omega \leq O(n^{-1} + m^{-1} + \beta^{1/2} + \eta^{1/2} + \eta\beta^{-1/2} + T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}).$$

The proof is completed. □

Before giving the detailed proof, we first give a useful lemma.

Lemma 26. *Let Assumption 1(ii), 2 (ii) and 3 (ii) hold for the empirical risk F_S, and F_S is μ-strongly convex, for Algorithm 2, we have*

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq \mathbb{E}_A[F_S(x_t)|\mathcal{F}_t] - \frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2 + L_f^2\eta_t\mathbb{E}_A[\|v_t - \nabla g_S(x_t)\|^2|\mathcal{F}_t] \\ + L_g^2 C_f^2 \eta_t \mathbb{E}_A[\|u_t - g_S(x_t)\|^2|\mathcal{F}_t] + \frac{L L_g^2 L_f^2 \eta_t^2}{2}.\end{aligned}$$

where $\mathbb{E}_A$ denotes the expectation taken with respect to the randomness of the algorithm, and $\mathcal{F}_t$ is the σ-field generated by $\{\omega_{j_0}, \ldots, \omega_{j_{t-1}}, \nu_{i_0}, \ldots, \nu_{i_{t-1}}\}$.

Proof. According to the smoothness of F_S, we have

$$\begin{aligned}F_S(x_{t+1}) &\leq F_S(x_t) + \langle \nabla F_S(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &\leq F_S(x_t) - \eta_t\langle \nabla F_S(x_t), \nabla g_S(x_t)\nabla f_S(g_S(x_t))\rangle + \frac{L\eta_t^2}{2}\|v_t \nabla f_{\nu_{j_t}}(u_t)\|^2 + \theta_t,\end{aligned}$$

where $\theta_t = \eta_t\langle \nabla F_S(x_t), g_S(x_t)\nabla f_S(g_S(x_t)) - v_t\nabla f_{\nu_{j_t}}(u_t)\rangle$. As for the term θ_t, we have

$$\begin{aligned}\mathbb{E}_A[\theta_t|\mathcal{F}_t] &= \eta_t\mathbb{E}_A[\langle \nabla F_S(x_t), \nabla g_S(x_t)\nabla f_S(g_S(x_t)) - v_t\nabla f_{\nu_{j_t}}(u_t)\rangle|\mathcal{F}_t] \\ &= \eta_t\langle \nabla F_S(x_t), \nabla g_S(x_t)\nabla f_S(g_S(x_t)) - v_t\nabla f_S(g_S(x_t))\rangle|\mathcal{F}_t] \\ &\quad + \eta_t\mathbb{E}_A[\langle \nabla F_S(x_t), v_t\nabla f_S(g_S(x_t)) - v_t\nabla f_S(u_t)\rangle|\mathcal{F}_t] + \eta_t\mathbb{E}_A[\langle \nabla F_S(x_t), v_t\nabla f_S(u_t) - v_t\nabla f_{\nu_{j_t}}(u_t)\rangle|\mathcal{F}_t] \\ &\leq \eta_t\|\nabla F_S(x_t)\| \cdot \|\nabla f_S(g_S(x_t))\| \cdot \|v_t - \nabla g_S(x_t)\| + \eta_t\|\nabla F_S(x_t)\| \cdot \|v_t\| \cdot \|\nabla f_S(g_S(x_t)) - \nabla f_S(u_t)\| \\ &\leq \eta_t L_f\|\nabla F_S(x_t)\| \cdot \|v_t - \nabla g_S(x_t)\| + \eta_t L_g C_f\|\nabla F_S(x_t)\| \cdot \|u_t - g_S(x_t)\| \\ &\leq \frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2 + L_f^2\eta_t\|v_t - \nabla g_S(x_t)\|^2 + L_g^2 C_f^2\eta_t\|u_t - g_S(x_t)\|^2,\end{aligned}$$

where the last inequality holds by Cauchy-Schwarz inequality. Combining above two inequalities, let $\mathcal{F}_t$ be the σ-field generated by $\{\omega_{j_0}, \ldots, \omega_{j_{t-1}}, \nu_{i_0}, \ldots, \nu_{i_{t-1}}\}$, we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq \mathbb{E}_A[F_S(x_t)|\mathcal{F}_t] - \frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2 + L_f^2\eta_t\mathbb{E}_A[\|v_t - \nabla g_S(x_t)\|^2|\mathcal{F}_t] \\ + L_g^2C_f^2\eta_t\mathbb{E}_A[\|u_t - g_S(x_t)\|^2|\mathcal{F}_t] + \frac{LL_g^2L_f^2\eta_t^2}{2}.\end{aligned}$$

Then we complete the proof. □

proof of Theorem 10. We begin to give the detailed proof of Theorem 10. Note that strong convexity implies the Polyak-Łojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla F_S(x)\|^2 \geq \mu(F_S(x) - F_S(x_*^S)), \quad \forall x.$$

Then according to Lemma 26 and PL condition, subtracting both sides with $F_S(x_*^S)$ we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1}) - F_S(x_*^S)] \leq (1-\mu\eta_t)\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + L_f^2\eta_t\mathbb{E}_A[\|v_t - \nabla g_S(x_t)\|^2|\mathcal{F}_t] \\ + L_g^2C_f^2\eta_t\mathbb{E}_A[\|u_t - g_S(x_t)\|^2|\mathcal{F}_t] + \frac{LL_g^2L_f^2\eta_t^2}{2}.\end{aligned}$$

Setting $\eta_t = \eta$ and $\beta_t = \beta$, using Lemma 23 and 24 , we have

$$\begin{aligned}&\mathbb{E}_A[F_S(x_{t+1}) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \frac{LL_g^2L_f^2\eta^2}{2} + L_g^2C_f^2\eta((\frac{c}{e})^cU(t\beta)^{-c} + 2\sigma_g^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}) \\ &\quad + L_f^2\eta((\frac{c}{e})^cV(t\beta)^{-c} + 2\sigma_{g'}^2\beta + \frac{2L_g^4L_f^2\eta^2}{\beta}).\end{aligned}$$

Telescoping the above inequality from 1 to $T-1$ we have

$$\begin{aligned}&\mathbb{E}[F_S(x_T) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)^{T-1}\mathbb{E}[F_S(x_1) - F_S(x_*^S)] + \frac{LL_g^2L_f^2\eta^2}{2}\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1} \\ &\quad + (\frac{c}{e})^c\beta^{-c}(L_g^2C_f^2\eta U + L_f^2\eta V)\sum_{t=1}^{T-1}t^{-c}(1-\mu\eta)^{T-t-1} \\ &\quad + (2L_g^2C_f^2\sigma_g^2\beta\eta + 2L_f^2\sigma_{g'}^2\beta\eta + \frac{2L_g^6C_f^2L_f^2\eta^3}{\beta} + \frac{2L_g^4L_f^4\eta^3}{\beta})\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1}.\end{aligned}$$

For $t = 0$, we have

$$\mathbb{E}_A[F_S(x_1) - F_S(x_*^S)] \leq (1-\mu\eta)\mathbb{E}_A[F_S(x_0) - F_S(x_*^S)] + L_g^2C_f^2\eta U + L_f^2\eta V + \frac{LL_g^2L_f^2\eta^2}{2}.$$

Then combining the above two inequality we have

$$\begin{aligned}&\mathbb{E}[F_S(x_T) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)^T\mathbb{E}_A[F_S(x_0) - F_S(x_*^S)] + \frac{LL_g^2L_f^2\eta^2}{2}\sum_{t=1}^{T}(1-\mu\eta)^{T-t} + (L_g^2C_f^2\eta U + L_f^2\eta V)(1-\mu\eta)^{T-1} \\ &\quad + (\frac{c}{e})^c\beta^{-c}(L_g^2C_f^2\eta U + L_f^2\eta V)\sum_{t=1}^{T-1}t^{-c}(1-\mu\eta)^{T-t-1} \\ &\quad + (2L_g^2C_f^2\sigma_g^2\beta\eta + 2L_f^2\sigma_{g'}^2\beta\eta + \frac{2L_g^6C_f^2L_f^2\eta^3}{\beta} + \frac{2L_g^4L_f^4\eta^3}{\beta})\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1}.\end{aligned}$$

According to the fact that $\sum_{t=1}^{T}(1-\mu\eta)^{T-t} \leq \frac{1}{\mu\eta}$, using Lemma 12, we have

$$\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1}t^{-c} \leq \frac{\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1}}{T-1}\sum_{t=1}^{T-1}t^{-c} \leq \frac{1}{T\mu\eta}\sum_{t=1}^{T-1}t^{-c}.$$

Then we can get

$$\begin{aligned}\mathbb{E}[F_S(x_T) - F_S(x_*^S)] \leq & (\frac{c}{e\mu})^c(\eta T)^{-c}D_x + \frac{LL_g^2L_f^2\eta}{2\mu} + (L_g^2C_f^2\eta U + L_f^2\eta V)(\frac{c}{e\mu})^c(\eta T)^{-c} \\ & + (\frac{c}{e})^c\beta^{-c}(L_g^2C_f^2\eta U + L_f^2\eta V)T^{-1}\mu^{-1}\sum_{t=1}^{T-1}t^{-c} \\ & + \frac{2L_g^2C_f^2\sigma_g^2\beta}{\mu} + \frac{2L_f^2\sigma_{g'}^2\beta}{\mu} + \frac{2L_g^6C_f^2L_f^2\eta^2}{\beta\mu} + \frac{2L_g^4L_f^4\eta^2}{\beta\mu}.\end{aligned} \tag{44}$$

According to $\sum_{t=1}^{T}t^{-z} = O(T^{1-z})$ for $z \in (-1,0)\cup(-\infty,-1)$ and $\sum_{t=1}^{T}t^{-1} = O(\log T)$, as long as $c \neq 1$ we have

$$\begin{aligned}\mathbb{E}[F_S(x_T) - F_S(x_*^S)] \leq O\Big(& D_x(\eta T)^{-c} + LL_g^2L_f^2\eta + (L_g^2C_f^2U + L_f^2V)(\eta T)^{-c}\eta \\ & + (L_g^2C_f^2\eta U + L_f^2\eta V)(\beta T)^{-c} + (L_g^2C_f^2\sigma_g^2 + L_f^2\sigma_{g'}^2)\beta + (L_g^6C_f^2L_f^2 + L_f^4L_g^4)\eta^2\beta^{-1}\Big).\end{aligned}$$

The proof is completed. □

proof of Theorem 11. Putting (42) into (41), we have

$$\begin{aligned}& \mathbb{E}_A[\|x_T - x_T^{k,\nu}\|] + 4\mathbb{E}_A[\|x_T - x_T^{l,\omega}\|] \\ \leq & 50L_gC_f\eta\sup_S(\frac{(L+\mu)\sigma_g\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{U}(L+\mu)}{L\mu\eta}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}) \\ & + 50L_f\eta\sup_S(\frac{(L+\mu)\sigma_{g'}\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{V}(L+\mu)}{L\mu\eta}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}) \\ & + 4\sqrt{\frac{2L_f^2L_g^2\eta(L+\mu)}{L\mu m}} + \frac{8(L+\mu)L_gL_f}{L\mu m} + \sqrt{\frac{2L_f^2L_g^2\eta(L+\mu)}{L\mu n}} + \frac{2(L+\mu)L_gL_f}{L\mu n}.\end{aligned}$$

From Theorem 20 we have

$$\begin{aligned}& \mathbb{E}[F(x_T) - F_S(x_T)] \\ \leq & 50C_fL_fL_g^2\eta\sup_S(\frac{(L+\mu)\sigma_g\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{U_u}(L+\mu)}{L\mu\eta}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}) \\ & + 50L_f^2L_g\eta\sup_S(\frac{(L+\mu)\sigma_{g'}\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{U_{\dot{v}}}(L+\mu)}{L\mu\eta}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}) \\ & + 4L_f^2L_g^2\sqrt{\frac{2\eta(L+\mu)}{L\mu m}} + \frac{8(L+\mu)L_g^2L_f^2}{L\mu m} + L_f^2L_g^2\sqrt{\frac{2\eta(L+\mu)}{L\mu n}} + \frac{2(L+\mu)L_g^2L_f^2}{L\mu n} \\ + & L_f\sqrt{m^{-1}\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(A(S)))]}.\end{aligned}$$

Combining (44) and above inequality, using $F_S(x_*^S) \leq F_S(x_*)$ we have

$$\begin{aligned}
&\mathbb{E}_{S,A}\left[F(A(S)) - F\left(x_*\right)\right] \\
&\leq 50C_f L_f L_g^2 \eta \sup_S\Big(\frac{(L+\mu)\sigma_g\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2 L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{U_u}(L+\mu)}{L\mu\eta}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}\Big) \\
&\quad + 50L_f^2 L_g \eta \sup_S\Big(\frac{(L+\mu)\sigma_{g'}\sqrt{2\beta}}{2L\mu\eta} + \frac{(L+\mu)\sqrt{2}L_g^2 L_f}{2L\mu\sqrt{\beta}} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{U_{\dot{v}}}(L+\mu)}{L\mu\eta}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}\Big) \\
&\quad + 4L_f^2 L_g^2\sqrt{\frac{2\eta(L+\mu)}{L\mu m}} + \frac{8(L+\mu)L_g^2 L_f^2}{L\mu m} + L_f^2 L_g^2\sqrt{\frac{2\eta(L+\mu)}{L\mu n}} + \frac{2(L+\mu)L_g^2 L_f^2}{L\mu n} \\
&\quad + L_f\sqrt{m^{-1}\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(A(S)))]} + (\frac{c}{e\mu})^c(\eta T)^{-c}D_x + \frac{LL_g^2 L_f^2\eta}{2\mu} + (L_g^2 C_f^2\eta U + L_f^2\eta V)(\frac{c}{e\mu})^c(\eta T)^{-c} \\
&\quad + (\frac{c}{e})^c\beta^{-c}(L_g^2 C_f^2\eta U + L_f^2\eta V)T^{-1}\mu^{-1}\sum_{t=1}^{T-1}t^{-c} + \frac{2L_g^2 C_f^2\sigma_g^2\beta}{\mu} + \frac{2L_f^2\sigma_{g'}^2\beta}{\mu} + \frac{2L_g^6 C_f^2 L_f^2\eta^2}{\beta\mu} + \frac{2L_g^4 L_f^4\eta^2}{\beta\mu}.
\end{aligned}$$

Setting $\eta = T^{-a}, \beta = T^{-b}$, since often we have $\eta \leq \min(\frac{1}{n}, \frac{1}{m})$, then we have

$$\begin{aligned}
\mathbb{E}_{S,A}\left[F(A(S)) - F\left(x_*\right)\right] \leq O(&T^{-\frac{b}{2}} + T^{\frac{b}{2}-a} + T^{\frac{c}{2}(b-1)} + m^{-\frac{1}{2}} + m^{-1} + T^{-c(1-a)} + T^{-c(1-b)} \\
&+ T^{-a} + T^{-b} + T^{b-a} + T^{b-2a} + m^{-\frac{1}{2}}T^{-\frac{a}{2}} + n^{-\frac{1}{2}}T^{-\frac{a}{2}}).
\end{aligned}$$

Setting $c = 3$, the dominating terms are $\mathcal{O}(T^{\frac{b}{2}-a}), \mathcal{O}(T^{-\frac{b}{2}}), \quad \mathcal{O}(T^{3(b-1)}), \mathcal{O}(T^{-\frac{a}{2}})$, and$\mathcal{O}(T^{6(a-1)})$.

Setting $a = b = \frac{6}{7}$, we have

$$\mathbb{E}_{S,A}\left[F(A(S)) - F\left(x_*\right)\right] = O\left(T^{-\frac{3}{7}}\right).$$

Setting $T = O(\max\{n^{7/6}, m^{7,6}\})$, we have the following

$$\mathbb{E}_{S,A}\left[F(A(S)) - F\left(x_*\right)\right] = O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right).$$

The proof is completed. □

E. K-level Stochastic Optimizations

Lemma 27 (lemma 6 in (Jiang et al., 2022))**.** *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold for the empirical risk. $x_t, u_t^{(i)}$ and $v_t^{(i)}$ are generated by Algorithm 3 for any $i \in [1, K]$, then we have*

$$\mathbb{E}[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2] \leq (1-\beta_t)\mathbb{E}[\|v_{t-1}^{(i)} - \nabla f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_J^2 + 2L_f^2\mathbb{E}[\|u_t^{(i-1)} - u_{t-1}^{(i-1)}\|^2] \tag{45}$$

Lemma 28 (lemma 6 in (Jiang et al., 2022))**.** *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold for the empirical risk. $x_t, u_t^{(i)}$ and $v_t^{(i)}$ are generated by Algorithm 3 for any $i \in [1, K]$, then we have*

$$\mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2] \leq (1-\beta_t)\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_f^2 + 2L_f^2\mathbb{E}[\|u_t^{(i-1)} - u_{t-1}^{(i-1)}\|^2] \tag{46}$$

Lemma 29 (lemma 7 in (Jiang et al., 2022))**.** *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold for the empirical risk. $x_t, u_t^{(i)}$ and $v_t^{(i)}$ are generated by Algorithm 3 for any $i \in [1, K]$. Then for any $P \in [1, K]$, we have*

$$\sum_{i=1}^P\mathbb{E}[\|u_{t+1}^{(i-1)} - u_t^{(i-1)}\|^2] \leq (\sum_{i=1}^P(2L_f^2)^{i-1})(\mathbb{E}[\|x_{t+1} - x_t\|^2] + 2\beta_{t+1}^2\sigma_f^2 P + 2\beta_{t+1}^2 P\sum_{i=1}^P\mathbb{E}[\|u_t^{(i)} - f_i(u_t^{(i-1)})\|^2]). \tag{47}$$

Lemma 30. *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold for the empirical risk.* $x_t, u_t^{(i)}$ *and* $v_t^{(i)}$ *are generated by Algorithm 3 for any* $i \in [1, K]$*, let* $0 < \eta_t = \eta < 1$ *and let* $0 < \beta_t = \beta < \max\{1, 1/(4K\sum_{i=1}^{K}(2L_f^2)^i\}$ *we have*

$$\begin{aligned}
&\sum_{i=1}^{K}\mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2] \\
&\leq \sum_{i=1}^{K}(\frac{c}{e})^c(\frac{t\beta}{2})^{-c}\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + 4\beta\sigma_f^2K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2L_f^K}{\beta}.
\end{aligned}$$

proof of Lemma 30. Now we give the detailed proof of Lemma 30. According to Lemma 28 and 29, we have

$$\begin{aligned}
&\sum_{i=1}^{K}\mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2] \\
&\leq \sum_{i=1}^{K}(1-\beta_t)\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_f^2K \\
&+ (\sum_{i=1}^{K}(2L_f^2)^i)(\mathbb{E}[\|x_t - x_{t-1}\|^2] + 2\beta_t^2\sigma_f^2K + 2\beta_t^2K\sum_{i=1}^{K}\mathbb{E}[\|u_{t-1}^{(i)} - f_i(u_{t-1}^{(i-1)})\|^2]).
\end{aligned} \tag{48}$$

According to the setting that $\beta_t \leq \max\{1, 1/(4K\sum_{i=1}^{K}(2L_f^2)^i\}$, we have

$$\begin{aligned}
&\sum_{i=1}^{K}\mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2] \\
&\leq \sum_{i=1}^{K}(1-\frac{\beta_t}{2})\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_f^2K + (\sum_{i=1}^{K}(2L_f^2)^i)(\mathbb{E}[\|x_t - x_{t-1}\|^2] + 2\beta_t^2\sigma_f^2K) \\
&\leq \sum_{i=1}^{K}(1-\frac{\beta_t}{2})\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_f^2K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \sum_{i=1}^{K}(2L_f^2)^i\eta_t^2L_f^K.
\end{aligned} \tag{49}$$

Then using Lemma 14, setting $\eta_t = \eta$ and $\beta_t = \beta$, similar to the proof of Lemma 23, we have

$$\begin{aligned}
&\sum_{i=1}^{K}\mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2] \\
&\leq \sum_{i=1}^{K}\prod_{j=1}^{t}(1-\frac{\beta_j}{2})\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + 4\beta\sigma_f^2K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2L_f^K}{\beta}.
\end{aligned} \tag{50}$$

Note that $\prod_{i=K}^{N} \leq \exp(-\sum_{i=K}^{N}\beta_i)$ for all $K \leq N$ and $\beta_i \geq 0$, then we have

$$\begin{aligned}
&\sum_{i=1}^{K}\mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2] \\
&\leq \sum_{i=1}^{K}\exp(-\frac{\beta i}{2})\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + 4\beta\sigma_f^2K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2L_f^K}{\beta} \\
&\leq \sum_{i=1}^{K}(\frac{c}{e})^c(\frac{t\beta}{2})^{-c}\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + 4\beta\sigma_f^2K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2L_f^K}{\beta}.
\end{aligned} \tag{51}$$

Then we finish the proof.

Lemma 31. *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold for the empirical risk. $x_t, u_t^{(i)}$ and $v_t^{(i)}$ are generated by Algorithm 3 for any $i \in [1, K]$, let $0 < \eta_t = \eta < 1$ and let $0 < \beta_t = \beta < \max\{1, 1/(8K\sum_{i=1}^K (2L_f^2)^i\}$ we have*

$$\begin{aligned}
&\sum_{i=1}^K \mathbb{E}[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2] \\
&\leq \sum_{i=1}^K (\frac{c}{e})^c (\frac{t\beta}{2})^{-c} (\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) + \frac{4(\sum_{i=1}^K (2L_f^2)^i)\eta^2 L_f^K}{\beta} \\
&+ 4\beta K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^K (2L_f^2)^i)).
\end{aligned}$$

proof of Lemma 31. Now we give the detailed proof of Lemma 31. According to Lemma 28, 29 and 27, we have

$$\begin{aligned}
&\sum_{i=1}^K (\mathbb{E}[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2] + \mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2]) \\
&\leq \sum_{i=1}^K (1-\beta_t)\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_f^2 K + \sum_{i=1}^K (1-\beta_t)\mathbb{E}[\|v_{t-1}^{(i)} - \nabla f_{i,S}(u_{t-1}^{(i-1)})\|^2] + 2\beta_t^2\sigma_J^2 K \\
&+ 2(\sum_{i=1}^K (2L_f^2)^i)(\mathbb{E}[\|x_t - x_{t-1}\|^2] + 2\beta_t^2\sigma_f^2 K + 2\beta_t^2 K \sum_{i=1}^K \mathbb{E}[\|u_{t-1}^{(i)} - f_i(u_{t-1}^{(i-1)})\|^2]).
\end{aligned}$$

According to the setting that $0 < \eta_t = \eta < 1$ and let $0 < \beta_t = \beta < \max\{1, 1/(8K\sum_{i=1}^K (2L_f^2)^i\}$ we have

$$\begin{aligned}
&\sum_{i=1}^K (\mathbb{E}[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2] + \mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2]) \\
&\leq \sum_{i=1}^K (1 - \frac{\beta_t}{2})\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] + \sum_{i=1}^K (1-\beta_t)\mathbb{E}[\|u_{t-1}^{(i)} - f_{i,S}(u_{t-1}^{(i-1)})\|^2] \\
&+ 2(\sum_{i=1}^K (2L_f^2)^i)\eta^2 L_f^K + 2\beta_t^2 K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^K (2L_f^2)^i)).
\end{aligned}$$

□

Using Lemma 14, we have

$$\begin{aligned}
&\sum_{i=1}^K (\mathbb{E}[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2] + \mathbb{E}[\|u_t^{(i)} - f_{i,S}(u_t^{(i-1)})\|^2]) \\
&\leq \sum_{i=1}^K \prod_{j=1}^t (1 - \frac{\beta_j}{2})(\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) + \frac{4(\sum_{i=1}^K (2L_f^2)^i)\eta^2 L_f^K}{\beta} \\
&+ 4\beta K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^K (2L_f^2)^i)).
\end{aligned}$$

Then we have

$$
\begin{aligned}
&\sum_{i=1}^{K}\mathbb{E}[\|v_t^{(i)}-\nabla f_{i,S}(u_t^{(i-1)})\|^2]\\
&\le \sum_{i=1}^{K}(\frac{c}{e})^{c}(\frac{t\beta}{2})^{-c}(\mathbb{E}[\|u_1^{(i)}-f_{i,S}(x_0)\|^2]+\mathbb{E}[\|v_1^{(i)}-\nabla f_{i,S}(x_0)\|^2])+\frac{4(\sum_{i=1}^{K}(2L_f^2)^i)\eta^2L_f^K}{\beta}\\
&+4\beta K(\sigma_f^2+\sigma_J^2+2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i)).
\end{aligned}
$$

This complete the proof. □

proof of Theorem 1.

$$
\begin{aligned}
&\mathbb{E}_{S,A}[F(A(S))-F_S(A(S))]\\
&=\mathbb{E}_{S,A}[\mathbb{E}_{\nu^{(K)}}[f_K^{\nu^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))]])]\\
&\qquad\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}}f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}\cdots(\frac{1}{n_1}\sum_{i_1=1}^{n_1}f_1^{\nu_{i_1}^{(1)}}(A(S))))]\\
&=\mathbb{E}_{S,A}[\mathbb{E}_{\nu^{(K)}}[f_K^{\nu^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))]])]\\
&\qquad\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])\\
&+\mathbb{E}_{S,A}[\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])\\
&\qquad\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}}f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])]\\
&\ \vdots\\
&+\mathbb{E}_{S,A}[\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}}f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])\\
&\qquad\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}}f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}\cdots(\frac{1}{n_1}\sum_{i_1=1}^{n_1}f_1^{\nu_{i_1}^{(1)}}(A(S))))].
\end{aligned}
\tag{52}
$$

Now we estimate the terms of the RHS. Define $S^{(i)}=\{\nu_1^{(1)},\cdots,\nu_{n_1}^{(1)},\cdots,\nu_1^{(i)'},\cdots,\nu_{n_i}^{(i)'},\cdots,\nu_1^{(K)},\cdots,\nu_{n_1}^{(K)}\}$, where $i\in[1,K]$.

For the first term, we have

$$
\begin{aligned}
&\mathbb{E}_{S,A}[\mathbb{E}_{\nu^{(K)}}[f_K^{\nu^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))]])]\\
&\qquad\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])\\
&\le\mathbb{E}_{S,A,S^{(K)}}\Big[\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S^{i,K}))])\\
&\qquad\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])\Big]\\
&\le L_f^K\|A(S^{i,K})-A(S)\|\\
&\le L_f^K\epsilon_K
\end{aligned}
$$

For the second term, we have

$$\begin{aligned}
&\mathbb{E}_{S,A}[\frac{1}{n_K}\sum_{i_K=1}^{n_K} f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))]) \\
&\qquad\qquad - \frac{1}{n_K}\sum_{i_K=1}^{n_K} f_K^{\nu_{i_K}^{(K)}}(\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}} f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))])] \\
&\leq L_f\mathbb{E}_{S,A}[\|f_{K-1}(f_{K-2}\circ\cdots\circ f_1(A(S))) - \frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}} f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S)))\|].
\end{aligned}$$

Besides,

$$\begin{aligned}
&f_{K-1}(f_{K-2}\circ\cdots\circ f_1(A(S))) - \frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}} f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S))) \\
&= \frac{1}{n_{K-1}}\sum_{j=1}^{n_{K-1}} \mathbb{E}_{v^{(K-1)},v_j^{(K-1)\prime}}[f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S))) - f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))] \\
&+ \frac{1}{n_{K-1}}\sum_{j=1}^{n_{K-1}} \mathbb{E}_{v_j^{(K-1)\prime}}[\mathbb{E}_{v^{(K-1)}}[f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))] \\
&\qquad\qquad - f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))] \\
&+ \frac{1}{n_{K-1}}\sum_{j=1}^{n_{K-1}} \mathbb{E}_{v_j^{(K-1)\prime}}[f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)}))) - f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S)))].
\end{aligned}$$

Note that S and $S^{j,(K-1)}$ differ by a single example. By the assumption on stability and Definition 1, we have

$$\begin{aligned}
&\mathbb{E}_{S,A}[\|f_{K-1}(f_{K-2}\circ\cdots\circ f_1(A(S))) - \frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}} f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S)))\|] \\
&\leq 2L_f^{K-1}\epsilon_{K-1} + \mathbb{E}_{S,A}[\frac{1}{n_{K-1}}\|\sum_{j=1}^{n_{K-1}} \mathbb{E}_{v_j^{(K-1)\prime}}[\mathbb{E}_{v^{(K-1)}}[f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))] \\
&\qquad\qquad - f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))]\|].
\end{aligned} \tag{53}$$

Next step, we need to estimate the second term of above inequality. We denote

$$\xi_j(S) = \mathbb{E}_{v_j^{(K-1)\prime}}[\mathbb{E}_{v^{(K-1)}}[f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))] - f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))].$$

Notice that

$$\mathbb{E}_{S,A}[\|\sum_{j=1}^{n_{K-1}}\xi_j(S)\|^2] = \mathbb{E}_{S,A}[\sum_{j=1}^{n_{K-1}}\|\xi_j(S)\|^2] + \sum_{j,i\in[n_{K-1}]:j\neq i}\mathbb{E}_{S,A}[\langle\xi_j(S),\xi_i(S)\rangle].$$

Using Cauchy-Schwartz inequality, we have

$$
\begin{aligned}
&\mathbb{E}_{S,A}[\sum_{j=1}^{n_{K-1}} \|\xi_j(S)\|^2] \\
&= \mathbb{E}_{S,A}[\sum_{j=1}^{n_{K-1}} \|\mathbb{E}_{v_j^{(K-1)'}}[\mathbb{E}_{v^{(K-1)}}[f_{K-1}^{v^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S^{j,(K-1)})))] \\
&\qquad\qquad\qquad\qquad\qquad\qquad - f_{K-1}^{v_j^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S^{j,(K-1)})))]\|^2] \\
&\leq \mathbb{E}_{S,A}[\sum_{j=1}^{n_{K-1}} \|\mathbb{E}_{v^{(K-1)}}[f_{K-1}^{v^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S^{j,(K-1)})))] - f_{K-1}^{v_j^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S^{j,(K-1)})))\|^2] \\
&= \mathbb{E}_{S,A}[\sum_{j=1}^{n_{K-1}} \|\mathbb{E}_{v^{(K-1)}}[f_{K-1}^{v^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S))] - f_{K-1}^{v_j^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S))\|^2] \\
&= n_{K-1}\mathbb{E}_{S,A}[\mathrm{Var}_{K-1}(A(S)],
\end{aligned}
$$

where $\mathrm{Var}_{K-1}(A(S) = \mathbb{E}_{v^{(K-1)}}[\|f_{K-1}(f_{K-2} \circ \cdots \circ f_1(A(S)) - f_{K-1}^{v^{(K-1)}}(f_{K-2} \circ \cdots \circ f_1(A(S))\|^2]$.

Next, we will estimate the term $\sum_{j,i\in[n_{K-1}]:j\neq i} \mathbb{E}_{S,A}[\langle \xi_j(S), \xi_i(S)\rangle]$. We first define

$$
\begin{aligned}
S^{i,K-1} &= \{\nu_1^{(1)}, \cdots, \nu_{n_1}^{(1)}, \cdots, \nu_{i-1}^{(K-1)}, \nu_i^{(K-1)'}, \nu_{i+1}^{(K-1)}, \cdots, \nu_{n_{K-1}}^{(K-1)}, \nu_1^{(K)}, \cdots, \nu_{n_K}^{(K)}\} \\
S^{i,j,K-1} &= \{\nu_1^{(1)}, \cdots, \nu_{n_1}^{(1)}, \cdots, \nu_{i-1}^{(K-1)}, \nu_i^{(K-1)'}, \nu_{i+1}^{(K-1)}, \cdots, \nu_{j-1}^{(K-1)}, \nu_j^{(K-1)'}, \nu_{j+1}^{(K-1)}, \cdots, \nu_{n_K}^{(K)}\}.
\end{aligned}
$$

Due to the symmetry between $\nu^{(K-1)}$ and $\nu_j^{(K-1)}$, we have

$$
\mathbb{E}_{\nu_j^{(K-1)}}[\xi_j(S)] = 0, \forall j \in [1, n_{K-1}].
$$

If $j \neq i$, then we have

$$
\begin{aligned}
\mathbb{E}_{S,A}[\langle \xi_j(S^{i,K-1}), \xi_i(S)\rangle] &= \mathbb{E}_{S,A}\mathbb{E}_{\nu_i^{(K-1)}}[\langle \xi_j(S^{i,K-1}), \xi_i(S)\rangle] \\
&= \mathbb{E}_{S,A}[\langle \xi_j(S^{i,K-1}), \mathbb{E}_{\nu_i^{(K-1)}}[\xi_i(S)]\rangle] = 0,
\end{aligned}
$$

In a similar way, we can get for any $j \neq i$

$$
\begin{aligned}
\mathbb{E}_{S,A}[\langle \xi_j(S), \xi_i(S^{j,K-1})\rangle] &= \mathbb{E}_{S,A}\mathbb{E}_{\nu_j^{(K-1)}}[\langle \xi_j(S), \xi_i(S^{j,K-1})\rangle] \\
&= \mathbb{E}_{S,A}[\langle \mathbb{E}_{\nu_j^{(K-1)}}[\xi_j(S)], \xi_i(S^{j,K-1})\rangle] = 0,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{S,A}[\langle \xi_j(S^{i,K-1}), \xi_i(S^{j,K-1})\rangle] &= \mathbb{E}_{S,A}\mathbb{E}_{\nu_j^{(K-1)}}[\langle \xi_j(S^{i,K-1}), \xi_i(S^{j,K-1})\rangle] \\
&= \mathbb{E}_{S,A}[\langle \mathbb{E}_{\nu_j^{(K-1)}}[\xi_j(S^{i,K-1})], \xi_i(S^{j,K-1})\rangle] = 0,
\end{aligned}
$$

Combining the above identities, we have for any $i \neq j$

$$
\begin{aligned}
&\mathbb{E}_{S,A}[\langle \xi_j(S), \xi_i(S)\rangle] \\
&= \mathbb{E}_{S,A}[\langle \xi_j(S) - \xi_j(S^{i,K-1}), \xi_i(S) - \xi_i(S^{j,K-1})\rangle] \\
&\leq \mathbb{E}_{S,A}[\|\xi_j(S) - \xi_j(S^{i,K-1})\| \cdot \|\xi_i(S) - \xi_i(S^{j,K-1})\|] \\
&\leq \frac{1}{2}\mathbb{E}_{S,A}[\|\xi_j(S) - \xi_j(S^{i,K-1})\|^2] + \frac{1}{2}\mathbb{E}_{S,A}[\|\xi_i(S) - \xi_i(S^{j,K-1})\|^2].
\end{aligned}
$$

Then

$$\begin{aligned}
&\mathbb{E}_{S,A}[\|\xi_j(S)-\xi_j(S^{i,K-1})\|^2]\\
&=2\mathbb{E}_{S,A}[\|f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))-f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{i,j,(K-1)})))\|^2]\\
&\quad+2\mathbb{E}_{S,A}[\|f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{i,j,(K-1)})))-f_{K-1}^{v_j^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S^{j,(K-1)})))\|^2]\\
&\le 4L_f^{K-1}\epsilon_{K-1}^2.
\end{aligned}$$

In a similar way, we can have

$$\mathbb{E}_{S,A}[\|\xi_i(S)-\xi_i(S^{j,K-1})\|^2]\le 4L_f^{K-1}\epsilon_{K-1}^2.$$

According to the above inequalities, we have

$$\sum_{j,i\in[n_{K-1}]:j\neq i}\mathbb{E}_{S,A}[\langle\xi_j(S),\xi_i(S)\rangle]\le 4(n_{K-1}-1)n_{K-1}L_f^{K-1}\epsilon_{K-1}^2,\forall j\neq i.$$

Then we have

$$\mathbb{E}_{S,A}[\|\sum_{j=1}^{n_{K-1}}\xi_j(S)\|^2]\le 4(n_{K-1}-1)n_{K-1}L_f^{K-1}\epsilon_{K-1}^2+n_{K-1}\mathbb{E}_{S,A}[\mathrm{Var}_{K-1}(A(S)].$$

Therefore

$$\mathbb{E}_{S,A}[\|\sum_{j=1}^{n_{K-1}}\xi_j(S)\|]\le 2n_{K-1}L_f^{K-1}\epsilon_{K-1}+\sqrt{n_{K-1}\mathbb{E}_{S,A}[\mathrm{Var}_{K-1}(A(S)]}. \tag{54}$$

Combining (53) and (54) we have

$$\begin{aligned}
&\mathbb{E}_{S,A}[\|f_{K-1}(f_{K-2}\circ\cdots\circ f_1(A(S)))-\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}}f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S)))\|]\\
&\le 4L_f^{K-1}\epsilon_{K-1}+\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-1}(A(S)]}{n_{K-1}}}.
\end{aligned}$$

Then the second term

$$\begin{aligned}
&\mathbb{E}_{S,A}[\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\mathbb{E}_{\nu^{(K-1)}}[f_{K-1}^{\nu^{(K-1)}}]\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))]])\\
&\qquad\qquad-\frac{1}{n_K}\sum_{i_K=1}^{n_K}f_K^{\nu_{i_K}^{(K)}}(\frac{1}{n_{K-1}}\sum_{i_{K-1}=1}^{n_{K-1}}f_{K-1}^{\nu_{i_{K-1}}^{(K-1)}}\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(A(S))]])]\\
&\le 4L_f^K\epsilon_{K-1}+L_f\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-1}(A(S)]}{n_{K-1}}},
\end{aligned}$$

where $\mathrm{Var}_{K-1}(A(S)=\mathbb{E}_{v^{(K-1)}}[\|f_{K-1}(f_{K-2}\circ\cdots\circ f_1(A(S))-f_{K-1}^{v^{(K-1)}}(f_{K-2}\circ\cdots\circ f_1(A(S))\|^2]$.

Similarly, we can get, for any $t\in[2,K]$, the t-th term of (52) is bounded by

$$4L_f^K\epsilon_{K-t+1}+L_f\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-t+1}(A(S)]}{n_{K-t+1}}},$$

where $\mathrm{Var}_{K-t+1}(A(S)=\mathbb{E}_{v^{(K-t+1)}}[\|f_{K-t+1}(f_{K-t}\circ\cdots\circ f_1(A(S))-f_{K-t+1}^{v^{(K-t+1)}}(f_{K-t}\circ\cdots\circ f_1(A(S))\|^2]$.

Then we can conclude that

$$\mathbb{E}_{S,A}[F(A(S))-F_S(A(S))]\le L_f^K\epsilon_K+4L_f^K\sum_{t=2}^{K}\epsilon_{K-t+1}+L_f\sum_{t=2}^{K}\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-t+1}(A(S)]}{n_{K-t+1}}},$$

where $\mathrm{Var}_{K-t+1}(A(S) = \mathbb{E}_{v^{(K-t+1)}}[\|f_{K-t+1}(f_{K-t} \circ \cdots \circ f_1(A(S)) - f_{K-t+1}^{v^{(K-t+1)}}(f_{K-t} \circ \cdots \circ f_1(A(S))\|^2]$.

This completes the proof.

□

E.1. Convex setting

proof of Theorem 4. Since changing one sample data can happen in any layer of the function, we define

$$S^{l,k} = \{\nu_1^{(1)}, \cdots, \nu_{n_1}^{(1)}, \cdots, \nu_1^{(k)}, \cdots, \nu_{l-1}^{(k)}, \nu_l^{(k)'}, \nu_{l+1}^{(k)}, \cdots, \nu_{n_k}^{(k)}, \cdots, \nu_1^{(K)}, \cdots, \nu_{n_K}^{(K)}\}.$$

Let $\{x_{t+1}\}$, $\{u_{t+1}^{(i)}\}$ and $\{v_{t+1}^{(i)}\}$ be produced by SVMR based on S, where $i \in [1, K]$ and represents an estimator of the function of layer i. $\{x_{t+1}^{l,k}\}$, $\{u_{t+1}^{(i),l,k}\}$ and$\{v_{t+1}^{(i),l,k}\}$ be produced by SVMR based on $S^{l,k}$. For any $l \in [1, n_k],\ k \in [1, K]$, let $x_0 = x_0^{l,k}$ be starting points in $\mathcal{X}$.

We begin with the estimation of the term $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,k}\|]$. For this purpose, we will consider two cases, $i_t \neq l$ and $i_t = l$.

Case 1($i_t \neq l$). We have

$$\begin{aligned} \|x_{t+1} - x_{t+1}^{l,k}\|^2 &\leq \|x_t - \eta_t \prod_{i=1}^{K} v_t^{(i)} - x_t^{l,k} + \eta_t \prod_{i=1}^{K} v_t^{(i),l,k}\|^2 \\ &\leq \|x_t - x_t^{l,k}\|^2 - 2\eta_t\langle \prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}, x_t - x_{t+1}^{l,k}\rangle + \eta_t^2 \| \prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}\|^2. \end{aligned} \tag{55}$$

Now we estimate the second term of above inequality.

$$\begin{aligned} &- 2\eta_t\langle \prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}, x_t - x_t^{l,k}\rangle \\ &= -2\eta_t\langle \prod_{i=1}^{K} v_t^{(i)} - \nabla f_{1,S}(x_t) \cdot \prod_{i=2}^{K} v_t^{(i)}, x_t - x_t^{l,k}\rangle \\ &- 2\eta_t\langle \nabla f_{1,S}(x_t) \cdot \prod_{i=2}^{K} v_t^{(i)} - \nabla f_{1,S}(x_t) \cdot \prod_{i=3}^{K} v_t^{(i)} \cdot \nabla f_{2,S}(u_t^{(1)}), x_t - x_t^{l,k}\rangle \\ &- 2\eta_t\langle \nabla f_{1,S}(x_t) \cdot \prod_{i=3}^{K} v_t^{(i)} \cdot \nabla f_{2,S}(u_t^{(1)}) - \nabla f_{1,S}(x_t) \cdot \prod_{i=3}^{K} v_t^{(i)} \cdot \nabla f_{2,S}(f_{1,S}(x_t)), x_t - x_t^{l,k}\rangle \\ &\vdots \\ &- 2\eta_t\langle \prod_{i=1}^{K} \nabla F_{i,S}(x_t) - \prod_{i=1}^{K} \nabla F_{i,S}(x_t^{l,k}), x_t - x_t^{l,k}\rangle \\ &- 2\eta_t\langle \prod_{i=1}^{K} \nabla F_{i,S}(x_t^{l,k}) - \prod_{i=2}^{K} \nabla F_{i,S}(x_t^{l,k}) \cdot v_t^{(1),l,k}, x_t - x_t^{l,k}\rangle \\ &- 2\eta_t\langle \prod_{i=2}^{K} \nabla F_{i,S}(x_t^{l,k}) \cdot v_t^{(1),l,k} - \prod_{i=3}^{K} \nabla F_{i,S}(x_t^{l,k}) \cdot v_t^{(1),l,k} \cdot \nabla f_{2,S}(u_t^{(1),l,k}), x_t - x_t^{l,k}\rangle \\ &\vdots \\ &- 2\eta_t\langle \prod_{i=1}^{K-1} v_t^{(i),l,k} \cdot \nabla f_{K,S}(u_t^{(K-1),l,k}) - \prod_{i=1}^{K} v_t^{(i),l,k}, x_t - x_t^{l,k}\rangle. \end{aligned}$$

From the above inequality, we decompose it to $K(K+1)+1 \leq K(K+2)$ terms, where K is the number of layers of the function. Using Assumption 3 (iii) we can get

$$\begin{aligned}
&- 2\eta_t\langle\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}, x_t - x_t^{l,k}\rangle \\
&\leq 2\eta_t L_f^{K-1}\|v_t^{(1)} - \nabla f_{1,S}(x_t)\| \cdot \|x_t - x_t^{l,k}\| \\
&\quad + (2\eta_t L_f^{K-1}\|v_t^{(2)} - \nabla f_{2,S}(u_t^{(1)})\| + 2\eta_t L_f^{K}\|u_t^{(1)} - f_{1,S}(x_t)\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad \vdots \\
&\quad + (2\eta_t L_f^{m_2}\|v_t^{(K)} - \nabla f_{K,S}(u_t^{(K-1)})\| + \cdots + 2\eta_t L_f^{K-1+(K-1)K/2}\|u_t^{(1)} - f_{1,S}(x_t)\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad - \frac{2\eta_t}{L}\|\prod_{i=1}^{K}\nabla F_{i,S}(x_t) - \prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k})\| \\
&\quad + (2\eta_t L_f^{K-1+(K-1)K/2}\|u_t^{(1),l,k} x_t^{l,k}\| + \cdots 2\eta_t L_f^{(K-1)K/2}\|\nabla f_{K,S}(u_t^{(K-1),l,k}) - \nabla v_t^{(K),l,k}\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad \vdots \\
&\quad + (2\eta_t L_f^{K}\|u_t^{(1),l,k} - f_{1,S}(x_t^{l,k})\| + 2\eta_t L_f^{K-1}\|v_t^{(2),l,k} - \nabla f_{2,S}(u_t^{(1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t L_f^{K-1}\|v_t^{(1),l,k} - \nabla f_{1,S}(x_t^{l,k})\| \cdot \|x_t - x_t^{l,k}\|.
\end{aligned} \tag{56}$$

Conclude above inequality, we have

$$\begin{aligned}
&- 2\eta_t\langle\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}, x_t - x_t^{l,k}\rangle \\
&\leq 2\eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}(\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\| + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}(\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\| + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad - \frac{2\eta_t}{L}\|\prod_{i=1}^{K}\nabla F_{i,S}(x_t) - \prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k})\|.
\end{aligned}$$

Now we consider the third term of (55). Similar to the (56), we have

$$\begin{aligned}
&\|\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}\| \\
&\leq L_f^{K-1}\|v_t^{(1)} - \nabla f_{1,S}(x_t)\| \cdot \|x_t - x_t^{l,k}\| \\
&\quad + (L_f^{K-1}\|v_t^{(2)} - \nabla f_{2,S}(u_t^{(1)})\| + L_f^{K}\|u_t^{(1)} - f_{1,S}(x_t)\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad \vdots \\
&\quad + (L_f^{m_2}\|v_t^{(K)} - \nabla f_{K,S}(u_t^{(K-1)})\| + \cdots + L_f^{K-1+(K-1)K/2}\|u_t^{(1)} - f_{1,S}(x_t)\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + \|\prod_{i=1}^{K} \nabla F_{i,S}(x_t) - \prod_{i=1}^{K} \nabla F_{i,S}(x_t^{l,k})\| \\
&\quad + (L_f^{K-1+(K-1)K/2}\|u_t^{(1),l,k} x_t^{l,k}\| + \cdots L_f^{(K-1)K/2}\|\nabla f_{K,S}(u_t^{(K-1),l,k}) - \nabla v_t^{(K),l,k}\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad \vdots \\
&\quad + (L_f^{K}\|u_t^{(1),l,k} - f_{1,S}(x_t^{l,k})\| + L_f^{K-1}\|v_t^{(2),l,k} - \nabla f_{2,S}(u_t^{(1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + L_f^{K-1}\|v_t^{(1),l,k} - \nabla f_{1,S}(x_t^{l,k})\| \cdot \|x_t - x_t^{l,k}\|.
\end{aligned}$$

Taking square on both sides of the above inequality, we have that

$$\begin{aligned}
&\eta_t^2\|\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}\|^2 \\
&\leq L_f^{2K-2}K(K+2)\eta_t^2\|v_t^{(1)} - \nabla f_{1,S}(x_t)\|^2 \\
&\quad + L_f^{2K-2}K(K+2)\eta_t^2\|v_t^{(2)} - \nabla f_{2,S}(u_t^{(1)})\|^2 + L_f^{2K}K(K+2)\eta_t^2\|u_t^{(1)} - f_{1,S}(x_t)\|^2 \\
&\quad \vdots \\
&\quad + L_f^{m}K(K+2)\eta_t^2\|v_t^{(K)} - \nabla f_{K,S}(u_t^{(K-1)})\|^2 + \cdots + L_f^{K-1+(K-1)K}K(K+2)\eta_t^2\|u_t^{(1)} - f_{1,S}(x_t)\|^2 \\
&\quad + K(K+2)\eta_t^2\|\prod_{i=1}^{K} \nabla F_{i,S}(x_t) - \prod_{i=1}^{K} \nabla F_{i,S}(x_t^{l,k})\|^2 \\
&\quad + L_f^{2K-2+(K-1)K}K(K+2)\eta_t^2\|u_t^{(1),l,k} x_t^{l,k}\|^2 + \cdots L_f^{(K-1)K}K(K+2)\eta_t^2\|\nabla f_{K,S}(u_t^{(K-1),l,k}) - \nabla v_t^{(K),l,k}\|^2\| \\
&\quad \vdots \\
&\quad + L_f^{2K}K(K+2)\eta_t^2\|u_t^{(1),l,k} - f_{1,S}(x_t^{l,k})\|^2 + L_f^{2K-2}K(K+2)\eta_t^2\|v_t^{(2),l,k} - \nabla f_{2,S}(u_t^{(1),l,k})\|^2 \\
&\quad + L_f^{2K-2}K(K+2)\eta_t^2\|v_t^{(1),l,k} - \nabla f_{1,S}(x_t^{l,k})\|^2.
\end{aligned}$$

where we have used the fact that $(\sum_{i=1}^{K} a_i)^2 \leq \sum_{i=1}^{K} K a_i^2$. Then we can conclude that

$$\begin{aligned}
&\eta_t^2\|\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}\|^2 \\
&\leq \eta_t^2 \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{2K-2j+(i-1)i}(\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2 + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2) \\
&\quad + \eta_t^2 \sum_{i=1}^{K} L_f^{2K-2i+(i-1)i}(\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2 + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2) \\
&\quad + (K+2)K\eta_t^2\|\prod_{i=1}^{K} \nabla F_{i,S}(x_t) - \prod_{i=1}^{K} \nabla F_{i,S}(x_t^{l,k})\|.
\end{aligned}$$

Putting above inequality into (55), according to $\eta_t \leq \frac{2}{LK(K+2)}$, we have

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,k}\|^2 \\
&\leq 2\eta_t \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i} (\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\| + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i} (\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\| + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + \eta_t^2 \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{2K-2j+(i-1)i} (\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2 + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2) \\
&\quad + \eta_t^2 \sum_{i=1}^{K} L_f^{2K-2i+(i-1)i} (\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2 + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2) + \|x_t - x_t^{l,k}\|^2.
\end{aligned}$$

Case 2 ($i_t = l$ **).** We have

$$\begin{aligned}
\|x_{t+1} - x_{t+1}^{l,k}\| &= \|x_t - \eta_t \prod_{i=1}^{K} v_t^{(i)} - x_t^{l,k} + \eta_t \prod_{i=1}^{K} v_t^{(i),l,k}\| \\
&\leq \|x_t - x_t^{l,k}\| + \eta_t \|\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}\| \leq \|x_t - x_t^{l,k}\| + 2\eta_t L_f^K.
\end{aligned} \tag{57}$$

Therefore, we have

$$\|x_{t+1} - x_{t+1}^{l,k}\|^2 \leq \|x_t - x_t^{l,k}\|^2 + 4\eta_t L_f^K \|x_t - x_t^{l,k}\| + 4\eta_t^2 L_f^{2K}.$$

Combining above two cases, we have

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,k}\|^2 \\
&\leq 2\eta_t \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i} (\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\| + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i} (\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\| + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + \eta_t^2 \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{2K-2j+(i-1)i} (\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2 + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2) \\
&\quad + \eta_t^2 \sum_{i=1}^{K} L_f^{2K-2i+(i-1)i} (\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2 + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2) + \|x_t - x_t^{l,k}\|^2 \\
&\quad + 4\eta_t L_f^K \|x_t - x_t^{l,k}\| \cdot \mathbf{1}_{[i_t = l]} + 4\eta_t^2 L_f^{2K} \cdot \mathbf{1}_{[i_t = l]}.
\end{aligned}$$

According to

$$\mathbb{E}_A[\|x_t - x_t^{l,k}\| \mathbf{1}_{[i_t = l]}] = \mathbb{E}_A[\|x_t - x_t^{l,k}\| \mathbb{E}_{i_t}[\mathbf{1}_{[i_t = l]}]] = \frac{1}{n_k} \mathbb{E}_A[\|x_t - x_t^{l,k}\|] \leq \frac{1}{n_k} (\mathbb{E}_A[\|x_t - x_t^{l,k}\|^2])^{1/2},$$

note that $\|x_0 - x_0^{l,k}\|^2 = 0$, we have

$$\begin{aligned}
&\mathbb{E}_A\|x_{t+1} - x_{t+1}^{l,k}\|^2 \\
&\leq 2\eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2} \\
&\quad + 2\eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2} \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2} \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2} \\
&\quad + \eta_t^2 \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{2K-2j+(i-1)i}(\mathbb{E}_A\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2 + \mathbb{E}_A\|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2) \\
&\quad + \eta_t^2 \sum_{i=1}^{K} L_f^{2K-2i+(i-1)i}(\mathbb{E}_A\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2 + \mathbb{E}_A\|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2) \\
&\quad + \mathbb{E}_A\|x_t - x_t^{l,k}\|^2 + \frac{4\eta_t L_f^K}{n_k}(\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2} + \frac{4\eta_t^2 L_f^{2K}}{n_k}.
\end{aligned}$$

Telescoping from 0 to $t-1$, according to $\|x_0 - x_0^{l,k}\|^2 = 0$, we have

$$\begin{aligned}
&\mathbb{E}_A\|x_t - x_t^{l,k}\|^2 \\
&\leq 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1} \eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1} \eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K} \eta_s L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K} \eta_s L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2} \cdot (\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + \sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1} \eta_s^2 L_f^{2K-2j+(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2 + \mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2) \\
&\quad + \sum_{s=1}^{t-1}\sum_{i=1}^{K} \eta_s^2 L_f^{2K-2i+(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2 + \mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2) \\
&\quad + \sum_{s=1}^{t-1} \frac{4\eta_s L_f^K}{n_k}(\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} + \sum_{s=1}^{t-1} \frac{4\eta_s^2 L_f^{2K}}{n_k}.
\end{aligned}$$

Similarly, for notational convenience, denote $u_t = (\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2}$, and letting

$$\begin{aligned}
S_t =& \sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s^2 L_f^{2K-2j+(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2 + \mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)\\
&+ \sum_{s=1}^{t-1}\sum_{i=1}^{K}\eta_s^2 L_f^{2K-2i+(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2 + \mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)\\
&+ \sum_{s=1}^{t-1}\frac{4\eta_s^2 L_f^{2K}}{n_k},\\
\alpha_s =& \frac{4\eta_s L_f^K}{n_k} + 2\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&+ 2\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2}\\
&+ 2\sum_{i=1}^{K}\eta_s L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}\\
&+ 2\sum_{i=1}^{K}\eta_s L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2}.
\end{aligned}$$

According to Lemma 14, we have

$$\begin{aligned}
u_t \le& \sqrt{S_t} + \sum_{s=1}^{t-1}\alpha_s\\
\le& \sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+(i-1)i/2}((\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} + (\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2})\\
&+ \sum_{s=1}^{t-1}\sum_{i=1}^{K}\eta_s L_f^{K-i+(i-1)i/2}((\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} + (\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2})\\
&+ (\sum_{s=1}^{t-1}\frac{4\eta_s^2 L_f^{2K}}{n_k})^{1/2} + \sum_{s=1}^{t-1}\frac{4\eta_s L_f^K}{n_k} + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&+ 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2}\\
&+ 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\eta_s L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}\\
&+ 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\eta_s L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2},
\end{aligned}$$

where the inequality holds by $(\sum_{i=1}^{K} a_i)^{1/2} \le \sum_{i=1}^{K}(a_i)^{1/2}$. Besides, if we let $\eta_t = \eta$, then it's easy to get

$$\begin{aligned}
&\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&\qquad\qquad \le \sup_S \eta\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}\eta_s L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k}-f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2}\\
&\qquad\qquad\le \sup_S \eta\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}.
\end{aligned}$$

This inequality is also true for $v_s^{(j)}$ and $v_s^{(j),l,k}$. Consequently, with T iterations, we obtain that

$$\begin{aligned}
u_T &\le 6\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&\quad+6\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}+(\sum_{s=1}^{T-1}\frac{4\eta_s^2L_f^{2K}}{n_k})^{1/2}+\frac{4\eta L_f^K T}{n_k}\\
&\le 6\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&\quad+6\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}+\frac{6\eta L_f^K T}{n_k},
\end{aligned}$$

where the last inequality holds by the fact that we often have $T\ge n_k$, for any $k\in[1,K]$. Besides

$$\mathbb{E}_A[\|x_T-X_T^{l,k}\|]\le u_T=(\mathbb{E}_A[\|x_T-X_T^{l,k}\|]^2)^{1/2}.$$

Then we can get the result for the k-th layer

$$\begin{aligned}
\mathbb{E}_A[\|x_T-X_T^{l,k}\|]=O\Big(&\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&+\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}+\frac{\eta L_f^K T}{n_k}\Big),
\end{aligned}$$

where $k\in[1,K]$. Then we have

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}_A[\|\epsilon_k\|]=O(&\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&+\sup_S \eta\sum_{s=1}^{T-1}\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}+\sum_{k=1}^{K}\frac{\eta L_f^K T}{n_k}).
\end{aligned} \tag{58}$$

This completes the proof. □

Corollary 5 (K-level Optimization)**.** *Consider SVMR in Algorithm 3 with* $0<\eta_t=\eta<2/LK(K+1)$ *and let* $0<\beta_t=\beta<\max\{1,\frac{1}{(4K\sum_{i=1}^{K}(2L_f^2)^i}\}$ *for any* $t\in[0,T-1]$*. With the output* $A(S)=x_T$*, then we have*

$$\sum_{k=1}^{K}\epsilon_k=O\Big(\eta T\big((\beta T)^{-\frac{c}{2}}+\beta^{1/2}+\eta\beta^{-1/2}\big)+\eta T\sum_{k=1}^{K}\frac{1}{n_k}\Big).$$

Now we give the proof of Corollary 5.

proof of Corollary 5. According to (58), we have

$$\begin{aligned}\sum_{k=1}^{K}\epsilon_k = O(\sup_S \eta \sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\ + \sup_S \eta \sum_{s=1}^{T-1}\sum_{i=1}^{K}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} + \sum_{k=1}^{K}\frac{6\eta L_f^K T}{n_k}).\end{aligned}$$

According to Lemma 30 and 31, we can get

$$\begin{aligned}\sum_{k=1}^{K}\epsilon_k &= O(\sum_{k=1}^{K}\frac{\eta T}{n_k} + \eta\sum_{s=1}^{T-1}((s\beta)^{-c/2} + \frac{\eta}{\sqrt{\beta}} + \sqrt{\beta})) \\ &= O(\sum_{k=1}^{K}\frac{\eta T}{n_k} + \eta T^{-c/2+1}\beta^{-c/2} + \eta^2\beta^{-\frac{1}{2}}T + \eta\beta^{1/2}T).\end{aligned}$$

This complete the proof. □

Before give the detailed proof of Theorem 5, we first give a useful lemma.

Lemma 32. *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold for the empirical risk F_S, for SVMR, we have for any $\gamma_t > 0$ and $\lambda_t > 0$ we have*

$$\begin{aligned}&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \\ &\le \|x_t - x_*^S\|^2 + L_f^K\eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \gamma_t\eta_t\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+\frac{1}{2}(i-1)i}\|x_t - x_*^S\|^2 \\ &\quad + \frac{\eta_t\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2|\mathcal{F}_t]}{\gamma_t} \\ &\quad + \frac{\eta_t\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2|\mathcal{F}_t]}{\lambda_t} + \lambda_t\eta_t\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}\|x_t - x_*^S\|^2.\end{aligned}$$

Proof. According to the update rule of SVMR, we have

$$\begin{aligned}\|x_{t+1} - x_*^S\|^2 &= \|x_t - \eta_t\prod_{i=1}^{K}v_t^{(i)} - x_*^S\|^2 \\ &\le \|x_t - x_*^S\|^2 + L_f^K\eta_t^2 - 2\eta_t\langle x_t - x_*^S, \prod_{i=1}^{K}\nabla F_{i,S}(x_t)\rangle + u_t,\end{aligned}$$

where $u_t = 2\eta_t\langle x_t - x_*^S, \prod_{i=1}^{K}\nabla F_{i,S}(x_t) - \prod_{i=1}^{K}v_t^{(i)}\rangle$. Let $\mathcal{F}_t$ be the σ field generated by S. Taking expectation with respect to the internal randomness of the algorithm and using Assumption 1 (iii), we have

$$\begin{aligned}&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2\|\mathcal{F}_t] \\ &\le \|x_t - x_*^S\|^2 + L_f^K\eta_t^2 - 2\eta_t\mathbb{E}_A[\langle x_t - x_*^S, \prod_{i=1}^{K}\nabla F_{i,S}(x_t)\rangle|\mathcal{F}_t] + \mathbb{E}_A[u_t|\mathcal{F}_t] \\ &= \|x_t - x_*^S\|^2 + L_f^K\eta_t^2 - 2\eta_t\langle x_t - x_*^S, \nabla F_S(x_t)\rangle + \mathbb{E}_A[u_t|\mathcal{F}_t] \\ &\le \|x_t - x_*^S\|^2 + L_f^K\eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \mathbb{E}_A[u_t|\mathcal{F}_t],\end{aligned}$$

where the last inequality comes from the convexity of F_S. Now we handle the term $\mathbb{E}_A[u_t|\mathcal{F}_t]$.

$$\begin{aligned}
u_t &= 2\eta_t\langle x_t - x_*^S, \prod_{i=1}^K \nabla F_{i,S}(x_t) - \prod_{i=1}^K v_t^{(i)}\rangle \\
&= 2\eta_t\langle \prod_{i=1}^K \nabla F_{i,S}(x_t) - \prod_{i=2}^K \nabla F_{i,S}(x_t)\cdot v_t^{(1)}, x_t - x_*^S\rangle \\
&+ 2\eta_t\langle \prod_{i=2}^K \nabla F_{i,S}(x_t)\cdot v_t^{(1)} - \prod_{i=3}^K \nabla F_{i,S}(x_t)\cdot v_t^{(1)}\cdot \nabla f_{2,S}(u_t^{(1)}), x_t - x_*^S\rangle \\
&+ 2\eta_t\langle \prod_{i=3}^K \nabla F_{i,S}(x_t)\cdot v_t^{(1)}\cdot \nabla f_{2,S}(u_t^{(1)}) - \prod_{i=3}^K \nabla F_{i,S}(x_t)\cdot v_t^{(1)}\cdot v_t^{(2)}, x_t - x_*^S\rangle \\
&\vdots \\
&+ 2\eta_t\langle \prod_{i=1}^{K-1} v_t^{(i)}\cdot \nabla f_{K,S}(u_t^{(K-1)}) - \prod_{i=1}^K v_t^{(i)}, x_t - x_*^S\rangle.
\end{aligned}$$

Conclude above inequality, we have

$$\begin{aligned}
&-2\eta_t\langle \prod_{i=1}^K v_t^{(i)} - \prod_{i=1}^K v_t^{(i),l,k}, x_t - x_t^{l,k}\rangle \\
&\leq 2\eta_t \sum_{i=1}^K \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i} \|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|\cdot\|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t \sum_{i=1}^K L_f^{K-i+\frac{1}{2}(i-1)i}\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|\cdot\|x_t - x_t^{l,k}\|.
\end{aligned}$$

Conclude above inequality, for any $\gamma_t > 0$ and $\lambda_t > 0$ we have

$$\begin{aligned}
&2\eta_t\langle x_t - x_*^S, \prod_{i=1}^K \nabla F_{i,S}(x_t) - \prod_{i=1}^K v_t^{(i)}\rangle \\
&\leq 2\eta_t \sum_{i=1}^K \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|\cdot\|x_t - x_*^S\| \\
&\quad + 2\eta_t \sum_{i=1}^K L_f^{K-i+\frac{1}{2}(i-1)i}\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|\cdot\|x_t - x_*^S\| \\
&\leq \frac{\eta_t \sum_{i=1}^K \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2}{\gamma_t} + \gamma_t\eta_t \sum_{i=1}^K \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\|x_t - x_*^S\|^2 \\
&\quad + \frac{\eta_t \sum_{i=1}^K L_f^{K-i+\frac{1}{2}(i-1)i}\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|}{\lambda_t} + \lambda_t\eta_t \sum_{i=1}^K L_f^{K-i+\frac{1}{2}(i-1)i}\|x_t - x_*^S\|^2.
\end{aligned}$$

Then we can get

$$
\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \\
&\leq \|x_t - x_*^S\|^2 + L_f^K \eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \gamma_t\eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\|x_t - x_*^S\|^2 \\
&\quad + \frac{\eta_t \sum_{i=1}^K \sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2|\mathcal{F}_t]}{\gamma_t} \\
&\quad + \frac{\eta_t \sum_{i=1}^K L_f^{K-i+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2|\mathcal{F}_t]}{\lambda_t} + \lambda_t\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}\|x_t - x_*^S\|^2.
\end{aligned}
$$

This completes the proof. □

Then we give the detailed proof of Theorem 5.

proof of Theorem 5. According to Lemma 32, setting $\eta_t = \eta$, $\beta_t = \beta$ and $\lambda_t = \gamma_t = \sqrt{\beta}$, we have

$$
\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2] \\
&\leq \mathbb{E}_A[\|x_t - x_*^S\|^2] + L_f^K\eta^2 - 2\eta\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \eta\sqrt{\beta}\sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|x_t - x_*^S\|^2] \\
&\quad + \frac{\eta\sum_{i=1}^K\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]}{\sqrt{\beta}} \\
&\quad + \frac{\eta\sum_{i=1}^K L_f^{K-i+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2]}{\sqrt{\beta}} + \sqrt{\beta}\eta\sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|x_t - x_*^S\|^2] \\
&\leq \mathbb{E}_A[\|x_t - x_*^S\|^2] + L_f^K\eta^2 - 2\eta\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\
&\quad + \frac{\eta K L_f^{m_1}\sum_{j=i}^{K-1}\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]}{\sqrt{\beta}} + \frac{\eta L_f^{m_2}\sum_{i=1}^K\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2]}{\sqrt{\beta}} \\
&\quad + \eta\sqrt{\beta}(K^2 L_f^{m_1} + K L_f^{m_2})\mathbb{E}_A[\|x_t - x_*^S\|^2],
\end{aligned}
$$

where $L_f^{m_1} = \max\{L_f^{K-j+\frac{1}{2}(i-1)i}\}$ for any $i, j \in [1, K]$ and $L_f^{m_2} = \max\{L_f^{K-i+\frac{1}{2}(i-1)i}\}$ for any $i \in [1, K]$. Using Lemma 30 and 31 we have

$$
\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2] \\
&\leq \mathbb{E}_A[\|x_t - x_*^S\|^2] + L_f^K\eta^2 - 2\eta\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\
&\quad + \eta K L_f^{m_1}\Big(\sum_{i=1}^{K}(\frac{c}{e})^c(\frac{t\beta}{2})^{-c}\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] \\
&\qquad\qquad + 4\beta\sigma_f^2 K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^K(2L_f^2)^i\eta^2 L_f^K}{\beta}\Big)/\sqrt{\beta} \\
&\quad + \eta L_f^{m_2}\Big(\sum_{i=1}^{K}(\frac{c}{e})^c(\frac{t\beta}{2})^{-c}(\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) \\
&\qquad\qquad + \frac{4(\sum_{i=1}^K(2L_f^2)^i)\eta^2 L_f^K}{\beta} + 4\beta K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i))\Big)/\sqrt{\beta} \\
&\quad + \eta\sqrt{\beta}(K^2 L_f^{m_1} + K L_f^{m_2})\mathbb{E}_A[\|x_t - x_*^S\|^2].
\end{aligned}
$$

Rearranging and telescoping the above inequality from 1 to T we have

$$
\begin{aligned}
&\sum_{t=1}^{T} 2\eta\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\
&\leq D_x + L_f^K\eta^2 T + \sum_{t=1}^{T}\eta K L_f^{m_1}\Big(\sum_{i=1}^{K}(\frac{c}{e})^c(\frac{t\beta}{2})^{-c}\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] \\
&\qquad\qquad + 4\beta\sigma_f^2 K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2 L_f^K}{\beta}\Big)/\sqrt{\beta} \\
&\quad + \sum_{t=1}^{T}\eta L_f^{m_2}\Big(\sum_{i=1}^{K}(\frac{c}{e})^c(\frac{t\beta}{2})^{-c}(\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) \\
&\qquad\qquad + \frac{4(\sum_{i=1}^{K}(2L_f^2)^i)\eta^2 L_f^K}{\beta} + 4\beta K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i))\Big)/\sqrt{\beta} \\
&\quad + \eta\sqrt{\beta}(K^2 L_f^{m_1} + K L_f^{m_2})D_x T.
\end{aligned}
$$

Then denote $L_f^m = \max\{L_f^{m_1}, L_f^{m_2}\}$ we can get

$$
\begin{aligned}
&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\
&\leq O\Big(D_x(\eta T)^{-1} + L_f^K\eta + (\sum_{i=1}^{K}U_i)L_f^m\beta^{-1/2-c}T^{-1}\sum_{t=1}^{T}t^{-c} + L_f^m\sigma_f^2((\sum_{i=1}^{K}(L_f^2)^i) + 1)\beta^{1/2} \\
&\quad + L_f^m\sum_{i=1}^{K}(L_f^2)^i\eta^2\beta^{-3/2} + (\sum_{i=1}^{K}U_i + V_i)L_f^m\beta^{-1/2-c}T^{-1}\sum_{t=1}^{T}t^{-c} + L_f^m(\sum_{i=1}^{K}(L_f^2)^i)\eta^2\beta^{-3/2} \\
&\quad + L_f^m(\sigma_f^2 + \sigma_J^2 + \sigma_f^2(\sum_{i=1}^{K}(L_f^2)^i))\beta^{1/2} + D_x L_f^m\beta^{1/2}\Big).
\end{aligned}
\tag{59}
$$

Noting that $\sum_{t=1}^{T}t^{-z} = O(T^{1-z})$ for $z \in (0,1)\cup(1,\infty)$ and $\sum_{t=1}^{T}t^{-1} = O(\log T)$, as long as $c > 2$ we get

$$
\begin{aligned}
&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\
&\leq O\Big(D_x(\eta T)^{-1} + L_f^K\eta + L_f^m(\sum_{i=1}^{K}U_i + V_i)\beta^{-1/2-c}T^{-c} \\
&\quad + (L_f^m(\sigma_f^2 + \sigma_J^2 + \sigma_f^2(\sum_{i=1}^{K}(L_f^2)^i)) + D_x L_f^m)\beta^{1/2} + L_f^m(\sum_{i=1}^{K}(L_f^2)^i)\eta^2\beta^{-3/2}\Big).
\end{aligned}
$$

This complete the proof.

□

proof of Theorem 6. According to (58), we have

$$
\begin{aligned}
\sum_{k=1}^{K} \|x_t - x_t^{l,k}\| \leq{}& 6K \sup_S \eta \sum_{s=1}^{t-1} \sum_{i=1}^{K} \sum_{j=1}^{i-1} L_f^{K-j+(i-1)i/2} (\mathbb{E}_A \|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\
&+ 6K \sup_S \eta \sum_{s=1}^{t-1} \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i} (\mathbb{E}_A \|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} + \sum_{k=1}^{K} \frac{6\eta L_f^K t}{n_k} \\
\leq{}& 6K^2 L_f^m \sup_S \eta \sum_{s=1}^{t-1} \sum_{j=1}^{K-1} (\mathbb{E}_A \|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\
&+ 6K L_f^m \sup_S \sum_{s=1}^{t-1} \sum_{i=1}^{K} (\mathbb{E}_A \|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} + \sum_{k=1}^{K} \frac{6\eta L_f^K t}{n_k}
\end{aligned}
$$

Then using Lemma 30 and 31 we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \|x_t - x_t^{l,k}\| \\
&\leq 6K^2 L_f^m \sup_S \eta \sum_{s=1}^{t-1} (\sum_{i=1}^{K} (\frac{c}{e})^c (\frac{s\beta}{2})^{-c} \mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + 4\beta\sigma_f^2 K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) \\
&\qquad + \frac{2\sum_{i=1}^{K}(2L_f^2)^i \eta^2 L_f^K}{\beta})^{1/2} + \sum_{k=1}^{K} \frac{6\eta L_f^K t}{n_k} \\
&\quad + 6K L_f^m \sup_S \sum_{s=1}^{t-1} (\sum_{i=1}^{K} (\frac{c}{e})^c (\frac{s\beta}{2})^{-c} (\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) \\
&\qquad + \frac{4(\sum_{i=1}^{K}(2L_f^2)^i)\eta^2 L_f^K}{\beta} + 4\beta K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i)))^{1/2}.
\end{aligned}
$$

Thus we get

$$
\begin{aligned}
&\sum_{k=1}^{K} \|x_t - x_t^{l,k}\| \\
&\leq 6K^2 L_f^m \eta \sqrt{\sum_{i=1}^{K} U_i(\frac{2c}{e})}^c \beta^{-c/2} \sum_{s=1}^{t} s^{-c/2} + 6K^2 L_f^m \sqrt{2\sum_{i=1}^{K}(2L_f^2)^i L_f^K} \cdot \eta^2 \beta^{-1/2} t + \sum_{k=1}^{K} \frac{6\eta L_f^K t}{n_k} \\
&\quad + 12K^2 L_f^m \sqrt{((\sigma_f^2 + \sigma_J^2) + 2(\sum_{i=1}^{K}(2L_f^2)^i) + \sigma_f^2)K} \cdot \beta^{1/2} t\eta + 6K L_f^m \eta \sqrt{\sum_{i=1}^{K}(U_i + V_i)(\frac{2c}{e})}^c \beta^{-c/2} \sum_{s=1}^{t} s^{-c/2}.
\end{aligned}
$$

According to Theorem 1, we have

$$
\begin{aligned}
&\mathbb{E}_{S,A}[F(x_t) - F_S(x_t))] \\
&\leq L_f^K \epsilon_K + 4L_f^K \sum_{k=1}^{K-1} \|x_t - x_t^{l,k}\| + L_f \sum_{k=1}^{K-1} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]}{n_k}} \\
&\leq 24K^2 L_f^{m+K} \eta \sqrt{\sum_{i=1}^{K} U_i(\frac{2c}{e})^c \beta^{-c/2} \sum_{s=1}^{t} s^{-c/2}} + 24K^2 L_f^{m+K} \sqrt{2\sum_{i=1}^{K}(2L_f^2)^i L_f^K \cdot \eta^2 \beta^{-1/2} t + \sum_{k=1}^{K} \frac{24\eta L_f^{2K} t}{n_k}} \\
&\quad + 48K^2 L_f^{m+K} \sqrt{((\sigma_f^2 + \sigma_J^2) + 2(\sum_{i=1}^{K}(2L_f^2)^i) + \sigma_f^2)K \cdot \beta^{1/2} t\eta} \\
&\quad + 24K^2 L_f^{m+K} \eta \sqrt{\sum_{i=1}^{K}(U_i + V_i)(\frac{2c}{e})^c \beta^{-c/2} \sum_{s=1}^{t} s^{-c/2}} + L_f \sum_{k=1}^{K-1} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]}{n_k}}.
\end{aligned}
$$

According to (59), we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{S,A}[F(x_t) - F_S(x_*^S))] \\
&\leq D_x \eta^{-1} + L_f^K \eta T + \sum_{t=1}^{T} K L_f^m (\sum_{i=1}^{K} (\frac{c}{e})^c (\frac{t\beta}{2})^{-c} \mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] \\
&\qquad\qquad + 4\beta\sigma_f^2 K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i \eta^2 L_f^K}{\beta})/\sqrt{\beta} \\
&\quad + \sum_{t=1}^{T} L_f^m (\sum_{i=1}^{K} (\frac{c}{e})^c (\frac{t\beta}{2})^{-c} (\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) \\
&\qquad\qquad + \frac{4(\sum_{i=1}^{K}(2L_f^2)^i)\eta^2 L_f^K}{\beta} + 4\beta((\sigma_f^2 + \sigma_J^2) + 2(\sum_{i=1}^{K}(2L_f^2)^i) + \sigma_f^2)K)/\sqrt{\beta} \\
&\quad + \sqrt{\beta} K^2 L_f^m D_x T + 24K^2 L_f^{m+K} \eta \sqrt{\sum_{i=1}^{K} U_i(\frac{2c}{e})^c \beta^{-c/2} \sum_{s=1}^{t} s^{-c/2}} \\
&\quad + 24K^2 L_f^{m+K} \sqrt{2\sum_{i=1}^{K}(2L_f^2)^i L_f^K \cdot \eta^2 \beta^{-1/2} t + \sum_{k=1}^{K} \frac{24\eta L_f^{2K} t}{n_k}} \\
&\quad + 48K^2 L_f^{m+K} \sqrt{((\sigma_f^2 + \sigma_J^2) + 2(\sum_{i=1}^{K}(2L_f^2)^i) + \sigma_f^2)K \cdot \beta^{1/2} t\eta} \\
&\quad + 24K^2 L_f^{m+K} \eta \sqrt{\sum_{i=1}^{K}(U_i + V_i)(\frac{2c}{e})^c \beta^{-c/2} \sum_{s=1}^{t} s^{-c/2}} + L_f \sum_{k=1}^{K-1} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]}{n_k}}.
\end{aligned}
$$

Noting that $\sum_{t=1}^{T} t^{-z} = O(T^{1-z})$ for $z \in (-1, 0) \cup (-\infty, -1)$ and $\sum_{t=1}^{T} t^{-1} = O(\log T)$, we have

$$
\sum_{t=1}^{T} \sum_{j=1}^{T} j^{-\frac{c}{2}} = O(\sum_{t=1}^{T} t^{1-\frac{c}{2}} (\log t)^{\mathbf{1}_{c=2}}) = O(T^{2-\frac{c}{2}} (\log T)^{\mathbf{1}_{c=2}}).
$$

Setting $\eta = T^{-a}$ and $\beta = T^{-b}$ we can get

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{S,A}[F(x_t)-F_S(x_*^S))]\\
&\leq O(T^{a}+T^{1-a}+T^{1-(1-b)c+\frac{b}{2}}(\log T)^{\mathbf{1}_{c=1}}+T^{-b/2}+T^{1+3b/2-2a}+T^{2-a}\sum_{k=1}^{K}n_k^{-1}+T^{1-b/2}+T^{2-a-b/2}\\
&\qquad+T^{2-a-c/2(1-b)}(\log T)^{\mathbf{1}_{c=1}}+T^{2-2a+1/2b}+T\sum_{k=1}^{K}n_k^{-1/2}).
\end{aligned}
$$

Dividing both side of above inequality with T, then from the choice of $A(S)$ we have

$$
\begin{aligned}
&\mathbb{E}_{S,A}[F(A(S))-F(x_*))]\\
&\leq O(T^{a-1}+T^{-a}+T^{-(1-b)c+\frac{b}{2}}(\log T)^{\mathbf{1}_{c=1}}+T^{-b/2-1}+T^{3b/2-2a}+T^{1-a}\sum_{k=1}^{K}n_k^{-1}+T^{-b/2}+T^{1-a-b/2}\\
&\qquad+T^{1-a-c/2(1-b)}(\log T)^{\mathbf{1}_{c=1}}+T^{1-2a+1/2b}+\sum_{k=1}^{K}n_k^{-1/2}).
\end{aligned}
$$

As long as we have $c>4$, the dominating terms are $O(T^{1-a-\frac{b}{2}})$, $O(T^{1+\frac{b}{2}-2a})$, $O(T^{1-a}\sum_{k=1}^{K}n_k^{-1})$, $O(T^{a-1})$, and $O(T^{\frac{3}{2}b-2a})$. Setting $a=b=\frac{4}{5}$, we have

$$
\mathbb{E}_{S,A}\Big[F(A(S))-F(x_*)\Big]=O(T^{-\frac{1}{5}}+T^{\frac{1}{5}}\sum_{k=1}^{K}n_k^{-1}+\sum_{k=1}^{K}n_k^{-1/2}).
$$

Letting $T=O(\max\{n_1^{2.5},\cdots,n_K^{2.5}\})$ we have the following

$$
\mathbb{E}_{S,A}\Big[F(A(S))-F(x_*)\Big]=O(\sum_{k=1}^{K}n_k^{-1/2}).
$$

This complete the proof. □

E.2. Strongly Convex setting

Similarly, since changing one sample data can happen in any layer of the function, we keep the same notations as in Section E.1.

Case 1($i_t\neq l$). We have

$$
\begin{aligned}
\|x_{t+1}-x_{t+1}^{l,k}\|^2&\leq\|x_t-\eta_t\prod_{i=1}^{K}v_t^{(i)}-x_t^{l,k}+\eta_t\prod_{i=1}^{K}v_t^{(i),l,k}\|^2\\
&\leq\|x_t-x_t^{l,k}\|^2-2\eta_t\langle\prod_{i=1}^{K}v_t^{(i)}-\prod_{i=1}^{K}v_t^{(i),l,k},x_t-x_{t+1}^{l,k}\rangle+\eta_t^2\|\prod_{i=1}^{K}v_t^{(i)}-\prod_{i=1}^{K}v_t^{(i),l,k}\|^2.
\end{aligned}
\tag{60}
$$

Now we estimate the second term of above inequality. we decompose it to $K(K+2)$ terms. According to the strongly convexity of $F_S(\cdot)$, we have

$$
\begin{aligned}
&\langle\prod_{i=1}^{K}\nabla F_{i,S}(x_t)-\prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k}),x_t-x_t^{l,k}\rangle\\
&\geq\frac{L\mu}{L+\mu}\|x_t-x_t^{l,k}\|^2+\frac{1}{L+\mu}\|\prod_{i=1}^{K}\nabla F_{i,S}(x_t)-\prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k})\|^2.
\end{aligned}
$$

Using Assumption 3 (iii) and strong convexity, similar to convex setting we can get

$$\begin{aligned}
&-2\eta_t\langle\prod_{i=1}^{K}v_t^{(i)}-\prod_{i=1}^{K}v_t^{(i),l,k},x_t-x_t^{l,k}\rangle\\
&\leq 2\eta_t L_f^{K-1}\|v_t^{(1)}-\nabla f_{1,S}(x_t)\|\cdot\|x_t-x_t^{l,k}\|\\
&\quad+(2\eta_t L_f^{K-1}\|v_t^{(2)}-\nabla f_{2,S}(u_t^{(1)})\|+2\eta_t L_f^{K}\|u_t^{(1)}-f_{1,S}(x_t)\|)\cdot\|x_t-x_t^{l,k}\|\\
&\quad\vdots\\
&\quad+(2\eta_t L_f^{m_2}\|v_t^{(K)}-\nabla f_{K,S}(u_t^{(K-1)})\|+\cdots+2\eta_t L_f^{K-1+(K-1)K/2}\|u_t^{(1)}-f_{1,S}(x_t)\|)\cdot\|x_t-x_t^{l,k}\|\\
&\quad-\frac{2\eta_t L\mu}{L+\mu}\|x_t-x_t^{l,k}\|^2-\frac{2\eta_t}{L+\mu}\|\prod_{i=1}^{K}\nabla F_{i,S}(x_t)-\prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k})\|^2\\
&\quad+(2\eta_t L_f^{K-1+(K-1)K/2}\|u_t^{(1),l,k}x_t^{l,k}\|+\cdots 2\eta_t L_f^{(K-1)K/2}\|\nabla f_{K,S}(u_t^{(K-1),l,k})-\nabla v_t^{(K),l,k}\|)\cdot\|x_t-x_t^{l,k}\|\\
&\quad\vdots\\
&\quad+(2\eta_t L_f^{K}\|u_t^{(1),l,k}-f_{1,S}(x_t^{l,k})\|+2\eta_t L_f^{K-1}\|v_t^{(2),l,k}-\nabla f_{2,S}(u_t^{(1),l,k})\|)\cdot\|x_t-x_t^{l,k}\|\\
&\quad+2\eta_t L_f^{K-1}\|v_t^{(1),l,k}-\nabla f_{1,S}(x_t^{l,k})\|\cdot\|x_t-x_t^{l,k}\|.
\end{aligned}\tag{61}$$

Conclude above inequality, we have

$$\begin{aligned}
&-2\eta_t\langle\prod_{i=1}^{K}v_t^{(i)}-\prod_{i=1}^{K}v_t^{(i),l,k},x_t-x_t^{l,k}\rangle\\
&\leq 2\eta_t\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{K-j+\frac{1}{2}(i-1)i}(\|u_t^{(j)}-f_{j,S}(u_t^{(j-1)})\|+\|u_t^{(j),l,k}-f_{j,S}(u_t^{(j-1),l,k})\|)\cdot\|x_t-x_t^{l,k}\|\\
&\quad+2\eta_t\sum_{i=1}^{K}L_f^{K-i+\frac{1}{2}(i-1)i}(\|v_t^{(i)}-\nabla f_{i,S}(u_t^{(i-1)})\|+\|v_t^{(i),k,l}-\nabla f_{i,S}(u_t^{(i-1),k,l})\|)\cdot\|x_t-x_t^{l,k}\|\\
&\quad-\frac{2\eta_t L\mu}{L+\mu}\|x_t-x_t^{l,k}\|^2-\frac{2\eta_t}{L+\mu}\|\prod_{i=1}^{K}\nabla F_{i,S}(x_t)-\prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k})\|^2.
\end{aligned}$$

Changing the assumption of convexity to strong convexity does not affect the third term on the right side of (60), so we have

$$\begin{aligned}
&\eta_t^2\|\prod_{i=1}^{K}v_t^{(i)}-\prod_{i=1}^{K}v_t^{(i),l,k}\|^2\\
&\leq\eta_t^2\sum_{i=1}^{K}\sum_{j=1}^{i-1}L_f^{2K-2j+(i-1)i}(\|u_t^{(j)}-f_{j,S}(u_t^{(j-1)})\|^2+\|u_t^{(j),l,k}-f_{j,S}(u_t^{(j-1),l,k})\|^2)\\
&\quad+\eta_t^2\sum_{i=1}^{K}L_f^{2K-2i+(i-1)i}(\|v_t^{(i)}-\nabla f_{i,S}(u_t^{(i-1)})\|^2+\|v_t^{(i),k,l}-\nabla f_{i,S}(u_t^{(i-1),k,l})\|^2)\\
&\quad+(K+2)K\eta_t^2\|\prod_{i=1}^{K}\nabla F_{i,S}(x_t)-\prod_{i=1}^{K}\nabla F_{i,S}(x_t^{l,k})\|.
\end{aligned}$$

By setting $\eta_t \leq \frac{2}{(L+\mu)(K+2)K}$ we have

$$
\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,k}\|^2 \\
&\leq (1 - \frac{2\eta_t L\mu}{L+\mu})\|x_t - x_t^{l,k}\|^2 \\
&\quad + 2\eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}(\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\| + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}(\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\| + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + \eta_t^2 \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{2K-2j+(i-1)i}(\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2 + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2) \\
&\quad + \eta_t^2 \sum_{i=1}^{K} L_f^{2K-2i+(i-1)i}(\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2 + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2).
\end{aligned}
$$

Case 2 ($i_t = l$). We have

$$
\begin{aligned}
\|x_{t+1} - x_{t+1}^{l,k}\| &= \|x_t - \eta_t \prod_{i=1}^{K} v_t^{(i)} - x_t^{l,k} + \eta_t \prod_{i=1}^{K} v_t^{(i),l,k}\| \\
&\leq \|x_t - x_t^{l,k}\| + \eta_t \|\prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} v_t^{(i),l,k}\| \leq \|x_t - x_t^{l,k}\| + 2\eta_t L_f^K.
\end{aligned}
\tag{62}
$$

Therefore, we have

$$
\|x_{t+1} - x_{t+1}^{l,k}\|^2 \leq \|x_t - x_t^{l,k}\|^2 + 4\eta_t L_f^K \|x_t - x_t^{l,k}\| + 4\eta_t^2 L_f^{2K}.
$$

Combining above two cases, and taking the expectation w.r.t. A we have

$$
\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,k}\|^2] \\
&\leq (1 - \frac{2\eta_t L\mu}{L+\mu})\|x_t - x_t^{l,k}\|^2 \\
&\quad + 2\eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}(\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\| + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + 2\eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}(\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\| + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|) \cdot \|x_t - x_t^{l,k}\| \\
&\quad + \eta_t^2 \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{2K-2j+(i-1)i}(\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2 + \|u_t^{(j),l,k} - f_{j,S}(u_t^{(j-1),l,k})\|^2) \\
&\quad + \eta_t^2 \sum_{i=1}^{K} L_f^{2K-2i+(i-1)i}(\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2 + \|v_t^{(i),k,l} - \nabla f_{i,S}(u_t^{(i-1),k,l})\|^2) + \|x_t - x_t^{l,k}\|^2 \\
&\quad + 4\eta_t L_f^K \|x_t - x_t^{l,k}\| \cdot \mathbf{1}_{[i_t = l]} + 4\eta_t^2 L_f^{2K} \cdot \mathbf{1}_{[i_t = l]}.
\end{aligned}
$$

Then setting $\eta_t = \eta$ and using Lemma 13 we can get

$$\begin{aligned}
&\mathbb{E}_A\|x_t - x_t^{l,k}\|^2 \\
&\le 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\cdot(\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2}\cdot(\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}\cdot(\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2}\cdot(\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} \\
&\quad + \sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2j+(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2 + \mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2) \\
&\quad + \sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2i+(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2 + \mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2) \\
&\quad + \sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\frac{4\eta L_f^K}{n_k}(\mathbb{E}_A\|x_s - x_s^{l,k}\|^2)^{1/2} + \sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\frac{4\eta^2 L_f^{2K}}{n_k}.
\end{aligned}$$

Similarly, setting $u_t = (\mathbb{E}_A\|x_t - x_t^{l,k}\|^2)^{1/2}$,

$$\begin{aligned}
S_t &= \sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2j+(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2 + \mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2) \\
&\quad + \sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2i+(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2 + \mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2) \\
&\quad + \sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\frac{4\eta^2 L_f^{2K}}{n_k},
\end{aligned}$$

and

$$\begin{aligned}
\alpha_s &= 2\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\
&\quad + 2\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2} \\
&\quad + 2\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} \\
&\quad + 2\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2} + \sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\frac{4\eta L_f^K}{n_k}.
\end{aligned}$$

Then according to Lemma 14, we have

$$\begin{aligned}
u_t &\leq \sqrt{S_t} + \sum_{s=1}^{t-1} \alpha_s \\
&\leq (\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2j+(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2 + \mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2))^{1/2} \\
&\quad + (\sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2i+(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2 + \mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2))^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j),l,k} - f_{j,S}(u_s^{(j-1),l,k})\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i)} - \nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2} \\
&\quad + 2\sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|v_s^{(i),k,l} - \nabla f_{i,S}(u_s^{(i-1),k,l})\|^2)^{1/2} \\
&\quad + \sqrt{\frac{2\eta L_f^{2K}(L+\mu)}{n_k L\mu}} + \frac{2L_f^K(L+\mu)}{n_k L\mu},
\end{aligned}$$

where the last inequality holds by

$$\sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\frac{4\eta L_f^K}{n_k} \leq \frac{4\eta L_f^K}{n_k}\cdot\frac{L+\mu}{2\eta L\mu} = \frac{2L_f^K(L+\mu)}{n_k L\mu}.$$

Next, we will discuss which one is the dominant one, $(\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2j+(i-1)i}\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}$ or $\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}$. According to Lemma 30 we have

$$\begin{aligned}
&(\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2 L_f^{2K-2j+(i-1)i}\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\
&\leq \sqrt{K}\eta L_f^m(\sum_{s=1}^{t-1}\sum_{j=1}^{K-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\mathbb{E}_A\|u_s^{(j)} - f_{j,S}(u_s^{(j-1)})\|^2)^{1/2} \\
&\leq \sqrt{K}\eta L_f^m(\sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}(\sum_{i=1}^{K}(\frac{2c}{e})^c(s\beta)^{-c}\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] \\
&\qquad\qquad + 4\beta\sigma_f^2 K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2 L_f^K}{\beta})^{1/2} \\
&\leq \sqrt{K}L_f^m\sqrt{(\frac{2c}{e})^c\sum_{i=1}^{K}U_i\frac{\sqrt{(L+\mu)\eta}}{\sqrt{2L\mu}}}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}} + 2\sigma_f K\sqrt{\frac{L_f^m(\sum_{i=1}^{K}(2L_f^2)^i) + 1)(L+\mu)\eta}{2L\mu}}\cdot\beta^{1/2} \\
&\quad + \sqrt{\frac{KL_f^m\sum_{i=1}^{K}(2L_f^2)^i L_f^K(L+\mu)}{L\mu}}\eta^{3/2}\beta^{-1/2},
\end{aligned}$$

where the inequality holds by Lemma 12. As for the later,

$$\begin{aligned}
&\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&\leq K\eta L_f^m\sum_{s=1}^{t-1}\sum_{j=1}^{K-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&\leq K\eta L_f^m\sum_{s=1}^{t-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}(\sum_{i=1}^{K}(\frac{2c}{e})^c(s\beta)^{-c}\mathbb{E}[\|u_1^{(i)}-f_{i,S}(x_0)\|^2]\\
&\qquad\qquad + 4\beta\sigma_f^2K((\sum_{i=1}^{K}(2L_f^2)^i)+1)+\frac{2\sum_{i=1}^{K}(2L_f^2)^i\eta^2L_f^K}{\beta})^{1/2}\\
&\leq KL_f^m\sqrt{(\frac{2c}{e})^c}\sum_{i=1}^{K}U_i\frac{(L+\mu)}{2L\mu}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}+2\sigma_fK^2L_f^m\sqrt{(\sum_{i=1}^{K}(2L_f^2)^i)+1)}\cdot\frac{(L+\mu)}{2L\mu}\beta^{1/2}\\
&\quad + KL_f^m\sqrt{2\sum_{i=1}^{K}(2L_f^2)^iL_f^K}\frac{(L+\mu)}{2L\mu}\eta\beta^{-1/2}.
\end{aligned} \tag{63}$$

Comparing the above results, we can find the dominant term is $\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+\frac{1}{2}(i-1)i}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}$. Take a similar action for several other items and we can get

$$\begin{aligned}
u_t &\leq \sqrt{S_t}+\sum_{s=1}^{t-1}\alpha_s\\
&\leq 6\sum_{s=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&\quad + 6\sum_{s=1}^{t-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-i+(i-1)i/2}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}\\
&\quad + \sqrt{\frac{2\eta L_f^{2K}(L+\mu)}{n_kL\mu}}+\frac{2L_f^K(L+\mu)}{n_kL\mu}.
\end{aligned} \tag{64}$$

Since often we have $\eta\leq\min\frac{1}{n_k}$ for any $k\in[1,K]$. Therefore, we have

$$\begin{aligned}
\epsilon_k \leq O\Big(&\sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&+\sum_{s=1}^{T-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K+(i-3)i/2}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}+\frac{L_f^K(L+\mu)}{L\mu n_k}\Big).
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\sum_{k=1}^{K}\epsilon_k \leq O\Big(&\sum_{s=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta L_f^{K-j+(i-1)i/2}(\mathbb{E}_A\|u_s^{(j)}-f_{j,S}(u_s^{(j-1)})\|^2)^{1/2}\\
&+(\sum_{s=1}^{T-1}\sum_{i=1}^{K}(1-\frac{2\eta L\mu}{L+\mu})^{t-s}\eta^2L_f^{K+(i-3)i/2}(\mathbb{E}_A\|v_s^{(i)}-\nabla f_{i,S}(u_s^{(i-1)})\|^2)^{1/2}+\sum_{k=1}^{K}\frac{L_f^K(L+\mu)}{L\mu n_k}\Big).
\end{aligned}$$

This completes the proof.

Corollary 6 (K-level Optimization)**.** *Consider Algorithm 2 with $0 < \eta_t = \eta < 2/(L+\mu)K(K+2)$ and let $0 < \beta_t = \beta < \max\{1, 1/(4K\sum_{i=1}^{K}(2L_f^2)^i\}$ for any $t \in [0, T-1]$ and the output $A(S) = x_T$. Then, we have the following results*

$$\sum_{k=1}^{K} \epsilon_k \leq O((T\beta)^{-\frac{c}{2}} + \beta^{\frac{1}{2}} + \eta\beta^{-\frac{1}{2}} + \sum_{k=1}^{K} n_k^{-1}).$$

Next, we give the proof of Corollary 6.

proof of corollary 6. Putting the result (63) into (64), since often we have $\eta \leq \min \frac{1}{n_k}$ for any $k \in [1, K]$. Therefore, we have

$$\sum_{k=1}^{K} \epsilon_k \leq O(T^{-\frac{c}{2}}\beta^{-\frac{c}{2}} + \beta^{1/2} + \eta\beta^{-1/2} + \sum_{k=1}^{K} n_k^{-1}).$$

This complete the proof. □

Before giving the proof of Theorem 10, we first give a useful lemma.

Lemma 33. *Let Assumption 1(iii), 2 (iii) and 3 (iii) hold, F_S is μ-strongly convex, then for SVMR, we have for any x*

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq \mathbb{E}_A[F_S(x_t)|\mathcal{F}_t] - \frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2 + \frac{\eta_t^2 L_f^K}{2} \\ + 4K^4 L_f^m \eta_t \sum_{j=1}^{K-1} \mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]|\mathcal{F}_t] \\ + 4K^2 L_f^m \eta_t \sum_{i=1}^{K-1} \mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2)|\mathcal{F}_t].\end{aligned}$$

where $\mathbb{E}_A\, A$ denotes the expectation taken with respect to the randomness of the algorithm, and $\mathcal{F}_t$ is the σ-field generated by S.

proof of Lemma 33. According to the Assumption 3 (iii) we have

$$\begin{aligned}F_S(x_{t+1}) &\leq F_S(x_t) + \langle \nabla F_S(x_t), x_{t+1} - x_t\rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2 \\ &\leq F_S(x_t) - \eta_t\langle \nabla F_S(x_t), \prod_{i=1}^{K} v_t^{(i)}\rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2 \\ &= F_S(x_t) - \eta_t\langle \nabla F_S(x_t), \prod_{i=1}^{K} \nabla F_{i,S}(x_t)\rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2 - u_t,\end{aligned}$$

where $u_t = \eta_t\langle \nabla F_S(x_t), \prod_{i=1}^{K} v_t^{(i)} - \prod_{i=1}^{K} \nabla F_{i,S}(x_t)\rangle$.

Let $\mathcal{F}_t$ be the σ-field generated by S. Taking expectation with respect to the randomness of the algorithm conditioned on $\mathcal{F}_t$, we have

$$\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq \mathbb{E}_A[F_S(x_t)|\mathcal{F}_t] - \eta_t\|\nabla F_S(x_t)\|^2 + \frac{\eta_t^2 L_f^K}{2} - \mathbb{E}_A[u_t|\mathcal{F}_t].$$

Now we bound the term $\mathbb{E}_A[u_t|\mathcal{F}_t]$.

$$\begin{aligned}
-\mathbb{E}_A[u_t|\mathcal{F}_t] &= \mathbb{E}_A[\eta_t\langle \nabla F_S(x_t), \prod_{i=1}^{K}\nabla F_{i,S}(x_t) - \prod_{i=1}^{K} v_t^{(i)}\rangle|\mathcal{F}_t] \\
&= \mathbb{E}_A[\eta_t\langle \prod_{i=1}^{K}\nabla F_{i,S}(x_t) - \prod_{i=2}^{K}\nabla F_{i,S}(x_t)\cdot v_t^{(1)}, \nabla F_S(x_t)\rangle|\mathcal{F}_t] \\
&+ \mathbb{E}_A[\eta_t\langle \prod_{i=2}^{K}\nabla F_{i,S}(x_t)\cdot v_t^{(1)} - \prod_{i=3}^{K}\nabla F_{i,S}(x_t)\cdot v_t^{(1)}\cdot \nabla f_{2,S}(u_t^{(1)}), \nabla F_S(x_t)\rangle|\mathcal{F}_t] \\
&+ \mathbb{E}_A[\eta_t\langle \prod_{i=3}^{K}\nabla F_{i,S}(x_t)\cdot v_t^{(1)}\cdot \nabla f_{2,S}(u_t^{(1)}) - \prod_{i=3}^{K}\nabla F_{i,S}(x_t)\cdot v_t^{(1)}\cdot v_t^{(2)}, \nabla F_S(x_t)\rangle|\mathcal{F}_t] \\
&\vdots \\
&+ \mathbb{E}_A[\eta_t\langle \prod_{i=1}^{K-1} v_t^{(i)}\cdot \nabla f_{K,S}(u_t^{(K-1)}) - \prod_{i=1}^{K} v_t^{(i)}, \nabla F_S(x_t)\rangle|\mathcal{F}_t].
\end{aligned}$$

Concluding the above inequality, using Assumption 1 (iii) we have

$$\begin{aligned}
-\mathbb{E}_A[u_t|\mathcal{F}_t] &\leq \eta_t \sum_{i=1}^{K}\sum_{j=1}^{i-1} L_f^{K-j+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|\cdot\|\nabla F_S(x_t)\||\mathcal{F}_t] \\
&+ \eta_t \sum_{i=1}^{K} L_f^{K-i+\frac{1}{2}(i-1)i}\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|\cdot\|\nabla F_S(x_t)\||\mathcal{F}_t] \\
&\leq K L_f^m \eta_t \sum_{j=1}^{K-1} \mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|\cdot\|\nabla F_S(x_t)\||\mathcal{F}_t] \\
&+ L_f^m \eta_t \sum_{i=1}^{K-1} \mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|\cdot\|\nabla F_S(x_t)\||\mathcal{F}_t].
\end{aligned}$$

According to Cauchy-Schwartz inequality, we have

$$\begin{aligned}
-\mathbb{E}_A[u_t|\mathcal{F}_t] &\leq K L_f^m \eta_t \sum_{j=1}^{K-1}(\frac{\|F_S(x_t)\|^2}{\gamma_t} + \gamma_t\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]|\mathcal{F}_t] \\
&+ L_f^m \eta_t \sum_{i=1}^{K-1}(\frac{\|F_S(x_t)\|^2}{\lambda_t} + \lambda_t\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2)|\mathcal{F}_t].
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] &\leq \mathbb{E}_A[F_S(x_t)|\mathcal{F}_t] - \eta_t\|\nabla F_S(x_t)\|^2 + \frac{\eta_t^2 L_f^K}{2} \\
&+ K L_f^m \eta_t (K\frac{\|F_S(x_t)\|^2}{\gamma_t} + \sum_{j=1}^{K-1}\gamma_t\mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]|\mathcal{F}_t] \\
&+ L_f^m \eta_t (K\frac{\|F_S(x_t)\|^2}{\lambda_t} + \sum_{i=1}^{K-1}\lambda_t\mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2)|\mathcal{F}_t].
\end{aligned}$$

Setting $\gamma_t = 4K^2 L_f^m$ and $\lambda_t = 4KL_f^m$, we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1})|\mathcal{F}_t] \leq \mathbb{E}_A[F_S(x_t)|\mathcal{F}_t] - \frac{\eta_t}{2}\|\nabla F_S(x_t)\|^2 + \frac{\eta_t^2 L_f^K}{2} \\ + 4K^4 L_f^m \eta_t \sum_{j=1}^{K-1} \mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]|\mathcal{F}_t] \\ + 4K^2 L_f^m \eta_t \sum_{i=1}^{K-1} \mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2)|\mathcal{F}_t].\end{aligned}$$

This complete the proof. □

Next, we will give the detailed proof of Theorem 10.

proof of Theorem 10. Note that strong convexity implies the Polyak-Łojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla F_S(x)\|^2 \geq \mu(F_S(x) - F_S(x_*^S)), \quad \forall x.$$

Then according to Lemma 33 and PL condition, subtracting both sides with $F_S(x_*^S)$ we have

$$\begin{aligned}\mathbb{E}_A[F_S(x_{t+1}) - F_S(x_*^S)] \leq (1-\mu\eta_t)\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \frac{\eta_t^2 L_f^K}{2} \\ + 4K^4 L_f^m \eta_t \sum_{j=1}^{K-1} \mathbb{E}_A[\|u_t^{(j)} - f_{j,S}(u_t^{(j-1)})\|^2]] + 4K^2 L_f^m \eta_t \sum_{i=1}^{K-1} \mathbb{E}_A[\|v_t^{(i)} - \nabla f_{i,S}(u_t^{(i-1)})\|^2)].\end{aligned}$$

By setting $\eta_t = \eta$, $\beta_t = \beta$, according to Lemma 30 and Lemma 31 we have

$$\begin{aligned}&\mathbb{E}_A[F_S(x_{t+1}) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)\mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] + \frac{\eta^2 L_f^K}{2} \\ &\quad + 4K^4 L_f^m \eta (\sum_{i=1}^{K} (\frac{c}{e})^c (\frac{t\beta}{2})^{-c} \mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + 4\beta\sigma_f^2 K((\sum_{i=1}^{K}(2L_f^2)^i) + 1) + \frac{2\sum_{i=1}^{K}(2L_f^2)^i \eta^2 L_f^K}{\beta}.) \\ &\quad + 4K^2 L_f^m \eta (\sum_{i=1}^{K} (\frac{c}{e})^c (\frac{t\beta}{2})^{-c} (\mathbb{E}[\|u_1^{(i)} - f_{i,S}(x_0)\|^2] + \mathbb{E}[\|v_1^{(i)} - \nabla f_{i,S}(x_0)\|^2]) + \frac{4(\sum_{i=1}^{K}(2L_f^2)^i)\eta^2 L_f^K}{\beta} \\ &\qquad + 4\beta K(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i))).\end{aligned}$$

Telescoping the above inequality from 1 to $T-1$, we have

$$\begin{aligned}&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\ &\leq (1-\mu\eta)^{T-1}\mathbb{E}_A[F_S(x_1) - F_S(x_*^S)] + \frac{\eta^2 L_f^K}{2}\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1} + 4K^4 L_f^m (\frac{2c}{e})^c \eta\beta^{-c} (\sum_{i=1}^{K} U_i) \sum_{t=1}^{T-1} t^{-c}(1-\mu\eta)^{T-t-1} \\ &\quad + 16K^5 L_f^m \sigma_f^2 ((\sum_{i=1}^{K}(2L_f^2)^i) + 1)\eta\beta \sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1} + \frac{8K^4 L_f^m \sum_{i=1}^{K}(2L_f^2)^i \eta^3}{\beta}\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1} \\ &\quad + 4K^2 L_f^m (\frac{2c}{e})^c \eta\beta^{-c} (\sum_{i=1}^{K}(U_i + V_i)) \sum_{t=1}^{T-1} t^{-c}(1-\mu\eta)^{T-t-1} + \frac{16K^2 L_f^m \sum_{i=1}^{K}(2L_f^2)^i \eta^3}{\beta}\sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1} \\ &\quad + 16K^3 L_f^m (\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i))\eta\beta \sum_{t=1}^{T-1}(1-\mu\eta)^{T-t-1}.\end{aligned}$$

For $t = 0$, we have

$$\begin{aligned}
&\mathbb{E}_A[F_S(x_1) - F_S(x_*^S)] \\
&\leq (1-\mu\eta)\mathbb{E}_A[F_S(x_0) - F_S(x_*^S)] + \frac{\eta^2 L_f^K}{2} + 4K^4 L_f^m \eta \sum_{j=1}^{K-1} U_i + 4K^2 L_f^m \eta \sum_{j=1}^{K-1} (U_i + V_i).
\end{aligned}$$

Then combining above two cases, we have

$$\begin{aligned}
&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\
&\leq (1-\mu\eta)^T \mathbb{E}_A[F_S(x_0) - F_S(x_*^S)] + \frac{\eta^2 L_f^K}{2} \sum_{t=1}^{T} (1-\mu\eta)^{T-t} + 4K^4 L_f^m (\frac{2c}{e})^c \eta \beta^{-c} (\sum_{i=1}^{K} U_i) \sum_{t=1}^{T-1} t^{-c} (1-\mu\eta)^{T-t-1} \\
&\quad + 16K^5 L_f^m \sigma_f^2 ((\sum_{i=1}^{K} (2L_f^2)^i) + 1) \eta\beta \sum_{t=1}^{T-1} (1-\mu\eta)^{T-t-1} + \frac{8K^4 L_f^m \sum_{i=1}^{K} (2L_f^2)^i \eta^3}{\beta} \sum_{t=1}^{T-1} (1-\mu\eta)^{T-t-1} \\
&\quad + 4K^2 L_f^m (\frac{2c}{e})^c \eta \beta^{-c} (\sum_{i=1}^{K} (U_i + V_i)) \sum_{t=1}^{T-1} t^{-c} (1-\mu\eta)^{T-t-1} + \frac{16K^2 L_f^m \sum_{i=1}^{K} (2L_f^2)^i \eta^3}{\beta} \sum_{t=1}^{T-1} (1-\mu\eta)^{T-t-1} \\
&\quad + 16K^3 L_f^m (\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2 (\sum_{i=1}^{K} (2L_f^2)^i)) \eta\beta \sum_{t=1}^{T-1} (1-\mu\eta)^{T-t-1} \\
&\quad + (4K^4 L_f^m \eta \sum_{j=1}^{K-1} U_i + 4K^2 L_f^m \eta \sum_{j=1}^{K-1} (U_i + V_i))(1-\mu\eta)^{T-1}.
\end{aligned}$$

Then from Lemma 12, we have

$$\sum_{t=1}^{T-1} (1-\mu\eta)^{T-t-1} t^{-c} \leq \frac{\sum_{t=1}^{T-1} (1-\mu\eta)^{T-t-1}}{T-1} \sum_{t=1}^{T-1} t^{-c} \leq \frac{1}{T\mu\eta} \sum_{t=1}^{T-1} t^{-c}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \\
&\leq (\frac{c}{e\mu})^c (\eta T)^{-c} D_x + \frac{\eta L_f^K}{\mu} \\
&\quad + \frac{4K^4 L_f^m (\frac{2c}{e})^c \beta^{-c} (\sum_{i=1}^{K} U_i)}{T\mu} \sum_{t=1}^{T-1} t^{-c} + \frac{16K^5 L_f^m \sigma_f^2 ((\sum_{i=1}^{K} (2L_f^2)^i) + 1)\beta}{\mu} \\
&\quad + \frac{8K^4 L_f^m \sum_{i=1}^{K} (2L_f^2)^i \eta^2}{\beta\mu} + \frac{4K^2 L_f^m (\frac{2c}{e})^c \beta^{-c} (\sum_{i=1}^{K} (U_i + V_i))}{T\mu} \sum_{t=1}^{T-1} t^{-c} \\
&\quad + \frac{16K^2 L_f^m \sum_{i=1}^{K} (2L_f^2)^i \eta^2}{\beta\mu} + \frac{16K^3 L_f^m (\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2 (\sum_{i=1}^{K} (2L_f^2)^i))\beta}{\mu} \\
&\quad + (4K^4 L_f^m \sum_{j=1}^{K-1} U_i + 4K^2 L_f^m \sum_{j=1}^{K-1} (U_i + V_i))(\frac{c}{e\mu})^c \eta (\eta T)^{-c}.
\end{aligned}$$

Moreover, note that $\sum_{t=1}^{T} t^{-z} = O(T^{1-z})$ for $z \in (0,1) \cup (1,\infty)$ and $\sum_{t=1}^{T} t^{-1} = O(\log T)$. As long as $c \neq 1$ we get

$$\begin{aligned}\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] = O\Big(&(\eta T)^{-c} D_x + \eta L_f^K + L_f^m (\sum_{i=1}^{K} U_i)(\beta T)^{-c} + L_f^m \sigma_f^2((\sum_{i=1}^{K}(L_f^2)^i) + 1)\beta \\ &+ L_f^m \sum_{i=1}^{K}(L_f^2)^i \eta^2 \beta^{-1} + L_f^m (\sum_{i=1}^{K}(U_i + V_i))(\beta T)^{-c} \\ &+ L_f^m \sum_{i=1}^{K}(L_f^2)^i \eta^2 \beta^{-1} + L_f^m (\sigma_f^2 + \sigma_J^2 + \sigma_f^2(\sum_{i=1}^{K}(L_f^2)^i))\beta \\ &+ (L_f^m \sum_{j=1}^{K-1} U_i + L_f^m \sum_{j=1}^{K-1}(U_i + V_i))\eta(\eta T)^{-c}\Big).\end{aligned}$$

By rearranging the above inequality, we can obtain

$$\begin{aligned}\mathbb{E}_A[F_S(x_T) - F_S(x_*^S)] \le O\Big(&(\eta T)^{-c} D_x + \eta L_f^K + L_f^m \sum_{i=1}^{K}(U_i + V_i)(\beta T)^{-c} \\ &+ L_f^m (\sigma_f^2 + \sigma_J^2 + \sigma_f^2(\sum_{i=1}^{K}(L_f^2)^i))\beta \\ &+ L_f^m \sum_{i=1}^{K}(L_f^2)^i \eta^2 \beta^{-1} + L_f^m \sum_{j=1}^{K-1}(U_i + V_i)\eta(\eta T)^{-c}\Big).\end{aligned}$$

The proof is completed. □

proof of Theorem 11. Combining Theorem 1, and Theorem 10 we have

$$\begin{aligned}&\mathbb{E}_{S,A}[F(x_T) - F_S(x_T)] \\ &\le L_f^K \epsilon_K + 4L_f^K \sum_{t=1}^{K-1} \epsilon_t + L_f \sum_{t=2}^{K} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-t+1}(A(S)]}{n_{K-t+1}}} \\ &\le 12L_f^K \sqrt{K L_f^{m+K} (\frac{2c}{e})^c \sum_{i=1}^{K} U_i \frac{(L+\mu)}{L\mu} T^{-\frac{c}{2}} \beta^{-\frac{c}{2}}} + L_f \sum_{t=2}^{K} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-t+1}(A(S)]}{n_{K-t+1}}} \\ &\quad + 24\sigma_f K L_f^K \sqrt{L_f^{m+K}(\sum_{i=1}^{K}(2L_f^2)^i) + 1)} \cdot \frac{(L+\mu)}{L\mu}\beta^{1/2} + 12L_f^K \sqrt{K L_f^{m+K} 2 \sum_{i=1}^{K}(2L_f^2)^i L_f^K \frac{(L+\mu)}{L\mu}} \eta\beta^{-1/2} \\ &\quad + 12L_f^K \sqrt{L_f^{m+K}(\frac{2c}{3})^c(\sum_{k=1}^{K}(U_i + v_1^{(i)}))} \frac{L+\mu}{L\mu} T^{-\frac{c}{2}}\beta^{-\frac{c}{2}} + 24L_f^K \sqrt{K L_f^{m+K} \sum_{i=1}^{K}(2L_f^2)^i L_f^K \frac{(L+\mu)}{L\mu}} \eta\beta^{-1/2} \\ &\quad + 24L_f^K \sqrt{(\sigma_f^2 + \sigma_J^2 + 2\sigma_f^2(\sum_{i=1}^{K}(2L_f^2)^i))} \frac{(L+\mu)}{L\mu}\beta^{1/2} + \sum_{k=1}^{K} \sqrt{\frac{2\eta L_f^{2K}(L+\mu)}{n_k L\mu}} + \sum_{k=1}^{K} \frac{2L_f^K(L+\mu)}{n_k L\mu},\end{aligned}$$

Then according to Theorem 10, we have

$$
\begin{aligned}
&\mathbb{E}_A[F(A(S)) - F(x_*)] \\
&\leq 12L_f^K\sqrt{KL_f^{m+K}(\frac{2c}{e})^c\sum_{i=1}^K U_i\frac{(L+\mu)}{L\mu}}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}} + L_f\sum_{t=2}^K\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_{K-t+1}(A(S))]}{n_{K-t+1}}} \\
&\quad + 24\sigma_f KL_f^K\sqrt{L_f^{m+K}(\sum_{i=1}^K(2L_f^2)^i)+1)\cdot\frac{(L+\mu)}{L\mu}}\beta^{1/2} + 12L_f^K\sqrt{KL_f^{m+K}2\sum_{i=1}^K(2L_f^2)^iL_f^K\frac{(L+\mu)}{L\mu}}\eta\beta^{-1/2} \\
&\quad + 12L_f^K\sqrt{L_f^{m+K}(\frac{2c}{3})^c(\sum_{k=1}^K(U_i+v_1^{(i)}))\frac{L+\mu}{L\mu}}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}} + 24L_f^K\sqrt{KL_f^{m+K}\sum_{i=1}^K(2L_f^2)^iL_f^K\frac{(L+\mu)}{L\mu}}\eta\beta^{-1/2} \\
&\quad + 24L_f^K\sqrt{(\sigma_f^2+\sigma_J^2+2\sigma_f^2(\sum_{i=1}^K(2L_f^2)^i))\frac{(L+\mu)}{L\mu}}\beta^{1/2} + \sum_{k=1}^K\sqrt{\frac{2\eta L_f^{2K}(L+\mu)}{n_kL\mu}} + \sum_{k=1}^K\frac{2L_f^K(L+\mu)}{n_kL\mu} \\
&\quad + (\eta T)^{-c}D_x + \eta L_f^K + L_f^m(\sum_{i=1}^K U_i)(\beta T)^{-c} + L_f^m\sigma_f^2((\sum_{i=1}^K(2L_f^2)^i)+1)\beta \\
&\quad + L_f^m\sum_{i=1}^K(L_f^2)^i\eta^2\beta^{-1} + L_f^m(\sum_{i=1}^K(U_i+V_i))(\beta T)^{-c} \\
&\quad + L_f^m\sum_{i=1}^K(L_f^2)^i\eta^2\beta^{-1} + L_f^m(\sigma_f^2+\sigma_J^2+\sigma_f^2(\sum_{i=1}^K(L_f^2)^i))\beta \\
&\quad + (L_f^m\sum_{j=1}^{K-1}U_i + L_f^m\sum_{j=1}^{K-1}(U_i+V_i))\eta(\eta T)^{-c}.
\end{aligned}
$$

Setting $\eta = T^{-a}\beta = T^{-b}$, we have

$$
\begin{aligned}
&\mathbb{E}_A[F(A(S)) - F(x_*)] \\
&\leq O(T^{\frac{c}{2}(b-1)} + T^{-\frac{b}{2}} + T^{\frac{b}{2}-a} + \sum_{i=1}^K n_k^{-1} + T^{-c(1-a)} + T^{-a} + +T^{-c(1-b)} + T^{-b} + T^{b-2a} + T^{-c(1-a)-a}).
\end{aligned}
$$

Setting $c = 3$, then the dominating terms are $O(T^{\frac{b}{2}-a})$, $O(T^{-\frac{b}{2}})$, $O(T^{\frac{c}{2}(b-1)})$, $O(T^{-\frac{a}{2}})$, and $O(T^{-c(1-a)})$. Then setting $a = b = \frac{6}{7}$ we have

$$\mathbb{E}_A[F(A(S)) - F(x_*)] = O(T^{-\frac{3}{7}}).$$

Then setting $T = O(\max\{n_1^{\frac{7}{6}}, \cdots, n_K^{\frac{7}{6}}\})$, we have

$$\mathbb{E}_A[F(A(S)) - F(x_*)] = O(\sum_{k=1}^K\frac{1}{\sqrt{n_k}}).$$

Then we complete the proof. □