Stability and Generalization of Stochastic Compositional Gradient Descent Algorithms

Ming Yang [*1] **Xiyuan Wei** [*2] **Tianbao Yang** [2] **Yiming Ying** [3]

Abstract

Many machine learning tasks can be formulated as a stochastic compositional optimization (SCO) problem such as reinforcement learning, AUC maximization and meta-learning, where the objective function involves a nested composition associated with an expectation. Although many studies have been devoted to studying the convergence behavior of SCO algorithms, there is little work on understanding their generalization, that is, how these learning algorithms built from training data would behave on future test examples. In this paper, we provide the stability and generalization analysis of stochastic compositional gradient descent algorithms in the framework of statistical learning theory. Firstly, we introduce a stability concept called *compositional uniform stability* and establish its quantitative relation with generalization for SCO problems. Then, we establish the compositional uniform stability results for two notable stochastic compositional gradient descent algorithms, namely SCGD and SCSC. Finally, we derive *dimension-independent* excess risk bounds for SCGD and SCSC by balancing stability results and optimization errors. To the best of our knowledge, these are the first-ever known results on stability and generalization analysis of stochastic compositional gradient descent algorithms.

1. Introduction

Recently, *stochastic compositional optimization (SCO)* has attracted considerable interest (Chen et al., 2021a;b; Dentcheva et al., 2017; Ghadimi et al., 2020; Hu et al., 2020; Tolstaya et al., 2018; Wang et al., 2017; 2016; Zhang & Lan, 2020) in machine learning. It has the following form:

$$\min_{x\in\mathcal{X}}\Big\{F(x) = f\circ g(x) = \mathbb{E}_\nu[f_\nu(\mathbb{E}_\omega[g_\omega(x)])]\Big\}, \quad (1)$$

where $f\circ g(x) = f(g(x))$ denotes the function composition, $f:\mathbb{R}^d\to\mathbb{R}$ and $g:\mathbb{R}^p\to\mathbb{R}^d$ are differentiable functions, ν,ω are random variables, and $\mathcal{X}$ is a convex domain in $\mathbb{R}^p$. SCO generalizes the classic (non-compositional) stochastic optimization where its objective function $F(\cdot)$ involves nested compositions of functions and each composition is associated with an expectation.

SCO problem (1) instantiates various learning problems. For example, reinforcement learning (Sutton & Barto, 2018; Szepesvári, 2010) aims to obtain a value function of the given policy that can be considered as an SCO problem (Wang et al., 2017). Model-agnostic meta-learning (MAML) (Finn et al., 2017) finds a common initialization for rapid adaptation to new tasks, which was essentially a SCO problem, as pointed out by Chen et al. (2021a). Portfolio optimization with risk aversion (Shapiro et al., 2021), bias-variance issues in supervised learning (Dentcheva et al., 2017; Tolstaya et al., 2018), and robust group distributional optimization (Qi et al., 2021a; Jiang et al., 2022b) can also be formulated in similar SCO forms. Likewise, other learning tasks, such as maximization of the area under precision-recall curves (AUCPRC), and other compositional performance measures, can be cast in a similar way (Yang, 2022).

There are a substantial number of studies devoted to studying the convergence behavior of stochastic compositional optimization algorithms for solving (1). Wang et al. (2017) pioneered the non-asymptotic analysis of the so-called stochastic compositional gradient decent algorithms (SCGD) which employed two-time scales with a slower stepsize to update the variable and a faster one used in the moving average sequence y_{t+1} to track the inner function $g(x_t)$. An accelerated version of SCGD was analyzed by Wang et al. (2016) and its adapted variant was studied by Tutunov et al. (2020). In particular, Chen et al. (2021a) proposed the stochasti-

[*]Equal contribution [1]Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY 12222, USA [2]Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA [3]The University of Sydney, School of Mathematics and Statistics, Sydney, NSW 2006, Australia. Correspondence to: Yiming Ying <yiming.ying@sydney.edu.au>.

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

cally corrected SCGD called SCSC which was shown to enjoy the same convergence rate as that of standard SGD in the non-compositional setting. Further extensions and their convergence analysis were investigated in different settings such as single timescale (Ghadimi et al., 2020; Ruszczynski, 2021), variance reduction techniques (Hu et al., 2019; Devraj & Chen, 2019; Lin et al., 2018), and applications to nonstandard learning tasks (Yang, 2022).

On the other important front, one crucial aspect of machine learning is the development of learning algorithms that can achieve strong generalization performance. Generalization refers to the ability of a learning algorithm to perform well on unseen or future test data, despite being trained on a limited set of historical training data. In the last couple of years, we have witnessed a large amount of work on addressing the generalization analysis of the vanilla stochastic gradient descent (SGD) with focus on the classical ERM formulation in the *non-compositional* setting. In particular, the stability and generalization of SGD have been studied using the uniform argument stability (Bassily et al., 2020; Charles & Papailiopoulos, 2018; Hardt et al., 2016; Kuzborskij & Lampert, 2018) and the on-average model stability (Lei & Ying, 2020). In Farnia & Ozdaglar (2021); Lei et al. (2021); Zhang et al. (2021), different stability and generalization measures are investigated for minimax optimization algorithms. However, to our knowledge, there is no work to understand the important stability and generalization properties of stochastic compositional optimization algorithms despite their increasing popularity in solving many machine learning tasks (Chen et al., 2021a; Dentcheva et al., 2017; Jiang et al., 2022b; Wang et al., 2017; Yang & Ying, 2022; Yang, 2022).

Our Contributions. In this paper, we are mainly interested in the stability and generalization of stochastic compositional optimization algorithms in the framework of Statistical Learning Theory(Vapnik, 1999; Bousquet et al., 2004). Our main contributions are summarized below.

- We introduce a stability concept called *compositional uniform stability* which is tailored to handle the composition structure in SCO problems. Furthermore, we show the qualitative connection between this stability concept and the generalization error for randomized SCO algorithms. Regarding technical contributions, we show that this connection can mainly be derived by estimating the stability terms involving the outer function f_ν and the vector-valued generalization term of the inner function g_ω, which will be further estimated using the sample splitting argument (Bousquet et al., 2020; Lei, 2022).
- More specifically, we establish the compositional uniform stability of SCGD and SCSC in the convex and smooth case. Our stability bound mainly involves two terms, that is, the empirical variance associated with the inner function g_ω and the convergence of the moving average sequence to track $g_S(x_t)$. Then we establish the excess risk bounds $\mathcal{O}(1/\sqrt{n}+1/\sqrt{m})$ for both SCGD and SCSC by balancing the stability results and optimization errors, where n and m denote the numbers of training data involving ν and ω, respectively. Our results demonstrate that to achieve the same excess risk rate of $\mathcal{O}(1/\sqrt{n}+1/\sqrt{m})$, SCGD requires a larger number of iterations, approximately $T \asymp \max(n^{3.5}, m^{3.5})$, while SCSC only needs $T \asymp \max(n^{2.5}, m^{2.5})$.
- We further extend the analysis of stability and generalization for SCGD and SCSC in the strongly convex and smooth case. Specifically, we show that SCGD requires approximately $T \asymp \max(n^{10/3}, m^{10/3})$ iterations, while SCSC only needs $T \asymp \max(n^{7/3}, m^{7/3})$ iterations to achieve the excess risk rate of $\mathcal{O}(1/n+1/\sqrt{m})$.

1.1. Related Work

In this section, we review related work on algorithmic stability and generalization analysis of stochastic optimization algorithms and algorithms for compositional problems.

Stochastic Compositional Optimization. The seminal work of Wang et al. (2017) introduced SCGD with two time scales, and Wang et al. (2016) presented an accelerated version. Lian et al. (2017) incorporated variance reduction, while Ghadimi et al. (2020) proposed a modified SCGD with a single timescale. Chen et al. (2021a) introduced SCSC, a stochastically corrected version with the same convergence rate as vanilla SGD. Ruszczynski (2021); Zhang & Lan (2020) explored problems with multiple levels of composition, and Wang & Yang (2022) proposed SOX for compositional problems. Recently, there has been a growing interest in applying stochastic compositional optimization algorithms to optimize performance measures in machine learning, such as AUC scores (Qi et al., 2021b; Lei & Ying, 2021; Yang, 2022). Most of these studies have focused mainly on convergence analysis.

Algorithmic Stability and Generalization for the Non-Compositional Setting. Uniform stability and generalization of ERM were established by Bousquet & Elisseeff (2002) in a strongly convex setting. Elisseeff et al. (2005) studied stability of randomized algorithms, and Feldman & Vondrak (2019); Bousquet et al. (2020) derived high-probability generalization bounds for uniformly stable algorithms. Hardt et al. (2016) established uniform argument stability and generalization of SGD in expectation for smooth convex functions. Kuzborskij & Lampert (2018) established data-dependent stability results for SGD. On-average model stability and generalization of SGD were derived in Lei & Ying (2020) for convex objectives in smooth and non-smooth settings. The stability and generalization of SGD with continuous convex and Lipschitz objectives were studied in Bassily et al. (2020). For the non-convex and smooth cases, the stability of SGD was investigated in Charles &

Papailiopoulos (2018); Lei & Ying (2020); Lei et al. (2022). Further extensions were made for SGD in pairwise learning (Shen et al., 2019; Yang et al., 2021), Markov chain SGD (Wang et al., 2022), and minimax optimization algorithms (Farnia & Ozdaglar, 2021; Lei et al., 2021). However, existing studies have primarily focused on SGD algorithms and their variants for the standard ERM problem in the noncompositional setting.

Recently, Hu et al. (2020) studied the generalization and uniform stability of the exact minimizer of the ERM counterpart for the SCO problem using the uniform convergence approach (Bartlett & Mendelson, 2002; Vapnik, 2013; Zhou, 2002). They also showed uniform stability of its ERM minimizer under the assumption of a Hölderian error bound condition that instantiates strong convexity. Their bounds are algorithm-independent. To the best of our knowledge, there is no existing work on stability and generalization for stochastic compositional optimization algorithms, despite their popularity in solving machine learning tasks.

Organization of the Paper. The paper is organized as follows. Section 2 formulates the learning problem and introduces the necessary stability concepts. Two popular stochastic compositional optimization algorithms, SCGD (Wang et al., 2017) and SCSC (Chen et al., 2021a), are presented to solve (1). The main results on stability and generalization for SCGD and SCSC algorithms are illustrated in Section 3. Finally, Section 4 concludes the article.

2. Problem Setting

In this section, we illustrate the objective of generalization analysis and the stability concept used in the framework of Statistical Learning Theory (Vapnik, 2013; Bousquet et al., 2004). Then, we describe two popular optimization schemes, i.e. SCGD and SCSC, for solving the SCO problems, as well as other necessary notation.

2.1. Target of Generalization Analysis

For simplicity, we are mainly concerned with the case that the random variables ν and ω are independent, which means that $g(\mathbf{x}) = \mathbb{E}[g_\omega(\mathbf{x})] = \mathbb{E}[g_\omega(\mathbf{x})|\nu]$ for any ν. This is the case that was considered in Wang et al. (2017). In particular, important learning tasks such as group distributionally robust optimization (DRO) (Qi et al., 2021a) and AUC maximization (Kar et al., 2013; Liu et al., 2018; Ying et al., 2016; Yang & Ying, 2022; Zhao et al., 2011) are two notable examples described below.

Specifically, DRO formulations are capable of handling noisy data, adversarial data, and imbalanced classification data and have received considerable attention in machine learning. In Qi et al. (2021a), a class of group DRO is formulated as

$$\min_{x\in\mathcal{X}} F_S(x) := \lambda \log\Big(\frac{1}{m}\sum_{j=1}^{m}\exp(\ell(x^T a_j, b_j)/\lambda)\Big),$$

which can be reduced to SCO problem with $f_{\nu_i}(y) = \lambda\log(y)$ for any ν_i, $g_{\omega_j}(x) = \exp(\ell(x^T a_j, b_j)/\lambda)$ with $\omega_j = \{a_j, b_j\}$ being an input/output pair. For AUC maximization, Lei & Ying (2021); Yang & Ying (2022) showed that maximizing the AUC score with the least squares loss can be considered an SCO problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}[(h_{\mathbf{w}}(\mathbf{x})) - a(\mathbf{w}))^2 \,|y=1] + \mathbb{E}[(h_{\mathbf{w}}(\mathbf{x}') - b(\mathbf{w}))^2 \,|y'=1] + (1 - a(\mathbf{w}) + b(\mathbf{w}))^2,$$

where $h_{\mathbf{w}}(\cdot)$ is the decision function, $a(\mathbf{w}) = \mathbb{E}[h_{\mathbf{w}}(\mathbf{x})|y=1]$, and $b(\mathbf{w}) = \mathbb{E}[h_{\mathbf{w}}(\mathbf{x}')|y'=-1]$. The above two examples fall into the setting where ν and ω are independent. Lastly, it is worth mentioning that in Appendix E we briefly discuss the case where the random variables ν and ω depend on each other.

In practice, we do not know the population distributions for ν and ω for SCO problem (1) but only have access to a set of training data $S = S_\nu \cup S_\omega$ where both $S_\nu = \{\nu_i : i = 1,\dots,n\}$ and $S_\omega = \{\omega_j : j = 1,\dots m\}$ are *distributed independently and identically (i.i.d.)*. As such, SCO problem (1) is reduced to the following nested empirical risk for SCO:

$$\min_{x\in\mathcal{X}}\Big\{F_S(x) := f_S(g_S(x)) = \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}\big(\frac{1}{m}\sum_{j=1}^{m} g_{\omega_j}(x)\big)\Big\}, \tag{2}$$

where $g_S : \mathbb{R}^p \to \mathbb{R}^d$ and $f_S : \mathbb{R}^d \to \mathbb{R}$ are the empirical versions of f and g in (1) and are defined, respectively, by $g_S(x) = \frac{1}{m}\sum_{j=1}^{m} g_{\omega_j}(x)$ and $f_S(y) = \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(y)$. We refer to $F(x)$ and $F_S(x)$ as the *(nested) true risk and empirical risk*, respectively, in this stochastic compositional setting.

Denote the least (nested) true and empirical risks, respectively, by $F(x_*) = \inf_{x\in\mathcal{X}} F(x)$ and $F(x_*^S) = \inf_{x\in\mathcal{X}} F_S(x)$. For a randomized algorithm A, denote by $A(S)$ its output model based on the training data S. Then, our ultimate goal is to analyze the *excess generalization error (i.e., excess risk)* of $A(S)$ which is given by $F(A(S)) - F(x_*)$. It can be decomposed as follows:

$$\begin{aligned}&\mathbb{E}_{S,A}[F(A(S)) - F(x_*)]\\ &= \mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] + \mathbb{E}_{S,A}[F_S(A(S)) - F_S(x_*)]\\ &\le \mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))]\\ &\quad + \mathbb{E}_{S,A}[F_S(A(S)) - F_S(x_*^S)],\end{aligned} \tag{3}$$

where we have used the fact that $F_S(x_*^S) \le F_S(x_*)$ by the definition of x_*^S. The first term on the right hand side of

(3) is called the *generalization (error) gap* (i.e., estimation error) and the second term is the optimization error. The optimization error (convergence analysis) in our study builds upon the analysis conducted in previous works such as Wang et al. (2017); Chen et al. (2021a). However, our main focus is on estimating the generalization gap using the algorithmic stability approach (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Lei & Ying, 2020). In order to achieve this, we introduce a proper definition of stability in the compositional setting, which will be outlined below.

2.2. Uniform Stability for SCO

Existing work of stability analysis (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Lei & Ying, 2020) focused on SGD algorithms in the non-compositional ERM setting. We will extend the algorithmic stability analysis to estimate the estimation error (i.e., generalization gap) for SCO problems.

In our new setting, when we consider neighboring training data sets differing in one single data point, the change of one data point can happen in either S_ν or S_ω. In particular, for any $i \in [1, n]$ and $j \in [1, m]$, let $S^{i,\nu}$ be the neighboring set of S where only i-th data point ν_i in S_ν is changed to ν_i' while S_ω remains the same. Likewise, denote by $S^{j,\omega}$ be the neighboring set of S where only j-th data point ω_ℓ in S_ω is changed to ω_j' while S_ν remains unchanged. Throughout the paper, we also denote by $S' = S_\nu' \cup S_\omega'$ the i.i.d. copy of S where $S_\nu' = \{\nu_1', \ldots, \nu_n'\}$ and $S_\omega' = \{\omega_1', \ldots, \omega_m'\}$.

Definition 2.1 (Compositional Uniform Stability)**.** We say that a randomized algorithm A is $(\epsilon_\nu, \epsilon_\omega)$-uniformly stable for SCO problem (1) if, any $i \in [1, n]$, $j \in [1, m]$, there holds

$$\mathbb{E}_A[\|A(S) - A(S^{i,\nu})\|] \leq \epsilon_\nu,$$
$$\text{and } \mathbb{E}_A[\|A(S) - A(S^{j,\omega})\|] \leq \epsilon_\omega, \tag{4}$$

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

We will show the relationship between the *compositional uniform stability* (i.e., Definition 2.1) and the generalization error (gap) which holds for any randomized algorithm. To do this, we need the following assumption.

Assumption 2.2. We assume that f_ν and g_ω are Lipschitz continuous with parameters L_f and L_g, respectively, i.e.,

(i) $\sup_\nu \|f_\nu(y) - f_\nu(\hat{y})\| \leq L_f\|y - \hat{y}\|$ for all $y, \hat{y} \in \mathbb{R}^d$.

(ii) $\sup_\omega \|g_\omega(x) - g_\omega(\hat{x})\| \leq L_g\|x - \hat{x}\|$ for all $x, \hat{x} \in \mathbb{R}^p$.

The Lipschitz continuous assumption on f_ν and g_ω is widely used in existing work (Wang et al., 2017; Chen et al., 2021a; Jiang et al., 2022b; Wang & Yang, 2022) on optimization error analysis. It is also imposed in Hu et al. (2020) on the generalization analysis of the exact risk minimizer of SCO.

The following theorem establishes the relationship between the stability of SCGD and its generalization.

Theorem 2.3. *If Assumption 2.2 is true and the randomized algorithm A is ϵ-uniformly stable then*

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

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

Remark 2.4. Theorem 1 describes the relationship between the compositional uniform stability and generalization (gap) for any randomized algorithm for SCO problems. It can be regarded as an extension of the counterpart for the non-compositional setting (Hardt et al., 2016). Indeed, if we let $g_\omega(x) = x$, then $g_S(x) = g_\omega(x) = x$ for any ω and S, the SCO problem is reduced to the standard non-compositional setting, i.e., $F(x) = \mathbb{E}_\nu[f_\nu(x)]$ and $F_S(x) = \frac{1}{n}\sum_{i=1}^n f_{\nu_i}(x)$. In this case, our result in Theorem 1 indicates, since there is no randomness w.r.t. ω, that $\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq L_f\epsilon_\nu$ which is exactly the case in the non-compositional setting (Hardt et al., 2016).

Remark 2.5. There are major technical challenges to deriving the relation between stability and generalization for SCO algorithms. First, it is not obvious to relate the generalization error to the compositional uniform stability as we defined. The decoupling between inner random variables and outer random variables is a key that allows us to conduct the analysis. Second, recall that, in the classical (noncompositional) setting, given the i.i.d. data $S = \{z_1, \ldots, z_n\}$, the empirical and population risks are given by $F_S(A(S)) = \frac{1}{n}\sum_{i=1}^n f(A(S); z_i)$ and $F(A(S)) = \mathbb{E}_z[f(A(S); z]$, respectively. Let $S^i = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$ be the neighboring set of S, but differ in the i-th data point. Using the symmetry between the i.i.d. datasets $S = \{z_1, \ldots, z_n\}$ and $S' = \{z_1', z_2', \ldots, z_n'\}$, one can observe that

$$\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] = \mathbb{E}_{S,A,S'}\big[\frac{1}{n}\sum_{i=1}^n f(A(S^i); z_i) - \frac{1}{n}\sum_{i=1}^n f(A(S); z_i)\big] \leq L_f\|A(S^i) - A(S)\|.$$

However, due to the compositional structure in our setting, one can see that

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] = \mathbb{E}_{S,A}\big[\mathbb{E}_\nu[f_\nu\big(g(A(S))\big)] - \frac{1}{n}\sum_{i=1}^n f_{\nu_i}\big(g(A(S))\big)\big] + \mathbb{E}_{S,A}\big[\frac{1}{n}\sum_{i=1}^n \big(f_{\nu_i}\big(g(A(S))\big) - f_{\nu_i}\big(\frac{1}{m}\sum_{j=1}^m g_{\omega_j}(A(S))\big)\big)\big].$$

Algorithm 1 *(Stochastically Corrected) Stochastic Compositional Gradient Descent*

1: **Inputs:** Training data $S_\nu = \{\nu_i : i = 1,\dots,n\}$, $S_\omega = \{\omega_j : j = 1,\dots,m\}$; Number of iterations T, parameters $\{\eta_t\}, \{\beta_t\}$
2: Initialize $x_0 \in \mathcal{X}$ and $y_0 \in \mathbb{R}^d$
3: **for** $t = 0$ to $T-1$ **do**
4: Randomly sample $j_t \in [1, m]$, obtain $g_{\omega_{j_t}}(x_t)$ and $\nabla g_{\omega_{j_t}}(x_t) \in \mathbb{R}^{p\times d}$
5: **SCGD update**: $y_{t+1} = (1-\beta_t)y_t + \beta_t g_{\omega_{j_t}}(x_t)$
6: **SCSC update**: $y_{t+1} = (1-\beta_t)y_t + \beta_t g_{\omega_{j_t}}(x_t) + (1-\beta_t)(g_{\omega_{j_t}}(x_t) - g_{\omega_{j_t}}(x_{t-1}))$
7: Randomly sample $i_t \in [1, n]$, obtain $\nabla f_{\nu_{i_t}}(y_{t+1}) \in \mathbb{R}^d$
8: **Update:**
9: $x_{t+1} = \Pi_{\mathcal{X}}\big(x_t - \eta_t \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\big)$
10: **end for**
11: **Outputs:** $A(S) = x_T$ or $x_\tau \sim \texttt{Unif}(\{x_t\}_{t=1}^T)$

The first term on the right-hand side of the above equality can be handled similarly as the non-compositional setting. The main challenge comes from the second term which, by the Lipschitz property of f_ν, involves a vector-valued generalization $\mathbb{E}_{S,A}\big[\big\|g(A(S)) - \frac{1}{m}\sum_{j=1}^m g_{\omega_j}(A(S))\big\|\big]$ because one can not interchange the expectation and the norm. We will overcome this obstacle using the sample-splitting argument (Bousquet et al., 2020; Lei, 2022).

2.3. SCO Optimization Algorithms

We will study two popular optimization algorithms for solving (2), i.e., SCGD (Wang et al., 2017) and SCSC (Chen et al., 2021a). Their pseudo-code is given in Algorithm 1. For SCGD, the updating sequence in Line 5 of Algorithm 1

$$y_{t+1} = (1-\beta_t)\, y_t + \beta_t g_{\omega_{j_t}}(x_t)$$

is used to track the expectation of $g_S(x_t) = \mathbb{E}_{j_t}[g_{\omega_{j_t}}(x_t)] = \frac{1}{m}\sum_{j=1}^m g_{\omega_j}(x_t)$. As shown in Wang et al. (2017), SCGD needs to choose a smaller stepsize η_t than the stepsize β_t to be convergent. This prevents the SCGD from choosing the same stepsize as SGD for the non-compositional stochastic problems. To address this issue, Chen et al. (2021a) proposed a stochastically corrected version of SCGD which is referred to as SCSC. In particular, the sequence y_{t+1} is given in Line 6 of Algorithm 1:

$$y_{t+1} = (1-\beta_t)(y_t + g_{\omega_{j_t}}(x_t) - g_{\omega_{j_t}}(x_{t-1})) + \beta_t g_{\omega_{j_t}}(x_t).$$

With y_{t+1} as an approximator for $g(x_t)$, the model parameter x is updated using the stochastic gradient descent step as given in Line 9 of Algorithm 1.

Below we list definitions about strong convexity and smoothness which will be used in subsequent sections.

Definition 2.6. A function $F : \mathbb{R}^p \to \mathbb{R}$ is σ-strongly convex with some $\sigma \geq 0$ if, for any $u, v \in \mathbb{R}^p$, we have $F(u) \geq F(v) + \langle \nabla F(v), u - v\rangle + \frac{\sigma}{2}\|u - v\|^2$. If $\sigma = 0$, we say that F is convex.

Definition 2.7. A function $F : \mathbb{R}^p \to \mathbb{R}$ is L-smooth if, for any $u, v \in \mathbb{R}^p$, we have $\|\nabla F(u) - \nabla F(v)\| \leq L\|u - v\|$.

In general, smoothness implies the gradient update of F cannot be overly expansive. Also the convexity and L-smooth of F implies that the gradients are co-coercive, hence we have

$$\langle \nabla F(u) - \nabla F(v), u - v\rangle \geq \frac{1}{L}\|\nabla F(u) - \nabla F(v)\|^2. \tag{5}$$

Note that if F is σ strongly convex, then $\varphi(x) = F(x) - \frac{\sigma}{2}\|x\|^2$ is convex with $(L-\sigma)$-smooth. Then, applying (5) to φ yields the following inequality:

$$\langle \nabla F(u) - \nabla F(v), u - v\rangle \geq \frac{L\sigma}{L+\sigma}\|u - v\|^2 + \frac{1}{L+\sigma}\|\nabla F(u) - \nabla F(v)\|^2. \tag{6}$$

3. Stability and Generalization

In this section, we will present our main results on estimating the stability bounds for SCGD and SCSC which subsequently can lead to estimation of their generalization gaps from Theorem 2.3. Then, we start from the error decomposition (3) to derive the bounds for their excess risks by trade-offing the bounds for the above generalization (error) gaps and optimization errors. We will present results in two different cases, i.e., convex and strongly convex settings, in different subsections. For brevity, we summarize our results for the excess risks for SCGD and SCSC in Table 1. Before illustrating our main results, we list some assumptions.

Assumption 3.1. We assume that the following conditions hold.

(i) With probability 1 w.r.t S, there holds $\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^m \big[\|g_{\omega_j}(x) - g_S(x)\|^2\big] \leq V_g$.

(ii) With probability 1 w.r.t. S, $\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^m [\|\nabla g_{\omega_j}(x) - \nabla g_S(x)\|^2] \leq C_g$.

(iii) With probability 1 w.r.t. ν, $f_\nu(\cdot)$ has Lipschitz continuous gradients, i.e. $\|\nabla f_\nu(y) - \nabla f_\nu(\bar{y})\| \leq C_f\|y - \bar{y}\|$ for all $y, \bar{y} \in \mathbb{R}^d$.

(iv) With probability 1 w.r.t. ν and S, $f_\nu(g_S(\cdot))$ is L-smooth, i.e., $\|\nabla g_S(x)\nabla f_\nu(g_S(x)) - \nabla g_S(x')\nabla f_\nu(g_S(x'))\| \leq L\|x - x'\|$ for any $x, x' \in \mathcal{X}$.

Remark 3.2. There are many practical applications aligning with our SCO problem and meeting Assumptions 2.2 and

3.1. In particular, for the distributionally robust optimization (Qi et al., 2021a) mentioned in Section 2.1, the loss $\ell(\cdot, b)$ being convex and smooth for any b. Since exp and log are non-decreasing functions, from the fact that the composition of a nondecreasing function and a convex function is convex, we conclude that $F_S(\cdot)$ is convex, a case that we focus mainly on in our paper. If $\mathcal{X}$ is a bounded domain, Lipschitz continuity and smoothness conditions hold true. The smoothness asusmptions are standard in the literature of stochastic compositional optimization, e.g., Wang et al. (2017); Chen et al. (2021a).

3.1. Convex Setting

In this subsection, we present our main results for SCGD and SCSC in the convex setting.

Stability Results. The following theorem establishes the *compositional uniform Stability* (See Definition 2.1) for SCGD and SCSC in the convex setting.

Theorem 3.3 (Stability, Convex)**.** *Suppose that Assumption 2.2 and 3.1 hold true and $f_\nu(g_S(\cdot))$ is convex. Consider Algorithm 1 with $\eta_t = \eta \le \frac{1}{2L}$, and $\beta_t = \beta \in (0,1)$ for any $t \in [0, T-1]$. Then, the outputs $A(S) = x_T$ of both SCGD and SCSC at iteration T are compositionally uniform stable with*

$$\epsilon_\nu + \epsilon_\omega = \mathcal{O}\Big(\frac{L_f L_g}{n}\eta T + \frac{L_f L_g}{m}\eta T + \sqrt{C_g} L_f \eta\sqrt{T} + C_f L_g \sup_S \sum_{j=0}^{T-1} \eta\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}\Big). \quad (7)$$

The proof for Theorem 3.3 is deferred to Appendix C.1.

Remark 3.4. In this remark, we discuss how the function composition plays a role in the stability analysis for SCGD and SCSC and then compare our results with that for SGD in the non-compositional setting (Hardt et al., 2016). To this end, considering the step sizes $\eta_t = \eta$ and $n = m$, then (7) is reduced to the following estimation:

$$\epsilon_\nu + \epsilon_\omega = \mathcal{O}\Big(\frac{\eta T}{n} + \sqrt{C_g}\eta\sqrt{T} + \eta \sup_S \sum_{j=0}^{T-1} \big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}\Big). \quad (8)$$

It was shown in Hardt et al. (2016) that the uniform stability for SGD with convex and smooth losses is of the order $\mathcal{O}\big(\frac{\eta T}{n}\big)$. By comparing these two results, we can see how the compositional structure plays a role in the stability analysis. Indeed, in contrast to the result for SGD, there are two extra terms in (8) for SCGD and SCSC, i.e., $\sqrt{C_g}\eta\sqrt{T}$ and $\eta \sup_S \sum_{j=0}^{T-1} \big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}$. Here, C_g is the (empirical) variance of the gradient of inner function, i.e., $\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^m \|\nabla g_{\omega_j}(x) - \nabla g_S(x)\|^2 \le C_g$ given in Assumption 3.1 and the other extra term arises when the moving-average sequence y_{t+1} is used to track $g_S(x_t)$. More importantly, our stability results indicate that, in order to boost generalization, it's vital to decrease the variance bound C_g (e.g., via larger minibatches for ω) and ensuring fast convergence of $\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]$. Notice that, if we let $g_\omega(x) = x$, then $g_S(x) = g_\omega(x) = x$ for any ω and S, then SCGD and SCSC reduce to the classical SGD, and our stability result (8) is the same as that of SGD since two extra terms mentioned above will be all zeros due to the fact that $y_{j+1} = g_S(x_j) = x_j$ and $C_g = 0$ in this case.

Combining (7) with the estimation for $\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]$ (Wang et al., 2017; Chen et al., 2021a) (see also Lemma A.1 and its self-contained proof in Appendix A), one can get the following explicit stability results.

Corollary 3.5. *Let Assumption 2.2 and 3.1 hold true and $f_\nu(g_S(\cdot))$ be convex. Consider Algorithm 1 with $\eta_t = \eta \le \frac{1}{2L}$, and $\beta_t = \beta \in (0,1)$ for any $t \in [0, T-1]$ and the output $A(S) = x_T$. Let c be an arbitrary constant. Then, we have the following results:*

- *SCGD is compositional uniformly stable with*

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

- *SCSC is compositional uniformly stable with*

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

Generalization results. Using the error decomposition (3), Corollary 3.5 and Theorem 2.3, we can derive the excess risk rates. To this end, we need the following results to estimate the optimization error, i.e., $F_S(A(S)) - F_S(x_*^S)$.

Theorem 3.6 (Optimization, Convex)**.** *Suppose Assumption 2.2 and 3.1 (i), (iii) hold for the empirical risk F_S and F_S is convex, $\mathbb{E}_A\|x_t - x_*^S\|^2$ is bounded by D_x for all $t \in [0, T-1]$ and $\mathbb{E}_A\|y_1 - g_S(x_0)\|^2$ is bounded by D_y. Let $A(S) = \frac{1}{T}\sum_{t=1}^T x_t$ be the solution produced by Algorithm 1 with SCGD or SCSC update, $\eta_t = \eta$ and $\beta_t = \beta$ for some $a, b \in (0,1]$. Let c be an arbitrary constant.*

- *For SCGD update, there holds*

$$\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)] = \mathcal{O}\Big(D_x(\eta T)^{-1} + L_f^2 L_g^2\eta + C_f D_y(\beta T)^{1-c}(\eta T)^{-1} + C_f V_g\beta^2\eta^{-1} + C_f L_f^2 L_g^3 D_x\eta\beta^{-1}\Big).$$

- *For SCSC update, there holds*

$$\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)] = \mathcal{O}\Big(D_x(\eta T)^{-1} + L_f^2 L_g^2\eta + C_f D_y(\beta T)^{-c}\beta^{-\frac{1}{2}} + C_f V_g\beta^{\frac{1}{2}} + C_f L_f^2 L_g^3\eta^2\beta^{-\frac{3}{2}} + C_f L_g^2 D_x\beta^{\frac{1}{2}}\Big).$$

Table 1. Number of Iterations T that Achieves Excess Risk for SCGD And SCSC Algorithm

setting \ algorithm		SCGD	SCSC
Convex F_S	# Iterations	$T \asymp \max(n^{3.5}, m^{3.5})$	$T \asymp \max(n^{2.5}, m^{2.5})$
	Excess risk	$\mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$
Strongly Convex F_S	# Iterations	$T \asymp \max(n^{10/3}, m^{10/3})$	$T \asymp \max(n^{7/3}, m^{7/3})$
	Excess risk	$\mathcal{O}\left(\frac{1}{n} + \frac{1}{\sqrt{m}}\right)$	$\mathcal{O}\left(\frac{1}{n} + \frac{1}{\sqrt{m}}\right)$

The boundedness assumptions are satisfied if the domain $\mathcal{X}$ is bounded in $\mathbb{R}^p$. The detailed proofs are given in Appendix C.2 and C.3. Note that the upper-bounds for the optimization error given in the above theorem hold true uniformly for any training data S.

Combining the above results with the stability bounds in Corollary 3.5 and Theorem 2.3, we can derive the following excess risk bounds for SCGD and SCSC.

Theorem 3.7 (Excess Risk Bound, Convex)**.** *Suppose Assumptions 2.2 and 3.1 hold true and $f_\nu(g_S(\cdot))$ is convex, $\mathbb{E}_A[\|x_t - x_*^S\|^2]$ is bounded by D_x for all $t \in [0, T-1]$ and $\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2]$ is bounded by D_y. Let $A(S) = \frac{1}{T}\sum_{t=1}^T x_t$ be a solution produced by Algorithm 1 with SCGD or SCSC update and $\eta = T^{-a}$ and $\beta = T^{-b}$ for some $a, b \in (0, 1]$.*

- *If we select $T \asymp \max(n^{3.5}, m^{3.5})$, $\eta = T^{-\frac{6}{7}}$ and $\beta = T^{-\frac{4}{7}}$, then, for the SCGD update, we have that $\mathbb{E}_{S,A}\big[F(A(S)) - F(x_*)\big] = \mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\big)$.*
- *If we select $T \asymp \max(n^{2.5}, m^{2.5})$, $\eta = T^{-\frac{4}{5}}$ and $\beta = T^{-\frac{4}{5}}$, then, for the SCSC update, there holds $\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}})$.*

Remark 3.8. In the appealing work (Hu et al., 2020), the uniform convergence using concentration inequalities and the number of coverage is used to study the generalization gap (estimation error) of the ERM minimizer related to the SCO problems. Applying their results to our case, they proved the following results: assuming that $\mathcal{X}$ is a bounded domain, f_ν and g_ω are both Lipschitz continuous and bounded, there holds, with high probability,

$$F(A(S)) - F_S(A(S)) \leq \sup_{x\in\mathcal{X}} |F(x) - F_S(x)| = \mathcal{O}\Big(\sqrt{\frac{p}{m+n}}\Big)$$

which is highly dependent on the dimension of the domain $\mathcal{X} \subseteq \mathbb{R}^p$. Comparing with their bounds, we can get excess risk bounds which is *dimension independent* of the optimization domain $\mathcal{X} \subseteq \mathbb{R}^p$. Dimension-independent generalization bounds were also provided in Hu et al. (2020) which requires the Hölder error bound condition (e.g., strong convexity). The proof there depends heavily on the property of the ERM minimizer of the SCO problem and does not apply to SCGD and SCSC.

Remark 3.9. Theorem 3.7 shows that the generalization error for SCGD can be achieved by the rate $\mathcal{O}\big(1/\sqrt{n} + 1/\sqrt{m}\big)$ in the convex case after appropriately selecting the iteration number T and step sizes η and β. Recall that in the noncompositional setting, Hardt et al. (2016); Lei & Ying (2020) established generalization error bounds $\mathcal{O}\big(1/\sqrt{n}\big)$ by choosing $T \asymp n$ for SGD in the convex and smooth case. To achieve a similar rate, our results indicate that SCGD and SCSC need more iterations to do that. The reason may be due to the usage of the moving average sequence y_{t+1} to track $g_S(x_t)$ and the (empirical) variance term for the inner function g_ω as mentioned in Remark 3.4.

Remark 3.10. Note that in Theorem 3.3 we present the stability result of the last iterate $A(S) = x_T$. While in Theorem 3.7 we present the generalization bound of $A(S) = \frac{1}{T}\sum_{t=1}^T x_t$, which is the average of the intermediate iterates $x_1, \ldots, x_T$. This stems from the fact that generalization is a combination of stability and optimization, and the main focus of optimization is the average of intermediate iterates in the convex setting (see e.g. Wang et al. (2017)).

3.2. Strongly Convex Setting

Stability Results. The following theorem establishes the *compositional uniform Stability* (See Definition 2.1) for SCGD and SCSC in the strongly convex setting.

Theorem 3.11 (Stability, Strongly Convex)**.** *Suppose that Assumption 2.2 and 3.1 hold true and $f_\nu(g_S(\cdot))$ is σ-strongly convex. Consider Algorithm 1 with $\eta_t = \eta \leq 1/\big(2L + 2\sigma\big)$ and $\beta_t = \beta \in (0, 1)$ for $t \in [0, T-1]$ and the output $A(S) = x_T$. Then, SCGD and SCSC are compositional uniform stable with*

$$\begin{aligned}
&\epsilon_\nu + \epsilon_\omega \\
&= \mathcal{O}\Big(\frac{L_g L_f(L+\sigma)}{\sigma L m} + \frac{L_g L_f(L+\sigma)}{\sigma L n} + \frac{L_f\sqrt{C_g(L+\sigma)\eta}}{\sqrt{\sigma L}} \\
&+ C_f L_g \eta \sup_S\Big\{\sum_{j=0}^{T-1}(1 - \eta\frac{L\sigma}{L+\sigma})^{T-j-1} \\
&\qquad \times \big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}\Big)\Big\}. \qquad (9)
\end{aligned}$$

The proof for Theorem 3.11 is given in Appendix D.1.

Remark 3.12. The stability for SGD with σ-strongly convex and smooth losses is of the order $\mathcal{O}(\frac{1}{\sigma n})$ which was established in Hardt et al. (2016). Comparing the result of SGD with our SCGD and SCSC, we have two extra terms if $n = m$, i.e., $\eta \sup_S \sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}$ and $\frac{L_f\sqrt{C_g(L+\sigma)\eta}}{\sqrt{\sigma L}}$, where C_g is the (empirical) variance of the gradient of inner function, i.e.$\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^m \|\nabla g_{\omega_j}(x) - \nabla g_S(x)\|^2 \leq C_g$. We can see that if $g_\omega(x) = x$, then $g_S(x) = g_\omega(x)$ for any ω and S. In this case, $\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]$ and C_g will be zeros. Therefore, our stability results in Theorem 3.11 match that of SGD in the non-compositional setting (Hardt et al., 2016).

Combining Theorem 3.11 with the estimation for $\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]$ in Lemma A.1 and using the Lemma A.4 which is given in Appendix A, we can derive the explicit stability bounds in the following corollary. Its detailed proof is given at the end of Section D.1 in the appendix.

Corollary 3.13. *Let Assumption 2.2 and 3.1 hold true and $f_\nu(g_S(\cdot))$ be σ-strongly convex. Consider Algorithm 1 with $\eta_t = \eta \leq 1/\big(2L + 2\sigma\big)$ and $\beta_t = \beta \in (0, 1)$ for $t \in [0, T-1]$ and the output $A(S) = x_T$. Let c be an arbitrary constant. Then, we have the following results:*

- *SCGD is compositional uniformly stable with*

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

- *SCSC is compositional uniformly stable with*

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

Generalization results. Using the error decomposition (3), Corollary 3.13 and Theorem 3.11, we can derive the excess risk rates. To this end, we need the following results to estimate the optimization error, i.e., $F_S(A(S)) - F_S(x_*^S)$.

Theorem 3.14 (Optimization, Strongly Convex)**.** *Suppose Assumption 2.2 and 3.1 (i), (iii) hold for the empirical risk F_S, and F_S is σ-strongly convex, and η, T is chosen such that $(\eta(T-1))^{-1} \leq \frac{\sigma}{2}$. Let $A(S) = \Big(\sum_{t=1}^T (1-\sigma\eta/2)^{T-t}x_t\Big) / \Big(\sum_{t=1}^T (1-\sigma\eta/2)^{T-t}\Big)$ be the solution produced by Algorithm 1 with SCGD or SCSC update and $\eta_t = \eta$ and $\beta_t = \beta$ for some $a, b \in (0, 1]$.*

- *For SCGD update, there holds*

$$\begin{aligned}&\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)]\\ &= \mathcal{O}\big(D_x(\eta T)^{-c} + L_f^2 L_g^2\eta + \frac{C_f^2 L_g^2 D_y}{\sigma}(\beta T)^{-c}\\ &\qquad + \frac{C_f^2 L_g^2 V_g}{\sigma}\beta + \frac{C_f^2 L_f^2 L_g^5}{\sigma}\eta^2\beta^{-2}\big).\end{aligned}$$

- *For SCSC update, there holds*

$$\begin{aligned}&\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)]\\ &= \mathcal{O}\big(D_x(\eta T)^{-c} + L_f^2 L_g^2\eta + \frac{C_f^2 L_g^2 D_y}{\sigma}(\beta T)^{-c}\\ &\qquad + \frac{C_f^2 L_g^2 V_g}{\sigma}\beta + \frac{C_f^2 L_f^2 L_g^5}{\sigma}\eta^2\beta^{-1}\big).\end{aligned}$$

Theorem 3.15 (Excess Risk Bound, Strongly Convex)**.** *Suppose Assumption 2.2 and 3.1 hold true, $f_\nu(g_S(\cdot))$ is σ-strongly convex, and η, T is chosen such that $(\eta(T-1))^{-1} \leq \frac{\sigma}{2}$. Denote $D_x := \mathbb{E}_A[F_S(x_0) - F_S(x_*^S)]$ and $D_y := \mathbb{E}_A[\|y_1 - g_S(x_0)\|^2]$. Let $A(S) = \Big(\sum_{t=1}^T (1-\sigma\eta/2)^{T-t}x_t\Big) / \Big(\sum_{t=1}^T (1-\sigma\eta/2)^{T-t}\Big)$ be a solution produced by Algorithm 1 with SCGD or SCSC update and $\eta = T^{-a}$ and $\beta = T^{-b}$ for some $a, b \in (0, 1]$.*

- *If we select $T \asymp \max(n^{\frac{10}{3}}, m^{\frac{10}{3}})$, $\eta = T^{-\frac{9}{10}}$ and $\beta = T^{-\frac{3}{5}}$, then, for the SCGD update, we have that $\mathbb{E}_{S,A}\big[F(A(S)) - F(x_*)\big] = \mathcal{O}\big(\frac{1}{n} + \frac{1}{\sqrt{m}}\big)$.*

- *If we select $T \asymp \max(n^{\frac{7}{3}}, m^{\frac{7}{3}})$, $\eta = \beta = T^{-\frac{6}{7}}$, then, for the SCSC update, there holds $\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}(\frac{1}{n} + \frac{1}{\sqrt{m}})$.*

Remark 3.16. Theorem 3.15 shows that the generalization error for SCGD can be achieved the rate $\mathcal{O}(1/n + 1/\sqrt{m})$ in the strongly convex case after carefully selecting the iteration number T and constant stepsize η and β. It is worthy of noting that, for achieving the rate $\mathcal{O}(1/n + 1/\sqrt{m})$, SCGD needs iteration $T \asymp \max(n^{10/3}, m^{10/3})$ in the strongly convex case while Theorem 3.7 shows that it needs more iterations, i.e., $T \asymp \max(n^{3.5}, m^{3.5})$ in the convex case. SCSC further improves the results as it only needs iteration $T \asymp \max(n^{7/3}, m^{7/3})$ in the strongly convex case.

Notice that our work focus on the challenge for convex and strongly convex problems first. It would be difficult to analyze non-convex objective without imposing conditions about the objective since we aim to analyze both the generalization error and optimization error to derive the excess risk bound. Some future work on stability analysis in the nonconvex case may be done under some assumption of Hölderian error bound condition as in the appealing work of Hu et al. (2020) or in the setting of shallow neural network structure for the inner function $g_\omega(\cdot)$ (e.g. Richards & Kuzborskij (2021)) where the convexity is relaxed to the weak convexity with enough large width for the neural network or in the setting of neural tagent kernel regime (e.g. Jacot et al. (2018); Nitanda & Suzuki (2020)). Also as a starting point, we only focus on the smooth setting in this work. Existing results showed that it is possible to extend our analysis to the non-smooth setting e.g., Lei & Ying (2020) removed the smoothness assumption in the stability

and generalization analysis of SGD for convex objectives using the approximate non-expansiveness of the gradient mapping. Hu et al. (2020) studied the generalization gap (estimation error) with high probability in the non-smooth setting for the minimizers of the ERM problem.

4. Conclusion

In this paper, we conduct a comprehensive study on the stability and generalization analysis of stochastic compositional optimization (SCO) algorithms. We introduce the concept of compositional uniform stability to handle the function composition structure inherent in SCO problems. By establishing the connection between stability and generalization error, we provide stability bounds for two popular SCO algorithms: SCGD and SCSC. In the convex case with standard smooth assumptions, we demonstrate that both SCGD and SCSC achieve an excess generalization error rate of $\mathcal{O}(1/\sqrt{n} + 1/\sqrt{m})$, with SCSC requiring fewer iterations than SCGD. Furthermore, we extend our analysis to the strongly convex case, where we show that SCGD and SCSC achieve the same rate of $\mathcal{O}(1/\sqrt{n} + 1/\sqrt{m})$ with even fewer iterations than in the convex case.

There are several directions for future research. Firstly, while our analysis only considers the convex and smooth cases, an interesting avenue for future research is to consider the case where the inner function and/or outer function are non-smooth and non-convex, e.g., neural networks with Rectified Linear Unit (ReLU) activation function. Secondly, it would be interesting to get optimal excess risk rates $\mathcal{O}(1/\sqrt{n} + 1/\sqrt{m})$ with linear time complexity $T = \mathcal{O}\big(\max(n, m)\big)$ for SCGD and SCSC. Thirdly, an important and interesting direction is multi-level analysis. In this setting, the stability result could involve the term $\sum_{n=1}^{N-1} \|y_n^t - f_n(y_{n-1}^t)\|^2$ where N is the number of levels

Impact Statement

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

Acknowledgements

We thank all reviewers for their valuable comments and suggestions. The work was partially supported by NSF grants under DMS-2110836, IIS-2103450, IIS-2110546, NSF Career Award 1844403.

References

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

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pp. 169–207, 2004.

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

Charles, Z. and Papailiopoulos, D. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pp. 744–753, 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, 2021a.

Chen, T., Sun, Y., and Yin, W. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021b.

Dentcheva, D., Penev, S., and Ruszczyński, A. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69:737–760, 2017.

Devraj, A. M. and Chen, J. Stochastic variance reduced primal dual algorithms for empirical composition optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Elisseeff, A., Evgeniou, T., and Pontil, M. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

Farnia, F. and Ozdaglar, A. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279, 2019.

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

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, 2016.

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.

Hu, Y., Chen, X., and He, N. Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jiang, W., Li, G., Wang, Y., Zhang, L., and Yang, T. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. *Advances in Neural Information Processing Systems*, 35:32499–32511, 2022a.

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, 2022b.

Kar, P., Sriperumbudur, B., Jain, P., and Karnick, H. On the generalization ability of online learning algorithms for pairwise loss functions. In *International Conference on Machine Learning*, pp. 441–449, 2013.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2820–2829, 2018.

Lei, Y. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. *arXiv preprint arXiv:2206.07082*, 2022.

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819, 2020.

Lei, Y. and Ying, Y. Stochastic proximal auc maximization. *The Journal of Machine Learning Research*, 22(1):2832–2876, 2021.

Lei, Y., Yang, Z., Yang, T., and Ying, Y. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pp. 6175–6186, 2021.

Lei, Y., Jin, R., and Ying, Y. Stability and generalization analysis of gradient methods for shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Lian, X., Wang, M., and Liu, J. Finite-sum composition optimization via variance reduced gradient descent. In *Artificial Intelligence and Statistics*, pp. 1159–1167. PMLR, 2017.

Lin, T., Fan, C., Wang, M., and Jordan, M. I. Improved oracle complexity for stochastic compositional variance reduced gradient. *arXiv preprint arXiv:1806.00458*, 2018.

Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic AUC maximization with O(1/n)-convergence rate. In *International Conference on Machine Learning*, pp. 3195–3203, 2018.

Nitanda, A. and Suzuki, T. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv preprint arXiv:2006.12297*, 2020.

Qi, Q., Guo, Z., Xu, Y., Jin, R., and Yang, T. An online method for a class of distributionally robust optimization with non-convex objectives. *Advances in Neural Information Processing Systems*, 34:10067–10080, 2021a.

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, 2021b.

Richards, D. and Kuzborskij, I. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems*, 34, 2021.

Ruszczynski, A. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

Schmidt, M., Roux, N., and Bach, F. Convergence rates of inexact proximal-gradient methods for convex optimization. *Advances in neural information processing systems*, 24, 2011.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Shen, W., Yang, Z., Ying, Y., and Yuan, X. Stability and optimization error of stochastic gradient descent for pairwise learning. *Analysis and Applications*, pp. 1–41, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Szepesvári, C. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Tolstaya, E., Koppel, A., Stump, E., and Ribeiro, A. Nonparametric stochastic compositional gradient descent for q-learning in continuous markov decision problems. In *2018 Annual American Control Conference (ACC)*, pp. 6608–6615. IEEE, 2018.

Tutunov, R., Li, M., Cowen-Rivers, A. I., Wang, J., and Bou-Ammar, H. Compositional adam: An adaptive compositional solver. *arXiv preprint arXiv:2002.03755*, 2020.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Vapnik, V. *The nature of statistical learning theory*. Springer, 2013.

Wang, B. and Yang, T. Finite-sum compositional stochastic optimization: Theory and applications. *arXiv preprint arXiv:2202.12396*, 2022.

Wang, M., Liu, J., and Fang, E. Accelerating stochastic composition optimization. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Wang, P., Lei, Y., Ying, Y., and Zhou, D.-X. Stability and generalization for markov chain stochastic gradient methods. *arXiv preprint arXiv:2209.08005*, 2022.

Yang, T. Algorithmic foundation of deep x-risk optimization. *arXiv preprint arXiv:2206.00439*, 2022.

Yang, T. and Ying, Y. Auc maximization in the era of big data and ai: A survey. *ACM Computing Surveys*, 55(8), 2022.

Yang, Z., Lei, Y., Wang, P., Yang, T., and Ying, Y. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021.

Ying, Y., Wen, L., and Lyu, S. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.

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.

Zhang, Z. and Lan, G. Optimal algorithms for convex nested stochastic composite optimization. *Mathematical programming*, 2020.

Zhao, P., Hoi, S. C., Jin, R., and Yang, T. Online AUC maximization. In *International Conference on Machine Learning*, pp. 233–240. Omnipress, 2011.

Zhou, D.-X. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

A. Technical Lemmas

Table 2. Notations

notations	meaning	mathematical language
L_f	L_f-Lipschitz continuous of $f_\nu(\cdot)$	$\sup_\nu \|f_\nu(y) - f_\nu(\hat{y})\| \leq L_f\|y-\hat{y}\|, \forall y, \hat{y} \in \mathbb{R}$
C_f	C_f-Lipschitz continuous of $\nabla f_\nu(\cdot)$	$\sup_\nu \|\nabla f_\nu(y) - \nabla f_\nu(\bar{y})\| \leq C_f\|y-\bar{y}\|, \forall y, \bar{y} \in \mathbb{R}$
L_g	L_g-Lipschitz continuous of $g_\omega(\cdot)$	$\sup_\omega \|g_\omega(x) - g_\omega(\hat{x})\| \leq L_g \|x-\hat{x}\|, \forall x, \hat{x} \in \mathcal{X}$
V_g	the empirical variance of the $g(\cdot)$	$\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^m \|g_{\omega_j}(x) - g_S(x)\|^2 \leq V_g$
C_g	the empirical variance of the $\nabla g(\cdot)$	$\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{j=1}^m \|\nabla g_{\omega_j}(x) - \nabla g_S(x)\|^2 \leq C_g$
L	L smooth of $f_\nu(g_S(\cdot))$	$\|g_S(u)\nabla f_\nu(g_S(u)) - g_S(v)\nabla f_\nu(g_S(v))\| \leq L\|u-v\|$
$\epsilon_\nu, \epsilon_\omega$	$(\epsilon_\nu, \epsilon_\omega)$-uniform stability	
n, m	n, m: the numbers of S_ν and S_ω, respectively	

First, we list some signal notations in Table 2 for our paper setting. To derive the stability and generalization bounds, we give the following lemmas.

The following lemma is directly adapted from Wang et al. (2017); Chen et al. (2021a) where both the population distribution for the random variables ν and ω are the uniform distributions over $S_\nu = \{\nu_1, \ldots, \nu_n\}$ and $S_\omega = \{\omega_1, \ldots, \omega_m\}$. It states that y_{t+1} behaves similarly to $g_S(x_t)$

Lemma A.1. *Let Assumption 2.2 and 3.1 (i) hold and (x_t, y_t) be generated by Algorithm 1. Let $\eta_t = \eta$, and $\beta_t = \beta$ for $\eta, \beta > 0$. Let $c > 0$ be an arbitrary constant.*

- *With SCGD update, we have*

$$\mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\|^2\big] \leq \Big(\frac{c}{e}\Big)^c (t\beta)^{-c}\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2] + L_f^2 L_g^3 \frac{\eta^2}{\beta^2} + 2V_g\beta.$$

- *With SCSC update we have*

$$\mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\|^2\big] \leq \Big(\frac{c}{e}\Big)^c (t\beta)^{-c}\mathbb{E}_A\big[\|y_1 - g_S(x_0)\|^2\big] + L_f^2 L_g^3 \frac{\eta^2}{\beta} + 2V_g\beta.$$

The next lemma was established in Schmidt et al. (2011) and this lemma was used in Wang et al. (2022).

Lemma A.2. *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:

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

Lemma A.3. *For any $\nu, c > 0$, we have*

$$e^{-\nu x} \leq \big(\frac{c}{\nu e}\big)^c x^{-c} \tag{10}$$

Lemma A.4. *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*

$$\frac{\sum_{i=1}^T a_i b_i}{\sum_{i=1}^T a_i} \leq \frac{\sum_{i=1}^T b_i}{T}. \tag{11}$$

Proof. To show (11), it suffices to show

$$\sum_{i=1}^T a_i b_i \sum_{j=1}^T 1 \leq \sum_{j=1}^T a_j \sum_{i=1}^T b_i.$$

Rearranging the summation, it suffices to show

$$\sum_{i=1}^{T}\sum_{j=1}^{T} a_i b_i - \sum_{i=1}^{T}\sum_{j=1}^{T} a_j b_i \leq 0.$$

The above inequality can be rewritten as

$$0 \geq \sum_{i=1}^{T}\sum_{j=1}^{T}(a_i - a_j)b_i = \sum_{i=1}^{T}\sum_{j=i+1}^{T}(a_i - a_j)(b_i - b_j),$$

where the last equality holds due to the symmetry between i and j. Since for $i < j$ we have $a_i \leq a_j$ and $b_i \geq b_j$, we know the above inequality holds, and thus (11) holds. Then we complete the proof. □

A.1. Proof of Lemma A.1

The proof of Lemma A.1 leverages the following results.

Lemma A.5 (Lemma 2 in Wang et al. (2017)). *Suppose Assumption 2.2 (ii) and 3.1 (i) hold for the empirical risk F_S. By running Algorithm 1 with SCGD update, we have*

$$\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2|\mathcal{F}_t] \leq (1-\beta_t)\|y_t - g_S(x_{t-1})\|^2 + \frac{L_g^2}{\beta_t}\|x_t - x_{t-1}\|^2 + 2V_g\beta_t^2 \tag{12}$$

Lemma A.6 (Lemma 1 in Chen et al. (2021a)). *Suppose Assumption 2.2 (ii) and 3.1 (i) hold for the empirical risk F_S. By running Algorithm 1 with SCSC update, we have*

$$\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2|\mathcal{F}_t] \leq (1-\beta_t)\|y_t - g_S(x_{t-1})\|^2 + L_g^2\|x_t - x_{t-1}\|^2 + 2V_g\beta_t^2 \tag{13}$$

Now we are ready to prove Lemma A.1.

Proof of Lemma A.1. We first present the proof for the SCGD update. Taking the expectation with respect to the internal randomness of the algorithm over (12) and noting that $\mathbb{E}_A[\|x_t - x_{t-1}\|^2] \leq L_f^2 L_g^2 \eta_{t-1}^2$, we get

$$\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2] \leq (1-\beta_t)\mathbb{E}_A[\|y_t - g_S(x_{t-1})\|^2] + \frac{L_f^2 L_g^3 \eta_{t-1}^2}{\beta_t} + 2V_g\beta_t^2.$$

Telescoping the above inequality from 1 to t yields

$$\begin{aligned}&\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2]\\ \leq&\prod_{i=1}^{t}(1-\beta_i)\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2] + L_f^2 L_g^3\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\beta_j)\frac{\eta_{i-1}^2}{\beta_i} + 2V_g\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\beta_j)\beta_i^2.\end{aligned}$$

Note that $\prod_{i=K}^{N}(1-\beta_i) \leq \exp(-\sum_{i=K}^{N}\beta_i)$ for all $K \leq N$ and $\beta_i > 0$, then setting $\eta_t = \eta, \beta_t = \beta$, thus we have

$$\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2] \leq \exp(-\beta t)\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2] + \sum_{i=1}^{t}(1-\beta)^{t-i}(L_g^3 L_f^2\frac{\eta^2}{\beta} + 2V_g\beta^2).$$

Using Lemma A.3 with $\nu = 1$, we get

$$\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2] \leq \left(\frac{c}{e}\right)^c(t\beta)^{-c}\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2] + L_g^3 L_f^2\frac{\eta^2}{\beta^2} + 2V_g\beta,$$

where the inequality holds for $\sum_{i=1}^{t}(1-\beta)^{t-i} \leq \frac{1}{\beta}$. Then we get the desired result for the SCGD update. Next we present the proof for the SCSC update. Taking the total expectation with respect to the internal randomness of the algorithm over (13) and noting that $\mathbb{E}_A[\|x_t - x_{t-1}\|^2] \leq L_f L_g \eta_{t-1}^2$, we get

$$\mathbb{E}_A[\|y_{t+1} - g_S(x_t)\|^2] \leq (1-\beta_t)\mathbb{E}_A[\|y_t - g_S(x_{t-1})\|^2] + L_f^2 L_g^3\eta_{t-1}^2 + 2V_g\beta_t^2.$$

Telescoping the above inequality from 1 to t yields

$$\begin{aligned}&\mathbb{E}_A[\|y_{t+1}-g_S(x_t)\|^2]\\ \leq&\prod_{i=1}^{t}(1-\beta_i)\mathbb{E}_A[\|y_1-g_S(x_0)\|^2]+L_f^2L_g^3\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\beta_j)\eta_{i-1}^2+2V_g\sum_{i=1}^{t}\prod_{j=i+1}^{t}(1-\beta_j)\beta_i^2.\end{aligned}$$

Note that $\prod_{i=K}^{N}(1-\beta_i)\leq\exp(-\sum_{i=K}^{N}\beta_i)$ for all $K\leq N$ and $\beta_i>0$, then setting $\eta_t=\eta,\beta_t=\beta$, thus we have

$$\begin{aligned}&\mathbb{E}_A[\|y_{t+1}-g_S(x_t)\|^2]\\ \leq&\exp(-t\beta)\mathbb{E}_A[\|y_1-g_S(x_0)\|^2]+\sum_{i=1}^{t}(1-\beta)^{t-i}(L_g^3L_f^2\eta^2+2V_g\beta^2).\end{aligned}$$

Using Lemma A.3 with $\nu=1$, we get

$$\mathbb{E}_A[\|y_{t+1}-g_S(x_t)\|^2]\leq\Big(\frac{c}{e}\Big)^c(t\beta)^{-c}\mathbb{E}_A[\|y_1-g_S(x_0)\|^2]+L_g^3L_f^2\frac{\eta^2}{\beta}+2V_g\beta,$$

where the inequality holds for $\sum_{i=1}^{t}(1-\beta)^{t-i}\leq\frac{1}{\beta}$. Then we get the desired result for the SCSC update. Then we complete the proof. □

B. Proof for Section 2

Proof of Theorem 2.3. Write

$$\begin{aligned}&\mathbb{E}_{S,A}\Big[F(A(S))-F_S(A(S))\Big]=\mathbb{E}_{S,A}\Big[\mathbb{E}_\nu[f_\nu\big(g(\mathbf{x})\big)]-\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}\big(\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(\mathbf{x})\big)\Big]\\ &=\mathbb{E}_{S,A}\Big[\mathbb{E}_\nu[f_\nu\big(g(A(S))\big)]-\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}\big(g(A(S))\big)\Big]\\ &+\mathbb{E}_{S,A}\Big[\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}\big(g(A(S))\big)-\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}\big(\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\big)\Big]\\ &\leq\mathbb{E}_{S,A}\Big[\mathbb{E}_\nu[f_\nu\big(g(A(S))\big)]-\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}\big(g(A(S))\big)\Big]\\ &+\mathbb{E}_{S,A}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(f_{\nu_i}\big(g(A(S))\big)-f_{\nu_i}\big(\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\big)\Big)\Big].\end{aligned}\tag{14}$$

Now we estimate the two terms on the right-hand side of (14). Define $S'^{,\nu}=\{\nu_1',\nu_2',...,\nu_n',\omega_1,\omega_2,...,\omega_m\}$. In particular, we have that

$$\begin{aligned}&\mathbb{E}_{S,A}\big[\mathbb{E}_\nu[f_\nu(g(A(S)))]-\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}(g(A(S)))\big]\\ &=\mathbb{E}_{S,A,S'^{,\nu}}\big[\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}(g(A(S^{i,\nu})))-\frac{1}{n}\sum_{i=1}^{n}f_{\nu_i}(g(A(S)))\big]\\ &=\mathbb{E}_{S,A,S'^{,\nu}}\big[\frac{1}{n}\sum_{i=1}^{n}\big(f_{\nu_i}(g(A(S^{i,\nu}))-f_{\nu_i}(g(A(S)))\big]\\ &\leq L_f\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,A,S'^{,\nu}}[\|g(A(S^{i,\nu}))-g(A(S))\|]\\ &\leq L_fL_g\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,A,S'^{,\nu}}[\|A(S^{i,\nu})-A(S)\|]\leq L_fL_g\epsilon_\nu.\end{aligned}\tag{15}$$

Furthermore,

$$\mathbb{E}_{S,A}\big[\frac{1}{n}\sum_{i=1}^{n}\big(f_{\nu_i}\big(g(A(S))\big)-f_{\nu_i}\big(\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\big)\big)\big]$$
$$\leq L_f\mathbb{E}_{S,A}\big[\big\|g(A(S))-\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\big\|\big]. \tag{16}$$

Now it is sufficient to estimate the term $\mathbb{E}_{S,A}\big[\big\|g(A(S))-\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\big\|\big]$. Note that, in general, g is a mapping from $\mathbb{R}^p$ to $\mathbb{R}^d$. To this end, we will use some ideas from Bousquet et al. (2020). To this end, we write

$$\begin{aligned}
&g(A(S))-\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\\
&=\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\omega,\omega_j'}\left[g_\omega(A(S))-g_\omega\left(A\left(S^{j,\omega}\right)\right)\right]+\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega\left[g_\omega\left(A\left(S^{j,\omega}\right)\right)\right]-g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right]\\
&+\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\omega_j'}\left[g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)-g_{\omega_j}(A(S))\right].
\end{aligned}$$

It then follows that:

$$\begin{aligned}
&\|g(A(S))-\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\|\leq\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\omega,\omega_j'}\|g_\omega(A(S))-g_\omega(A(S^{j,\omega}))\|\\
&+\frac{1}{m}\|\sum_{j=1}^{m}\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega[g_\omega(A(S^{j,\omega}))]-g_{\omega_j}(A(S^{j,\omega}))]\|+\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\omega_j'}\|g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)-g_{\omega_j}(A(S))\|.
\end{aligned}$$

Note S and $S^{j,\omega}$ differ by a single example. By the assumption on stability and Definition 2.1, we further get

$$\begin{aligned}
&\mathbb{E}_{S,A}[\|g(A(S))-\frac{1}{m}\sum_{j=1}^{m}g_{\omega_j}(A(S))\|]\\
&\leq\mathbb{E}_{S,A}[\frac{1}{m}\|\sum_{j=1}^{m}\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega[g_\omega(A(S^{j,\omega}))]-g_{\omega_j}(A(S^{j,\omega}))]\|]+2L_g\epsilon_\omega.
\end{aligned}\tag{17}$$

Next step, we need to estimate $\|\sum_{j=1}^{m}\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega\left[g_\omega\left(A\left(S^{j,\omega}\right)\right)\right]-g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)]\|$.
Using a similar proof technique in paper (Lei, 2022), we can set $\xi_j(S)$ as a function of S as follows

$$\xi_j(S)=\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega\left[g_\omega\left(A\left(S^{j,\omega}\right)\right)\right]-g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)].$$

Notice that:

$$\mathbb{E}_{S,A}[\|\sum_{j=1}^{m}\xi_j(S)\|^2]=\mathbb{E}_{S,A}[\sum_{j=1}^{m}\|\xi_j(S)\|^2]+\sum_{j,i\in[m]:j\neq i}\mathbb{E}_{S,A}[\langle\xi_j(S),\xi_i(S)\rangle]. \tag{18}$$

According to the definition of $\xi_j(S)$ and the Cauchy-Schwarz inequality, we know

$$\begin{aligned}
&\mathbb{E}_{S,A}[\sum_{j=1}^{m}\|\xi_j(S)\|^2]=\sum_{j=1}^{m}\mathbb{E}_{S,A}[\|\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega[g_\omega(A(S^{j,\omega}))]-g_{\omega_j}(A(S^{j,\omega}))]\|^2]\\
&\leq\sum_{j=1}^{m}\mathbb{E}_{S,A}[\|\mathbb{E}_\omega[g_\omega(A(S^{j,\omega}))]-g_{\omega_j}(A(S^{j,\omega}))\|^2]\\
&=\sum_{j=1}^{m}\mathbb{E}_{S,A}[\|\mathbb{E}_\omega[g_\omega(A(S))]-g_{\omega_j'}(A(S))\|^2]=m\mathbb{E}_{S,A}\left[\mathrm{Var}_\omega(g_\omega(A(S)))\right],
\end{aligned}\tag{19}$$

where the variance term $\mathrm{Var}_\omega(g_\omega(A(S))) = \mathbb{E}_\omega\left[\|g(A(S)) - g_\omega(A(S))\|^2\right]$.
Next, we will estimate the second term on the right-hand side of (18). To this end, we define

$$S^{i,\omega} = \{\omega_1, \ldots, \omega_{i-1}, \omega_i', \omega_{i+1}, \ldots, \omega_m, \nu_1, \ldots, \nu_n\};$$
$$S^{i,j,\omega} = \left\{\omega_1, \ldots, \omega_{i-1}, \omega_i', \omega_{i+1}, \ldots, \omega_{j-1}, \omega_j', \omega_{j+1}, \ldots, \omega_m, \nu_1, \ldots, \nu_n\right\}.$$

Due to the symmetry between ω and ω_j, we can have

$$\mathbb{E}_{w_j}\left[\xi_j(S)\right] = 0, \forall j \in [m] \tag{20}$$

If $j \neq i$,we have

$$\begin{aligned}\mathbb{E}_{S,A}\left[\left\langle \xi_j\left(S^{i,\omega}\right), \xi_i(S)\right\rangle\right] &= \mathbb{E}_{S,A}\mathbb{E}_{\omega_i}\left[\left\langle \xi_j\left(S^{i,\omega}\right), \xi_i(S)\right\rangle\right] \\ &= \mathbb{E}_{S,A}\left[\left\langle \xi_j\left(S^{i,\omega}\right), \mathbb{E}_{\omega_i}\left[\xi_i(S)\right]\right\rangle\right] = 0,\end{aligned}$$

where the second equality holds since the $\xi_j\left(S^{i,\omega}\right)$ is independent of ω_i and the last identity follows from $\mathbb{E}_{w_i}\left[\xi_i(S)\right] = 0$ due to (20) . In a similar way, we can get the following equations for $j \neq i$

$$\begin{aligned}\mathbb{E}_{S,A}\left[\left\langle \xi_j(S), \xi_i\left(S^{j,\omega}\right)\right\rangle\right] &= \mathbb{E}_{S,A}\mathbb{E}_{\omega_j}\left[\left\langle \xi_j(S), \xi_i\left(S^{j,\omega}\right)\right\rangle\right] \\ &= \mathbb{E}_{S,A}\left[\left\langle \mathbb{E}_{\omega_j}\left[\xi_j(S)\right], \xi_i\left(S^{j,\omega}\right)\right\rangle\right] = 0,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{S,A}\left[\left\langle \xi_j\left(S^{i,\omega}\right), \xi_i\left(S^{j,\omega}\right)\right\rangle\right] &= \mathbb{E}_{S,A}\mathbb{E}_{\omega_j}\left[\left\langle \xi_j\left(S^{i,\omega}\right), \xi_i\left(S^{j,\omega}\right)\right\rangle\right] \\ &= \mathbb{E}_{S,A}\left[\left\langle \mathbb{E}_{\omega_j}\left[\xi_j\left(S^{i,\omega}\right)\right], \xi_i\left(S^{j,\omega}\right)\right\rangle\right] = 0.\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}_{S,A}\left[\langle \xi_j(S), \xi_i(S)\rangle\right] = \mathbb{E}_{S,A}\left[\left\langle \xi_j(S) - \xi_j\left(S^{i,\omega}\right), \xi_i(S) - \xi_i\left(S^{j,\omega}\right)\right\rangle\right] \\ &\leq \mathbb{E}_{S,A}\left[\left\|\xi_j(S) - \xi_j\left(S^{i,\omega}\right)\right\| \cdot \left\|\xi_i(S) - \xi_i\left(S^{j,\omega}\right)\right\|\right] \\ &\leq \frac{1}{2}\mathbb{E}_{S,A}\left[\left\|\xi_j(S) - \xi_j\left(S^{i,\omega}\right)\right\|^2\right] + \frac{1}{2}\mathbb{E}_{S,A}\left[\left\|\xi_i(S) - \xi_i\left(S^{j,\omega}\right)\right\|^2\right],\end{aligned} \tag{21}$$

where the third inequality use $ab \leq \frac{1}{2}\left(a^2 + b^2\right)$. With the definition of $\xi_j(S)$, $S^{i,\omega}$ and $S^{i,j,\omega}$, we can have the following identity for $j \neq i$

$$\begin{aligned}&\mathbb{E}_{S,A}\left[\left\|\xi_j(S) - \xi_j\left(S^{i,\omega}\right)\right\|^2\right] \\ &= \mathbb{E}_{S,A}\left[\left\|\mathbb{E}_{\omega_j'}[\mathbb{E}_\omega\left[g_\omega\left(A\left(S^{j,\omega}\right)\right)\right] - g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)] - \mathbb{E}_{\omega_j'}\left[\mathbb{E}_\omega\left[g_\omega\left(A\left(S^{i,j,\omega}\right)\right)\right] - g_{\omega_j}\left(A\left(S^{i,j,\omega}\right)\right)\right]\right\|^2\right] \\ &= \mathbb{E}_{S,A}\left[\left\|\mathbb{E}_{\omega_j'}\mathbb{E}_\omega\left[g_\omega\left(A\left(S^{j,\omega}\right)\right) - g_\omega\left(A\left(S^{i,j,\omega}\right)\right)\right] + \mathbb{E}_{\omega_j'}\left[g_{\omega_j}\left(A\left(S^{i,j,\omega}\right)\right) - g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)\right]\right\|^2\right].\end{aligned}$$

Then using the elementary inequality $(a+b)^2 \leq 2\left(a^2 + b^2\right)$ and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}&\mathbb{E}_{S,A}\left[\left\|\xi_j(S) - \xi_j\left(S^{i,\omega}\right)\right\|^2\right] \\ &\leq 2\mathbb{E}_{S,A}\left[\left\|g_\omega\left(A\left(S^{j,\omega}\right)\right) - g_\omega\left(A\left(S^{i,j,\omega}\right)\right)\right\|^2\right] + 2\mathbb{E}_{S,A}\left[\left\|g_{\omega_j}\left(A\left(S^{i,j,\omega}\right)\right) - g_{\omega_j}\left(A\left(S^{j,\omega}\right)\right)\right\|^2\right] \\ &\leq 2\mathbb{E}_{S,A}\left[L_g^2\left\|A\left(S^{j,\omega}\right) - A\left(S^{i,j,\omega}\right)\right\|^2\right] + 2\mathbb{E}_{S,A}\left[L_g^2\left\|A\left(S^{i,j,\omega}\right) - A\left(S^{j,\omega}\right)\right\|^2\right].\end{aligned}$$

Since $S^{i,\omega}$ and $S^{i,j,\omega}$ differ by one example, it follows from the definition of stability, we can have

$$\mathbb{E}_{S,A}\left[\left\|\xi_j(S) - \xi_j\left(S^{i,\omega}\right)\right\|^2\right] \leq 4L_g^2\epsilon_\omega^2, \forall j \neq i.$$

In a similar way, we can have

$$\mathbb{E}_{S,A}\left[\left\|\xi_i(S)-\xi_i\left(S^{j,\omega}\right)\right\|^2\right]\leq 4L_g^2\epsilon_\omega^2, \forall j\neq i.$$

Combining above two inequalities into (21), we get

$$\sum_{j,i\in[m]:j\neq i}\mathbb{E}_{S,A}\left[\langle\xi_j(S),\xi_i(S)\rangle\right]\leq 4m(m-1)L_g^2\epsilon_\omega^2, \forall j\neq i. \tag{22}$$

Then combining the (22) and (19) into (18), we can have

$$\mathbb{E}_{S,A}\big[\|\sum_{j=1}^m\xi_j(S)\|^2\big]=m\mathbb{E}_{S,A}\left[\mathrm{Var}_\omega(g_\omega(A(S)))\right]+4m(m-1)L_g^2\epsilon_\omega^2.$$

Then we get

$$\mathbb{E}_{S,A}\big[\|\sum_{j=1}^m\xi_j(S)\|\big]\leq(\mathbb{E}_{S,A}\big[\|\sum_{j=1}^m\xi_j(S)\|^2\big])^{1/2}\leq\sqrt{m\mathbb{E}_{S,A}\left[\mathrm{Var}_\omega(g_\omega(A(S)))\right]}+2mL_g\epsilon_\omega,$$

plugging the above inequality back into (17), we get

$$\mathbb{E}_{S,A}\big[\|g(A(S))-\frac{1}{m}\sum_{j=1}^m g_{\omega_j}(A(S))\|\big]\leq\sqrt{m^{-1}\mathbb{E}_{S,A}\left[\mathrm{Var}_\omega(g_\omega(A(S)))\right]}+4L_g\epsilon_\omega. \tag{23}$$

Using the result (23) into (16) and then combining with the result (15) into (14), we get final result

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

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

C. Proof for the Convex Setting

C.1. Stability

Proof of Theorem 3.3. For any $k\in[n]$, define $S^{k,\nu}=\{\nu_1,...,\nu_{k-1},\nu_k',\nu_{k+1},...,\nu_n,\omega_1,...,\omega_m\}$ as formed from S_ν by replacing the k-th element. For any $l\in[m]$, define $S^{l,\omega}=\{\nu_1,...,\nu_n,\omega_1,...,\omega_{l-1},\omega_l',\omega_{l+1},...,\omega_m\}$ as formed from S_ω by replacing the l-th element. Let $\{x_{t+1}\}$ and $\{y_{t+1}\}$ be produced by Algorithm 1 based on S, $\{x_{t+1}^{k,\nu}\}$ and $\{y_{t+1}^{k,\nu}\}$ be produced by Algorithm 1 based on $S^{k,\nu}$, $\{x_{t+1}^{l,\omega}\}$ and $\{y_{t+1}^{l,\omega}\}$ be produced by Algorithm 1 based on $S^{l,\omega}$. Let $x_0=x_0^{k,\nu}$ and $x_0=x_0^{l,\omega}$ be starting points in $\mathcal{X}$. Since changing one sample data can happen in either S_ν or S_ω, we estimate $\mathbb{E}_A\big[\|x_{t+1}-x_{t+1}^{k,\nu}\|\big]$ and $\mathbb{E}_A\big[\|x_{t+1}-x_{t+1}^{l,\omega}\|\big]$ as follows.

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

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

Case 1 ($i_t\neq k$). If $i_t\neq k$, we have

$$\begin{aligned}&\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\leq\|x_t-\eta_t\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-x_t^{k,\nu}+\eta_t\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2\\&=\|x_t-x_t^{k,\nu}\|^2-2\eta_t\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}),x_t-x_t^{k,\nu}\rangle\\&+\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2.\end{aligned} \tag{24}$$

Taking the expectation w.r.t j_t on the both sides of (24) implies that

$$\begin{aligned}&\mathbb{E}_{j_t}\big[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\big]\\&\leq\mathbb{E}_{j_t}\big[\|x_t-x_t^{k,\nu}\|^2\big]-2\eta_t\mathbb{E}_{j_t}\big[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}),x_t-x_t^{k,\nu}\rangle\big]\\&+\eta_t^2\mathbb{E}_{j_t}\big[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2\big].\end{aligned} \tag{25}$$

We first estimate the second term on the right hand side of (25). It can be decomposed as

$$\begin{aligned}
&-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}),x_t-x_t^{k,\nu}\rangle\right]\\
&=-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{k,\nu}\rangle\right]\\
&\quad-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{k,\nu}\rangle\right]\\
&\quad-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})),x_t-x_t^{k,\nu}\rangle\right]\\
&\quad-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})),x_t-x_t^{k,\nu}\rangle\right]\\
&\quad-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}),x_t-x_t^{k,\nu}\rangle\right].
\end{aligned}\tag{26}$$

Now we estimate the terms on the right hand side of (26) one by one. To this end, noticing that j_t is independent of i_t and x_t, then $\mathbb{E}_{j_t}\left[\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\right]=\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))$ holds true. Consequently,

$$\begin{aligned}
&-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{k,\nu}\rangle\right]=0,\\
&-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})),x_t-x_t^{k,\nu}\rangle\right]=0.
\end{aligned}\tag{27}$$

Then by Part (iv) of Assumption 3.1, we know $f_\nu(g_S(\cdot))$ is L-smooth. Combining this with the convexity of $f_\nu(g_S(\cdot))$ and inequality (5), we get

$$\begin{aligned}
&\langle\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})),x_t-x_t^{k,\nu}\rangle\\
&\geq\frac{1}{L}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2.
\end{aligned}\tag{28}$$

Furthermore, noticing that x_t is independent of j_t, we get

$$\begin{aligned}
&-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{k,\nu}\rangle\right]\\
&\leq 2\eta_t\mathbb{E}_{j_t}\left[\left|\langle\nabla g_{\omega_{j_t}}(x_t)(\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla f_{\nu_{i_t}}(g_S(x_t))),x_t-x_t^{k,\nu}\rangle\right|\right]\\
&\leq 2\eta_t\mathbb{E}_{j_t}\left[\|\nabla g_{\omega_{j_t}}(x_t)(\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla f_{\nu_{i_t}}(g_S(x_t)))\|\|x_t-x_t^{k,\nu}\|\right]\\
&\leq 2\eta_t\mathbb{E}_{j_t}\left[\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla f_{\nu_{i_t}}\left(g_S(x_t)\right)\|\|x_t-x_t^{k,\nu}\|\right]\\
&\leq C_fL_g2\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}-g_S(x_t)\|\right]\|x_t-x_t^{k,\nu}\|,
\end{aligned}\tag{29}$$

where the last inequality holds by L_g Lipschitz continuity of g_ω in Assumption 2.2(ii) and the C_f Lipschitz continuous gradients of f_ν in Assumption 3.1(iii). Analogous to (29), we get

$$\begin{aligned}
&-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}),x_t-x_t^{k,\nu}\rangle\right]\\
&\leq 2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|\right]\|x_t-x_t^{k,\nu}\|.
\end{aligned}\tag{30}$$

Putting (27), (28), (29) and (30) into (26), we get that

$$\begin{aligned}
&-2\eta_t\mathbb{E}_{j_t}\left[\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}),x_t-x_t^{k,\nu}\rangle\right]\\
&\leq 2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}-g_S(x_t)\|\right]\|x_t-x_t^{k,\nu}\|+2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|\right]\|x_t-x_t^{k,\nu}\|\\
&-2\eta_t\frac{1}{L}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2.
\end{aligned}\tag{31}$$

We estimate the third term on the right hand side of (25) as follows:

$$\begin{aligned}
&\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|\\
&\leq\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\|\\
&+\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\|\\
&+\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|\\
&+\|\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|\\
&+\|\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))-\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|.
\end{aligned}$$

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

$$\begin{aligned}
&\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2 \\
\leq& 4\eta_t^2 C_f^2\|\nabla g_{\omega_{j_t}}(x_t)(y_{t+1} - g_S(x_t))\|^2 + 4\eta_t^2 C_f^2\|\nabla g_{\omega_{j_t}}(x_t)(g_S(x_t^{k,\nu}) - y_{t+1}^{k,\nu})\|^2 \\
&+ 8\eta_t^2\|(\nabla g_{\omega_{j_t}}(x_t) - \nabla g_S(x_t))\nabla f_{\nu_{i_t}}(g_S(x_t))\|^2 \\
&+ 8\eta_t^2\|(\nabla g_{\omega_{j_t}}(x_t^{k,\nu}) - \nabla g_S(x_t^{k,\nu}))\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2 \\
&+ 4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2, \qquad (32)
\end{aligned}$$

where we have used the fact that $(\sum_{i=1}^5 a_i)^2 \leq 4a_1^2 + 4a_2^2 + 4a_3^2 + 8a_4^2 + 8a_5^2$ and part (iii) of Assumption 3.1, i.e., C_f-Lipschitz continuity of ∇f_ν. Taking the expectation w.r.t. j_t on both sides of (32), there holds

$$\begin{aligned}
&\mathbb{E}_{j_t}\left[\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2\right] \\
\leq& 4\eta_t^2 C_f^2\mathbb{E}_{j_t}\left[\|\nabla g_{\omega_{j_t}}(x_t)\|^2\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2\mathbb{E}_{j_t}\left[\|\nabla g_{\omega_{j_t}}(x_t)\|^2\|g_S(x_t^{k,\nu}) - y_{t+1}^{k,\nu}\|^2\right] \\
&+ 8\eta_t^2\mathbb{E}_{j_t}\left[\|\nabla g_{\omega_{j_t}}(x_t) - \nabla g_S(x_t)\|^2\|\nabla f_{\nu_{i_t}}(g_S(x_t))\|^2\right] \\
&+ 8\eta_t^2\mathbb{E}_{j_t}\left[\|\nabla g_{\omega_{j_t}}(x_t^{k,\nu}) - \nabla g_S(x_t^{k,\nu})\|^2\|\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2\right] \\
&+ 4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\| \\
\leq& 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|^2\right] + 16\eta_t^2 L_f^2 C_g \\
&+ 4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2, \qquad (33)
\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of f_ν and g_ω according to Assumption 2.2 as well as part (ii) of Assumption 3.1.

Putting (31) and (33) back into (25) implies that

$$\begin{aligned}
&\mathbb{E}_{j_t}\left[\|x_{t+1} - x_{t+1}^{k,\nu}\|^2\right] \\
&\leq \|x_t - x_t^{k,\nu}\|^2 + 2C_f L_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1} - g_S\|\right]\|x_t - x_t^{k,\nu}\| \\
&+ 2C_f L_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|\right]\|x_t - x_t^{k,\nu}\| \\
&+ (4\eta_t^2 - 2\eta_t\frac{1}{L})\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2 \\
&+ 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|^2\right] + 16\eta_t^2 L_f^2 C_g \\
&\leq \|x_t - x_t^{k,\nu}\|^2 + 2C_f L_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1} - g_S(x_t)\|\right]\|x_t - x_t^{k,\nu}\| \\
&+ 2C_f L_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|\right]\|x_t - x_t^{k,\nu}\| \\
&+ 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|^2\right] + 16\eta_t^2 L_f^2 C_g,
\end{aligned}$$

where in the second inequality we have used the fact that $\eta_t \leq \frac{1}{2L}$.

Case 2 ($i_t = k$). If $i_t = k$, we have

$$\begin{aligned}
\|x_{t+1} - x_{t+1}^{k,\nu}\| &= \|x_t - \eta_t\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - x_t^{k,\nu} + \eta_t\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu'_{i_t}}(y_{t+1}^{k,\nu})\| \\
&\leq \|x_t - x_t^{k,\nu}\| + \eta_t\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu'_{i_t}}(y_{t+1}^{k,\nu})\| \\
&\leq \|x_t - x_t^{k,\nu}\| + \eta_t\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(y_{t+1})\| + \eta_t\|\nabla g_{\omega_{j_t}}(x_t^{k,\nu})\|\|\nabla f_{\nu'_{i_t}}(y_{t+1}^{k,\nu})\| \\
&\leq \|x_t - x_t^{k,\nu}\| + 2L_g L_f\eta_t,
\end{aligned}$$

where in the third inequality we have used Assumption 2.2, i.e., the Lipschitz continuity of f_ν and g_ω. Taking the square of the terms on both sides of the above inequality and taking the expectation w.r.t. j_t yield that

$$\mathbb{E}_{j_t}\left[\|x_{t+1} - x_{t+1}^{k,\nu}\|^2\right] \leq \|x_t - x_t^{k,\nu}\|^2 + 4L_g L_f\eta_t\|x_t - x_t^{k,\nu}\| + 4L_g^2 L_f^2\eta_t^2. \qquad (34)$$

Combining **Case 1** and **Case 2** together, we have that

$$\begin{aligned}
&\mathbb{E}_{j_t}\big[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\big] \le \|x_t - x_t^{k,\nu}\|^2 + 2C_f L_g \eta_t \mathbb{E}_{j_t}\big[\|y_{t+1}-g_S(x_t)\|\big]\|x_t - x_t^{k,\nu}\| \\
&+ 2C_f L_g \eta_t \mathbb{E}_{j_t}\big[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|\big]\|x_t - x_t^{k,\nu}\| \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_{j_t}\big[\|y_{t+1}-g_S(x_t)\|^2\big] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_{j_t}\big[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\big] + 16\eta_t^2 L_f^2 C_g \\
&+ 4L_g L_f \eta_t \|x_t - x_t^{k,\nu}\|\mathbb{I}_{[i_t=k]} + 4L_g^2 L_f^2 \eta_t^2 \mathbb{I}_{[i_t=k]}. \qquad (35)
\end{aligned}$$

Taking the expectation w.r.t. A on both sides of (35), we get that

$$\begin{aligned}
&\mathbb{E}_A\big[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\big] \\
\le&\mathbb{E}_A\big[\|x_t - x_t^{k,\nu}\|^2\big] + 2C_f L_g \eta_t \mathbb{E}_A\big[\mathbb{E}_{j_t}[\|y_{t+1}-g_S(x_t)\|]\|x_t - x_t^{k,\nu}\|\big] \\
&+ 2C_f L_g \eta_t \mathbb{E}_A\big[\mathbb{E}_{j_t}[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|]\|x_t - x_t^{k,\nu}\|\big] \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}-g_S(x_t)\|^2\big] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\big] + 16\eta_t^2 L_f^2 C_g \\
&+ 4L_f L_g \eta_t \mathbb{E}_A[\|x_t - x_t^{k,\nu}\|\mathbb{I}_{[i_t=k]}] + 4L_g^2 L_f^2 \eta_t^2 \mathbb{E}_A[\mathbb{I}_{[i_t=k]}] \\
\le&\mathbb{E}_A\big[\|x_t - x_t^{k,\nu}\|^2\big] + 2C_f L_g \eta_t (\mathbb{E}_A\big[\|y_{t+1}-g_S(x_t)\|^2\big])^{1/2}(\mathbb{E}_A\big[\|x_t - x_t^{k,\nu}\|^2\big])^{1/2} \\
&+ 2C_f L_g \eta_t (\mathbb{E}_A\big[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\big])^{1/2}(\mathbb{E}_A\big[\|x_t - x_t^{k,\nu}\|^2\big])^{1/2} \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}-g_S(x_t)\|^2\big] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\big] + 16\eta_t^2 L_f^2 C_g \\
&+ 4L_f L_g \eta_t \mathbb{E}_A[\|x_t - x_t^{k,\nu}\|\mathbb{I}_{[i_t=k]}] + 4L_g^2 L_f^2 \eta_t^2 \mathbb{E}_A[\mathbb{I}_{[i_t=k]}], \qquad (36)
\end{aligned}$$

where the second inequality holds by the Cauchy-Schwarz inequality. Observe that

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

Note that $\|x_0 - x_0^{k,\nu}\|^2 = 0$. Combining above observation with (36) implies that

$$\begin{aligned}
&\mathbb{E}_A\big[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\big] \\
&\le 2C_f L_g \sum_{j=1}^{t} \eta_j (\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big])^{1/2}\big(\mathbb{E}_A\big[\|x_j - x_j^{k,\nu}\|^2\big]\big)^{1/2} \\
&+ 2C_f L_g \sum_{j=1}^{t} \eta_j (\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big])^{1/2}\big(\mathbb{E}_A\big[\|x_j - x_j^{k,\nu}\|^2\big]\big)^{1/2} \\
&+ 4C_f^2 L_g^2 \sum_{j=0}^{t} \eta_j^2 \mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big] + 4C_f^2 L_g^2 \sum_{j=0}^{t} \eta_j^2 \mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big] \\
&+ 16L_f^2 C_g \sum_{j=0}^{t} \eta_j^2 + \frac{4L_g L_f}{n}\sum_{j=1}^{t} \eta_j (\mathbb{E}_A\big[\|x_j - x_j^{k,\nu}\|^2\big])^{1/2} + \frac{4L_f^2 L_g^2}{n}\sum_{j=0}^{t}\eta_j^2. \qquad (37)
\end{aligned}$$

For notational convenience, we denote by $u_t = (\mathbb{E}_A\big[\|x_t - x_t^{k,\nu}\|^2\big])^{1/2}$. Using this notation, from (37) we get that

$$\begin{aligned}
u_t^2 \le& 2C_f L_g \sum_{j=1}^{t-1} \eta_j (\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big])^{1/2} u_j + 2C_f L_g \sum_{j=1}^{t-1} \eta_j (\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big])^{1/2} u_j \\
&+ 4C_f^2 L_g^2 \sum_{j=0}^{t-1} \eta_j^2 \mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big] + 4C_f^2 L_g^2 \sum_{j=0}^{t-1} \eta_j^2 \mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big] \\
&+ 16L_f^2 C_g \sum_{j=0}^{t-1} \eta_j^2 + \frac{4L_g L_f}{n}\sum_{j=1}^{t-1} \eta_j u_j + \frac{4L_f^2 L_g^2}{n}\sum_{j=0}^{t-1}\eta_j^2.
\end{aligned}$$

We will apply Lemma A.2 to get the desired estimation from the above recursive inequality. To this end, we define

$$S_t = 4C_f^2L_g^2\sum_{j=0}^{t-1}\eta_j^2\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big] + 4C_f^2L_g^2\sum_{j=0}^{t-1}\eta_j^2\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big] + \frac{4L_f^2L_g^2}{n}\sum_{j=0}^{t-1}\eta_j^2 + 16L_f^2C_g\sum_{j=0}^{t-1}\eta_j^2,$$

$$\alpha_j = 2C_fL_g\eta_j(\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big])^{1/2} + 2C_fL_g\eta_j(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big])^{1/2} + \frac{4L_gL_f}{n}\eta_j.$$

Now applying Lemma A.2 with u_t, S_t and α_j defined above, we get

$$\begin{aligned}
u_t &\le \sqrt{S_t} + \sum_{j=1}^{t-1}\alpha_j \\
&\le \big(4C_f^2L_g^2\sum_{j=0}^{t-1}\eta_j^2\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big]\big)^{1/2} + \big(4C_f^2L_g^2\sum_{j=0}^{t-1}\eta_j^2\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big]\big)^{1/2} \\
&+ \big(\frac{4L_f^2L_g^2}{n}\sum_{j=0}^{t-1}\eta_j^2\big)^{1/2} + (16L_f^2C_g\sum_{j=0}^{t-1}\eta_j^2)^{1/2} + 2C_fL_g\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big])^{1/2} \\
&+ 2C_fL_g\sum_{j=1}^{t-1}\eta_j(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big])^{1/2} + \frac{4L_fL_g}{n}\sum_{j=1}^{t-1}\eta_j \\
&\le 4C_fL_g\sum_{j=0}^{t-1}\eta_j(\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big])^{1/2} + 4C_fL_g\sum_{j=0}^{t-1}\eta_j(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big])^{1/2} \\
&+ 4L_f\sqrt{C_g}(\sum_{j=0}^{t-1}\eta_j^2)^{1/2} + \big(\frac{4L_f^2L_g^2}{n}\sum_{j=0}^{t-1}\eta_j^2\big)^{1/2} + \frac{4L_fL_g}{n}\sum_{j=0}^{t-1}\eta_j,
\end{aligned} \tag{38}$$

where the second inequality uses the fact that $(\sum_{i=1}^4 a_i)^{1/2} \le \sum_{i=1}^4 (a_i)^{1/2}$ and the last inequality holds by the fact that

$$\big(4C_f^2L_g^2\sum_{j=0}^{t-1}\eta_j^2\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big]\big)^{1/2} \le 2C_fL_g\sum_{j=0}^{t-1}\eta_j(\mathbb{E}_A\big[\|y_{j+1}-g_S(x_j)\|^2\big])^{1/2},$$

$$\big(4C_f^2L_g^2\sum_{j=0}^{t-1}\eta_j^2\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big]\big)^{1/2} \le 2C_fL_g\sum_{j=0}^{t-1}\eta_j(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\big])^{1/2}.$$

Furthermore, if $\eta_t = \eta$, it is easy to see that $\sum_{j=0}^{T-1}(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2])^{1/2} \le \sup_S \eta\sum_{j=0}^{T-1}(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2])^{1/2}$ and $\sum_{j=0}^{T-1}(\mathbb{E}_A[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2])^{1/2} \le \sup_S \eta\sum_{j=0}^{T-1}(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2])^{1/2}$. Consequently, with T iterations, we obtain that

$$\begin{aligned}
u_T \le& 8C_fL_g\sup_S\eta\sum_{j=0}^{T-1}(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2])^{1/2} + 4L_f\sqrt{C_g}(\sum_{j=0}^{T-1}\eta^2)^{1/2} + \big(\frac{4L_f^2L_g^2}{n}\sum_{j=0}^{T-1}\eta^2\big)^{1/2} \\
&+ \frac{4L_gL_f}{n}\sum_{j=0}^{T-1}\eta \\
\le& 8C_fL_g\sup_S\eta\sum_{j=0}^{T-1}(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2])^{1/2} + 4L_f\sqrt{C_g}\eta\sqrt{T} + \frac{6L_gL_f}{n}\eta T,
\end{aligned}$$

where the last inequality holds by the fact that $\big(\frac{4L_f^2L_g^2}{n}\sum_{j=0}^{T-1}\eta^2\big)^{1/2} = \frac{2L_gL_f}{\sqrt{n}}\eta\sqrt{T} \le \frac{2L_fL_g}{n}\eta T$ because often we have $T \ge n$.

Since $\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|\big] \leq u_T = (\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|^2\big])^{1/2}$, we further get

$$\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|\big] \leq 8C_f L_g \sup_S \eta \sum_{j=0}^{T-1} (\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2])^{1/2} + 4L_f\sqrt{C_g}\eta\sqrt{T} + \frac{6L_f L_g}{n}\eta T. \tag{39}$$

We got the following desired result for Case 1:

$$\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|\big] = \mathcal{O}\Big(L_f L_g \frac{T\eta}{n} + L_f\sqrt{C_g}\eta\sqrt{T} + C_f L_g \sup_S \sum_{j=0}^{T-1} \eta\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}\Big).$$

Next we move on to the estimation of $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|]$.

Estimation of $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|]$. We will estimate it by considering two cases, i.e., $j_t \neq l$ and $j_t = l$.

Case 1 ($j_t \neq l$). If $j_t \neq l$, we have

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - \eta_t \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - x_t^{l,\omega} + \eta_t \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|^2 \\
&= \|x_t - x_t^{l,\omega}\|^2 - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega}), x_t - x_t^{l,\omega}\rangle \\
&+ \eta_t^2 \|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|^2.
\end{aligned} \tag{40}$$

We will estimate the second term and the third one on the right hand side of (40) as follows. First, we estimate the second term. To this end, using similar arguments in (26), it can be decomposed as

$$\begin{aligned}
&-2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega}), x_t - x_t^{l,\omega}\rangle \\
&= -2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)(\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla f_{\nu_{i_t}}(g_S(x_t))), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t^{l,\omega})(\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})), x_t - x_t^{l,\omega}\rangle.
\end{aligned} \tag{41}$$

Using the convexity of $f_\nu(g_S(\cdot))$, part (iv) of Assumption 3.1 and inequality (5), we have

$$\begin{aligned}
&\langle \nabla g_S\big(x_t\big)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\geq \frac{1}{L}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2.
\end{aligned} \tag{42}$$

Furthermore, using part (ii) of Assumption 2.2 and part (iii) of Assumption 3.1, we get

$$\begin{aligned}
&- 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)(\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla f_{\nu_{i_t}}(g_S(x_t))), x_t - x_t^{l,\omega}\rangle \\
&\leq 2\eta_t\|\nabla g_{\omega_{j_t}}(x_t)\big(\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla f_{\nu_{i_t}}\big(g_S(x_t)\big)\big\|\|x_t - x_t^{l,\omega}\| \\
&\leq 2\eta_t\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla f_{\nu_{i_t}}\big(g_S(x_t)\big)\|\|x_t - x_t^{l,\omega}\| \\
&\leq 2\eta_t C_f L_g\|y_{t+1} - g_S(x_t)\|\|x_t - x_t^{l,\omega}\|.
\end{aligned} \tag{43}$$

Likewise,

$$\begin{aligned}
&- 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t^{l,\omega})(\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\leq 2\eta_t C_f L_g\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|\|x_t - x_t^{l,\omega}\|.
\end{aligned} \tag{44}$$

Putting (42), (43) and (44) into (41) yields that

$$
\begin{aligned}
&-2\eta_t\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega}),x_t-x_t^{l,\omega}\rangle\\
&\le 2\eta_t C_f L_g\|y_{t+1}-g_S(x_t)\|\|x_t-x_t^{l,\omega}\|+2\eta_t C_f L_g\|y_{t+1}^{l,\omega}-g_S(x_t^{l,\omega})\|\|x_t-x_t^{l,\omega}\|\\
&-2\eta_t\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{l,\omega}\rangle\\
&-2\eta_t\frac{1}{L}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2\\
&-2\eta_t\langle\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))-\nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})),x_t-x_t^{l,\omega}\rangle. \qquad (45)
\end{aligned}
$$

Next we will estimate the third term on the right hand side of (40). In analogy to the argument in (32), one can show that

$$
\begin{aligned}
&\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-\nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|^2\\
&\le 4\eta_t^2C_f^2\|\nabla g_{\omega_{j_t}}(x_t)(y_{t+1}-g_S(x_t))\|^2+4\eta_t^2C_f^2\|\nabla g_{\omega_{j_t}}(x_t)(g_S(x_t^{l,\omega})-y_{t+1}^{l,\omega})\|^2\\
&+8\eta_t^2\|(\nabla g_{\omega_{j_t}}(x_t)-\nabla g_S(x_t))\nabla f_{\nu_{i_t}}(g_S(x_t))\|^2\\
&+8\eta_t^2\|(\nabla g_{\omega_{j_t}}(x_t^{l,\omega})-\nabla g_S(x_t^{l,\omega}))\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2\\
&+4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2\\
&\le 4\eta_t^2C_f^2L_g^2\|y_{t+1}-g_S(x_t)\|^2+4\eta_t^2C_f^2L_g^2\|\nabla g_S(x_t^{l,\omega})-y_{t+1}^{l,\omega}\|^2\\
&+8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)-\nabla g_S(x_t)\|^2+8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t^{l,\omega})-\nabla g_S(x_t^{l,\omega})\|^2\\
&+4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2, \qquad (46)
\end{aligned}
$$

where, in the second inequality, we have used Assumption 2.2.

Putting the results (45) and (46) into (40) implies that

$$
\begin{aligned}
&\|x_{t+1}-x_{t+1}^{l,\omega}\|^2\\
&\le\|x_t-x_t^{l,\omega}\|^2+2\eta_tC_fL_g\|y_{t+1}-g_S(x_t)\|\|x_t-x_t^{l,\omega}\|+2\eta_tC_fL_g\|y_{t+1}^{l,\omega}-g_S(x_t^{l,\omega})\|\|x_t-x_t^{l,\omega}\|\\
&-2\eta_t\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{l,\omega}\rangle\\
&-2\eta_t\langle\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))-\nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})),x_t-x_t^{l,\omega}\rangle\\
&+(4\eta_t^2-2\eta_t\frac{1}{L})\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2\\
&+4\eta_t^2C_f^2L_g^2\|y_{t+1}-g_S(x_t)\|^2+4\eta_t^2C_f^2L_g^2\|g_S(x_t^{l,\omega})-y_{t+1}^{l,\omega}\|^2\\
&+8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)-\nabla g_S(x_t)\|^2+8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t^{l,\omega})-\nabla g_S(x_t^{l,\omega})\|^2\\
&\le\|x_t-x_t^{l,\omega}\|^2+2\eta_tC_fL_g\|y_{t+1}-g_S(x_t)\|\|x_t-x_t^{l,\omega}\|+2\eta_tC_fL_g\|y_{t+1}^{l,\omega}-g_S(x_t^{l,\omega})\|\|x_t-x_t^{l,\omega}\|\\
&-2\eta_t\langle\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)),x_t-x_t^{l,\omega}\rangle\\
&-2\eta_t\langle\nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))-\nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})),x_t-x_t^{l,\omega}\rangle\\
&+4\eta_t^2C_f^2L_g^2\|y_{t+1}-g_S(x_t)\|^2+4\eta_t^2C_f^2L_g^2\|g_S(x_t^{l,\omega})-y_{t+1}^{l,\omega}\|^2\\
&+8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)-\nabla g_S(x_t)\|^2+8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t^{l,\omega})-\nabla g_S(x_t^{l,\omega})\|^2, \qquad (47)
\end{aligned}
$$

where we have used the fact that $\eta_t\le\frac{1}{2L}$ in the second inequality.

Case 2 ($j_t=l$). If $j_t=l$, from Assumption 2.2 we have that

$$
\begin{aligned}
\|x_{t+1}-x_{t+1}^{l,\omega}\|&=\|x_t-\eta_t\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-x_t^{l,\omega}+\eta_t\nabla g_{\omega'_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|\\
&\le\|x_t-x_t^{l,\omega}\|+\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})+\nabla g_{\omega'_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|\\
&\le\|x_t-x_t^{l,\omega}\|+\eta_t\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(y_{t+1})\|+\eta_t\|\nabla g_{\omega'_{j_t}}(x_t^{l,\omega})\|\|\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|\\
&\le\|x_t-x_t^{l,\omega}\|+2\eta_tL_gL_f.
\end{aligned}
$$

Therefore,

$$\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - x_t^{l,\omega}\|^2 + 4L_g L_f \eta_t \|x_t - x_t^{l,\omega}\| + 4\eta_t^2 L_g^2 L_f^2. \tag{48}$$

Combining **Case 1** and **Case 2** together, we obtain

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \\
&\leq \|x_t - x_t^{l,\omega}\|^2 + 2C_f L_g \eta_t \left\|y_{t+1} - g_S\left(x_t\right)\right\| \left\|x_t - x_t^{l,\omega}\right\| + 2C_f L_g \eta_t \|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\| \|x_t - x_t^{l,\omega}\| \\
&- 2\eta_t \langle \nabla g_{\omega_{j_t}}\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right) - \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right), x_t - x_t^{l,\omega} \rangle \mathbb{I}_{[j_t \neq l]} \\
&- 2\eta_t \langle \nabla g_S(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega} \rangle \mathbb{I}_{[j_t \neq l]} \\
&+ 8L_f^2 \eta_t^2 \left\|\nabla g_{\omega_{j_t}}\left(x_t\right) - \nabla g_S\left(x_t\right)\right\|^2 + 8L_f^2 \eta_t^2 \|\nabla g_{\omega_{j_t}}(x_t^{l,\omega}) - \nabla g_S(x_t^{l,\omega})\|^2 \\
&+ 4\eta_t^2 C_f^2 L_g^2 \|y_{t+1} - g_S(x_t)\|^2 + 4\eta_t^2 C_f^2 L_g^2 \|g_S(x_t^{l,\omega}) - y_{t+1}^{l,\omega}\|^2 \\
&+ 4\eta_t L_g L_f \|x_t - x_t^{l,\omega}\| \mathbb{I}_{[j_t = l]} + 4\eta_t^2 L_g^2 L_f^2 \mathbb{I}_{[j_t = l]}.
\end{aligned} \tag{49}$$

Taking the expectation w.r.t. A on both sides of (49) yields that

$$\begin{aligned}
&\mathbb{E}_A\left[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2\right] \\
&\leq \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|^2\right] + 2C_f L_g \eta_t \mathbb{E}_A\left[\|y_{t+1} - g_S(x_t)\| \|x_t - x_t^{l,\omega}\|\right] \\
&+ 2C_f L_g \eta_t \mathbb{E}_A\left[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\| \|x_t - x_t^{l,\omega}\|\right] \\
&- 2\eta_t \mathbb{E}_A\left[\langle \nabla g_{\omega_{j_t}}\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right) - \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right), x_t - x_t^{l,\omega} \rangle \mathbb{I}_{[j_t \neq l]}\right] \\
&- 2\eta_t \mathbb{E}_A\left[\langle \nabla g_S(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega} \rangle \mathbb{I}_{[j_t \neq l]}\right] \\
&+ 8L_f^2 \eta_t^2 \mathbb{E}_A\left[\|\nabla g_{\omega_{j_t}}(x_t) - \nabla g_S(x_t)\|^2\right] + 8L_f^2 \eta_t^2 \mathbb{E}_A\left[\|\nabla g_{\omega_{j_t}}(x_t^{l,\omega}) - \nabla g_S(x_t^{l,\omega})\|^2\right] \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|g_S(x_t^{l,\omega}) - y_{t+1}^{l,\omega}\|^2\right] \\
&+ 4\eta_t L_f L_g \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\| \mathbb{I}_{[j_t = l]}\right] + 4\eta_t^2 L_g^2 L_f^2 \mathbb{E}_A\left[\mathbb{I}_{[j_t = l]}\right].
\end{aligned} \tag{50}$$

We will estimate the terms on the right hand side of the above inequality. To this end, denote

$$\begin{aligned}
T_1 &:= \langle \nabla g_{\omega_{j_t}}\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right) - \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right), x_t - x_t^{l,\omega} \rangle, \\
T_2 &:= \langle \nabla g_S(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega} \rangle.
\end{aligned}$$

Taking the expectation w.r.t. A on both sides of the above identity, we have

$$\begin{aligned}
\mathbb{E}_A[T_1] &= \mathbb{E}_A[\langle \nabla g_{\omega_{j_t}}\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right) - \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right), x_t - x_t^{l,\omega} \rangle] \\
&= \mathbb{E}_A[\langle \mathbb{E}_{j_t}[\nabla g_{\omega_{j_t}}\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right)] - \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right), x_t - x_t^{l,\omega} \rangle] \\
&= \mathbb{E}_A[\langle \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right) - \nabla g_S\left(x_t\right) \nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right), x_t - x_t^{l,\omega} \rangle] \\
&= 0,
\end{aligned} \tag{51}$$

where the second identity holds true since j_t is independent of i_t and x_t. Therefore,

$$\begin{aligned}
-2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t \neq l]}] &= -2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t \neq l]}] + 2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t = l]}] - 2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t = l]}] \\
&= (-2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t \neq l]}] - 2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t = l]}]) + 2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t = l]}] \\
&= -2\eta_t \mathbb{E}_A[T_1] + 2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t = l]}] \\
&= 2\eta_t \mathbb{E}_A[T_1 \mathbb{I}_{[j_t = l]}].
\end{aligned} \tag{52}$$

We further get the following estimation

$$
\begin{aligned}
&- 2\eta_t \mathbb{E}_A\left[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega}\rangle \mathbb{I}_{[j_t\neq l]}\right] \\
&= 2\eta_t \mathbb{E}_A\left[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega}\rangle \mathbb{I}_{[j_t= l]}\right] \\
&\leq 2\eta_t \mathbb{E}_A\left[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\|\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right] \\
&\leq 2\eta_t \mathbb{E}_A\left[(\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(g_S(x_t))\| + \|\nabla g_S(x_t)\|\|\nabla f_{\nu_{i_t}}(g_S(x_t))\|)\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right] \\
&\leq 4\eta_t L_g L_f \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right], \qquad (53)
\end{aligned}
$$

where the last inequality holds true due to Assumption 2.2. Similar to estimations of (51) , (52) and (53), one can show that

$$
\begin{aligned}
&- 2\eta_t \mathbb{E}_A[T_2 \mathbb{I}_{[j_t\neq l]}] \\
&= -2\eta_t \mathbb{E}_A[\langle \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \mathbb{I}_{[j_t\neq l]}] \\
&= 2\eta_t \mathbb{E}_A\left[\langle \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \mathbb{I}_{[j_t= l]}\right] \\
&\leq 4\eta_t L_g L_f \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right]. \qquad (54)
\end{aligned}
$$

Substituting (53) and (54) into (50) and noting that C_g represents the empirical variance associated with the gradient of the inner function as given in part (ii) of Assumption 3.1, we obtain

$$
\begin{aligned}
&\mathbb{E}_A\left[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2\right] \\
&\leq \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|^2\right] + 2C_f L_g \eta_t \mathbb{E}_A\left[\|y_{t+1} - g_S(x_t)\|\|x_t - x_t^{l,\omega}\|\right] \\
&\quad + 2C_f L_g \eta_t \mathbb{E}_A\left[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|\|x_t - x_t^{l,\omega}\|\right] \\
&\quad + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\right] \\
&\quad + 16\eta_t^2 L_f^2 C_g + 12\eta_t L_g L_f \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right] + 4\eta_t^2 L_g^2 L_f^2 \mathbb{E}_A\left[\mathbb{I}_{[j_t= l]}\right] \\
&\leq \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|^2\right] + 2C_f L_g \eta_t (\mathbb{E}_A\left[\|y_{t+1} - g_S(x_t)\|^2\right])^{1/2} (\mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|^2\right])^{1/2} \\
&\quad + 2C_f L_g \eta_t (\mathbb{E}_A\left[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\right])^{1/2} (\mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|^2\right])^{1/2} \\
&\quad + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{t+1} - g_S(x_t)\|^2\right] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\right] \\
&\quad + 16\eta_t^2 L_f^2 C_g + 12\eta_t L_g L_f \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right] + 4\eta_t^2 L_g^2 L_f^2 \mathbb{E}_A\left[\mathbb{I}_{[j_t= l]}\right], \qquad (55)
\end{aligned}
$$

where the second inequality holds by the Cauchy-Schwarz inequality. Observe that

$$
\begin{aligned}
\mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\mathbb{I}_{[j_t= l]}\right] &= \mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\mathbb{E}_{j_t}[\mathbb{I}_{[j_t= l]}]\right] \\
&= \frac{1}{m}\mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|\right] \leq \frac{1}{m}(\mathbb{E}_A\left[\|x_t - x_t^{l,\omega}\|^2\right])^{1/2}.
\end{aligned}
$$

Note that $\|x_0 - x_0^{l,\omega}\|^2 = 0$. Combining the above two estimations together implies that

$$
\begin{aligned}
&\mathbb{E}_A\left[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2\right] \leq 2C_f L_g \sum_{i=1}^{t} \eta_i (\mathbb{E}_A\left[\|y_{i+1} - g_S(x_i)\|^2\right])^{1/2} (\mathbb{E}_A\left[\|x_i - x_i^{l,\omega}\|^2\right])^{1/2} \\
&+ 2C_f L_g \sum_{i=1}^{t} \eta_i (\mathbb{E}_A\left[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\right])^{1/2} (\mathbb{E}_A\left[\|x_i - x_i^{l,\omega}\|^2\right])^{1/2} \\
&+ 4\sum_{i=0}^{t} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{i+1} - g_S(x_i)\|^2\right] + 4\sum_{i=0}^{t} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\left[\|y_{i+1}^{l,\omega} - g_S\left(x_i^{l,\omega}\right)\|^2\right] \\
&+ 16L_f^2 C_g \sum_{i=0}^{t} \eta_i^2 + \frac{12L_g L_f}{m} \sum_{i=1}^{t} \eta_i (\mathbb{E}_A\left[\|x_i - x_i^{l,\omega}\|^2\right])^{1/2} + \frac{4L_g^2 L_f^2}{m} \sum_{i=0}^{t} \eta_i^2.
\end{aligned}
$$

Again, for notational convenience, let $u_t = (\mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\|^2\big])^{1/2}$. The above estimation can be rewritten as

$$\begin{aligned}
u_t^2 \leq\ & 2C_f L_g \sum_{i=1}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2} u_i + 2C_f L_g \sum_{i=1}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2} u_i \\
& + 4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1} - g_S\left(x_i\right)\|^2\big] + 4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big] \\
& + 16 L_f^2 C_g \sum_{i=0}^{t-1} \eta_i^2 + \frac{12 L_f L_g}{m} \sum_{i=1}^{t-1} \eta_i u_i + \frac{4 L_g^2 L_f^2}{m} \sum_{i=0}^{t-1} \eta_i^2. \qquad (56)
\end{aligned}$$

We will use Lemma A.2 to get the desired estimation. For this purpose, define

$$\begin{aligned}
S_t =\ & 4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1} - g_S\left(x_i\right)\|^2\big] + 4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big] \\
& + 16 L_f^2 C_g \sum_{i=0}^{t-1} \eta_i^2 + \frac{4 L_g^2 L_f^2}{m} \sum_{i=0}^{t-1} \eta_i^2, \\
\alpha_i =\ & 2C_f L_g \eta_i (\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2} + 2C_f L_g \eta_i (\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2} + \frac{12 L_g L_f}{m} \eta_i.
\end{aligned}$$

Now applying Lemma A.2 with u_t, S_t and α_i define as above to (56), we get

$$\begin{aligned}
u_t \leq\ & \sqrt{S_t} + \sum_{i=1}^{t-1} \alpha_i \\
\leq\ & 2C_f L_g \sum_{i=1}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2} + 2C_f L_g \sum_{i=1}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2} \\
& + (4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1} - g_S\left(x_i\right)\|^2\big])^{1/2} + (4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2} \\
& + (16 L_f^2 C_g \sum_{i=0}^{t-1} \eta_i^2)^{1/2} + \frac{12 L_f L_g}{m} \sum_{i=1}^{t-1} \eta_i + (\frac{4 L_g L_f}{m} \sum_{i=0}^{t-1} \eta_i^2)^{1/2} \\
\leq\ & 4C_f L_g \sum_{i=0}^{t-1} \eta_j (\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2} + 4C_f L_g \sum_{i=0}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2} \\
& + (16 L_f^2 C_g \sum_{i=0}^{t-1} \eta_i^2)^{1/2} + \frac{12 L_g L_f}{m} \sum_{i=0}^{t-1} \eta_i + (\frac{4 L_g^2 L_f^2}{m} \sum_{i=0}^{t-1} \eta_i^2)^{1/2}, \qquad (57)
\end{aligned}$$

where the second inequality uses the fact that $(\sum_{i=1}^4 a_i)^{1/2} \leq \sum_{i=1}^4 (a_i)^{1/2}$ and the last inequality holds by the fact that $\big(4C_f^2 L_g^2 \sum_{i=0}^{t-1} \eta_i^2 \mathbb{E}_A\big[\|y_{i+1} - g_S\left(x_i\right)\|^2\big]\big)^{1/2} \leq 2C_f L_g \sum_{i=0}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2}$ and $(4\sum_{i=0}^{t-1} \eta_i^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2} \leq 2C_f L_g \sum_{i=0}^{t-1} \eta_i (\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2}$.

If $\eta_i = \eta$, note that $\eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{j+1} - g_S(x_i)\|^2])^{1/2} \leq \sup_S \eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2])^{1/2}$ and $\eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2])^{1/2} \leq \sup_S \eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2])^{1/2}$. Consequently, with T iterations,

we further obtain that

$$\begin{aligned}u_T \leq & 8C_f L_g \sup_S \eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2])^{1/2} + (16L_f^2 C_g \sum_{i=0}^{T-1} \eta^2)^{1/2} + \frac{12L_f L_g}{m} \sum_{i=0}^{T-1} \eta \\ & + (\frac{4L_g^2 L_f^2}{m} \sum_{i=0}^{T-1} \eta^2)^{1/2} \\ \leq & 8C_f L_g \sup_S \eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2])^{1/2} + 4L_f \sqrt{C_g} \eta \sqrt{T} + \frac{14 L_g L_f}{m} \eta T.\end{aligned}$$

where the last inequality holds by the fact that $\big(\frac{4L_f^2 L_g^2}{m} \sum_{i=0}^{T-1} \eta^2\big)^{1/2} = \frac{2L_f L_g}{\sqrt{m}} \eta \sqrt{T} \leq \frac{2L_f L_g}{m} \eta T$ because often we have $T \geq m$. Noting that $\mathbb{E}_A\big[\|x_T - x_T^{l,\omega}\|\big] \leq u_T = (\mathbb{E}_A\big[\|x_T - x_T^{l,\omega}\|^2\big])^{1/2}$, we further get

$$\begin{aligned}\mathbb{E}_A\big[\|x_T - x_T^{l,\omega}\|\big] \leq & 8C_f L_g \sup_S \eta \sum_{i=0}^{T-1} (\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2])^{1/2} \\ & + 4L_f \sqrt{C_g} \eta \sqrt{T} + \frac{14 L_f L_g}{m} \eta T. \end{aligned} \tag{58}$$

Equivalently,

$$\mathbb{E}_A\big[\|x_T - x_T^{l,\omega}\|\big] = \mathcal{O}\Big(\frac{L_f L_g}{m} \eta T + L_f \sqrt{C_g} \eta \sqrt{T} + \sup_S C_f L_g \sum_{i=0}^{T-1} \eta \big(\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2]\big)^{\frac{1}{2}}\Big).$$

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

$$\epsilon_\nu + \epsilon_\omega = \mathcal{O}\Big(\frac{L_f L_g}{n} \eta T + \frac{L_f L_g}{m} \eta T + L_f \sqrt{C_g} \eta \sqrt{T} + C_f L_g \sup_S \sum_{j=0}^{T-1} \eta \big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}}\Big). \tag{59}$$

The proof is completed. □

Next we move on to the proof of Corollary 3.5

Proof of Corollary 3.5. Considering the constant step size $\eta_t = \eta$, and with the result of the SCGD update in Lemma A.1, we have

$$\begin{aligned}\epsilon_\nu + \epsilon_\omega &= \mathcal{O}(\eta T n^{-1} + \eta T m^{-1} + \eta T^{\frac{1}{2}} + \eta \sum_{j=1}^{T-1} (j^{-c/2} \beta^{-c/2} + \eta/\beta + \beta^{1/2})) \\ &= \mathcal{O}(\eta T n^{-1} + \eta T m^{-1} + \eta T^{\frac{1}{2}} + \eta T^{-c/2+1} \beta^{-c/2} + \eta^2 \beta^{-1} T + \eta \beta^{1/2} T).\end{aligned}$$

With the result of the SCSC update in Lemma A.1, we have

$$\begin{aligned}\epsilon_\nu + \epsilon_\omega &= \mathcal{O}(\eta T n^{-1} + \eta T m^{-1} + \eta T^{\frac{1}{2}} + \eta \sum_{j=1}^{T-1} (j^{-c/2} \beta^{-c/2} + \eta \beta^{-\frac{1}{2}} + \beta^{1/2})) \\ &= \mathcal{O}(\eta T n^{-1} + \eta T m^{-1} + \eta T^{\frac{1}{2}} + \eta T^{-c/2+1} \beta^{-c/2} + \eta^2 \beta^{-\frac{1}{2}} T + \eta \beta^{1/2} T).\end{aligned}$$

□

C.2. Optimization

Lemma C.1. *Suppose Assumptions 2.2 and 3.1 (iii) holds for the empirical risk F_S, By running Algorithm 1, we have for any $\gamma_t > 0$*

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \leq & \big(1 + \frac{C_f L_g^2 \eta_t}{\gamma_t}\big)\|x_t - x_*^S\|^2 + L_f^2 L_g^2 \eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S)) \\ & + \gamma_t C_f \eta_t \mathbb{E}_A[\|g_S(x_t) - y_{t+1}\|^2|\mathcal{F}_t]. \end{aligned} \tag{60}$$

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

The proof of Lemma C.1 is deferred to the end of this subsection. Now we are ready to prove the convergence of Algorithm 1 for the convex case.

Proof of Theorem 3.6. We first present the proof for the SCGD update. Taking the total expectation with respect to the internal randomness of A on both sides of (60), we get

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2] \leq & \mathbb{E}_A[\|x_t - x_*^S\|^2] + L_f^2 L_g^2 \eta_t^2 - 2\eta_t \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\ & + \gamma_t C_f \eta_t \mathbb{E}_A[\|g_S(x_t) - y_{t+1}\|^2] + \frac{C_f L_g^2 \eta_t}{\gamma_t}\mathbb{E}_A[\|x_t - x_*^S\|^2]. \end{aligned} \tag{61}$$

Setting $\eta_t = \eta, \beta_t = \beta$ and $\gamma_t = \frac{\beta_t}{\eta_t} = \frac{\beta}{\eta}$, plugging Lemma A.1 into (61), we have

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2] \leq & \mathbb{E}_A[\|x_t - x_*^S\|^2] + L_f^2 L_g^2 \eta^2 - 2\eta \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\ & + C_f \beta \left((\frac{c}{e})^c D_y (t\beta)^{-c} + L_g^3 L_f^2 \frac{\eta^2}{\beta^2} + 2V_g \beta \right) + C_f L_g^2 \mathbb{E}_A[\|x_t - x_*^S\|^2]\frac{\eta^2}{\beta}. \end{aligned}$$

Setting $\eta = T^{-a}, \beta = T^{-b}$, telescoping the above inequality for $t = 1, \cdots, T$, and noting that $\mathbb{E}_A[\|x_t - x_*^S\|^2]$ is bounded by D_x, we get

$$\begin{aligned}2\eta \sum_{t=1}^T \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \leq & D_x + L_f^2 L_g^2 \eta^2 T + (\frac{c}{e})^c C_f D_y \beta^{1-c} \sum_{t=1}^T t^{-c} + 2C_f V_g \beta^2 T \\ & + C_f L_f^2 L_g^3 \eta^2 \beta^{-1} T + C_f L_g^2 D_x \eta^2 \beta^{-1} T. \end{aligned} \tag{62}$$

From the choice of $A(S)$ and the convexity of F_S, noting that $\sum_{t=1}^T t^{-z} = \mathcal{O}(T^{1-z})$ for $z \in (0,1) \cup (1, \infty)$ and $\sum_{t=1}^T t^{-1} = \mathcal{O}(\log T)$, as long as $c \neq 1$ we get

$$\begin{aligned}& \mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)] \\ =& \mathcal{O}\Big(D_x (\eta T)^{-1} + L_f^2 L_g^2 \eta + C_f D_y (\beta T)^{1-c} (\eta T)^{-1} + C_f V_g \beta^2 \eta^{-1} + C_f L_f^2 L_g^3 D_x \eta \beta^{-1} \Big). \end{aligned}$$

Then we get the desired result for the SCGD update. Next we present the proof for the SCSC update. Setting $\eta_t = \eta, \beta_t = \beta$ and $\gamma_t = \frac{1}{\sqrt{\beta_t}} = \frac{1}{\sqrt{\beta}}$, plugging Lemma A.1 into (61), we have

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2] \leq & \mathbb{E}_A[\|x_t - x_*^S\|^2] + L_f^2 L_g^2 \eta^2 - 2\eta \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \\ & + C_f \frac{\eta}{\sqrt{\beta}} \left((\frac{c}{e})^c D_y (t\beta)^{-c} + L_g^3 L_f^2 \frac{\eta^2}{\beta} + 2V_g \beta \right) + C_f L_g^2 \mathbb{E}_A[\|x_t - x_*^S\|^2]\eta\sqrt{\beta}. \end{aligned}$$

Setting $\eta = T^{-a}, \beta = T^{-b}$, telescoping the above inequality for $t = 1, \cdots, T$, and noting that $\mathbb{E}_A[\|x_t - x_*^S\|^2]$ is bounded by D_x, we get

$$\begin{aligned}2\eta \sum_{t=1}^T \mathbb{E}_A[F_S(x_t) - F_S(x_*^S)] \leq & D_x + L_f^2 L_g^2 \eta^2 T + (\frac{c}{e})^c C_f D_y \eta \beta^{-\frac{1}{2}-c} \sum_{t=1}^T t^{-c} + 2C_f V_g \eta \beta^{\frac{1}{2}} T \\ & + C_f L_f^2 L_g^3 \eta^3 \beta^{-\frac{3}{2}} T + C_f L_g^2 D_x \eta \beta^{\frac{1}{2}} T. \end{aligned} \tag{63}$$

From the choice of $A(S)$ and the convexity of F_S, noting that $\sum_{t=1}^T t^{-z} = \mathcal{O}(T^{1-z})$ for $z \in (0,1)\cup(1,\infty)$ and $\sum_{t=1}^T t^{-1} = \mathcal{O}(\log T)$, as long as $c > 2$ we get

$$\begin{aligned}&\mathbb{E}_A[F_S(A(S)) - F_S(x_*^S)]\\ =&\mathcal{O}\Big(D_x(\eta T)^{-1} + L_f^2 L_g^2 \eta + C_f D_y(\beta T)^{-c}\beta^{-\frac{1}{2}} + C_f V_g \beta^{\frac{1}{2}} + C_f L_f^2 L_g^3 \eta^2 \beta^{-\frac{3}{2}} + C_f L_g^2 D_x \beta^{\frac{1}{2}}\Big).\end{aligned}$$

We have completed the proof. □

Proof of Lemma C.1. From Algorithm 1 we have

$$\begin{aligned}&\|x_{t+1} - x_*^S\|^2\\ \le&\|x_t - \eta_t \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - x_*^S\|^2\\ =&\|x_t - x_*^S\|^2 + \eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2 - 2\eta_t\langle x_t - x_*^S, \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\rangle\\ =&\|x_t - x_*^S\|^2 + \eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2 - 2\eta_t\langle x_t - x_*^S, \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\rangle + u_t,\end{aligned}$$

where

$$u_t := 2\eta_t\langle x_t - x_*^S, \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\rangle.$$

Let $\mathcal{F}_t$ be the σ-field generated by $\{\omega_{j_0},\dots,\omega_{j_{t-1}},\nu_{i_0},\dots,\nu_{i_{t-1}}\}$. Taking the expectation with respect to the internal randomness of the algorithm and using Assumption 2.2, we have

$$\begin{aligned}&\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t]\\ \le&\|x_t - x_*^S\|^2 + L_f^2 L_g^2\eta_t^2 - 2\eta_t\mathbb{E}_A[\langle x_t - x_*^S, \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\rangle|\mathcal{F}_t] + \mathbb{E}_A[u_t|\mathcal{F}_t]\\ =&\|x_t - x_*^S\|^2 + L_f^2 L_g^2\eta_t^2 - 2\eta_t\langle x_t - x_*^S, \nabla F_S(x_t)\rangle + \mathbb{E}_A[u_t|\mathcal{F}_t]\\ \le&\|x_t - x_*^S\|^2 + L_f^2 L_g^2\eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S)) + \mathbb{E}_A[u_t|\mathcal{F}_t],\end{aligned} \tag{64}$$

where the last inequality comes from the convexity of F_S. From the Cauchy-Schwarz inequality, Young's inequality, Assumption 2.2 (ii) and 3.1 (iii) we have, for all $\gamma_t > 0$, that

$$\begin{aligned}&2\eta_t\langle x_t - x_*^S, \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\rangle\\ \le&2\eta_t\|x_t - x_*^S\|\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla f_{\nu_{i_t}}(y_{t+1})\|\\ \le&2C_f\eta_t\|x_t - x_*^S\|\|\nabla g_{\omega_{j_t}}(x_t)\|\|g_S(x_t) - y_{t+1}\|\\ \le&2C_f\eta_t\Big(\frac{\|x_t - x_*^S\|^2\|\nabla g_{\omega_{j_t}}(x_t)\|^2}{2\gamma_t} + \frac{\gamma_t}{2}\|g_S(x_t) - y_{t+1}\|^2\Big)\\ \le&\frac{C_f L_g^2\eta_t}{\gamma_t}\|x_t - x_*^S\|^2 + \gamma_t C_f\eta_t\|g_S(x_t) - y_{t+1}\|^2.\end{aligned} \tag{65}$$

Substituting (65) into (64), we get

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1} - x_*^S\|^2|\mathcal{F}_t] \le&\Big(1 + \frac{C_f L_g^2\eta_t}{\gamma_t}\Big)\|x_t - x_*^S\|^2 + L_f^2 L_g^2\eta_t^2 - 2\eta_t(F_S(x_t) - F_S(x_*^S))\\ &+ \gamma_t C_f\eta_t\mathbb{E}_A[\|g_S(x_t) - y_{t+1}\|^2|\mathcal{F}_t].\end{aligned} \tag{66}$$

The proof is completed. □

C.3. Excess Generalization

Proof of Theorem 3.7. We first present the proof for the SCGD update. Setting $\eta_t = \eta, \beta_t = \beta$ for $\eta, \beta > 0$, from (39) and (58) we get for all t

$$\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] \le 8C_f L_g \sup_S \eta\sum_{j=0}^{t-1}(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2])^{1/2} + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t. \tag{67}$$

and

$$\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \leq 8C_f L_g \sup_S \eta \sum_{i=0}^{t-1} (\mathbb{E}_A[\|y_{i+1} - g_S(x_i)\|^2])^{1/2} + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{14L_f L_g}{m}\eta t. \tag{68}$$

Plugging Lemma A.1 with SCGD update into (67) and (68), then we have

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] \leq & 8C_f L_g \eta \sum_{j=1}^{t-1} \sqrt{\left(\frac{c}{e}\right)^c D_y (j\beta)^{-c} + L_f L_g^2 \frac{\eta^2}{\beta^2} + 2V_g\beta} \\ & + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + 8C_f L_g D_y \eta.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \leq & 8C_f L_g \eta \sum_{j=1}^{t-1} \sqrt{\left(\frac{c}{e}\right)^c D_y (j\beta)^{-c} + L_f L_g^2 \frac{\eta^2}{\beta^2} + 2V_g\beta} \\ & + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{14L_f L_g}{m}\eta t + 8C_f L_g D_y \eta.\end{aligned}$$

From the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we get

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] \leq & 8C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y}\eta\beta^{-\frac{c}{2}} \sum_{j=1}^{t-1} j^{-\frac{c}{2}} + 8C_f\sqrt{L_f}L_g^2\frac{\eta^2}{\beta}t + 8C_f L_g\sqrt{2V_g}\eta\sqrt{\beta}t \\ & + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + 8C_f L_g D_y\eta.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \leq & 8C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y}\eta\beta^{-\frac{c}{2}} \sum_{j=1}^{t-1} j^{-\frac{c}{2}} + 8C_f\sqrt{L_f}L_g^2\frac{\eta^2}{\beta}t + 8C_f L_g\sqrt{2V_g}\eta\sqrt{\beta}t \\ & + 4L_g\sqrt{C_g}\eta\sqrt{t} + \frac{14L_f L_g}{m}\eta t + 8C_f L_g D_y\eta.\end{aligned}$$

Thus we get

$$\begin{aligned}& \mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] + 4\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \\ \leq & 40C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y}\eta\beta^{-\frac{c}{2}} \sum_{j=1}^{t} j^{-\frac{c}{2}} + 40C_f\sqrt{L_f}L_g^2\frac{\eta^2}{\beta}t + 40C_f L_g\sqrt{2V_g}\eta\sqrt{\beta}t \\ & + 20L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + \frac{56L_f L_g}{m}\eta t + 40C_f L_g D_y\eta.\end{aligned}$$

Using Theorem 2.3, we have

$$\begin{aligned}& \mathbb{E}_{S,A}\left[F(x_t) - F_S(x_t)\right] \\ \leq & 40C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} L_f L_g \eta\beta^{-\frac{c}{2}} \sum_{j=1}^{t} j^{-\frac{c}{2}} + 40C_f\sqrt{L_f}L_f L_g^3\frac{\eta^2}{\beta}t + 40C_f\sqrt{2V_g}L_f L_g^2\eta\sqrt{\beta}t \\ & + 20\sqrt{C_g}L_f^2 L_g\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + \frac{56L_f L_g}{m}\eta t + 40C_f L_g D_y\eta + L_f\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}}.\end{aligned} \tag{69}$$

From (62) we get

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{S,A}[F_S(x_t) - F_S(x_*^S)] \leq & D_x\eta^{-1} + L_f L_g\eta T + (\frac{c}{e})^c C_f D_y \eta^{-1}\beta^{1-c}\sum_{t=1}^{T} t^{-c} + 2C_f V_g\eta^{-1}\beta^2 T \\ & + C_f L_f L_g^2\eta\beta^{-1}T + C_f L_g D_x\eta\beta^{-1}T.\end{aligned} \tag{70}$$

Setting $\eta = T^{-a}$ and $\beta = T^{-b}$ in (69) with $a, b \in (0, 1]$ and telescoping from $t = 1, \ldots, T$, then adding the result with (70), and using the fact $F_S(x_*^S) \leq F_S(x_*)$, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{S,A}[F(x_t) - F(x_*)] \\
\leq& 40C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} L_f L_g T^{-a+\frac{bc}{2}} \sum_{t=1}^{T}\sum_{j=1}^{t} j^{-\frac{c}{2}} + 40C_f\sqrt{L_f} L_f L_g^3 T^{b-2a}\sum_{t=1}^{T} t \\
&+ 40C_f\sqrt{2V_g} L_f L_g^2 T^{-a-\frac{b}{2}} \sum_{t=1}^{T} t + 20\sqrt{C_g} L_f^2 L_g T^{-a}\sum_{t=1}^{T}\sqrt{t} + \frac{6L_f^2 L_g^2}{n} T^{-a}\sum_{t=1}^{T} t \\
&+ \frac{56L_f^2 L_g^2}{m} T^{-a}\sum_{t=1}^{T} t + 40C_f L_g D_y T^{1-a} + L_f \sum_{t=1}^{T}\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}} \\
&+ D_x T^a + L_f L_g T^{1-a} + (\frac{c}{e})^c C_f D_y T^{-b(1-c)+a}\sum_{t=1}^{T} t^{-c} \\
&+ 2C_f V_g T^{1-2b+a} + C_f L_f L_g^2 T^{1+b-a} + C_f L_g D_x T^{1+b-a}. \qquad (71)
\end{aligned}$$

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

$$\sum_{t=1}^{T}\sum_{j=1}^{T} j^{-\frac{c}{2}} = \mathcal{O}\left(\sum_{t=1}^{T} t^{1-\frac{c}{2}} (\log t)^{\mathbb{I}_{c=2}}\right) = \mathcal{O}\left(T^{2-\frac{c}{2}} (\log T)^{\mathbb{I}_{c=2}}\right).$$

With the same derivation we can get the bounds on other terms on the right hand side of (71). Then we get

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

Dividing both sides of (71) with T, then from the choice of $A(S)$ we get

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

Since $a, b \in (0, 1]$, as long as we have $c > 2$, the dominating terms are

$$\mathcal{O}(T^{1-a-\frac{b}{2}}), \quad \mathcal{O}(T^{1+b-2a}), \quad \mathcal{O}(n^{-1}T^{1-a}), \quad \mathcal{O}(m^{-1}T^{1-a}), \quad \mathcal{O}(T^{a-1}), \quad \mathcal{O}(T^{a-2b}).$$

Setting $a = \frac{6}{7}$ and $b = \frac{4}{7}$ yields

$$\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}(T^{-\frac{1}{7}} + \frac{T^{\frac{1}{7}}}{n} + \frac{T^{\frac{1}{7}}}{m} + \frac{1}{\sqrt{m}}).$$

Setting $T = \mathcal{O}(\max\{n^{3.5}, m^{3.5}\})$ yields the following bound

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

Then we get the desired result for the SCGD update. Next we present the proof for the SCSC update. Plugging Lemma A.1 with SCSC update into (67) and (68), then we have

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] \leq & 8C_f L_g \eta \sum_{j=1}^{t-1} \sqrt{\left(\frac{c}{e}\right)^c D_y (j\beta)^{-c} + L_f L_g^2 \frac{\eta^2}{\beta} + 2V_g \beta} \\ & + 4L_f \sqrt{C_g} \eta \sqrt{t} + \frac{6L_f L_g}{n} \eta t + 8C_f L_g D_y \eta.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \leq & 8C_f L_g \eta \sum_{j=1}^{t-1} \sqrt{\left(\frac{c}{e}\right)^c D_y (j\beta)^{-c} + L_f L_g^2 \frac{\eta^2}{\beta} + 2V_g \beta} \\ & + 4L_f \sqrt{C_g} \eta \sqrt{t} + \frac{14L_f L_g}{m} \eta t + 8C_f L_g D_y \eta.\end{aligned}$$

From the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we get

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] \leq & 8C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} \eta \beta^{-\frac{c}{2}} \sum_{j=1}^{t-1} j^{-\frac{c}{2}} + 8C_f \sqrt{L_f} L_g^2 \frac{\eta^2}{\sqrt{\beta}} t + 8C_f L_g \sqrt{2V_g} \eta \sqrt{\beta} t \\ & + 4L_f \sqrt{C_g} \eta \sqrt{t} + \frac{6L_f L_g}{n} \eta t + 8C_f L_g D_y \eta.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \leq & 8C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} \eta \beta^{-\frac{c}{2}} \sum_{j=1}^{t-1} j^{-\frac{c}{2}} + 8C_f \sqrt{L_f} L_g^2 \frac{\eta^2}{\sqrt{\beta}} t + 8C_f L_g \sqrt{2V_g} \eta \sqrt{\beta} t \\ & + 4L_g \sqrt{C_g} \eta \sqrt{t} + \frac{14L_f L_g}{m} \eta t + 8C_f L_g D_y \eta.\end{aligned}$$

Thus we get

$$\begin{aligned}& \mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] + 4\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \\ \leq & 40C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} \eta \beta^{-\frac{c}{2}} \sum_{j=1}^{t} j^{-\frac{c}{2}} + 40C_f \sqrt{L_f} L_g^2 \frac{\eta^2}{\sqrt{\beta}} t + 40C_f L_g \sqrt{2V_g} \eta \sqrt{\beta} t \\ & + 20L_f \sqrt{C_g} \eta \sqrt{t} + \frac{6L_f L_g}{n} \eta t + \frac{56L_f L_g}{m} \eta t + 40C_f L_g D_y \eta.\end{aligned}$$

Using Theorem 2.3, we have

$$\begin{aligned}& \mathbb{E}_{S,A}\left[F(x_t) - F_S(x_t)\right] \\ \leq & 40C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} L_f L_g \eta \beta^{-\frac{c}{2}} \sum_{j=1}^{t} j^{-\frac{c}{2}} + 40C_f \sqrt{L_f} L_f L_g^3 \frac{\eta^2}{\sqrt{\beta}} t + 40C_f \sqrt{2V_g} L_f L_g^2 \eta \sqrt{\beta} t \\ & + 20\sqrt{C_g} L_f^2 L_g \eta \sqrt{t} + \frac{6L_f L_g}{n} \eta t + \frac{56L_f L_g}{m} \eta t + 40C_f L_g D_y \eta + L_f \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}}. \qquad (72)\end{aligned}$$

From (63) we get

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{S,A}[F_S(x_t) - F_S(x_*^S)] \leq & D_x \eta^{-1} + L_f L_g \eta T + (\frac{c}{e})^c C_f D_y \beta^{-\frac{1}{2}-c} \sum_{t=1}^{T} t^{-c} + 2C_f V_g \beta^{\frac{1}{2}} T \\ & + C_f L_f L_g^2 \eta^2 \beta^{-\frac{3}{2}} T + C_f L_g D_x \beta^{\frac{1}{2}} T. \qquad (73)\end{aligned}$$

Setting $\eta = T^{-a}$ and $\beta = T^{-b}$ in (69) with $a, b \in (0, 1]$ and telescoping from $t = 1, \ldots, T$, then adding the result with (70), and using the fact $F_S(x_*^S) \leq F_S(x_*)$, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{S,A}[F(x_t) - F(x_*)] \leq 40 C_f L_g \sqrt{\left(\frac{c}{e}\right)^c D_y} L_f L_g T^{-a+\frac{bc}{2}} \sum_{t=1}^{T} \sum_{j=1}^{t} j^{-\frac{c}{2}} \\
&+ 40 C_f \sqrt{L_f} L_f L_g^3 T^{\frac{b}{2}-2a} \sum_{t=1}^{T} t + 40 C_f \sqrt{2V_g} L_f L_g^2 T^{-a-\frac{b}{2}} \sum_{t=1}^{T} t \\
&+ 20\sqrt{C_g} L_f^2 L_g T^{-a} \sum_{t=1}^{T} \sqrt{t} + \frac{6L_f^2 L_g^2}{n} T^{-a} \sum_{t=1}^{T} t + \frac{56 L_f^2 L_g^2}{m} T^{-a} \sum_{t=1}^{T} t + 40 C_f L_g D_y T^{1-a} \\
&+ L_f \sum_{t=1}^{T} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}} + D_x T^a + L_f L_g T^{1-a} + (\frac{c}{e})^c C_f D_y T^{b(\frac{1}{2}+c)} \sum_{t=1}^{T} t^{-c} \\
&+ 2 C_f V_g T^{1-\frac{b}{2}} + C_f L_f L_g^2 T^{1+\frac{3}{2}b-2a} + C_f L_g D_x T^{1-\frac{b}{2}}. \qquad (74)
\end{aligned}$$

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

$$\sum_{t=1}^{T} \sum_{j=1}^{T} j^{-\frac{c}{2}} = \mathcal{O}\left(\sum_{t=1}^{T} t^{1-\frac{c}{2}} (\log t)^{\mathbb{I}_{c=2}}\right) = \mathcal{O}\left(T^{2-\frac{c}{2}} (\log T)^{\mathbb{I}_{c=2}}\right).$$

With the same derivation for estimating other terms on the right hand side of (74), we get

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

Dividing both sides of (74) with T, then from the choice of $A(S)$ we get

$$\begin{aligned}
&\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}\Big(T^{1-a-\frac{c(1-b)}{2}} (\log T)^{\mathbb{I}_{c=2}} + T^{1+\frac{b}{2}-2a} + T^{1-a-\frac{b}{2}} + T^{\frac{1}{2}-a} \\
&+ n^{-1} T^{1-a} + m^{-1} T^{1-a} + T^{-a} + m^{-\frac{1}{2}} + T^{a-1} + T^{-a} + T^{-(1-b)c+\frac{b}{2}} (\log T)^{\mathbb{I}_{c=1}} \\
&+ T^{-\frac{b}{2}} + T^{\frac{3}{2}b-2a} + T^{-\frac{b}{2}}\Big). \qquad (75)
\end{aligned}$$

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

$$\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}(T^{-\frac{1}{5}} + \frac{T^{\frac{1}{5}}}{n} + \frac{T^{\frac{1}{5}}}{m} + \frac{1}{\sqrt{m}}). \qquad (76)$$

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

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

Therefore, we get the desired result for the SCSC update. The proof is complete. □

D. Proof for the Strongly Convex Setting

D.1. Stability

Proof of Theorem 3.11. The proof is analogous to the convex case. For any $k \in [n]$, define $S^{k,\nu} = \{\nu_1, ..., \nu_{k-1}, \nu_k', \nu_{k+1}, ..., \nu_n, \omega_1, ..., \omega_m\}$ as formed from S_ν by replacing the k-th element. For any $l \in [m]$, define

$S^{l,\omega} = \{\nu_1, ..., \nu_n, \omega_1, ..., \omega_{l-1}, \omega_l', \omega_{l+1}, ..., \omega_m\}$ as formed from S_ω by replacing the l-th element. Let $\{x_{t+1}\}$ and $\{y_{t+1}\}$ be produced by Algorithm 1 based on S, $\{x_{t+1}^{k,\nu}\}$ and $\{y_{t+1}^{k,\nu}\}$ be produced by Algorithm 1 based on $S^{k,\nu}$, $\{x_{t+1}^{l,\omega}\}$ and $\{y_{t+1}^{l,\omega}\}$ be produced by Algorithm 1 based on $S^{l,\omega}$. Let $x_0 = x_0^{k,\nu}$ and $x_0 = x_0^{l,\omega}$ be starting points in $\mathcal{X}$. Since changing one sample data can happen in either S_ν or S_ω, we need to consider the $\mathbb{E}_A\big[\|x_{t+1} - x_{t+1}^{k,\nu}\|\big]$ and $\mathbb{E}_A\big[\|x_{t+1} - x_{t+1}^{l,\omega}\|\big]$.

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

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

Case 1 ($i_t \neq k$). If $i_t \neq k$, we have

$$\begin{aligned}\|x_{t+1} - x_{t+1}^{k,\nu}\|^2 &\leq \|x_t - \eta_t \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - x_t^{k,\nu} + \eta_t \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2 \\ &= \|x_t - x_t^{k,\nu}\|^2 - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}), x_t - x_t^{k,\nu}\rangle \\ &\quad + \eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2. \end{aligned} \tag{77}$$

Taking the expectation w.r.t. j_t on the both sides of (77) implies that

$$\begin{aligned}&\mathbb{E}_{j_t}\big[\|x_{t+1} - x_{t+1}^{k,\nu}\|^2\big] \\ &\leq \mathbb{E}_{j_t}\big[\|x_t - x_t^{k,\nu}\|^2\big] - 2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}), x_t - x_t^{k,\nu}\rangle\big] \\ &+ \eta_t^2\mathbb{E}_{j_t}\big[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|\big]. \end{aligned} \tag{78}$$

We first estimate the second term on the right hand side of (78). It can be decomposed as

$$\begin{aligned}&-2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}), x_t - x_t^{k,\nu}\rangle\big] \\ &= -2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{k,\nu}\rangle\big] \\ &\quad -2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{k,\nu}\rangle\big] \\ &\quad -2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})), x_t - x_t^{k,\nu}\rangle\big] \\ &\quad -2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})), x_t - x_t^{k,\nu}\rangle\big] \\ &\quad -2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu})) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}), x_t - x_t^{k,\nu}\rangle\big]. \end{aligned} \tag{79}$$

We will estimate the terms on the right hand side of the above equality. Indeed, from part (iv) of Assumption 3.1, we know that $f_\nu(g_S(\cdot))$ is L-smooth. This combined with the strongly convexity of $f_\nu(g_S(\cdot))$ and inequality (6) implied that

$$\begin{aligned}&\langle \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}), x_t - x_t^{k,\nu}\rangle \\ &\geq \frac{L\sigma}{L+\sigma}\|x_t - x_t^{k,\nu}\|^2 + \frac{1}{L+\sigma}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2. \end{aligned} \tag{80}$$

Substituting (27), (29), (30) and (80) into (79), we get that

$$\begin{aligned}&-2\eta_t\mathbb{E}_{j_t}\big[\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu}), x_t - x_t^{k,\nu}\rangle\big] \\ &\leq 2C_f L_g\eta_t\mathbb{E}_{j_t}\big[\|y_{t+1} - g_S(x_t)\|\big]\|x_t - x_t^{k,\nu}\| - \frac{2L\eta_t\sigma}{L+\sigma}\|x_t - x_t^{k,\nu}\|^2 \\ &-2\eta_t\frac{1}{L+\sigma}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2 \\ &+2C_f L_g\eta_t\mathbb{E}_{j_t}\big[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|\big]\|x_t - x_t^{k,\nu}\|. \end{aligned} \tag{81}$$

Furthermore, similar to the argument for (33), we take the expectation w.r.t. j_t of the third term on the right hand side of (77) and then obtain that

$$\begin{aligned}&\mathbb{E}_{j_t}[\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(y_{t+1}^{k,\nu})\|^2] \\ &\leq 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\big[\|y_{t+1} - g_S(x_t)\|^2\big] + 4\eta_t^2 C_f^2 L_g^2\mathbb{E}_{j_t}\big[\|y_{t+1}^{k,\nu} - g_S(x_t^{k,\nu})\|^2\big] + 16\eta_t^2 L_f^2 C_g \\ &+ 4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2. \end{aligned} \tag{82}$$

Putting (81) and (82) back into (78) implies that

$$\begin{aligned}
&\mathbb{E}_{j_t}\left[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\right]\\
&\leq (1-\frac{2L\sigma\eta_t}{L+\sigma})\|x_t-x_t^{k,\nu}\|^2+2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}-g_S(x_t)\|\right]\|x_t-x_t^{k,\nu}\|\\
&+2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|\right]\|x_t-x_t^{k,\nu}\|\\
&+(4\eta_t^2-2\eta\frac{1}{L+\sigma})\|\nabla g_S\left(x_t\right)\nabla f_{\nu_{i_t}}\left(g_S\left(x_t\right)\right)-\nabla g_S(x_t^{k,\nu})\nabla f_{\nu_{i_t}}(g_S(x_t^{k,\nu}))\|^2\\
&+4\eta_t^2C_f^2L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}-g_S\left(x_t\right)\|^2\right]+4\eta_t^2C_f^2L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\right]+16\eta_t^2L_f^2C_g\\
&\leq (1-\frac{2L\sigma\eta_t}{L+\sigma})\|x_t-x_t^{k,\nu}\|^2+2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}-g_S(x_t)\|\right]\|x_t-x_t^{k,\nu}\|\\
&+2C_fL_g\eta_t\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|\right]\|x_t-x_t^{k,\nu}\|\\
&+4\eta_t^2C_f^2L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}-g_S\left(x_t\right)\|^2\right]+4\eta_t^2C_f^2L_g^2\mathbb{E}_{j_t}\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\right]+16\eta_t^2L_f^2C_g,
\end{aligned}$$

where in the second inequality we have used the fact that $\eta_t\leq\frac{1}{2(L+\sigma)}$.

Case 2 ($i_t=k$). If $i_t=k$, in analogy to the argument in (34), we have

$$\mathbb{E}_{j_t}\left[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2\right]\leq\|x_t-x_t^{k,\nu}\|^2+4L_gL_f\eta_t\|x_t-x_t^{k,\nu}\|+4L_g^2L_f^2\eta_t^2. \tag{83}$$

Combining the results of **Case 1** and **Case 2** and taking the expectation w.r.t. A, we have that

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2]\\
&\leq\big(1-2\eta_t\frac{L\sigma}{L+\sigma}+\frac{2\eta_tL\sigma}{n(L+\sigma)}\big)\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|^2\right]\\
&+2C_fL_g\eta_t(\mathbb{E}_A\left[\|y_{t+1}-g_S(x_t)\|^2\right])^{1/2}\big(\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|^2\right]\big)^{1/2}\\
&+2C_fL_g\eta_t(\mathbb{E}_A\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\right])^{1/2}\big(\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|^2\right]\big)^{1/2}\\
&+4\eta_t^2C_f^2L_g^2\mathbb{E}_A\left[\|y_{t+1}-g_S\left(x_t\right)\|^2\right]+4\eta_t^2C_f^2L_g^2\mathbb{E}_A\left[\|y_{t+1}^{k,\nu}-g_S(x_t^{k,\nu})\|^2\right]\\
&+16\eta_t^2L_f^2C_g+4\eta_tL_gL_f\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|\mathbb{I}_{[i_t=k]}\right]+4\eta_t^2L_f^2L_g^2\mathbb{E}_A\left[\mathbb{I}_{[i_t=k]}\right].
\end{aligned}\tag{84}$$

Note that $\eta_t\frac{L\sigma}{L+\sigma}\geq\frac{2\eta_tL\sigma}{n(L+\sigma)}$ as $n\geq2$. We further get that $1-2\eta_t\frac{L\sigma}{L+\sigma}+\frac{2\eta_tL\sigma}{n(L+\sigma)}\leq1-\eta_t\frac{L\sigma}{L+\sigma}$. Observe that $\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|\mathbb{I}_{[i_t=k]}\right]=\frac{1}{n}\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|\right]\leq\frac{1}{n}(\mathbb{E}_A\left[\|x_t-x_t^{k,\nu}\|^2\right])^{1/2}$. If $\eta_t=\eta$, combining the above observations with (84) implies that

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1}-x_{t+1}^{k,\nu}\|^2]\\
&\leq2C_fL_g\sum_{j=1}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta(\mathbb{E}_A\left[\|y_{j+1}-g_S(x_j)\|^2\right])^{1/2}\big(\mathbb{E}_A\left[\|x_j-x_j^{k,\nu}\|^2\right]\big)^{1/2}\\
&+2C_fL_g\sum_{j=1}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta(\mathbb{E}_A\left[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\right])^{1/2}\big(\mathbb{E}_A\left[\|x_j-x_j^{k,\nu}\|^2\right]\big)^{1/2}\\
&+4C_f^2L_g^2\sum_{j=0}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta^2\mathbb{E}_A\left[\|y_{j+1}-g_S(x_j)\|^2\right]\\
&+4C_f^2L_g^2\sum_{j=0}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta^2\mathbb{E}_A\left[\|y_{j+1}^{k,\nu}-g_S(x_j^{k,\nu})\|^2\right]\\
&+16L_f^2C_g\sum_{j=0}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta^2+\frac{4L_gL_f}{n}\sum_{j=1}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta(\mathbb{E}_A\left[\|x_j-x_j^{k,\nu}\|^2\right])^{1/2}\\
&+\frac{4L_f^2L_g^2}{n}\sum_{j=0}^{t}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta^2.
\end{aligned}\tag{85}$$

Again, for notatioanl convenience, let $u_t = (\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|^2])^{1/2}$. The above estimation can be equivalently rewritten as

$$
\begin{aligned}
u_t^2 \leq\ & 2C_f L_g \sum_{j=1}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta(\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big])^{1/2}u_j \\
& + 2C_f L_g \sum_{j=1}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} - g_S(x_j^{k,\nu})\|^2\big])^{1/2}u_j \\
& + 4C_f^2 L_g^2 \sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta^2\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big] \\
& + 4C_f^2 L_g^2 \sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta^2\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} - g_S(x_j^{k,\nu})\|^2\big] + 16L_f^2 C_g\eta^2\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1} \\
& + \frac{4L_g L_f}{n}\eta\sum_{j=1}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}u_j + \frac{4L_f^2 L_g^2}{n}\eta^2\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}. \qquad (86)
\end{aligned}
$$

Note that $16L_f^2 C_g\eta^2\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1} \leq 16L_f^2 C_g\eta^2\frac{L+\sigma}{L\eta\sigma} = 16L_f^2 C_g\frac{L+\sigma}{L\sigma}\eta$ and $\frac{4L_f^2L_g^2}{n}\eta^2\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1} \leq \frac{4L_f^2L_g^2}{n}\eta^2\frac{L+\sigma}{L\eta\sigma} = \frac{4L_f^2L_g^2}{n}\frac{L+\sigma}{L\sigma}\eta$. Furthermore, define

$$
\begin{aligned}
S_t =\ & 4C_f^2 L_g^2 \sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta^2\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big] \\
& + 4C_f^2 L_g^2 \sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta^2\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} - g_S(x_j^{k,\nu})\|^2\big] + 16L_f^2 C_g\frac{L+\sigma}{L\sigma}\eta + \frac{4L_f^2L_g^2}{n}\frac{L+\sigma}{L\sigma}\eta, \\
\alpha_j =\ & 2C_f L_g(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta(\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big])^{1/2} \\
& + 2C_f L_g(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} - g_S(x_j^{k,\nu})\|^2\big])^{1/2} + \frac{4L_g L_f}{n}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j}\eta.
\end{aligned}
$$

Now applying Lemma A.2 with u_t, S_t and α_j defined above to (86), we get

$$
\begin{aligned}
u_t \leq\ & \sqrt{S_t} + \sum_{j=1}^{t-1}\alpha_j \\
\leq\ & 2C_f L_g(\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta^2\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big])^{1/2} \\
& + 2C_f L_g(\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta^2\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} - g_S(x_j^{k,\nu})\|^2\big])^{1/2} \\
& + 2C_f L_g\sum_{j=1}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta(\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big])^{1/2} \\
& + 2C_f L_g\sum_{j=1}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\eta(\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} - g_S(x_j^{k,\nu})\|^2\big])^{1/2} + 4L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta} \\
& + 2L_g L_f\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} + \frac{4L_g L_f(L+\sigma)}{nL\sigma}
\end{aligned}
$$

where the last inequality uses the fact that $(\sum_{i=1}^4 a_i)^{1/2} \leq \sum_{i=1}^4 (a_i)^{1/2}$ and we use the fact that $\frac{4L_g L_f}{n}\eta\sum_{j=1}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1} \leq \frac{4L_g L_f(L+\sigma)}{nL\sigma}$. Note that $\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big] \leq \sup_S \mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]$ and $\mathbb{E}_A\big[\|y_{j+1}^{k,\nu} -$

$g_S(x_j^{k,\nu})\|^2\big] \leq \sup_S \mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]$. Consequently, with T iterations, since $\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|\big] \leq u_T = (\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|^2\big])^{1/2}$, we further obtain

$$\begin{aligned}
\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|\big] &\leq 4C_f L_g \eta \sup_S \Big(\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big]\Big)^{1/2} \\
&+ 4C_f L_g \eta \sup_S \sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{1/2} + 4L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta} \\
&+ 2L_f L_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} + \frac{4L_g L_f(L+\sigma)}{nL\sigma}. \qquad (87)
\end{aligned}$$

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

Likewise, we will estimate $\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|]$ by considering two cases, i.e., $j_t \neq l$ and $j_t = l$.

Case 1 ($j_t \neq l$). If $j_t \neq l$, we have

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - \eta_t \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - x_t^{l,\omega} + \eta_t \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|^2 \\
&= \|x_t - x_t^{l,\omega}\|^2 - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega}), x_t - x_t^{l,\omega}\rangle \\
&+ \eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|^2. \qquad (88)
\end{aligned}$$

We first estimate the second term on the right hand side of (88). It can be decomposed as

$$\begin{aligned}
&-2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega}), x_t - x_t^{l,\omega}\rangle \\
&= -2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)(\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla f_{\nu_{i_t}}(g_S(x_t))), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\quad - 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t^{l,\omega})(\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})), x_t - x_t^{l,\omega}\rangle. \qquad (89)
\end{aligned}$$

From the strongly convexity of $f_\nu(g_S(\cdot))$, part (iv) of Assumption 3.1 and inequality (6), we have

$$\begin{aligned}
&\langle \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle \\
&\geq \frac{L\sigma}{L+\sigma}\|x_t - x_t^{l,\omega}\|^2 + \frac{1}{L+\sigma}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2. \qquad (90)
\end{aligned}$$

Plugging (43), (44) and (90) into (89) implies that

$$\begin{aligned}
&-2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega}), x_t - x_t^{l,\omega}\rangle \\
&\leq 2\eta_t C_f L_g\|y_{t+1} - g_S(x_t)\|\|x_t - x_t^{l,\omega}\| + 2\eta_t C_f L_g\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|\|x_t - x_t^{l,\omega}\| \\
&- 2\eta_t\langle \nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega}\rangle \\
&- 2\eta_t\frac{L\sigma}{L+\sigma}\|x_t - x_t^{l,\omega}\|^2 - 2\eta_t\frac{1}{L}\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2 \\
&- 2\eta_t\langle \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega}\rangle. \qquad (91)
\end{aligned}$$

Next we estimate the last term on the right hand side of (88). Using arguments similar to that for (46), we have

$$\begin{aligned}
&\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1}) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(y_{t+1}^{l,\omega})\|^2 \\
&\leq 4\eta_t^2 C_f^2 L_g^2\|y_{t+1} - g_S(x_t)\|^2 + 4L_g^2\eta_t^2 C_f^2\|g_S(x_t^{l,\omega}) - y_{t+1}^{l,\omega}\|^2 \\
&+ 8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t) - \nabla g_S(x_t)\|^2 + 8L_f^2\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t^{l,\omega}) - \nabla g_S(x_t^{l,\omega})\|^2 \\
&+ 4\eta_t^2\|\nabla g_S(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t^{l,\omega})\nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega}))\|^2. \qquad (92)
\end{aligned}$$

Putting (91) and (92) into (88) and noting that $\eta_t \leq \frac{1}{2(L+\sigma)}$, we get

$$\begin{aligned}
&\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq (1 - \frac{2L\sigma\eta_t}{L+\sigma})\|x_t - x_t^{l,\omega}\|^2 + 2\eta_t C_f L_g \|y_{t+1} - g_S(x_t)\| \|x_t - x_t^{l,\omega}\| \\
&+ 2\eta_t C_f L_g \|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\| \|x_t - x_t^{l,\omega}\| \\
&- 2\eta_t \langle \nabla g_{\omega_{j_t}}(x_t) \nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t) \nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega} \rangle \\
&- 2\eta_t \langle \nabla g_S(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega} \rangle. \\
&+ 4\eta_t^2 C_f^2 L_g^2 \|y_{t+1} - g_S(x_t)\|^2 + 4\eta_t^2 C_f^2 L_g^2 \|g_S(x_t^{l,\omega}) - y_{t+1}^{l,\omega}\|^2 \\
&+ 8L_f^2 \eta_t^2 \|\nabla g_{\omega_{j_t}}(x_t) - \nabla g_S(x_t)\|^2 + 8L_f^2 \eta_t^2 \|\nabla g_{\omega_{j_t}}(x_t^{l,\omega}) - \nabla g_S(x_t^{l,\omega})\|^2. \qquad (93)
\end{aligned}$$

Case 2 ($j_t = l$). If $j_t = l$, using the argument similar to (48), it is easy to see that

$$\|x_{t+1} - x_{t+1}^{l,\omega}\|^2 \leq \|x_t - x_t^{l,\omega}\|^2 + 4L_g L_f \eta_t \|x_t - x_t^{l,\omega}\| + 4\eta_t^2 L_g^2 L_f^2. \qquad (94)$$

Combining **Case 1** and **Case 2** and taking the expectation w.r.t. A on both sides and together with part (ii) of Assumption 3.1 , we have

$$\begin{aligned}
&\mathbb{E}_A\big[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2\big] \\
&\leq (1 - \frac{2L\sigma\eta_t}{L+\sigma} + \frac{2\eta_t L\sigma}{m(L+\sigma)})\mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\|^2\big] + 2C_f L_g \eta_t \mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\| \|x_t - x_t^{l,\omega}\|\big] \\
&+ 2C_f L_g \eta_t \mathbb{E}_A\big[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\| \|x_t - x_t^{l,\omega}\|\big] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\|^2\big] \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\big] + 16C_g L_f^2 \eta_t^2 \\
&- 2\eta_t \mathbb{E}_A\big[\langle \nabla g_{\omega_{j_t}}(x_t) \nabla f_{\nu_{i_t}}(g_S(x_t)) - \nabla g_S(x_t) \nabla f_{\nu_{i_t}}(g_S(x_t)), x_t - x_t^{l,\omega} \rangle \mathbb{I}_{[j_t \neq l]}\big] \\
&- 2\eta_t \mathbb{E}_A\big[\langle \nabla g_S(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})) - \nabla g_{\omega_{j_t}}(x_t^{l,\omega}) \nabla f_{\nu_{i_t}}(g_S(x_t^{l,\omega})), x_t - x_t^{l,\omega} \rangle \mathbb{I}_{[j_t \neq l]}\big] \\
&+ 4\eta_t L_f L_g \mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\| \mathbb{I}_{[j_t = l]}\big] + 4\eta_t^2 L_g^2 L_f^2 \mathbb{E}_A\big[\mathbb{I}_{[j_t = l]}\big]. \qquad (95)
\end{aligned}$$

Note that $\eta_t \frac{L\sigma}{L+\sigma} \geq \frac{2\eta_t L\sigma}{m(L+\sigma)}$ as $m \geq 2$. We further get that $1 - 2\eta_t \frac{L\sigma}{L+\sigma} + \frac{2\eta_t L\sigma}{m(L+\sigma)} \leq 1 - \eta_t \frac{L\sigma}{L+\sigma}$. Plugging (53) and (54) into (95) implies that

$$\begin{aligned}
&\mathbb{E}_A\big[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2\big] \\
&\leq (1 - \frac{L\sigma\eta_t}{L+\sigma})\mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\|^2\big] + 2C_f L_g \eta_t \mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\| \|x_t - x_t^{l,\omega}\|\big] \\
&+ 2C_f L_g \eta_t \mathbb{E}_A\big[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\| \|x_t - x_t^{l,\omega}\|\big] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\|^2\big] \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\big] + 16C_g L_f^2 \eta_t^2 \\
&+ 12\eta_t L_f L_g \mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\| \mathbb{I}_{[j_t = l]}\big] + 4\eta_t^2 L_g^2 L_f^2 \mathbb{E}_A\big[\mathbb{I}_{[j_t = l]}\big] \\
&\leq (1 - \frac{L\sigma\eta_t}{L+\sigma})\mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\|^2\big] + 2C_f L_g \eta_t (\mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\|^2\big])^{1/2} (\mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\|^2\big])^{1/2} \\
&+ 2C_f L_g \eta_t (\mathbb{E}_A\big[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\big])^{1/2} (\mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\|^2\big])^{1/2} \\
&+ 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1} - g_S(x_t)\|^2\big] + 4\eta_t^2 C_f^2 L_g^2 \mathbb{E}_A\big[\|y_{t+1}^{l,\omega} - g_S(x_t^{l,\omega})\|^2\big] + 16C_g L_f^2 \eta_t^2 \\
&+ 12\eta_t L_f L_g \mathbb{E}_A\big[\|x_t - x_t^{l,\omega}\| \mathbb{I}_{[j_t = l]}\big] + 4\eta_t^2 L_g^2 L_f^2 \mathbb{E}_A\big[\mathbb{I}_{[j_t = l]}\big],
\end{aligned}$$

where the second inequality holds by the Cauchy-Schwarz inequality. In addition, observe that

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

If $\eta_t = \eta$, using the above observations, noting $\|x_0 - x_0^{l,\omega}\|^2 = 0$, we can obtain

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1} - x_{t+1}^{l,\omega}\|^2] \\
&\leq 2C_f L_g \sum_{i=1}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta(\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2}\big(\mathbb{E}_A\big[\|x_i - x_i^{l,\omega}\|^2\big]\big)^{1/2} \\
&+ 2C_f L_g \sum_{i=1}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta(\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2}\big(\mathbb{E}_A\big[\|x_i - x_i^{l,\omega}\|^2\big]\big)^{1/2} \\
&+ 4C_f^2 L_g^2 \sum_{i=0}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta^2\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big] \\
&+ 4C_f^2 L_g^2 \sum_{i=0}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta^2\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big] \\
&+ 16L_f^2 C_g \sum_{i=0}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta^2 + \frac{12L_g L_f}{m}\sum_{i=1}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta(\mathbb{E}_A\big[\|x_i - x_i^{l,\omega}\|^2\big])^{1/2} \\
&+ \frac{4L_f^2 L_g^2}{m}\sum_{i=0}^{t}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i}\eta^2. \qquad (96)
\end{aligned}$$

For notional convenience, let $u_t = (\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|^2])^{1/2}$. Therefore, (96) can be equivalently rewritten as

$$\begin{aligned}
u_t^2 &\leq 2C_f L_g \sum_{i=1}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big])^{1/2}u_i \\
&+ 2C_f L_g \sum_{i=1}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big])^{1/2}u_i \\
&+ 4C_f^2 L_g^2 \sum_{i=0}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2\mathbb{E}_A\big[\|y_{i+1} - g_S(x_i)\|^2\big] \\
&+ 4C_f^2 L_g^2 \sum_{i=0}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2\mathbb{E}_A\big[\|y_{i+1}^{l,\omega} - g_S(x_i^{l,\omega})\|^2\big] + 16L_f^2 C_g \sum_{i=0}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2 \\
&+ \frac{12L_g L_f}{m}\sum_{i=1}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta u_i + \frac{4L_f^2 L_g^2}{m}\sum_{i=0}^{t-1}(1 - \frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2. \qquad (97)
\end{aligned}$$

We will use Lemma A.2 to get the desired result. To this end, notice that

$$\begin{aligned}
16L_f^2 C_g\eta^2\sum_{i=0}^{t-1}(1 - \eta\frac{L\sigma}{L+\sigma})^{t-i-1} &\leq 16L_f^2 C_g\eta^2\frac{L+\sigma}{L\eta\sigma} = 16L_f^2 C_g\frac{L+\sigma}{L\sigma}\eta, \\
\frac{4L_f^2 L_g^2}{m}\eta^2\sum_{i=0}^{t-1}(1 - \eta\frac{L\sigma}{L+\sigma})^{t-i-1} &\leq \frac{4L_f^2 L_g^2}{m}\eta^2\frac{L+\sigma}{L\eta\sigma} = \frac{4L_f^2 L_g^2}{m}\frac{L+\sigma}{L\sigma}\eta.
\end{aligned}$$

Moreover, we define

$$
\begin{aligned}
S_t &= 4C_f^2L_g^2\sum_{i=0}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big] \\
&+ 4C_f^2L_g^2\sum_{i=0}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2\mathbb{E}_A\big[\|y_{i+1}^{l,\omega}-g_S(x_i^{l,\omega})\|^2\big] \\
&+ 16L_f^2C_g\frac{L+\sigma}{L\sigma}\eta+\frac{4L_f^2L_g^2}{m}\frac{L+\sigma}{L\sigma}\eta, \\
\alpha_i &= 2C_fL_g(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big])^{1/2} \\
&+ 2C_fL_g(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1}^{l,\omega}-g_S(x_i^{l,\omega})\|^2\big])^{1/2}+\frac{12L_gL_f}{m}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta.
\end{aligned}
$$

Applying Lemma A.2 with u_t, S_t and α_i defined as above to (97), we get

$$
\begin{aligned}
u_t &\le \sqrt{S_t}+\sum_{i=1}^{t-1}\alpha_i \\
&\le (4C_f^2L_g^2\sum_{i=0}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big])^{1/2} \\
&+ (4C_f^2L_g^2\sum_{i=0}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta^2\mathbb{E}_A\big[\|y_{i+1}^{l,\omega}-g_S(x_i^{l,\omega})\|^2\big])^{1/2} \\
&+ 2C_fL_g\sum_{i=1}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big])^{1/2} \\
&+ 2C_fL_g\sum_{i=1}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1}^{l,\omega}-g_S(x_i^{l,\omega})\|^2\big])^{1/2} \\
&+ 4L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta}+2L_fL_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}}+\frac{12L_gL_f(L+\sigma)}{mL\sigma},
\end{aligned}
$$

where we have used the fact that $(\sum_{i=1}^4 a_i)^{1/2}\le\sum_{i=1}^4(a_i)^{1/2}$ and $\frac{12L_gL_f}{m}\sum_{i=0}^{t-1}(1-\frac{L\sigma\eta}{L+\sigma})^{t-i-1}\eta\le\frac{12L_gL_f(L+\sigma)}{mL\sigma}$.

Note that $\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big]\le\sup_S\mathbb{E}_A[\|y_{i+1}-g_S(x_i)\|^2]$ and $\mathbb{E}_A\big[\|y_{i+1}^{l,\omega}-g_S(x_i^{l,\omega})\|^2\big]\le\sup_S\mathbb{E}_A[\|y_{i+1}-g_S(x_i)\|^2]$. Consequently, with T iterations, since $\mathbb{E}_A\big[\|x_T-x_T^{l,\omega}\|\big]\le u_T=(\mathbb{E}_A\big[\|x_T-x_T^{l,\omega}\|^2\big])^{1/2}$, we further obtain

$$
\begin{aligned}
&\mathbb{E}_A\big[\|x_T-x_T^{l,\omega}\|\big] \\
&\le 4C_fL_g\eta\sup_S(\sum_{i=0}^{T-1}(1-\frac{L\sigma\eta}{L+\sigma})^{T-i-1}\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big])^{1/2} \\
&+ 4C_fL_g\eta\sup_S\sum_{i=0}^{T-1}(1-\frac{L\sigma\eta}{L+\sigma})^{T-i-1}\eta(\mathbb{E}_A\big[\|y_{i+1}-g_S(x_i)\|^2\big])^{1/2} \\
&+ 4L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta}+2L_fL_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}}+\frac{12L_gL_f(L+\sigma)}{mL\sigma}.
\end{aligned}
\tag{98}
$$

Combining the estimations for $\mathbb{E}_A\big[\|x_T - x_T^{k,\nu}\|\big]$ and $\mathbb{E}_A\big[\|x_T - x_T^{l,\omega}\|\big]$, we obtain

$$\begin{aligned}
\epsilon_\nu + \epsilon_\omega \leq\ & 8C_f L_g \eta \sup_S \Big(\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big]\Big)^{1/2} \\
& + 8C_f L_g \eta \sup_S \sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{1/2} \\
& + 8L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta} + 2L_f L_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} + \frac{4L_g L_f(L+\sigma)}{nL\sigma} \\
& + 2L_f L_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} + \frac{12L_g L_f(L+\sigma)}{mL\sigma} \\
\leq\ & 16C_f L_g \eta \sup_S \sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}} \\
& + 8L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta} + 2L_f L_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} + \frac{4L_g L_f(L+\sigma)}{nL\sigma} \\
& + 2L_f L_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} + \frac{12L_g L_f(L+\sigma)}{mL\sigma}. \qquad (99)
\end{aligned}$$

Next we will verify why the second inequality of (99) holds true. With the result of SCGD update in Lemma A.1, we have

$$\begin{aligned}
& \eta\Big(\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\mathbb{E}_A\big[\|y_{j+1} - g_S(x_j)\|^2\big]\Big)^{1/2} \\
& \leq \eta\Big(\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}((\frac{c}{e})^c(j\beta)^{-c}\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2] + L_f^2 L_g^3\frac{\eta^2}{\beta^2} + 2V_g\beta)\Big)^{1/2} \\
& \leq \eta\Big(\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}(L_f^2 L_g^3\frac{\eta^2}{\beta^2} + 2V_g\beta)\Big)^{1/2} + \eta((\frac{c}{e})^c D_y\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}(j\beta)^{-c})^{1/2} \\
& \leq \frac{L_f L_g\sqrt{L_g(L+\sigma)}}{\sqrt{L\sigma}}\frac{\eta^{3/2}}{\beta} + \sqrt{\frac{2V_g(L+\sigma)}{L\sigma}}\sqrt{\eta\beta} + (\frac{c}{e})^{\frac{c}{2}}\sqrt{D_y}\frac{\sqrt{(L+\sigma)\eta}}{\sqrt{L\sigma}}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}, \qquad (100)
\end{aligned}$$

where the last inequality holds by the fact that $\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1} \leq \frac{L+\sigma}{\eta L\sigma}$ and Lemma A.4. To see this, $(\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}(j\beta)^{-c})^{1/2} \leq (\frac{\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\sum_{j=1}^{T-1}(j\beta)^{-c}}{T})^{1/2} \leq (\frac{T^{-c+1}\beta^{-c}(L+\sigma)}{T\eta L\sigma})^{1/2} = \frac{T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}\sqrt{L+\sigma}}{\sqrt{\eta L\sigma}}$. And

$$\begin{aligned}
& \eta\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\big(\mathbb{E}_A[\|y_{j+1} - g_S(x_j)\|^2]\big)^{\frac{1}{2}} \\
& \leq \eta\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}((\frac{c}{e})^c(j\beta)^{-c}\mathbb{E}_A[\|y_1 - g_S(x_0)\|^2] + L_f^2 L_g^3\frac{\eta^2}{\beta^2} + 2V_g\beta)^{1/2} \\
& \leq \eta\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}(\sqrt{L_g}L_g L_f\frac{\eta}{\beta} + \sqrt{2V_g}\sqrt{\beta}) + (\frac{c}{e})^{\frac{c}{2}}\sqrt{D_y}\eta\sum_{j=1}^{T-1}(1-\eta\sigma)^{T-j-1}(j\beta)^{-\frac{c}{2}} \\
& \leq \frac{\sqrt{L_g}L_g L_f(L+\sigma)}{L\sigma}\frac{\eta}{\beta} + \frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta} + (\frac{c}{e})^{\frac{c}{2}}\frac{\sqrt{D_y}(L+\sigma)}{L\sigma}T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}, \qquad (101)
\end{aligned}$$

where the last inequality holds by the fact that $\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1} \leq \frac{L+\sigma}{\eta L\sigma}$ and Lemma A.4. To see this $\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}(j\beta)^{-\frac{c}{2}} \leq \frac{\sum_{j=1}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\sum_{j=1}^{T-1}(j\beta)^{-\frac{c}{2}}}{T} \leq \frac{T^{-\frac{c}{2}}\beta^{-\frac{c}{2}}(L+\sigma)}{\eta L\sigma}$. Comparing the result (100) and (101), the dominating terms are (101). We can show that with result of SCSC update in Lemma A.1, the dominating term is

$\eta\sum_{j=0}^{T-1}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j-1}\big(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2]\big)^{\frac{1}{2}}$.

Since often we have $\eta \leq \min(\frac{1}{n},\frac{1}{m})$, then $\frac{\sqrt{\eta}}{\sqrt{n}} \leq \frac{1}{n}$. Consequently, we get that $\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} \leq \frac{(L+\sigma)}{nL\sigma}$. And $\frac{\sqrt{\eta}}{\sqrt{m}} \leq \frac{1}{m}$, $\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} \leq \frac{(L+\sigma)}{mL\sigma}$. We further get the final stability result for σ-strongly convex setting which holds for SCGD and SCSC in Theorem 3.11

$$\epsilon_\nu+\epsilon_\omega = \mathcal{O}\Big(\frac{L_gL_f(L+\sigma)}{\sigma Lm}+\frac{L_gL_f(L+\sigma)}{\sigma Ln}+\frac{L_f\sqrt{C_g(L+\sigma)\eta}}{\sqrt{\sigma L}} + C_fL_g\eta\sup_S\sum_{j=1}^{T}(1-\eta\frac{L\sigma}{L+\sigma})^{T-j}\big(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2]\big)^{\frac{1}{2}}\Big). \tag{102}$$

This completes the proof. □

Next we move on to the Corollary 3.13

Proof of Corollary 3.13. Putting the result (101) to (102), we get stability result of SCGD for strongly convex problems

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

With SCSC update in Lemma A.1, with a same progress, we have stability result of SCSC for strongly convex problems

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

□

D.2. Optimization

Lemma D.1. *Suppose Assumptions 2.2 (ii) and 3.1 (iii) holds and F_S is σ-strongly convex. By running Algorithm 1, we have for any $x \in \mathcal{X}$*

$$\mathbb{E}_A[\|x_{t+1}-x\|^2|\mathcal{F}_t] \leq (1-\frac{\sigma\eta_t}{2})\|x_t-x\|^2+\eta_t^2\mathbb{E}_A[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2|\mathcal{F}_t] - 2\eta_t(F_S(x_t)-F_S(x))+2C_f^2L_g^2\frac{\eta_t}{\sigma}\mathbb{E}_A[\|g_S(x_t)-y_{t+1}\|^2|\mathcal{F}_t]. \tag{103}$$

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 $\{\omega_{j_0},\ldots,\omega_{j_{t-1}},\nu_{i_0},\ldots,\nu_{i_{t-1}}\}$.

The proof of Lemma D.1 is deferred to the end of this subsection. Now we are ready to prove the convergence of Algorithm 1 for strongly convex problems.

Proof of Theorem 3.14. We first present the proof for the SCGD update. Taking full expectation over (103) with $x=x_*^S$ and using Assumption 2.2, we get

$$\mathbb{E}_A[\|x_{t+1}-x_*^S\|^2] \leq (1-\frac{\sigma\eta_t}{2})\mathbb{E}_A[\|x_t-x_*^S\|^2]+L_f^2L_g^2\eta_t^2-2\eta_t\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)] + 2C_f^2L_g^2\frac{\eta_t}{\sigma}\mathbb{E}_A[\|g_S(x_t)-y_{t+1}\|^2]. \tag{104}$$

Setting $\eta_t=\eta$ and $\beta_t=\beta$, plugging Lemma A.1 into (104), and letting $D_y := \mathbb{E}_A[\|y_1-g_S(x_0)\|^2]$, we have

$$\mathbb{E}_A[\|x_{t+1}-x_*^S\|^2] \leq (1-\frac{\sigma\eta}{2})\mathbb{E}_A[\|x_t-x_*^S\|^2]+L_f^2L_g^2\eta^2-2\eta\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)] + \frac{2C_f^2L_g^2\eta}{\sigma}\left((\frac{c}{e})^cD_y(t\beta)^{-c}+L_g^3L_f^2\frac{\eta^2}{\beta^2}+2V_g\beta\right).$$

Multiplying the above inequality with $\left(1-\frac{\sigma\eta}{2}\right)^{T-t}$ and telescoping for $t=1,\ldots,T$, we get

$$\begin{aligned}
&2\eta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)]\\
\leq&\left(1-\frac{\sigma\eta}{2}\right)^{T}\mathbb{E}_A[\|x_1-x_*^S\|^2]+L_f^2L_g^2\eta^2\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\\
&+\frac{2C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\eta\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}\\
&+\frac{4C_f^2L_g^2V_g}{\sigma}\eta\beta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^3}{\beta^2}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}.
\end{aligned}$$

Note that we have

$$\mathbb{E}_A[\|x_1-x_*^S\|^2]\leq\mathbb{E}_A[\|x_0-x_*^S-\eta\nabla g_{\omega_{j_0}}(x_0)\nabla f_{\nu_{i_0}}(y_1)\|^2]\leq 2\mathbb{E}_A[\|x_0-x_*^S\|^2]+2L_f^2L_g^2\eta_t^2$$

Combining the above two inequalities yields

$$\begin{aligned}
&2\eta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)]\\
\leq&2\left(1-\frac{\sigma\eta}{2}\right)^{T}\mathbb{E}_A[\|x_0-x_*^S\|^2]+2L_f^2L_g^2\eta^2\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\\
&+\frac{2C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\eta\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}\\
&+\frac{4C_f^2L_g^2V_g}{\sigma}\eta\beta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^3}{\beta^2}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}.
\end{aligned}$$

From Lemma A.3 we know $(1-\frac{\sigma\eta}{2})^T\leq\exp(-\frac{\sigma\eta T}{2})\leq(\frac{2c}{e\sigma})^c(\eta T)^{-c}$. Also we have $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t}=\frac{1-(1-\frac{\sigma\eta}{2})^{T-1}}{1-(1-\frac{\sigma\eta}{2})}\leq\frac{2}{\sigma\eta}$. Dividing both sides of the above inequality by 2η, and letting $D_x:=\mathbb{E}_A[\|x_0-x_*^S\|^2]$, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)]\\
\leq&\left(\frac{2c}{e\sigma}\right)^c D_x\eta^{-c-1}T^{-c}+\frac{2L_f^2L_g^2}{\sigma}+\frac{C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}\\
&+\frac{4C_f^2L_g^2V_g}{\sigma^2}\frac{\beta}{\eta}+\frac{2C_f^2L_f^2L_g^5}{\sigma^2}\frac{\eta}{\beta^2}.
\end{aligned}\tag{105}$$

Dividing both sides of (105) by $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t}$, noting that for $(\eta(T-1))^{-1}\leq\frac{\sigma}{2}$ we have $(1-\frac{\sigma\eta}{2})^{T-1}\leq\exp(-\frac{\sigma\eta(T-1)}{2})\leq\frac{1}{2}$, and thus $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t}=\frac{1-(1-\frac{\sigma\eta}{2})^{T-1}}{1-(1-\frac{\sigma\eta}{2})}\geq\frac{1}{\sigma\eta}$, from the choice of $A(S)$ and convexity of F_S we get

$$\begin{aligned}
\mathbb{E}_A[F_S(A(S))-F_S(x_*^S)]\leq&(\frac{2c}{e\sigma})^{c-1}D_x(\eta T)^{-c}+\frac{C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}\frac{\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}}{\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}}\\
&+2L_f^2L_g^2\eta+\frac{4C_f^2L_g^2V_g}{\sigma}\beta+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^2}{\beta^2}.
\end{aligned}\tag{106}$$

Note that $\left(1-\frac{\sigma\eta}{2}\right)^{T-t}$ is non-decreasing with respect to t and for $c>0$, t^{-c} is non-increasing with respect to t. Then from Lemma A.4 we have

$$\frac{\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}}{\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}}\leq\frac{\sum_{t=1}^{T}t^{-c}}{T}$$

Thus (106) simplifies to

$$\begin{aligned}\mathbb{E}_A[F_S(A(S))-F_S(x_*^S)]\leq&(\frac{2c}{e\sigma})^{c-1}D_x(\eta T)^{-c}+2L_f^2L_g^2\eta+\frac{C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}T^{-1}\sum_{t=1}^{T}t^{-c}\\&+\frac{4C_f^2L_g^2V_g}{\sigma}\beta+\frac{2C_f^2L_f^2L_f^5}{\sigma}\frac{\eta^2}{\beta^2}.\end{aligned}$$

Note that $\sum_{t=1}^{T}t^{-z}=\mathcal{O}(T^{1-z})$ for $z\in(0,1)\cup(1,\infty)$ and $\sum_{t=1}^{T}t^{-1}=\mathcal{O}(\log T)$. As long as $c\neq1$ we get

$$\begin{aligned}&\mathbb{E}_A[F_S(A(S))-F_S(x_*^S)]\\=&\mathcal{O}\left(D_x(\eta T)^{-c}+2L_f^2L_g^2\eta+\frac{C_f^2L_g^2D_y}{\sigma}(\beta T)^{-c}+\frac{C_f^2L_g^2V_g}{\sigma}\beta+\frac{C_f^2L_f^2L_g^5}{\sigma}\eta^2\beta^{-2}\right).\end{aligned}$$

Then we get the desired result for the SCGD update. Next we present the proof for the SCSC update. Setting $\eta_t=\eta$ and $\beta_t=\beta$. Plugging Lemma A.1 into (104), and letting $D_y:=\mathbb{E}_A[\|y_1-g_S(x_0)\|^2]$, we have

$$\begin{aligned}\mathbb{E}_A[\|x_{t+1}-x_*^S\|^2]\leq&(1-\frac{\sigma\eta}{2})\mathbb{E}_A[\|x_t-x_*^S\|^2]+L_f^2L_g^2\eta^2-2\eta\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)]\\&+\frac{2C_f^2L_g^2\eta}{\sigma}\left((\frac{c}{e})^cD_y(t\beta)^{-c}+L_g^3L_f^2\frac{\eta^2}{\beta}+2V_g\beta\right).\end{aligned}$$

Telescoping the above inequality for $t=1,\cdots,T$, and rearranging the terms, we get

$$\begin{aligned}&2\eta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)]\\\leq&\left(1-\frac{\sigma\eta}{2}\right)^{T}\mathbb{E}_A[\|x_1-x_*^S\|^2]+L_f^2L_g^2\eta^2\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\\&+\frac{2C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\eta\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}\\&+\frac{4C_f^2L_g^2V_g}{\sigma}\eta\beta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^3}{\beta}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\\\leq&2\left(1-\frac{\sigma\eta}{2}\right)^{T}\mathbb{E}_A[\|x_0-x_*^S\|^2]+2L_f^2L_g^2\eta^2\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\\&+\frac{2C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\eta\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}\\&+\frac{4C_f^2L_g^2V_g}{\sigma}\eta\beta\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^3}{\beta}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\end{aligned}$$

From Lemma A.3 we know $(1-\frac{\sigma\eta}{2})^T\leq\exp(-\frac{\sigma\eta T}{2})\leq(\frac{2c}{e\sigma})^c(\eta T)^{-c}$. Also we have $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t}=\frac{1-(1-\frac{\sigma\eta}{2})^{T-1}}{1-(1-\frac{\sigma\eta}{2})}\leq$

$\frac{2}{\sigma\eta}$. Dividing both sides of the above inequality by 2η, and letting $D_x := \mathbb{E}_A[\|x_0 - x_*^S\|^2]$, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_A[F_S(x_t)-F_S(x_*^S)]\\
\leq&\left(\frac{2c}{e\sigma}\right)^c D_x\eta^{-c-1}T^{-c}+\frac{2L_f^2L_g^2}{\sigma}+\frac{C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}\\
&+\frac{4C_f^2L_g^2V_g}{\sigma^2}\frac{\beta}{\eta}+\frac{2C_f^2L_f^2L_g^5}{\sigma^2}\frac{\eta}{\beta}.
\end{aligned}\tag{107}$$

Dividing both sides of (107) by $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t}$, noting that for $(\eta(T-1))^{-1}\leq\frac{\sigma}{2}$ we have $(1-\frac{\sigma\eta}{2})^{T-1}\leq\exp(-\frac{\sigma\eta(T-1)}{2})\leq\frac{1}{2}$, and thus $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t}=\frac{1-(1-\frac{\sigma\eta}{2})^{T-1}}{1-(1-\frac{\sigma\eta}{2})}\geq\frac{1}{\sigma\eta}$, from the choice of $A(S)$ and convexity of F_S we get

$$\begin{aligned}
&\mathbb{E}_A[F_S(A(S))-F_S(x_*^S)]\\
\leq&(\frac{2c}{e\sigma})^{c-1}D_x(\eta T)^{-c}+2L_f^2L_g^2\eta+\frac{C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}\frac{\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c}}{\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}}\\
&+\frac{4C_f^2L_g^2V_g}{\sigma}\beta+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^2}{\beta}\\
\leq&(\frac{2c}{e\sigma})^{c-1}D_x(\eta T)^{-c}+2L_f^2L_g^2\eta+\frac{C_f^2L_g^2D_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}T^{-1}\sum_{t=1}^{T}t^{-c}\\
&+\frac{4C_f^2L_g^2V_g}{\sigma}\beta+\frac{2C_f^2L_f^2L_g^5}{\sigma}\frac{\eta^2}{\beta},
\end{aligned}$$

where the last inequality comes from Lemma A.4. Noting that $\sum_{t=1}^{T}t^{-z}=\mathcal{O}(T^{1-z})$ for $z\in(0,1)\cup(1,\infty)$ and $\sum_{t=1}^{T}t^{-1}=\mathcal{O}(\log T)$, as long as $c\neq1$ we get

$$\begin{aligned}
&\mathbb{E}_A[F_S(A(S))-F_S(x_*^S)]\\
=&\mathcal{O}\left(D_x(\eta T)^{-c}+2L_f^2L_g^2\eta+\frac{C_f^2L_g^2D_y}{\sigma}(\beta T)^{-c}+\frac{C_f^2L_g^2V_g}{\sigma}\beta+\frac{C_f^2L_f^2L_g^5}{\sigma}\eta^2\beta^{-1}\right).
\end{aligned}$$

Then we get the desired result for the SCSC update. Then we complete the proof. □

Proof of Lemma D.1. From Algorithm 1 we have for any $x\in\mathcal{X}$

$$\begin{aligned}
&\|x_{t+1}-x\|^2\\
\leq&\|x_t-\eta_t\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})-x\|^2\\
=&\|x_t-x\|^2+\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2-2\eta_t\langle x_t-x,\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\rangle\\
=&\|x_t-x\|^2+\eta_t^2\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2-2\eta_t\langle x_t-x,\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\rangle+u_t,
\end{aligned}$$

where

$$u_t := 2\eta_t\langle x_t-x,\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\rangle.$$

Let $\mathcal{F}_t$ be the σ-field generated by $\{\omega_{j_0},\ldots,\omega_{j_{t-1}},\nu_{i_0},\ldots,\nu_{i_{t-1}}\}$. Taking expectation with respect to the randomness of

the algorithm conditioned on $\mathcal{F}_t$, we have

$$\begin{aligned}
&\mathbb{E}_A[\|x_{t+1}-x\|^2|\mathcal{F}_t]\\
\leq&\|x_t-x\|^2+\eta_t^2\mathbb{E}_A[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2|\mathcal{F}_t]\\
&-2\eta_t\mathbb{E}_A[\langle x_t-x,\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(g_S(x_t))\rangle|\mathcal{F}_t]+\mathbb{E}_A[u_t|\mathcal{F}_t]\\
=&\|x_t-x\|^2+\eta_t^2\mathbb{E}_A[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2|\mathcal{F}_t]-2\eta_t\langle x_t-x,\nabla F_S(x_t)\rangle+\mathbb{E}_A[u_t|\mathcal{F}_t]\\
\leq&\|x_t-x\|^2+\eta_t^2\mathbb{E}_A[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2|\mathcal{F}_t]-2\eta_t(F_S(x_t)-F_S(x)+\frac{\sigma}{2}\|x_t-x\|^2)\\
&+\mathbb{E}_A[u_t|\mathcal{F}_t],
\end{aligned}\tag{108}$$

where the last inequality comes from the strong convexity of F_S. Note that from Cauchy-Schwartz inequality, Young's inequality, Assumption 2.2 (ii) and 3.1 (iii) we have

$$\begin{aligned}
u_t\leq&2\eta_t\|x_t-x\|\|\nabla g_{\omega_{j_t}}(x_t)\|\|\nabla f_{\nu_{i_t}}(g_S(x_t))-\nabla f_{\nu_{i_t}}(y_{t+1})\|\\
\leq&2C_f\eta_t\|x_t-x\|\|\nabla g_{\omega_{j_t}}(x_t)\|\|g_S(x_t)-y_{t+1}\|\\
\leq&2C_f\eta_t\left(\frac{\|x_t-x\|^2\|\nabla g_{\omega_{j_t}}(x_t)\|^2}{2\gamma}+\frac{\gamma}{2}\|g_S(x_t)-y_{t+1}\|^2\right)\\
\leq&\frac{C_fL_g^2\eta_t}{\gamma}\|x_t-x\|^2+\gamma C_f\eta_t\|g_S(x_t)-y_{t+1}\|^2
\end{aligned}\tag{109}$$

for any $\gamma>0$. Substituting (109) into (108), we get

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1}-x\|^2|\mathcal{F}_t]\leq&\left(1+\frac{C_fL_g^2\eta_t}{\gamma}-\sigma\eta_t\right)\|x_t-x\|^2+\eta_t^2\mathbb{E}_A[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2|\mathcal{F}_t]\\
&-2\eta_t(F_S(x_t)-F_S(x))+\gamma C_f\eta_tE[\|g_S(x_t)-y_{t+1}\|^2|\mathcal{F}_t].
\end{aligned}$$

Setting $\gamma=\frac{2C_fL_g^2}{\sigma}$, we have

$$\begin{aligned}
\mathbb{E}_A[\|x_{t+1}-x\|^2|\mathcal{F}_t]\leq&(1-\frac{\sigma\eta_t}{2})\|x_t-x\|^2+\eta_t^2\mathbb{E}_A[\|\nabla g_{\omega_{j_t}}(x_t)\nabla f_{\nu_{i_t}}(y_{t+1})\|^2|\mathcal{F}_t]\\
&-2\eta_t(F_S(x_t)-F_S(x))+2C_f^2L_g^2\frac{\eta_t}{\sigma}\mathbb{E}_A[\|g_S(x_t)-y_{t+1}\|^2|\mathcal{F}_t].
\end{aligned}$$

Then we complete the proof. □

D.3. Generalization

Proof of Theorem 3.15. We first present the proof for the SCGD update. From the stability results (87), (98) and (99) we get

$$\begin{aligned}
&\mathbb{E}_A[\|x_t-x_t^{k,\nu}\|]+4\mathbb{E}_A[\|x_t-x_t^{l,\omega}\|]\\
\leq&40C_fL_g\eta\sup_S\sum_{j=0}^{t-1}(1-\eta\frac{L\sigma}{L+\sigma})^{t-j-1}\left(\mathbb{E}_A[\|y_{j+1}-g_S(x_j)\|^2]\right)^{\frac{1}{2}}+20L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta}\\
&+2L_fL_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}}+\frac{4L_gL_f(L+\sigma)}{nL\sigma}+8L_fL_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}}+\frac{48L_gL_f(L+\sigma)}{mL\sigma}.
\end{aligned}$$

Plugging (101) into the above inequality, we get

$$\begin{aligned}
&\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] + 4\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \\
\leq& 40C_f L_g \frac{\sqrt{L_g} L_g L_f (L+\sigma)}{L\sigma} \frac{\eta}{\beta} + 40C_f L_g \frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta} \\
&+ 40C_f L_g (\frac{c}{e})^{\frac{c}{2}} \frac{D_y(L+\sigma)}{L\sigma} t^{-\frac{c}{2}} \beta^{-\frac{c}{2}} + 20L_f \sqrt{C_g \frac{L+\sigma}{L\sigma}}\sqrt{\eta} + 2L_f L_g \sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} \\
&+ \frac{4L_g L_f (L+\sigma)}{nL\sigma} + 8L_f L_g \sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} + \frac{48L_g L_f (L+\sigma)}{mL\sigma}.
\end{aligned}$$

Using Theorem 2.3, we have

$$\begin{aligned}
&\mathbb{E}_{S,A}\left[F(x_t) - F_S(x_t)\right] \\
\leq& 40C_f \frac{\sqrt{L_g} L_g^3 L_f^2 (L+\sigma)}{L\sigma} \frac{\eta}{\beta} + 40C_f L_g^2 L_f \frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta} \\
&+ 40C_f L_g^2 L_f (\frac{c}{e})^{\frac{c}{2}} \frac{D_y(L+\sigma)}{L\sigma} t^{-\frac{c}{2}} \beta^{-\frac{c}{2}} + 20L_f^2 L_g \sqrt{C_g \frac{L+\sigma}{L\sigma}}\sqrt{\eta} + 2L_f^2 L_g^2 \sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} \\
&+ \frac{4L_g^2 L_f^2 (L+\sigma)}{nL\sigma} + 8L_f^2 L_g^2 \sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} + \frac{48L_g^2 L_f^2 (L+\sigma)}{mL\sigma} + L_f \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}}.
\end{aligned} \tag{110}$$

From (105) we get

$$\begin{aligned}
&\sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \mathbb{E}_{S,A}[F_S(x_t) - F_S(x_*^S)] \\
&\leq \left(\frac{2c}{e\sigma}\right)^c D_x \eta^{-c-1} T^{-c} + \frac{2L_f L_g}{\sigma} + \frac{C_f^2 L_g D_y}{\sigma} \left(\frac{c}{e}\right)^c \beta^{-c} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} t^{-c} \\
&\quad + \frac{4C_f^2 L_g V_g}{\sigma^2} \frac{\beta}{\eta} + \frac{2C_f^2 L_f L_g^3}{\sigma^2} \frac{\eta}{\beta^2}.
\end{aligned} \tag{111}$$

Multiplying both sides of (110) with $\left(1 - \frac{\sigma\eta}{2}\right)^{T-t}$, telescoping from $t = 1, \ldots, T$, then adding the result with (111), and using the fact $F_S(x_*^S) \leq F_S(x_*)$, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \mathbb{E}_{S,A}[F(x_t) - F(x_*)] \\
\leq& 40C_f \frac{\sqrt{L_g} L_g^3 L_f^2 (L+\sigma)}{L\sigma} \frac{\eta}{\beta} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} + 40C_f L_g^2 L_f \frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \\
&+ 40C_f L_g^2 L_f (\frac{c}{e})^{\frac{c}{2}} \frac{D_y(L+\sigma)}{L\sigma} \beta^{-\frac{c}{2}} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} t^{-\frac{c}{2}} + \frac{4L_f^2 L_g^2 (L+\sigma)}{L\sigma n} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \\
&+ 20L_f^2 L_g \sqrt{C_g \frac{L+\sigma}{L\sigma}}\sqrt{\eta} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} + 2L_f^2 L_g^2 \sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \\
&+ \frac{48L_g^2 L_f^2 (L+\sigma)}{L\sigma m} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} + 8L_f^2 L_g^2 \sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \\
&+ L_f \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}} + \left(\frac{2c}{e\sigma}\right)^c D_x \eta^{-c-1} T^{-c} + \frac{2L_f L_g}{\sigma} \\
&+ \frac{C_f^2 L_g D_y}{\sigma} \left(\frac{c}{e}\right)^c \beta^{-c} \sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} t^{-c} + \frac{4C_f^2 L_g V_g}{\sigma^2} \frac{\beta}{\eta} + \frac{2C_f^2 L_f L_g^3}{\sigma^2} \frac{\eta}{\beta^2}.
\end{aligned}$$

Dividing both sides of the above inequality by $\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}$, and setting $\eta = T^{-a}$ and $\beta = T^{-b}$ with $a, b \in (0,1]$, then from the choice of $A(S)$ and convexity of F and Lemma A.4, noting that $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t} = \frac{1-(1-\frac{\sigma\eta}{2})^{T-1}}{1-(1-\frac{\sigma\eta}{2})} \geq \frac{1}{\sigma\eta}$ for $(\eta(T-1))^{-1} \leq \frac{\sigma}{2}$, we get

$$
\begin{aligned}
&\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] \\
\leq& 40C_f\frac{\sqrt{L_g}L_g^3L_f^2(L+\sigma)}{L\sigma}T^{b-a} + 40C_fL_g^2L_f\frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}T^{-\frac{b}{2}} + \frac{4L_g^2L_f^2(L+\sigma)}{nL\sigma} \\
&+ 40C_fL_g^2L_f(\frac{c}{e})^{\frac{c}{2}}\frac{D_y(L+\sigma)}{L\sigma}T^{\frac{bc}{2}-1}\sum_{t=1}^{T}t^{-\frac{c}{2}} + 2L_f^2L_g^2\sqrt{\frac{L+\sigma}{L\sigma}}\frac{1}{\sqrt{n}}T^{-\frac{a}{2}} \\
&+ 20L_f^2L_g\sqrt{C_g\frac{L+\sigma}{L\sigma}}T^{-\frac{a}{2}} + 8L_f^2L_g^2\sqrt{\frac{L+\sigma}{L\sigma}}\frac{1}{\sqrt{m}}T^{-\frac{a}{2}} + \frac{48L_g^2L_f^2(L+\sigma)}{mL\sigma} \\
&+ L_f\left(\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}}\right)\Big/\left(\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\right) \\
&+ (\frac{2c}{e\sigma})^{c-1}D_xT^{-c(1-a)} + 2L_fL_gT^{-a} + \frac{C_f^2L_gD_y}{\sigma}\left(\frac{c}{e}\right)^cT^{bc-1}\sum_{t=1}^{T}t^{-c} + \frac{4C_f^2L_gV_g}{\sigma}T^{-b} \\
&+ \frac{2C_f^2L_fL_g^3}{\sigma}T^{2b-2a}.
\end{aligned}
$$

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

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

Since $a, b \in (0,1]$, setting $c = 3$, the dominating terms are

$$
\mathcal{O}(\frac{1}{\sqrt{m}}), \quad \mathcal{O}(T^{b-a}), \quad \mathcal{O}(T^{-\frac{b}{2}}), \quad \mathcal{O}(T^{\frac{3}{2}(b-1)}), \quad \mathcal{O}(T^{-\frac{a}{2}}), \quad \mathcal{O}(T^{3(a-1)}).
$$

Setting $a = \frac{9}{10}$ and $b = \frac{3}{5}$ yields

$$
\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}(T^{-\frac{3}{10}} + \frac{1}{\sqrt{m}}). \tag{112}
$$

Setting $T = \mathcal{O}(\max\{n^{\frac{10}{3}}, m^{\frac{10}{3}}\})$ yields the following bound

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

Then we get the desired result for the SCGD update. Next we present the proof for the SCSC update. With the same derivation as the SCGD case, we get

$$
\begin{aligned}
&\mathbb{E}_A[\|x_t - x_t^{k,\nu}\|] + 4\mathbb{E}_A[\|x_t - x_t^{l,\omega}\|] \leq 40C_fL_g\frac{\sqrt{L_g}L_gL_f(L+\sigma)}{L\sigma}\frac{\eta}{\sqrt{\beta}} \\
&+ 40C_fL_g\frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta} + 40C_fL_g(\frac{c}{e})^{\frac{c}{2}}\frac{D_y(L+\sigma)}{L\sigma}t^{-\frac{c}{2}}\beta^{-\frac{c}{2}} \\
&+ 20L_f\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta} + 2L_fL_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} + \frac{4L_gL_f(L+\sigma)}{nL\sigma} + 8L_fL_g\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} \\
&+ \frac{48L_gL_f(L+\sigma)}{mL\sigma}.
\end{aligned}
$$

Using Theorem 2.3, we have

$$\begin{aligned}
&\mathbb{E}_{S,A}\left[F(x_t) - F_S(x_t)\right] \leq 40C_f\frac{\sqrt{L_g}L_g^3L_f^2(L+\sigma)}{L\sigma}\frac{\eta}{\sqrt{\beta}} + 40C_fL_g^2L_f\frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta} \\
&+ 40C_fL_g^2L_f(\frac{c}{e})^{\frac{c}{2}}\frac{D_y(L+\sigma)}{L\sigma}t^{-\frac{c}{2}}\beta^{-\frac{c}{2}} + 20L_f^2L_g\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta} + 2L_f^2L_g^2\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}} \\
&+ \frac{4L_g^2L_f^2(L+\sigma)}{nL\sigma} + 8L_f^2L_g^2\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}} + \frac{48L_g^2L_f^2(L+\sigma)}{mL\sigma} + L_f\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}}. \qquad (113)
\end{aligned}$$

From (107) we get

$$\begin{aligned}
&\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_{S,A}[F_S(x_t) - F_S(x_*^S)] \\
&\leq \left(\frac{2c}{e\sigma}\right)^c D_x\eta^{-c-1}T^{-c} + \frac{2L_fL_g}{\sigma} + \frac{C_f^2L_gD_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c} \\
&+ \frac{4C_f^2L_gV_g}{\sigma^2}\frac{\beta}{\eta} + \frac{2C_f^2L_fL_g^3}{\sigma^2}\frac{\eta}{\beta}. \qquad (114)
\end{aligned}$$

Multiplying both sides of (113) with $\left(1-\frac{\sigma\eta}{2}\right)^{T-t}$, telescoping from $t=1,\ldots,T$, then adding the result with (114), and using the fact $F_S(x_*^S) \leq F_S(x_*)$, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\mathbb{E}_{S,A}[F(x_t) - F(x_*)] \\
\leq& 40C_f\frac{\sqrt{L_g}L_g^3L_f^2(L+\sigma)}{L\sigma}\frac{\eta}{\sqrt{\beta}}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} + 40C_fL_g^2L_f\frac{\sqrt{2V_g}(L+\sigma)}{L\sigma}\sqrt{\beta}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} \\
&+ 40C_fL_g^2L_f(\frac{c}{e})^{\frac{c}{2}}\frac{D_y(L+\sigma)}{L\sigma}\beta^{-\frac{c}{2}}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-\frac{c}{2}} + \frac{4L_f^2L_g^2(L+\sigma)}{L\sigma n}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} \\
&+ 20L_f^2L_g\sqrt{C_g\frac{L+\sigma}{L\sigma}}\sqrt{\eta}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} + 2L_f^2L_g^2\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{n}}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} \\
&+ \frac{48L_g^2L_f^2(L+\sigma)}{L\sigma m}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} + 8L_f^2L_g^2\sqrt{\frac{L+\sigma}{L\sigma}}\sqrt{\frac{\eta}{m}}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t} \\
&+ L_f\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}\sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}} + \left(\frac{2c}{e\sigma}\right)^c D_x\eta^{-c-1}T^{-c} + \frac{2L_fL_g}{\sigma} \\
&+ \frac{C_f^2L_gD_y}{\sigma}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}t^{-c} + \frac{4C_f^2L_gV_g}{\sigma^2}\frac{\beta}{\eta} + \frac{2C_f^2L_fL_g^3}{\sigma^2}\frac{\eta}{\beta}.
\end{aligned}$$

Dividing both sides of the above inequality by $\sum_{t=1}^{T}\left(1-\frac{\sigma\eta}{2}\right)^{T-t}$, and setting $\eta = T^{-a}$ and $\beta = T^{-b}$ with $a, b \in (0,1]$, then from the choice of $A(S)$ and convexity of F and Lemma A.4, noting that $\sum_{t=1}^{T}(1-\frac{\sigma\eta}{2})^{T-t} = \frac{1-(1-\frac{\sigma\eta}{2})^{T-1}}{1-(1-\frac{\sigma\eta}{2})} \geq \frac{1}{\sigma\eta}$

for $(\eta(T-1))^{-1} \leq \frac{\sigma}{2}$, we get

$$\begin{aligned}
&\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] \\
\leq& 40C_f \frac{\sqrt{L_g} L_g^3 L_f^2 (L+\sigma)}{L\sigma} T^{\frac{b}{2}-a} + 40C_f L_g^2 L_f \frac{\sqrt{2V_g}(L+\sigma)}{L\sigma} T^{-\frac{b}{2}} + \frac{4L_g^2 L_f^2 (L+\sigma)}{nL\sigma} \\
&+ 40C_f L_g^2 L_f (\frac{c}{e})^{\frac{c}{2}} \frac{D_y(L+\sigma)}{L\sigma} T^{\frac{bc}{2}-1} \sum_{t=1}^{T} t^{-\frac{c}{2}} + 2L_f^2 L_g^2 \sqrt{\frac{L+\sigma}{L\sigma}} \frac{1}{\sqrt{n}} T^{-\frac{a}{2}} \\
&+ 20L_f^2 L_g \sqrt{C_g \frac{L+\sigma}{L\sigma}} T^{-\frac{a}{2}} + 8L_f^2 L_g^2 \sqrt{\frac{L+\sigma}{L\sigma}} \frac{1}{\sqrt{m}} T^{-\frac{a}{2}} + \frac{48L_g^2 L_f^2 (L+\sigma)}{mL\sigma} \\
&+ L_f \left(\sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_\omega(g_\omega(x_t))]}{m}} \right) / \left(\sum_{t=1}^{T} \left(1 - \frac{\sigma\eta}{2}\right)^{T-t} \right) \\
&+ (\frac{2c}{e\sigma})^{c-1} D_x T^{-c(1-a)} + 2L_f L_g T^{-a} + \frac{C_f^2 L_g D_y}{\sigma} \left(\frac{c}{e}\right)^c T^{bc-1} \sum_{t=1}^{T} t^{-c} + \frac{4C_f^2 L_g V_g}{\sigma} T^{-b} \\
&+ \frac{2C_f^2 L_f L_g^3}{\sigma} T^{b-2a}.
\end{aligned}$$

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

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

Since $a, b \in (0, 1]$, setting $c = 6$, the dominating terms are

$$\mathcal{O}(\frac{1}{\sqrt{m}}), \quad \mathcal{O}(T^{\frac{b}{2}-a}), \quad \mathcal{O}(T^{-\frac{b}{2}}), \quad \mathcal{O}(T^{3(b-1)}), \quad \mathcal{O}(T^{-\frac{a}{2}}), \quad \mathcal{O}(T^{6(a-1)}).$$

Setting $a = b = \frac{6}{7}$ yields

$$\mathbb{E}_{S,A}\Big[F(A(S)) - F(x_*)\Big] = \mathcal{O}(T^{-\frac{3}{7}} + \frac{1}{\sqrt{m}}). \tag{115}$$

Setting $T = \mathcal{O}(\max\{n^{\frac{7}{3}}, m^{\frac{7}{3}}\})$ yields the following bound

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

Then we get the desired result for the SCSC update. We have completed the proof. □

E. Stability and Generalization of Coupled Compositional Stochastic Optimization Algorithms

This Appendix explores extensions of the dependent case, where the random variables ν and ω exhibit interdependence. We provide an intuitive approach to achieving stability and generalization results for Coupled Compositional Stochastic Optimization Algorithms. Detailed technical proofs are deferred to future research.

Coupled Stochastic Optimization(CSO) problems which has gained more general interest (Qi et al., 2021b; Jiang et al., 2022a; Wang & Yang, 2022). The CSO problems can be represented as:

$$\min_{x \in \mathcal{X}} \Big\{ F(x) = f \circ g(x) = \mathbb{E}_\nu[f_\nu(\mathbb{E}_{\omega|\nu}[g_\omega(x, \nu)])] \Big\}, \tag{116}$$

Similarity to the independent case, with the training data $S = S_\nu \bigcup S_\omega$ where $S_\nu = \{\nu_i : i = 1, \ldots, i = n\}$ and $S_\omega = \big\{\omega_{ij} : i = 1, \ldots, n, j = 1, \ldots m\big\}$, CSO problem (116) can be reduced to the following nested empirical risk for CSO:

$$\min_{x\in\mathcal{X}}\{F_S(x) = f_S(g_S(x)) = \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\frac{1}{m}\sum_{j=1}^{m} g_{\omega_{ij}}(x,\nu_i))\}, \tag{117}$$

where $g_S : \mathbb{R}^p \to \mathbb{R}^d$ and $f_S : \mathbb{R}^d \to \mathbb{R}$ are the empirical versions of f and g in (116) and are defined, respectively, by $g_S(x) = \frac{1}{m}\sum_{j=1}^{m} g_{\omega_{ij}}(x,\nu_i)$ and $f_S(y) = \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(y)$. We refer to $F(x)$ and $F_S(x)$ as the *(nested) true risk and empirical risk*, respectively, in this stochastic compositional setting.

To analyze the excess generalization error *(that is, excess risk)* of $A(S)$ given by $F(A(S)) - F(x_*)$, we still focus on the estimation error and the optimization error from (3).

Estimation Error. We introduce uniform stability for CSO problems to study the estimation error. Unlike the independent case, for any $k \in [n]$, let $S^k = S_\nu^k \cup S_\omega^k$. When we replace the k-th sample ν_k in $\{\nu_i\}_{i=1}^n$ to ν_k', the corresponding sample $\{\omega_{kj}\}_{j=1}^m$ in S_ω changes to $\{\omega_{kj}'\}_{j=1}^m$. From this, we can introduce the following uniform stability concept in the following way.

Definition E.1. We say that a randomized algorithm A is ϵ-uniformly stable for CSO problems (116) if, any $k \in [1,n]$, there holds

$$\mathbb{E}[\|A(S) - A(S^k)\|] \le \epsilon$$

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

On the basis of our stability concept, we can build the connection between the uniform stability and the generalization error for CSO problems. One can show that,

$$\begin{aligned}
&\mathbb{E}_{A,S}[F(A(S)) - F_S(A(S)] = \mathbb{E}_{A,S}[F(A(S)) - \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\frac{1}{m}\sum_{j=1}^{m} g_{\omega_{ij}}(x,\nu_i))]\\
&= \mathbb{E}_{A,S}[F(A(S)) - \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\mathbb{E}_{\omega|\nu_i}[g_\omega(A(S^i),\nu_i)])]\\
&+ \mathbb{E}_{A,S}[\frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\mathbb{E}_{\omega|\nu_i}[g_\omega(A(S^i),\nu_i)]) - \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\frac{1}{m}\sum_{j=1}^{m} g_{\omega_{ij}}(A(S^i),\nu_i))]\\
&+ \mathbb{E}_{A,S}[\frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\frac{1}{m}\sum_{j=1}^{m} g_{\omega_{ij}}(A(S^i),\nu_i)) - \frac{1}{n}\sum_{i=1}^{n} f_{\nu_i}(\frac{1}{m}\sum_{j=1}^{m} g_{\omega_{ij}}(A(S),\nu_i))].
\end{aligned} \tag{118}$$

It is easy to estimate the first term on the right-hand side of (118) and the third term can be bounded using the Lipschitz continuity of $f_\nu(\cdot)$ and $g_\omega(\cdot)$. The main challenge comes from the second term. We can use some ideas from (Hu et al., 2020) to estimate this term, where the key step is to use the independence between $\{\omega_{ij}\}_{j=1}^m$ being independent with $A(S^i)$ and L smoothing of $f_\nu(\cdot)$.

Optimization Error. Based on the independent case, to analyze the optimization error for CSO problems, we can extend the SCGD update (line 5 in Algorithm 1) and the SCSC update (line 6 in Algorithm 1) to the dependent cases that were studied in the two latest algorithms (SOX (Wang & Yang, 2022), MSVR (Jiang et al., 2022a)). We list the main updates of these two algorithms in the following.

The update of estimator $y_{i_t}^{t+1}$ which estimates the inner empirical risk of $g_S(x_t;\nu_{i_t})$ is shown below for SOX:

$$y_{i_t}^{t+1} = (1-\beta)y_{i_t}^t + \beta \cdot g_{\omega_{i_t j_t}}(x^t;\nu_{i_t}).$$

As mentioned in (Jiang et al., 2022a), with $y_{i_t}^{t+1}$, the gradient estimator computed by exponential moving average. A useful technique for achieving a better optimization error is by using a variance reduction techniques which given as following(MSVR):

$$y_{i_t}^{t+1} = (1-\beta)y_{i_t}^t + \beta \cdot g_{\omega_{i_t j_t}}(x^t;\nu_{i_t}) + \gamma(g_{\omega_{i_t j_t}}(x^t;\nu_{i_t}) - g_{\omega_{i_t j_t}}(x^{t-1};\nu_{i_t})).$$

For the update of model parameter, SOX and MSVR use $y_{i_t}^t$ instead of $y_{i_t}^{t+1}$ in $\nabla f_{\nu_{i_t}}(\cdot)$ to get better optimization errors due to the independence between $y_{i_t}^t$ and the random variable ν_{i_t}:

$$x^{t+1} = \Pi_{\mathcal{X}}\left(x^t - \eta \nabla g_{\omega_{i_t j_t}}(x^t; \nu_{i_t}) \nabla f_{\nu_{i_t}}(y_{i_t}^t)\right).$$

From the above update of the model parameter, we can use a similar technique in Appendix C to get the stability results of SOX and MSVR(without momentum update). Since we use $y_{i_t}^t$ here, we will estimate the term $\sup_S \sum_{j=0}^{T-1}(\mathbb{E}_A[\frac{1}{n}\sum_{i=1}^n \|y_i^j - g_S(x^j;\nu_i)\|^2])^{1/2}$ in the dependent case compared to the independent case in Theorem 3.3. Then we can combine the estimation error and the optimization error of the existing optimization paper to get the final excess risk.