Differentially Private Algorithms for the Stochastic Compositional Optimization Problem

Zhuanghua Liu
School of Computing, National University of Singapore
CNRS@CREATE LTD, 1 Create Way, #08-01 CREATE Tower, Singapore 138602

Weida Li **Xiaokui Xiao** **Bryan Kian Hsiang Low**
School of Computing, National University of Singapore

Abstract

In this paper, we study the stochastic compositional optimization problem under the constraint of differential privacy. We first introduce two private algorithms: noisy stochastic compositional gradient descent (NSCGD) and the noisy stochastically corrected stochastic compositional gradient (NSCSC) method. We use the algorithmic stability approach to establish bounds on the excess population loss of both methods in strongly convex and convex cases. However, these methods require gradient computations that are super-linear in the number of training samples. To address this, we propose a class of output perturbation-based randomized algorithms by exploiting the stability of the compositional empirical risk minimizer under the privacy constraint. These algorithms achieve comparable excess population risk with significantly reduced gradient computations.

1 INTRODUCTION

We consider the following stochastic compositional optimization (SCO) problem

$$\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) \coloneqq f(g(\mathbf{x})), \tag{1}$$

where $\mathcal{X}$ is a convex and compact set. The outer and inner functions $f\colon \mathbb{R}^p \to \mathbb{R}$ and $g\colon \mathbb{R}^d \to \mathbb{R}^p$ have the forms of

$$f(\mathbf{y}) \coloneqq \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{P}}[F(\mathbf{y};\boldsymbol{\xi})] \quad \text{and} \quad \mathrm{g}(\mathbf{x}) \coloneqq \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{Q}}[\mathrm{G}(\mathbf{x};\boldsymbol{\zeta})].$$

The random variables $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are independent samples from some unknown distributions $\mathcal{P}$ and $\mathcal{Q}$. This formulation is ubiquitous in many real-world applications, including risk management (Dentcheva et al., 2017), statistical learning (Wang et al., 2017a), reinforcement learning (Sutton, 2018), and meta-learning (Finn et al., 2017). On the other hand, the dependence of modern machine learning applications on large datasets has raised concerns about user privacy. In this context, differential privacy (DP) (Dwork, 2006) has emerged as a gold standard for designing privacy-preserving algorithms. Therefore, it is essential to develop algorithms for the SCO problem under the DP constraints (DP-SCO).

Most of the existing work studied randomized algorithms for the stochastic convex optimization problem $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \coloneqq \mathbb{E}[F(\mathbf{x},\boldsymbol{\xi})]$ under (ϵ,δ)-DP guarantees (Bassily et al., 2020, 2019; Feldman et al., 2020). In particular, Bassily et al. (2019) established the optimal excess population loss bound of $\mathcal{O}(1/\sqrt{n}+\sqrt{d\log(1/\delta)}/(n\epsilon))$ of the noisy SGD (NSGD) method with $\mathcal{O}\big(\min(n^{\frac{3}{2}}, n^{\frac{5}{2}}/d)\big)$ gradient computations, where n is the number of samples and d is the dimension of the problem. Later, Feldman et al. (2020) proposed the phased-SGD method that achieves the optimal population bound with linear time. However, it is unclear whether these methods can be extended to the more complex stochastic compositional optimization problems.

In this paper, we propose the noisy stochastic compositional gradient descent (NSCGD) and noisy stochastically corrected stochastic compositional gradient (NSCSC) methods, which extend the classical SCGD (Wang et al., 2017a) and its enhanced variant SCSC

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

Table 1: We compare the utility bounds and gradient complexity of stochastic first-order methods for solving the stochastic optimization problem in the strongly convex case. The fourth column indicates whether the method satisfies the DP constraint, while the fifth column specifies whether the objective function has a compositional structure. $\hat{\varkappa}$ is a variant of condition number.

Methods	population loss	Gradient Complexity	DP	Comp.
`Phased-SGD` (Feldman et al.)	$\tilde{\mathcal{O}}\left(\frac{1}{\mu n}+\frac{d}{\epsilon^2\mu n^2}\right)$	$\mathcal{O}(n)$	✓	✗
`SCGD` (Yang et al.)[a]	$\mathcal{O}\left(\frac{1}{\mu n}+\frac{1}{\mu\sqrt{m}}\right)$	$\mathcal{O}\big(\min\{n^{\frac{10}{3}}, m^{\frac{5}{3}}\}\big)$	✗	✓
`SCSC` (Yang et al.)[b]	$\mathcal{O}\left(\frac{1}{\mu n}+\frac{1}{\mu\sqrt{m}}\right)$	$\mathcal{O}\big(\min\{n^{\frac{7}{3}}, m^{\frac{7}{6}}\}\big)$	✗	✓
`NSCGD` (Thm 4.2)	$\tilde{\mathcal{O}}\left(\frac{1}{\mu n}+\frac{1}{\mu\sqrt{m}}+\frac{\sqrt{d}}{\epsilon}\left(\frac{1}{\mu n}+\frac{1}{\mu m}\right)\right)$	$\mathcal{O}\big(\min\{n^{\frac{14}{3}}, m^{\frac{7}{3}}\}\big)$	✓	✓
`NSCSC` (Thm 4.3)	$\tilde{\mathcal{O}}\left(\frac{1}{\mu n}+\frac{1}{\mu\sqrt{m}}+\frac{\sqrt{d}}{\epsilon}\left(\frac{1}{\mu n}+\frac{1}{\mu m}\right)\right)$	$\mathcal{O}\big(\min\{n^{\frac{7}{3}}, m^{\frac{7}{6}}\}\big)$	✓	✓
`OP` (Thm 5.2)	$\tilde{\mathcal{O}}\left(\frac{1}{\mu^2 n}+\frac{\sqrt{d}}{\mu\sqrt{m}}+\frac{d}{\epsilon^2}\left(\frac{1}{\mu^2 n^2}+\frac{1}{\mu m}\right)\right)$	$\tilde{\mathcal{O}}(n+m+\hat{\varkappa}^3)$	✓	✓

[a,b] Yang et al. (2024) reported gradient complexities of $\mathcal{O}(\max\{n^{\frac{10}{3}}, m^{\frac{10}{3}}\})$ and $\mathcal{O}(\max\{n^{\frac{7}{3}}, m^{\frac{7}{3}}\})$ for the SCGD and SCSC methods, respectively. However, a careful inspection of Eqs. (112) and (115) in their manuscript suggests that the claimed population loss can be achieved with a lower gradient complexity.

(Chen et al., 2021), respectively, to the differentially private setting. For both methods, differential privacy is ensured by adding Gaussian noise to the gradient estimators. We further establish bounds on the excess population loss for both methods using algorithmic stability techniques (Bousquet and Elisseeff, 2002; Yang et al., 2024). Given the sample size of the inner function m and that of the outer function n in the compositional empirical risk minimization (ERM) problem, we rigorously show that both methods achieve the excess population loss of $\tilde{\mathcal{O}}(1/(\mu n)+1/(\mu\sqrt{m})+(1/(\mu n)+1/(\mu m))\sqrt{d}/\epsilon)$ with at least $\Theta\big(\min\{n^{\frac{7}{3}}, m^{\frac{7}{6}}\}\big)$ gradient computations for the μ-strongly convex function and that of $\tilde{\mathcal{O}}(1/\sqrt{n}+1/\sqrt{m}+(1/n+1/m)\sqrt{d}/\epsilon)$ with at least $\Theta\big(\min\{n^4, m^4\}\big)$ gradient computations for the convex function. We note that these methods require gradient computations that are super-linear in the number of training samples.

To overcome this limitation, we introduce a class of output perturbation–based randomized algorithms to further reduce gradient computations in the DP-SCO problem. While NSCGD and NSCSC leverage the algorithmic stability of the underlying non-private SCGD and SCSC methods, more sophisticated algorithms, such as compositional SVRG and accelerated SCGD (Lian et al., 2017; Wang et al., 2017b), cannot be directly applied due to the lack of corresponding stability guarantees. For the strongly convex case, we propose the output perturbation (OP) method that leverages the stability of the empirical loss minimizer, thereby avoiding algorithm-specific stability requirements. Specifically, if an algorithm's output is sufficiently close to the empirical solution, its stability is automatically controlled. The OP method achieves an excess population loss of $\tilde{\mathcal{O}}\left(1/(\mu^2 n)+\sqrt{d}/(\mu\sqrt{m})+\big(1/(\mu^2 n^2)+1/(\mu m)\big)\,d/\epsilon^2\right)$ with $\tilde{\mathcal{O}}(m+n+\hat{\varkappa}^3)$ gradient computations where $\hat{\varkappa}$ is a variant of the condition number. In the convex case, output perturbation cannot be directly applied because the empirical loss minimizer need not be unique. To address this issue, we introduce the Phased output perturbation (Phased OP) method, which leverages the phased regularization framework (Feldman et al., 2020) to iteratively solve a sequence of regularized strongly convex ERM subproblems using output perturbation. This algorithm achieves an excess population loss of $\tilde{\mathcal{O}}\left(1/\sqrt{n}+\sqrt{d}/\sqrt{m}+(1/n+1/\sqrt{m})\sqrt{d}/\epsilon\right)$ with $\tilde{\mathcal{O}}(m+n^{\frac{3}{2}})$ gradient computations. While the classical phased regularization framework (Feldman et al., 2020) was designed for nonsmooth DP-stochastic convex optimization with quadratic gradient complexity, our results demonstrate that it can be adapted to the more challenging DP-stochastic compositional optimization setting, achieving even lower gradient costs under an additional smoothness assumption. We summarize the theoretical results of the proposed methods alongside existing work in Table 1 and 2, corresponding to the strongly convex case and the convex case, respectively.

Paper Organization In Section 2, we provide a literature review for non-private stochastic convex compositional optimization and differentially private stochastic convex optimization. In Section 3, we formalize the notations and assumptions of our problem and introduce the background of differential privacy. In Section 4, we propose the NSCGD and NSCSC methods, and we provide their privacy guarantee and excess population

loss bounds of these algorithms. In Section 5, we introduce a class of output perturbation-based algorithms by adding Gaussian noise to the output of ERM problems. Finally, we conclude this work in Section 6.

2 RELATED WORK

In this section, we review related work of stochastic compositional optimization for convex and strongly convex functions, stability and generalization results of the stochastic compositional optimization, and randomized algorithms for solving differentially private stochastic convex optimization.

2.1 Stochastic Compositional Optimization

In a pioneer work, Wang et al. (2017a) investigated the non-asymptotic convergence of stochastic compositional optimization by introducing the stochastic compositional gradient descent (SCGD) method. This approach leverages two sequences of stepsizes of different time scales to update the variable and track the inner function value, respectively. Wang et al. (2017b) proposed an accelerated version of SCGD by incorporating an extrapolation-smoothing scheme. Later, Lian et al. (2017); Liu et al. (2018) developed variance-reduced randomized algorithms that achieve a linear convergence rate for the strongly convex objective function. Chen et al. (2021) proposed the stochastically corrected stochastic compositional gradient (SCSC) method, which incorporates an additional stochastic correction term on top of the SCGD approach, achieving the same convergence rate as standard SGD in the non-compositional setting. Liu et al. (2024) studied the non-asymptotic convergence analysis of nonconvex nonsmooth stochastic compositional optimization.

2.2 Stability and Generalization of Stochastic Compositional Optimization

Existing research (Bassily et al., 2020; Bousquet and Elisseeff, 2002; Hardt et al., 2016; Kuzborskij and Lampert, 2018; Lei and Ying, 2020) has primarily focused on the stability and generalization of the SGD algorithm and its variations for solving non-compositional stochastic convex optimization. In contrast, research on compositional settings has been more limited. Hu et al. (2020) studied the generalization and sample complexity of the exact minimizer of the compositional ERM problem using the uniform convergence approach (Bartlett and Mendelson, 2002; Vapnik, 2013). Recently, Yang et al. (2024) studied the stability and generalization of the SCGD and SCSC methods for solving the SCO problem by introducing the notion of compositional uniform stability in the non-private setting. Building on this, Pan et al. (2024) extended the stability and generalization results to K-level stochastic optimization with $K \geq 3$.

2.3 Differentially Private Stochastic Convex Optimization

Stochastic convex optimization under the constraint of differential privacy have been thoroughly studied for over a decade (Bassily et al., 2020, 2014; Chaudhuri and Monteleoni, 2008; Chaudhuri et al., 2011; Feldman et al., 2020). Bassily et al. (2014) showed that the optimal excess empirical risk is $\tilde{\mathcal{O}}(\sqrt{d}/\epsilon n)$. Later, Bassily et al. (2019) presented an asymptotically tight bound of $\tilde{\mathcal{O}}(1/\sqrt{n} + \sqrt{d}/\epsilon n)$ on the excess population loss using the noisy stochastic gradient descent. Under mild smoothness assumptions, Feldman et al. (2020) proposed randomized algorithms that achieve optimal excess population loss with gradient computations linear in the sample size. Recently, Asi et al. (2021); Bassily et al. (2021) studied the private stochastic convex optimization problem in non-Euclidean settings. Beyond stochastic convex optimization, the stochastic minimax optimization problem under DP constraint has been investigated in recent works (Boob and Guzmán, 2024; Yang et al., 2022; Zhang et al., 2022). However, to the best of our knowledge, there is no existing work that addresses stochastic compositional optimization under DP constraints.

3 PRELIMINARIES

In this section, we formalize notations and assumptions throughout this paper, then we provide some background on differential privacy.

3.1 Notations

For any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we denote their inner product as $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$. We use $\|\cdot\|$ to represent the Euclidean norm of the vector. A function $h\colon \mathbb{R}^d \to \mathbb{R}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that $h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$. We say a function h is μ-strongly convex if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds that $h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2$. For the convex function, we can infer that $\mu = 0$.

Usually, the underlying distributions $\mathcal{P}$ and $\mathcal{Q}$ in the objective function (1) are unknown, and the SCO problem cannot be solved exactly. However, it is generally assumed that samples from these unknown distributions are accessible. Formally, given a set of samples $S \coloneqq S_{\boldsymbol{\xi}} \cup S_{\boldsymbol{\zeta}}$ composed of i.i.d. samples $S_{\boldsymbol{\xi}} = \{\boldsymbol{\xi}_i \sim \mathcal{P}\colon i = 1, \ldots, n\}$ and $S_{\boldsymbol{\zeta}} = \{\boldsymbol{\zeta}_j \sim \mathcal{Q}\colon j = 1, \ldots, m\}$, we consider the following compositional empirical risk

Table 2: We compare the utility bounds and gradient complexity of proposed stochastic first-order methods for solving the stochastic optimization problem in the convex case. The fourth column indicates whether the method satisfies the DP constraint, while the fifth column specifies whether the objective function has a compositional structure.

Methods	population loss	Gradient Complexity	DP	Comp.
NSGD (Bassily et al.)	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d}}{\epsilon n}\right)$	$\mathcal{O}\big(\min(n^{\frac{3}{2}}, n^{\frac{5}{2}}/d)\big)$	✓	✗
Phased-SGD (Feldman et al.)	$\mathcal{O}\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d}}{\epsilon n}\right)$	$\mathcal{O}(n)$	✓	✗
SCGD (Yang et al.)	$\mathcal{O}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\right)$	$\mathcal{O}\big(\min\{n^{\frac{7}{2}}, m^{\frac{7}{2}}\}\big)$	✗	✓
SCSC (Yang et al.)	$\mathcal{O}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\right)$	$\mathcal{O}\big(\min\{n^{\frac{5}{2}}, m^{\frac{5}{2}}\}\big)$	✗	✓
NSCGD (Thm 4.4)	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}+\frac{\sqrt{d}}{\epsilon}\left(\frac{1}{n}+\frac{1}{m}\right)\right)$	$\mathcal{O}\big(\min\{n^6, m^6\}\big)$	✓	✓
NSCSC (Thm 4.5)	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}+\frac{\sqrt{d}}{\epsilon}\left(\frac{1}{n}+\frac{1}{m}\right)\right)$	$\mathcal{O}\big(\min\{n^4, m^4\}\big)$	✓	✓
Phased OP (Thm 5.4)	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d}}{\sqrt{m}}+\frac{\sqrt{d}}{\epsilon}\left(\frac{1}{n}+\frac{1}{\sqrt{m}}\right)\right)$	$\tilde{\mathcal{O}}\big(n^{\frac{3}{2}}+m\big)$	✓	✓

minimization (ERM) problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x}) := f_S(g_S(\mathbf{x})), \tag{2}$$

where the empirical outer function and inner functions are

$$f_S(\mathbf{y}) := \frac{1}{n}\sum_{i=1}^{n} F(\mathbf{y};\boldsymbol{\xi}_i), \quad g_S(\mathbf{x}) := \frac{1}{m}\sum_{j=1}^{m} G(\mathbf{x};\boldsymbol{\zeta}_i).$$

We denote the optimal solutions $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$ and $\mathbf{x}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x})$. For any $i \in [n]$ and $j \in [m]$, let $S^{i,\boldsymbol{\xi}}$ be the same copy of S except the i-th data point $\boldsymbol{\xi}_i$ in $S_{\boldsymbol{\xi}}$ is changed to $\boldsymbol{\xi}_i'$ while $S_{\boldsymbol{\zeta}}$ remains the same. Similarly, let $S^{j,\boldsymbol{\zeta}}$ be the same copy of S except the j-th data point $\boldsymbol{\zeta}_j$ in $S_{\boldsymbol{\zeta}}$ is changed to $\boldsymbol{\zeta}_j'$ while $S_{\boldsymbol{\xi}}$ remains the same.

3.2 Assumptions

In the remainder of this paper, we always suppose the following assumptions are satisfied.

Assumption 3.1. *The function $F(\cdot;\boldsymbol{\xi})$ is L_f-Lipschitz continuous and $G(\cdot;\boldsymbol{\zeta})$ is L_g-Lipschitz continuous for any given $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$, respectively.*

Assumption 3.2. *We assume the following conditions hold true.*

- *For any $\boldsymbol{\zeta}$, it holds that $\sup_{\mathbf{x}\in\mathcal{X}} \|G(\mathbf{x},\boldsymbol{\zeta})\| \le M_g$.*
- *Given any random variable $\boldsymbol{\zeta} \sim \mathcal{Q}$, it holds that $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\|G(\mathbf{x},\boldsymbol{\zeta}) - g(\mathbf{x})\|^2] \le V_g$.*
- *Given any random variable $\boldsymbol{\xi}$, the function $F(\cdot;\boldsymbol{\xi})$ has C_f-Lipschitz continuous gradients such that*

$$\|\nabla F(\mathbf{x};\boldsymbol{\xi}) - \nabla F(\mathbf{y};\boldsymbol{\xi})\| \le C_f \|\mathbf{x}-\mathbf{y}\|$$

for any $\mathbf{x},\mathbf{y} \in \mathbb{R}^p$.

- *Given any random variable $\boldsymbol{\zeta}$, the function $G(\cdot;\boldsymbol{\zeta})$ has C_g-Lipschitz continuous gradients such that*

$$\|\nabla G(\mathbf{x};\boldsymbol{\zeta}) - \nabla G(\mathbf{y};\boldsymbol{\zeta})\| \le C_g \|\mathbf{x}-\mathbf{y}\|$$

for any $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$.

We remark that Assumption 3.1 and 3.2 are common in existing literature of the stochastic compositional optimization problem (Lian et al., 2017; Wang et al., 2017a,b; Yang et al., 2024).

Assumption 3.3. *The feasible set $\mathcal{X}$ is convex and compact, and its diameter is bounded by D_X.*

3.3 Differential Privacy

Differential privacy (DP) introduced by Dwork (2006) measures the privacy leakage of an algorithm, which is defined as follows.

Definition 3.4. *For two datasets $S = \{\boldsymbol{\xi}_i\}_{i=1}^n$ and $S' = \{\boldsymbol{\xi}_i'\}_{i=1}^n$, we say a pair (S, S') is neighboring if $\max\{|S \setminus S'|, |S' \setminus S|\} = 1$ and we denote the neighboring datasets with $S \sim S'$. For an algorithm $\mathcal{A}$, some privacy parameters $\epsilon > 0$ and $\delta \in (0,1)$, we say $\mathcal{A}$ satisfies (ϵ,δ)-differential privacy if $\mathcal{P}(\mathcal{A}(S) \in O) \le e^{\epsilon}\mathcal{P}(\mathcal{A}(S') \in O) + \delta$ for all $S \sim S'$ and any subset O is in the range of $\mathcal{A}$.*

Our privacy guarantees leverage the well-known Gaussian mechanism, which ensures privacy by adding Gaussian noise to an algorithm with bounded sensitivity.

Definition 3.5 (Feldman et al. (2020))**.** *Let $\mathcal{A}$ be some randomized algorithm operating on S and outputting a vector in $\mathbb{R}^d$. If $\mathcal{A}$ has sensitivity $\Delta_{\mathcal{A}} :=$*

Algorithm 1: Gradient Perturbation

Inputs: Dataset $S = S_{\boldsymbol{\xi}} \cup S_{\boldsymbol{\zeta}}$, privacy budget ϵ, δ, number of iterations T, parameters $\{\eta_t\}$, $\{\beta_t\}$, σ, μ
Initialize $\mathbf{x}_0 = \mathbf{x}_{-1} \in \mathcal{X}$ and $\mathbf{y}_0 \in \mathbb{R}^p$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $j_t \in [m]$
  Let $\mathbf{z}_t = (1-\beta_t)(G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t}) - G(\mathbf{x}_{t-1}; \boldsymbol{\zeta}_{j_t}))$
  SCGD: $\mathbf{y}_{t+1} = (1-\beta_t)\mathbf{y}_t + \beta_t G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t})$
  SCSC: $\mathbf{y}_{t+1} = (1-\beta_t)\mathbf{y}_t + \beta_t G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t}) + \mathbf{z}_t$
  Sample $i_t \in [n]$
  Sample independent noise $\nu_t \sim \mathcal{N}(0, \sigma^2 I)$
  Let $\tilde{\mathbf{g}}_t = \nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t}) \nabla F(\mathbf{y}_{t+1}; \boldsymbol{\xi}_{i_t}) + \nu_t$
  Update $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta_t \tilde{\mathbf{g}}_t)$
end
Return: $\tilde{\mathbf{x}} = \frac{\sum_{t=1}^{T}(1-\mu\eta_t/2)^{T-t}\mathbf{x}_t}{\sum_{t=1}^{T}(1-\mu\eta_t/2)^{T-t}}$.

$\sup_{S\sim S'} \|\mathcal{A}(S) - \mathcal{A}(S')\|$ *with probability at least* $1-\delta$*, then the Gaussian mechanism outputs* $\mathcal{A}(S) + \mathcal{N}(0, (\Delta_{\mathcal{A}}\sqrt{2\log(1.25/\delta)}/\epsilon)^2 I_d)$ *and it achieves* $(\epsilon, 2\delta)$*-DP.*

4 NOISY STOCHASTIC METHODS

In this section, we propose noisy stochastic compositional gradient descent (NSCGD) and noisy stochastically corrected stochastic compositional gradient (NSCSC) for solving the DP-SCO problem. Then we provide the theoretical analysis of both methods in the strongly convex and convex cases.

4.1 The Algorithm

We present the details of the NSCGD and NSCSC method in Algorithm 1. At each iteration of the algorithm, it leverages the gradient estimator of the standard SCGD or the SCSC method to approximate the gradient of the compositional function $\Phi(\mathbf{x}_t)$. Then, a Gaussian noise is added to the gradient estimator, which leads to the noisy gradient estimator $\tilde{\mathbf{g}}_t$. After that, the model parameter is updated with projected gradient descent using the noisy gradient estimator. For the strongly convex case, the output of the algorithm is the weighted average of iterates across all iterations as follows,

$$\tilde{\mathbf{x}} = \frac{\sum_{t=1}^{T}(1-\mu\eta_t/2)^{T-t}\mathbf{x}_t}{\sum_{t=1}^{T}(1-\mu\eta_t/2)^{T-t}}.$$

For the convex case, the strong convexity parameter μ of the objective function is 0. It implies that the output of the algorithm is the average of all iterates such that $\tilde{\mathbf{x}} = \sum_{t=1}^{T}\mathbf{x}_t/T$.

By leveraging the moments accountant method stated in Lemma A.9, we can characterize the privacy guarantee of the NSCGD and NSCSC with the following theorem.

Theorem 4.1. *There exist some constants* c_1*,* c_1'*,* c_2 *and* c_2' *so that for any* $\epsilon \le T\min\{c_1' n^{-2}, c_2' m^{-2}\}$ *and* $0 < \delta \le \frac{1}{2}$*, Algorithm 1 is* (ϵ, δ)*-differentially private if we choose*

$$\sigma = \frac{c_1 L_f L_g \sqrt{T\log(1/\delta)}}{n\epsilon} + \frac{c_2 L_f L_g \sqrt{T\log(1/\delta)}}{m\epsilon}.$$

Given the privacy guarantee of the proposed methods, we will present the excess population loss of both methods in the strongly convex and convex cases in the next subsection.

4.2 Utility Analysis

We first consider the utility analysis of the proposed methods in the strongly convex case. The following theorem shows the excess population loss bound for solving the DP-SCO problem by Algorithm 1 with the SCGD update in the strongly convex setting.

Theorem 4.2. *Suppose that the functions* $\Phi(\cdot)$ *and* $F(G_S(\cdot); \boldsymbol{\xi})$ *are* μ*-strongly convex for any* $\boldsymbol{\xi}$ *and* S*. By choosing the following parameters*

$$T = \Theta\big(\min\{n^{\frac{14}{3}}, m^{\frac{7}{3}}\}\big), \quad \beta_t = \beta = T^{-\frac{3}{7}},$$
$$\eta_t = \eta = \frac{1}{\max\big(T^{\frac{13}{14}}, T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1} + n^{-1}\epsilon^{-1})\big)},$$

the NSCGD method (Algorithm 1 with the SCGD update) is (ϵ, δ)*-DP and its output* $\tilde{\mathbf{x}}$ *satisfies*

$$\mathbb{E}_{S,\mathcal{A}}\left[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)\right] = \mathcal{O}\left(\frac{1}{\mu n} + \frac{1}{\mu\sqrt{m}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{\mu n} + \frac{1}{\mu m}\right)\right).$$

The SCSC method has been demonstrated to achieve the same population loss with improved gradient complexity over the SCGD method in the context of the non-private SCO problem. It is therefore of interest to investigate whether the NSCSC method can similarly outperform the NSCGD method in terms of gradient complexity under DP constraints. To this end, we present the following bound on the excess population loss for solving the DP-SCO problem by Algorithm 1 with the SCSC update.

Theorem 4.3. *Suppose that the functions* $\Phi(\cdot)$ *and* $F(G_S(\cdot); \boldsymbol{\xi})$ *are* μ*-strongly convex for any* $\boldsymbol{\xi}$ *and* S*. By choosing the following parameters*

$$T = \Theta\big(\min\{n^{\frac{7}{3}}, m^{\frac{7}{6}}\}\big), \quad \beta_t = \beta = T^{-\frac{6}{7}},$$

$$\eta_t = \eta = \frac{1}{\max\left(T^{\frac{13}{14}}, T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1} + n^{-1}\epsilon^{-1})\right)},$$

the NSCSC method (Algorithm 1 with the SCSC update) is (ϵ, δ)-DP and its output $\tilde{\mathbf{x}}$ satisfies

$$\mathbb{E}_{S,\mathcal{A}}\left[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)\right] = \mathcal{O}\left(\frac{1}{\mu n} + \frac{1}{\mu\sqrt{m}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{\mu n} + \frac{1}{\mu m}\right)\right).$$

Our proof strategy decomposes the error term $\mathbb{E}_{S,\mathcal{A}}\left[\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}^*)\right]$ into generalization, optimization, and approximation errors: $\mathbb{E}_{S,\mathcal{A}}\left[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)\right]$, $\mathbb{E}_{S,\mathcal{A}}\left[\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)\right]$, and $\mathbb{E}_{S,\mathcal{A}}\left[\Phi_S(\mathbf{x}^*) - \Phi(\mathbf{x}^*)\right]$, respectively. The generalization error is analyzed via algorithmic stability, the optimization error is handled using classical optimization theory, and the approximation error is bounded under Assumption 3.2. This decomposition allows us to separately quantify the contributions of data sampling and optimization steps to the overall error. Complete proofs of Theorems 4.2 and 4.3 are provided in Appendix B.1.1.

We now present the utility analysis of the NSCGD and NSCSC methods for the convex case. The following theorem shows the excess population loss bound for solving the DP-SCO problem by Algorithm 1 with the SCGD update in the convex setting.

Theorem 4.4. *Suppose that the functions $\Phi(\cdot)$ and $F(G_S(\cdot); \boldsymbol{\xi})$ are convex for any $\boldsymbol{\xi}$ and S. By choosing*

$$T = \Theta\big(\min\{n^6, m^6\}\big), \quad \beta_t = \beta = T^{-\frac{1}{3}},$$
$$\eta_t = \eta = \frac{1}{\max\left(T^{\frac{11}{12}}, T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1} + n^{-1}\epsilon^{-1})\right)},$$

the NSCGD method (Algorithm 1 with the SCGD update) is (ϵ, δ)-DP and its output $\tilde{\mathbf{x}}$ satisfies

$$\mathbb{E}_{S,\mathcal{A}}\left[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)\right] = \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

Then we establish the utility guarantee for solving the DP-SCO problem by Algorithm 1 with the SCSC update. This result demonstrates that incorporating the stochastic correction term $\mathbf{z}_t$ improves the gradient complexity of the NSCGD method.

Theorem 4.5. *Suppose that the functions $\Phi(\cdot)$ and $F(G_S(\cdot); \boldsymbol{\xi})$ are convex for any $\boldsymbol{\xi}$ and S. By choosing*

$$T = \Theta\big(\min\{n^4, m^4\}\big), \quad \beta_t = \beta = T^{-\frac{1}{2}},$$
$$\eta_t = \eta = \frac{1}{\max\left(T^{\frac{7}{8}}, T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1} + n^{-1}\epsilon^{-1})\right)},$$

Algorithm 2: Output Perturbation

1 **Inputs:** data $S = S_{\boldsymbol{\xi}} \cup S_{\boldsymbol{\zeta}}$, privacy budget ϵ, δ, strong convexity parameter μ, and noise scaling parameter σ, a randomized algorithm $\mathcal{A}$.

2 Run the algorithm $\mathcal{A}$ on the strongly convex smooth finite-sum problem $\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n F(\frac{1}{m}\sum_{j=1}^m G(\mathbf{x}; \boldsymbol{\zeta}_j); \boldsymbol{\xi}_i)$ to obtain output $\mathcal{A}(S)$ such that $\|\mathcal{A}(S) - \mathbf{x}_S^*\| \leq \frac{L_f L_g}{\mu n} + \sqrt{\frac{L_f M_g}{\mu m}}$ with probability at least $1 - \delta/4$, where the variable $\mathbf{x}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x})$ is the optimal solution of the empirical loss.

3 **Return:** $\tilde{\mathbf{x}} = \Pi_{\mathcal{X}}(\mathcal{A}(S) + \mathcal{N}(0, \sigma^2 I_d))$.

the NSCSC method (Algorithm 1 with the SCSC update) is (ϵ, δ)-DP and its output $\tilde{\mathbf{x}}$ satisfies

$$\mathbb{E}_{S,\mathcal{A}}\left[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)\right] = \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

The complete proof of Theorem 4.4 and 4.5 is presented in Appendix B.2.1. We remark that the excess population loss of NSCGD and NSCSC methods matches that of their non-private counterpart (Yang et al., 2024) in both strongly convex and convex settings by taking $\epsilon, \delta \to \infty$.

5 QUERY-EFFICIENT STOCHASTIC METHODS

We have shown that NSCGD and NSCSC methods require super-linear gradient queries to solve the DP-SCO problem. In this section, we propose a class of output perturbation-based algorithms for solving the DP-SCO problem with improved gradient complexity.

5.1 Strongly Convex Case

Firstly, we consider the case when the objective function is μ-strongly convex. The utility analysis of the NSCGD and NSCSC methods in Section 4 heavily relies on the stability results of the SCGD and SCSC methods. Such stability results are lacking for more sophisticated non-private randomized algorithms including variance-reduced and accelerated methods (Lian et al., 2017; Wang et al., 2017b). To sidestep the algorithm-specific stability analysis, we leverage the stability of the optimal solution to the compositional ERM problem. In particular, we provide the following theorem to show the optimal solution to the problem (2) has bounded

sensitivity, which leads to a bounded generalization error.

Lemma 5.1. *Let the function $F(G_S(\cdot), \boldsymbol{\xi})$ be μ-strongly convex for any $\boldsymbol{\xi}$ and S. Then for any neighboring datasets $S \sim S'$, it holds that*

$$\|\mathbf{x}_S^* - \mathbf{x}_{S'}^*\| \leq \frac{2L_f L_g}{\mu n} + 2\sqrt{\frac{L_f M_g}{\mu m}}.$$

where we denote $\mathbf{x}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x})$ and $\mathbf{x}_{S'}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_{S'}(\mathbf{x})$. Given some scalar $v \in (0,1)$, the stability result leads to the bound of the generalization error as follows.

$$\begin{aligned}&\mathbb{E}[\Phi(\mathbf{x}_S^*) - \Phi(\mathbf{x}^*)]\\ \leq&\frac{2L_f^2 L_g^2}{\mu n} + \frac{cL_f M_g}{\sqrt{m}}\sqrt{d\log\left(\frac{D_X}{v}\right)} + \frac{2L_f V_g}{\sqrt{m}} \\ &+ 2vL_f L_g.\end{aligned} \tag{3}$$

where $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$ is the optimal solution and c is some constant.

Lemma 5.1 ensures that if the output of an algorithm $\mathcal{A}$ is sufficiently close to the optimal solution of the compositional ERM, its sensitivity is bounded. Thus, applying the Gaussian mechanism to the output of $\mathcal{A}$ yields differential privacy. The procedure is summarized in Algorithm 2, and the following theorem establishes its utility guarantee, with the proof deferred to Appendix C.2.1.

Theorem 5.2. *Let the function $\Phi(\cdot)$ and $F(G_S(\cdot); \boldsymbol{\xi})$ be μ-strongly convex for any $\boldsymbol{\xi}$ and S. By choosing the parameter*

$$\sigma = \left(\frac{4L_f L_g}{\mu n} + 4\sqrt{\frac{L_f M_g}{\mu m}}\right)\frac{\sqrt{2\log(2.5/\delta)}}{\epsilon},$$

Algorithm 2 is (ϵ, δ)-DP and its output $\tilde{\mathbf{x}}$ satisfies that

$$\begin{aligned}&\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)]\\ =&\tilde{\mathcal{O}}\left(\frac{1}{\mu^2 n} + \frac{\sqrt{d}}{\mu\sqrt{m}} + \frac{d\log(1/\delta)}{\epsilon^2}\left(\frac{1}{\mu^2 n^2} + \frac{1}{\mu m}\right)\right).\end{aligned} \tag{4}$$

We note that the second term in the generalization bound (4) exhibits a $\sqrt{d}$ dependency, which originates from the corresponding term in the generalization error of the optimal compositional ERM solution (3). However, this dependency of $\sqrt{d}$ in equation (3) is unimprovable, as it aligns with the sample complexity established in the prior work (Hu et al., 2020).

For the gradient complexity, Algorithm 2 requires the output of the algorithm $\mathcal{A}$ satisfies $\|\mathcal{A}(S) - \mathbf{x}_S^*\| \leq L_f L_g/(\mu n) + \sqrt{L_f M_g/(\mu m)}$ with probability $1 - \delta/4$. However, the convergence result of most existing work (Lian et al., 2017; Wang et al., 2017a,b) for solving the strongly convex composite ERM problem (2) is measured by the suboptimality of the function value gap such that $\mathbb{E}[\Phi_S(\mathcal{A}(S)) - \Phi_S(\mathbf{x}_S^*)] \leq \iota$. The strong convexity of $\Phi_S(\cdot)$ and Jensen's inequality imply that $\mathbb{E}[\|\mathcal{A}(S) - \mathbf{x}_S^*\|] \leq \sqrt{2\iota/\mu}$. Applying Markov's inequality, we can obtain that

$$\begin{aligned}&\mathbb{P}\left(\|\mathcal{A}(S) - \mathbf{x}_S^*\| > \frac{L_f L_g}{\mu n} + \sqrt{\frac{L_f M_g}{\mu m}}\right)\\ \leq&\sqrt{2\iota}\left(\frac{L_f L_g}{\sqrt{\mu} n} + \sqrt{\frac{L_f M_g}{m}}\right)^{-1}.\end{aligned}$$

By taking $\iota = \delta^2 L_f^2 L_g^2/(32\mu n^2) + \delta^2 L_f M_g/(32m)$, we can derive that above probability is less than $\delta/4$. Consequently, the convergence criteria for $\mathcal{A}$ in Algorithm 2 is satisfied. Among existing randomized algorithms for the finite-sum strongly convex SCO problem, the best-known convergence rate is achieved by the compositional SVRG method (Lian et al., 2017). If compositional SVRG is used as the subroutine algorithm $\mathcal{A}$, the total sample complexity of Algorithm 2 becomes $\mathcal{O}((m + n + \hat{\varkappa}^3)\log(m/(\delta L_f M_g))$ where

$$\hat{\varkappa} := \max\left\{\frac{C_f}{\mu}, \frac{L_f C_g + L_g^2 C_f}{\mu}, \frac{L_f C_g + L_g^2 C_f}{C_f}\right\}.$$

We emphasize that our output perturbation framework is compatible with other stochastic methods for solving the SCO problem, and a detailed comparison of convergence rates for existing methods is provided in Table 3.

5.2 Convex Case

We proceed to consider the case when the objective function is convex. The output perturbation method (Algorithm 2) cannot be directly applied to convex problems, which generally admit non-unique solutions, in contrast to the strongly convex case, which ensures uniqueness. A straightforward approach to solving the problem is to transform the convex problem into a strongly convex one by adding an ℓ_2 regularizer $\mu\|\mathbf{x}\|^2/2$ to the objective function. Algorithm 2 is then applied to solve the regularized problem with a carefully chosen parameter μ. This method yields an excess population loss bound of

$$\mathcal{O}\left(\left(\frac{1}{\sqrt{n}} + \frac{1}{m^{\frac{1}{3}}}\right)\frac{\sqrt{d\log(1/\delta)}}{\epsilon} + \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\sqrt{m}} + \frac{1}{m^{\frac{1}{3}}}\right),$$

and the proof is deferred to Appendix C.3.1.

Algorithm 3: Phased Output Perturbation

Inputs: data $S = S_{\boldsymbol{\xi}} \cup S_{\boldsymbol{\zeta}}$, privacy budget ϵ, δ, number of iterations K, some parameter μ, a sequence $\{\sigma_k\}$
Initialize $\tilde{\mathbf{x}}_0 = \mathbf{x}_0 \in \mathcal{X}$ and $K = \log n$, $\bar{n} = n/K$, $\bar{m} = m/K$
for $k = 1, \ldots, K$ **do**
  Set $\mu_k = \mu \cdot 2^k$
  Run the algorithm $\mathcal{A}$ on the regularized finite-sum minimization problem $\hat{\Phi}_{S,k}(\mathbf{x})$ defined in (5) to obtain the output $\mathbf{x}_k$ such that $\|\mathbf{x}_k - \hat{\mathbf{x}}_k^*\| \leq \frac{L_f L_g}{\mu_k \bar{n}} + \sqrt{\frac{L_f M_g}{\mu_k \bar{m}}}$ with probability at least $1 - \delta/4$, where $\hat{\mathbf{x}}_k^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \hat{\Phi}_{S,k}(\mathbf{x})$ is the empirical optimal solution.
  $\tilde{\mathbf{x}}_k = \Pi_{\mathcal{X}}(\mathbf{x}_k + \mathcal{N}(0, \sigma_k^2 I_d))$
end
Return: $\tilde{\mathbf{x}}_K$.

Building on the phased regularization approach (Feldman et al., 2020) for nonsmooth stochastic convex optimization, we propose a multi-phase output perturbation method for the DP-SCO problem, which achieves an improved excess population loss bound compared with the single-regularizer approach. We first introduce the intuition of our algorithm. We divide the whole procedure into K phases and at each phase k, we use some randomized algorithm $\mathcal{A}$ to obtain an approximate solution $\mathbf{x}_k$ of the regularized problem

$$\hat{\Phi}_{S,k}(\mathbf{x}) := \frac{1}{\bar{n}} \sum_{j=(k-1)\bar{n}+1}^{k\bar{n}} F\left(\frac{1}{\bar{m}} \sum_{j=(k-1)\bar{m}+1}^{k\bar{m}} G(\mathbf{x}; \boldsymbol{\zeta}_j); \boldsymbol{\xi}_i\right) + \frac{\mu_k}{2} \|\mathbf{x} - \tilde{\mathbf{x}}_{k-1}\|^2, \tag{5}$$

where we choose $\bar{m} = m/\log(n)$, $\bar{n} = n/\log(n)$, and $\{\mu_k = \mu \cdot 2^k\}$ is an increasing sequence of regularization parameters. We then add a Gaussian noise with the scaling parameter σ_k on the output $\mathbf{x}_k$ to ensure differential privacy. The complete procedure, referred to as the phased output perturbation (phased OP) method, is outlined in Algorithm 3.

To establish the privacy and utility guarantees of the algorithm, we first present the following stability and generalization results. This lemma extends Lemma 5.1 by removing the smoothness assumption on the regularizer Ψ, which introduces additional technical challenges. The full proof is provided in Appendix C.1.1.

Lemma 5.3. *Let the function $F(G_S(\cdot), \boldsymbol{\xi})$ be convex for any $\boldsymbol{\xi}$ and S, and $\Psi(\cdot)$ be any μ-strongly convex function. For any neighboring datasets $S \sim S'$, it holds that*

$$\|\hat{\mathbf{x}}_{S'}^* - \hat{\mathbf{x}}_S^*\| \leq \frac{2L_f L_g}{\mu n} + 2\sqrt{\frac{L_f M_g}{\mu m}}.$$

where we denote $\hat{\mathbf{x}}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\Phi_S(\mathbf{x}) + \Psi(\mathbf{x})\}$ and $\hat{\mathbf{x}}_{S'}^* = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\Phi_{S'}(\mathbf{x}) + \Psi(\mathbf{x})\}$.

Given some scalar $v \in (0, 1)$, the stability result leads to the bound of the generalization error as follows,

$$\begin{aligned}&\mathbb{E}[\Phi(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*) - \Phi(\hat{\mathbf{x}}^*) - \Psi(\hat{\mathbf{x}}^*)]\\ \leq &\frac{2L_f^2 L_g^2}{\mu n} + \frac{cL_f M_g}{\sqrt{m}}\sqrt{d\log\left(\frac{D_X}{v}\right)} + \frac{2L_f V_g}{\sqrt{m}} + 2vL_f L_g.\end{aligned}$$

where $\hat{\mathbf{x}}^* = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\Phi(\mathbf{x}) + \Psi(\mathbf{x})\}$ is the optimal solution, and c is some constant.

The above lemma implies that the algorithm $\mathcal{A}$ at each phase has bounded sensitivity if its output is sufficiently close to the minimizer of the regularized ERM problem. Adding Gaussian noise with an appropriate choice of scaling parameter σ_k ensures differential privacy of $\tilde{\mathbf{x}}_k$. Since this method only accesses a disjoint partition of the whole dataset S during each phase, we can apply the parallel composition theorem (Lemma A.8) to ensure the privacy guarantee of the whole dataset.

We proceed to consider the utility bound of excess population loss. The following theorem shows the utility guarantee for solving the DP-SCO problem by Algorithm 3 in the convex case, with the proof deferred to Appendix C.3.2.

Theorem 5.4. *Let the function $\Phi(\cdot)$ and $F(G_S(\cdot); \boldsymbol{\xi})$ be convex for any $\boldsymbol{\xi}$ and S. By choosing the parameter*

$$\sigma_k = \left(\frac{4L_f L_g}{\mu_k \bar{n}} + 4\sqrt{\frac{L_f M_g}{\mu_k \bar{m}}}\right)\frac{\sqrt{2\log(2.5/\delta)}}{\epsilon},$$

$$\mu = \max\left\{\frac{L_f L_g}{D_X\sqrt{n}}, \frac{L_f L_g}{D_X n}\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon}\right\},$$

Algorithm 3 is (ϵ, δ)-DP and its output $\tilde{\mathbf{x}}_K$ satisfies that

$$\begin{aligned}&\mathbb{E}[\Phi(\tilde{\mathbf{x}}_K) - \Phi(\mathbf{x}^*)]\\ =&\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\sqrt{m}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{n} + \frac{1}{\sqrt{m}}\right)\right).\end{aligned}$$

For the gradient complexity of Algorithm 3 with the compositional SVRG as its subroutine $\mathcal{A}$, we can show that phase k requires $\mathcal{O}\big((\bar{m} + \bar{n} + \mu_k^{-3})\log(\bar{m}/\delta)\big)$ gradient computations following a similar analysis for the strongly convex case. By summing up k from 1 to K, it leads to a total gradient complexity of $\mathcal{O}\big((m + n^{\frac{3}{2}})\log(m/\delta)\big)$.

6 CONCLUSION

In this paper, we introduce NSCGD and NSCSC methods for solving the stochastic compositional optimization problem under the DP constraint. We present the privacy guarantee and the excess population loss of the proposed methods for both strongly convex and convex cases. In addition, we propose a class of output perturbation methods for solving the DP-SCO problem with an improved gradient complexity.

In future work, it is interesting to study the lower bound of excess population loss to solve the DP-SCO problem. It is also an intriguing research direction to design randomized algorithms for solving the DP-SCO problem in the nonconvex setting.

Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2023-08-043T-J). This research is part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-029). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Asi, H., Feldman, V., Koren, T., and Talwar, K. (2021). Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR.

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

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. (2020). Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. (2019). Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32.

Bassily, R., Guzmán, C., and Nandi, A. (2021). Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE.

Boob, D. and Guzmán, C. (2024). Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming*, 204(1):255–297.

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

Chaudhuri, K. and Monteleoni, C. (2008). Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

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

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

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Feldman, V., Koren, T., and Talwar, K. (2020). Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449.

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

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradi-

ent descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

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

Kulkarni, J., Lee, Y. T., and Liu, D. (2021). Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*.

Kuzborskij, I. and Lampert, C. (2018). Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR.

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

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

Liu, L., Liu, J., and Tao, D. (2018). Duality-free methods for stochastic composition optimization. *IEEE transactions on neural networks and learning systems*, 30(4):1205–1217.

Liu, Z., Luo, L., and Low, B. K. H. (2024). Gradient-free methods for nonconvex nonsmooth stochastic compositional optimization. *Advances in Neural Information Processing Systems*, 37:45438–45461.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30.

Pan, X., Li, X., Liu, J., Sun, T., Sun, K., Chen, L., and Qu, Z. (2024). Stability and generalization for stochastic recursive momentum-based algorithms for (strongly-) convex one to k-level stochastic optimizations. *arXiv preprint arXiv:2407.05286*.

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

Sutton, R. S. (2018). Reinforcement learning: An introduction. *A Bradford Book*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

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

Wang, M., Liu, J., and Fang, E. X. (2017b). Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18(105):1–23.

Yang, M., Wei, X., Yang, T., and Ying, Y. (2024). Stability and generalization of stochastic compositional gradient descent algorithms. In *International Conference on Machine Learning*, pages 56542–56593. PMLR.

Yang, Z., Hu, S., Lei, Y., Vashney, K. R., Lyu, S., and Ying, Y. (2022). Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*, pages 2192–2202. PMLR.

Zhang, L., Thekumparampil, K. K., Oh, S., and He, N. (2022). Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]
(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Differentially Private Algorithms for the Stochastic Compositional Optimization Problem: Supplementary Materials

The appendix is organized as follows. Section A introduces several key lemmas for the analysis of stability and generalization of NSCGD and NSCSC methods. Section B presents the proof of privacy guarantee and the excess population loss of both NSCGD and NSCSC methods for solving the DP-SCO problem for strongly convex and convex cases. Section C presents the proof of the excess population loss of the output perturbation methods introduced in Section 5. Section D provides the convergence analysis of the SCSC method for solving the strongly convex SCO problem.

A Supporting Lemmas

In this section, we introduce several key lemmas which are essential for the analysis of NSCGD and NSCSC methods. We first show the smoothness constant of the compositional function $F(g_S(\cdot);\boldsymbol{\xi})$.

Lemma A.1. *The function $F(g_S(\cdot);\boldsymbol{\xi})$ is $L_fC_g + L_g^2C_f$ smooth.*

Proof. Assumption 3.1, 3.2 and the triangle inequality imply that:

$$\begin{aligned}
&\|\nabla g_S(\mathbf{x})\nabla F(g_S(\mathbf{x});\boldsymbol{\xi}) - \nabla g_S(\mathbf{y})\nabla F(g_S(\mathbf{y});\boldsymbol{\xi})\| \\
\leq& \|\nabla g_S(\mathbf{x})\nabla F(g_S(\mathbf{x});\boldsymbol{\xi}) - \nabla g_S(\mathbf{y})\nabla F(g_S(\mathbf{x});\boldsymbol{\xi})\| + \|\nabla g_S(\mathbf{y})\nabla F(g_S(\mathbf{x});\boldsymbol{\xi}) - \nabla g_S(\mathbf{y})\nabla F(g_S(\mathbf{y});\boldsymbol{\xi})\| \\
\leq& \|\nabla g_S(\mathbf{x}) - \nabla g_S(\mathbf{y})\| \|\nabla F(g_S(\mathbf{x});\boldsymbol{\xi})\| + \|\nabla g_S(\mathbf{y})\| \|\nabla F(g_S(\mathbf{x});\boldsymbol{\xi}) - \nabla F(g_S(\mathbf{y});\boldsymbol{\xi})\| \\
\leq& L_fC_g \|\mathbf{x}-\mathbf{y}\| + L_g^2C_f \|\mathbf{x}-\mathbf{y}\|.
\end{aligned}$$

□

Next we present the stability for SCO introduced in Yang et al. (2024) for the stability analysis of proposed algorithms.

Definition A.2 (Yang et al. (2024))**.** *We say that a randomized algorithm $\mathcal{A}$ is $(\epsilon_{\boldsymbol{\xi}}, \epsilon_{\boldsymbol{\zeta}})$-uniformly stable for ERM problem (2) if for any $i \in [n]$ and $j \in [m]$, it holds that*

$$\mathbb{E}_{\mathcal{A}}[\|\mathcal{A}(S) - \mathcal{A}(S^{i,\boldsymbol{\xi}})\|] \leq \epsilon_{\boldsymbol{\xi}} \quad \text{and} \quad \mathbb{E}_{\mathcal{A}}[\|\mathcal{A}(\mathrm{S}) - \mathcal{A}(\mathrm{S}^{\mathrm{j},\boldsymbol{\zeta}})\|] \leq \epsilon_{\boldsymbol{\zeta}},$$

where the expectation is taken w.r.t. the internal randomness of the algorithm A.

The stability result implies the following generalization property.

Theorem A.3 (Yang et al. (2024))**.** *Under Assumption 3.1, a randomized algorithm $\mathcal{A}$ is $(\epsilon_{\boldsymbol{\xi}}, \epsilon_{\boldsymbol{\zeta}})$-uniformly stable, then it holds that*

$$\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathcal{A}(S)) - \Phi_S(\mathcal{A}(S))] \leq L_fL_g\epsilon_{\boldsymbol{\xi}} + 4L_fL_g\epsilon_{\boldsymbol{\zeta}} + L_f\sqrt{\frac{\mathbb{E}_{S,\mathcal{A}}[\mathrm{Var}(\mathrm{G}(\mathcal{A}(\mathrm{S});\boldsymbol{\zeta}))]}{m}},$$

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

Furthermore, Yang et al. (2024) introduced the following lemmas for the utility analysis of non-private SCGD and SCSC methods, which are useful for our analysis.

Lemma A.4 (Yang et al. (2024))**.** *For any $\iota, c > 0$, we have*

$$e^{-\iota x} \leq \left(\frac{c}{\iota e}\right)^c x^{-c}.$$

Lemma A.5 (Yang et al. (2024)). *Let $\{a_i\}_{i=1}^T$, $\{b_i\}_{i=1}^T$ be two sequences of positive real numbers such that $a_i \le a_{i+1}$ and $b_i \ge b_{i+1}$ for all i. Then we have*

$$\frac{\sum_{i=1}^T a_i b_i}{\sum_{i=1}^T a_i} \le \frac{\sum_{i=1}^T b_i}{T}.$$

The following two lemmas track the difference between the variable $\mathbf{y}_{t+1}$ and inner function value $g_S(\mathbf{x}_t)$ for the SCGD and SCSC methods.

Lemma A.6 (Wang et al. (2017a)). *Under Assumption 3.1 and 3.2, running Algorithm 1 with SCGD update, we have*

$$\|\mathbf{y}_{t+1} - g_S(\mathbf{x}_t)\|^2 \le (1-\beta_t)\|\mathbf{y}_t - g_S(\mathbf{x}_{t-1})\|^2 + \frac{L_g^2}{\beta_t}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 2V_g\beta_t^2.$$

Lemma A.7 (Chen et al. (2021)). *Under Assumption 3.1 and 3.2, running Algorithm 1 with SCSC update, we have*

$$\|\mathbf{y}_{t+1} - g_S(\mathbf{x}_t)\|^2 \le (1-\beta_t)\|\mathbf{y}_t - g_S(\mathbf{x}_{t-1})\|^2 + L_g^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 2V_g\beta_t^2.$$

Finally, we present some established results in the DP literature. Given a single Gaussian mechanism defined in Definition 3.5, there are some ways of combining them to design more sophisticated privacy-preserving algorithms. One fundamental composition result states that if disjoint data batches are used at each iteration, the composition will retain the largest privacy parameter among its building blocks. We give the formal definition below.

Lemma A.8 (McSherry (2009)). *Given a dataset S and its disjoint partition $S = \cup_{k=1}^K S_k$, we define the mechanisms as $\mathcal{A}_1 = \mathcal{A}_1(S_1)$, $\mathcal{A}_2 = \mathcal{A}_2(S_2; \mathcal{A}_1), \ldots, \mathcal{A}_K = \mathcal{A}_K(S_K; \mathcal{A}_{K-1})$. Suppose each mechanism $\mathcal{A}_k(S_k; \mathcal{A}_{k-1})$ is (ϵ_k, δ_k)-DP with respect to the dataset S_k for $k = 1, \ldots, K$, then the composition $\mathcal{A}_K$ is $(\max_k \epsilon_k, \max_k \delta_k)$-DP with respect to the full dataset S.*

Some of the algorithms proposed in this work leverage sampling with replacement to select data used in each iteration, then the privacy of such algorithms can be obtained with the moment accountant method Abadi et al. (2016). It is formally defined as follows.

Lemma A.9 (Abadi et al. (2016); Kulkarni et al. (2021)). *Let $D \in \mathcal{Y}^n$ be a dataset over some domain $\mathcal{Y}$, and let $h_1, \ldots, h_T \colon \mathcal{Y} \to \mathbb{R}^d$ be a series of functions with $\|h_t(\mathbf{y})\| \le \lambda$ for any $\mathbf{y} \in \mathcal{Y}$ and $t \in [T]$. Let $\epsilon < c_1 b^2 T/n^2, \delta \in (0, 1/2]$ where c_1 is some constant and $b \le n/10$ is the minibatch size. Let $\sigma_t = c_2 b\lambda\sqrt{T\log(1/\delta)}/n\epsilon$ given some constant c_2. Then the algorithm which samples batches $B_1, \ldots, B_T$ of size b uniformly from D and outputs $\frac{1}{b}\sum_{\mathbf{y}\in B_t} h_t(\mathbf{y}) + \nu_t$ for all $t \in [T]$ where $\nu_t \sim \mathcal{N}(0, \sigma_t^2 I)$, is (ϵ, δ)-DP.*

B Privacy and Utility Analysis of NSCGD and NSCSC methods

The following concept of ℓ_2 stability is widely used in the DP literature Dwork et al. (2014).

Definition B.1. *Given a function $g \colon \mathcal{Z}^n \to \mathbb{R}^d$, we say g has $\Delta(g)$ ℓ_2-sensitivity if for any neighboring dataset S, S', we have*

$$\|g(S) - g(S')\| \le \Delta(g).$$

Now we can show both NSCGD and NSCSC methods are (ϵ, δ)-DP under an appropriate choice of hyperparameters.

Proof of Theorem 4.1

Proof. At iteration t, we focus on the update $\mathbf{g}_t(S) = \nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1}(S); \boldsymbol{\xi}_{i_t}) + \nu_t$ where $\nu_t \sim \mathcal{N}(0, \sigma^2)$. Assumption 3.1 and Cauchy–Schwarz inequality implies that

$$\|\nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1}(S); \boldsymbol{\xi}_{i_t})\| \le L_f L_g.$$

Considering the neighboring datasets S and S' and the only different index of the datasets are $\boldsymbol{\xi}_{i_t}$ and $\boldsymbol{\xi}'_{i_t}$. By Lemma A.9, to guarantee the algorithm be (ϵ, δ)-differentially private w.r.t. such neighboring datasets, it suffices to choose

$$\sigma \geq \frac{c_1 L_f L_g \sqrt{T \log(1/\delta)}}{n\epsilon}. \tag{6}$$

for some explicit constants c_1.

Now consider the neighboring datasets S and S' where the only different indexes of the datasets are $\boldsymbol{\zeta}_j$ and $\boldsymbol{\zeta}'_j$. By Lemma A.9, to guarantee Algorithm 1 be (ϵ, δ)-differentially private, it suffices to choose

$$\sigma \geq \frac{c_2 L_f L_g \sqrt{T \log(1/\delta)}}{m\epsilon}, \tag{7}$$

for some explicit constants c_2. Combining Eq.(6) and (7) with adaptive composition, we can set

$$\sigma = \frac{c_1 L_f L_g \sqrt{T \log(1/\delta)}}{n\epsilon} + \frac{c_2 L_f L_g \sqrt{T \log(1/\delta)}}{m\epsilon}.$$

□

We proceed to consider the excess population risk of both NSCGD and NSCSC methods. Note that the excess population loss of the SCO problem at each iteration can be decomposed into three error terms,

$$\begin{aligned} &\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}^*)] \\ \leq &\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)] + \mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)] + \mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}^*) - \Phi(\mathbf{x}^*)], \end{aligned} \tag{8}$$

where the inequality follows from the optimality of $\mathbf{x}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x})$. Our basic idea is to use algorithmic stability to study the generalization error term $\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)]$ and use optimization theory to study the optimization error $\mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)]$, and the third error can be upper bounded by applying Assumption 3.1.

In addition, we need the following key theorem to derive the optimization error bounds for the NSCGD and NSCSC methods.

Theorem B.2. *Under Assumption 3.1 and 3.2, running Algorithm 1 with $\eta_t = \eta$ and $\beta_t = \beta$.*

- *With the SCGD update, it holds that*

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{t+1} - g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t] \\ \leq &\left(\frac{c}{e}\right)^c (t\beta)^{-c} \mathbb{E}[\|\mathbf{y}_1 - g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + \frac{L_f^2 L_g^4 \eta^2}{\beta^2} + \frac{d L_g^2 \sigma^2 \eta^2}{\beta^2} + 2V_g\beta. \end{aligned}$$

- *With the SCSC update, it holds that*

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{t+1} - g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t] \\ \leq &\left(\frac{c}{e}\right)^c (t\beta)^{-c} \mathbb{E}[\|\mathbf{y}_1 - g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + \frac{L_f^2 L_g^4 \eta^2}{\beta} + \frac{d L_g^2 \sigma^2 \eta^2}{\beta} + 2V_g\beta. \end{aligned}$$

Proof. We first present the proof for the SCGD update. Lemma A.6 implies that

$$\begin{aligned} &\|\mathbf{y}_{t+1} - g_S(\mathbf{x}_t)\|^2 \\ \leq &(1-\beta_t)\|\mathbf{y}_t - g_S(\mathbf{x}_{t-1})\|^2 + \frac{L_g^2}{\beta_t}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 2V_g\beta_t^2 \\ = &(1-\beta_t)\|\mathbf{y}_t - g_S(\mathbf{x}_{t-1})\|^2 + \frac{L_g^2\eta_{t-1}^2}{\beta_t}\left\|\nabla G(\mathbf{x}_{t-1};\boldsymbol{\zeta}_{j_{t-1}})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_{t-1}}) + \nu_t\right\|^2 + 2V_g\beta_t^2 \\ = &(1-\beta_t)\|\mathbf{y}_t - g_S(\mathbf{x}_{t-1})\|^2 + \frac{L_g^2\eta_{t-1}^2}{\beta_t}\left\|\nabla G(\mathbf{x}_{t-1};\boldsymbol{\zeta}_{j_{t-1}})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_{t-1}})\right\|^2 + \frac{L_g^2\eta_{t-1}^2}{\beta_t}\|\nu_t\|^2 \end{aligned}$$

$$+ \frac{2L_g^2\eta_{t-1}^2}{\beta_t}\langle \nabla G(\mathbf{x}_{t-1};\boldsymbol{\zeta}_{j_{t-1}})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_{t-1}}),\nu_t\rangle + 2V_g\beta_t^2.$$

Take expectations on both sides of the inequality, we have

$$\begin{aligned}&\mathbb{E}_A[\|\mathbf{y}_{t+1}-g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t]\\ \le&(1-\beta_t)\mathbb{E}[\|\mathbf{y}_t-g_S(\mathbf{x}_{t-1})\|^2] + \frac{L_f^2L_g^4\eta_{t-1}^2}{\beta_t} + \frac{dL_g^2\eta_{t-1}^2\sigma^2}{\beta_t} + 2V_g\beta_t^2.\end{aligned}$$

Telescoping the above inequality from 1 to t yields

$$\begin{aligned}&\mathbb{E}_A[\|\mathbf{y}_{t+1}-g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t]\\ \le&\prod_{i=1}^t(1-\beta_i)\mathbb{E}[\|\mathbf{y}_1-g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + (L_f^2L_g^4+dL_g^2\sigma^2)\sum_{i=1}^t\prod_{j=i+1}^t(1-\beta_j)\frac{\eta_{i-1}^2}{\beta_i}\\ &+2V_g\sum_{i=1}^t\prod_{j=i+1}^t(1-\beta_j)\beta_i^2.\end{aligned}$$

Note that $1-\beta_i \le \exp(-\beta_i)$ for all $\beta_i > 0$, then setting $\eta_t = \eta$, $\beta_t = \beta$, we have

$$\begin{aligned}&\mathbb{E}_A[\|\mathbf{y}_{t+1}-g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t]\\ \le&\exp(-\beta t)\mathbb{E}[\|\mathbf{y}_1-g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + \sum_{i=1}^t(1-\beta)^{t-i}\left(\frac{L_f^2L_g^4\eta^2}{\beta} + \frac{dL_g^2\sigma^2\eta^2}{\beta} + 2V_g\beta^2\right).\end{aligned}$$

Using Lemma A.4 with $\iota = 1$, we have

$$\begin{aligned}&\mathbb{E}_A[\|\mathbf{y}_{t+1}-g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t]\\ \le&\left(\frac{c}{e}\right)^c(t\beta)^{-c}\mathbb{E}[\|\mathbf{y}_1-g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + \sum_{i=1}^t(1-\beta)^{t-i}\left(\frac{L_f^2L_g^4\eta^2}{\beta} + \frac{dL_g^2\sigma^2\eta^2}{\beta} + 2V_g\beta^2\right)\\ \le&\left(\frac{c}{e}\right)^c(t\beta)^{-c}\mathbb{E}[\|\mathbf{y}_1-g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + \frac{L_f^2L_g^4\eta^2}{\beta^2} + \frac{dL_g^2\sigma^2\eta^2}{\beta^2} + 2V_g\beta.\end{aligned}$$

Using a similar derivation as for the SCGD update, we arrive at the following result for the SCSC update.

$$\begin{aligned}&\mathbb{E}_A[\|\mathbf{y}_{t+1}-g_S(\mathbf{x}_t)\|^2 \mid \mathcal{F}_t]\\ \le&\left(\frac{c}{e}\right)^c(t\beta)^{-c}\mathbb{E}[\|\mathbf{y}_1-g_S(\mathbf{x}_0)\|^2 \mid \mathcal{F}_0] + \frac{L_f^2L_g^4\eta^2}{\beta} + \frac{dL_g^2\sigma^2\eta^2}{\beta} + 2V_g\beta.\end{aligned}$$

□

B.1 Strongly Convex Function

In this subsection, we present the utility analysis of the NSCGD and NSCSC methods for the strongly convex case. Firstly, we review the uniform stability results for the non-private SCGD and SCSC methods as follows.

Lemma B.3 (Yang et al. (2024, Theorem 5))**.** *Let $F(g_S(\cdot);\boldsymbol{\xi})$ be a μ-strongly convex function for any $\boldsymbol{\xi}$ and L is the smoothness constant of $\Phi_S(\cdot)$. Let $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$ be the sequence of iterates produced by non-private SCGD or SCSC method on dataset S, $\{\mathbf{x}_t^{i,\boldsymbol{\xi}}\}$ and $\{\mathbf{y}_t^{i,\boldsymbol{\xi}}\}$ be the sequence of iterated based on the dataset $S^{i,\boldsymbol{\xi}}$, $\{\mathbf{x}_t^{j,\boldsymbol{\zeta}}\}$ and $\{\mathbf{y}_t^{j,\boldsymbol{\zeta}}\}$ be the sequence of iterates based on the dataset $S^{j,\boldsymbol{\zeta}}$. Running the non-private SCGD or SCSC method with $\eta_t = \eta$ and $\beta_t = \beta \in (0,1)$ for any $t \in [T-1]$, then it holds that*

$$\begin{aligned}\mathbb{E}_A\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right] \le& 8C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\eta\frac{L\mu}{L+\mu}\right)^{t-k-1}\left(\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2]\right)^{\frac{1}{2}}\\ &+4L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}} + 2L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}} + \frac{4L_fL_g(L+\mu)}{nL\mu}\end{aligned}$$

and

$$\mathbb{E}_A\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right]\leq 8C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\eta\frac{L\mu}{L+\mu}\right)^{t-k-1}\left(\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2]\right)^{\frac{1}{2}}$$
$$+4L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{12L_fL_g(L+\mu)}{mL\mu}.$$

Based on the above result, we can derive the uniform stability of the NSCGD and NSCSC methods.

Lemma B.4. *Suppose that the function $F(g_S(\cdot);\boldsymbol{\xi})$ is μ-strongly convex. Let $r_k=\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|$, $\{\mathbf{x}_t\}$, $\{\mathbf{x}_t^{i,\boldsymbol{\xi}}\}$ and $\{\mathbf{x}_t^{j,\boldsymbol{\zeta}}\}$ be the sequences generated by Algorithm 1 w.r.t. S, $S^{i,\boldsymbol{\xi}}$, and $S^{j,\boldsymbol{\zeta}}$, respectively. Fix $\eta_t=\eta$, then it holds that*

$$\mathbb{E}_A\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right]\leq 8C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\frac{L\mu\eta}{L+\mu}\right)^{t-k-1}\left(\mathbb{E}_{\mathcal{A}}\left[r_k^2\right]\right)^{\frac{1}{2}}$$
$$+4L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_fL_g\sqrt{\frac{\eta(L+\mu)}{nL\mu}}+\frac{4L_fL_g(L+\mu)}{nL\mu},$$
$$\mathbb{E}_A\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right]\leq 8C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\frac{L\mu\eta}{L+\mu}\right)^{t-k-1}\left(\mathbb{E}_{\mathcal{A}}[r_k^2]\right)^{\frac{1}{2}}$$
$$+4L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_fL_g\sqrt{\frac{\eta(L+\mu)}{mL\mu}}+\frac{12L_fL_g(L+\mu)}{mL\mu},$$

where $L=L_fC_g+L_g^2C_f$.

Proof. The smoothness constant can be verified with Lemma A.1.

Let $\{\tilde{\mathbf{x}}_t\}$, $\{\tilde{\mathbf{y}}_t\}$ be the sequence produced by the non-private NSCGD or NSCSC method based on the dataset S, $\{\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}}\}$, $\{\tilde{\mathbf{y}}_t^{i,\boldsymbol{\xi}}\}$ be the sequence based on $S^{i,\boldsymbol{\xi}}$, $\{\tilde{\mathbf{x}}_t^{j,\boldsymbol{\zeta}}\}$, $\{\tilde{\mathbf{y}}_t^{j,\boldsymbol{\zeta}}\}$ be the sequence based on $S^{j,\boldsymbol{\zeta}}$.

We begin with the estimation of $\mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}_{t+1}^{i,\boldsymbol{\xi}}\right\|\right]$. We denote the set $S'_{\boldsymbol{\xi}}=\{\boldsymbol{\xi}'_k:k=1,\ldots,n\}$ where $\boldsymbol{\xi}'_k=\boldsymbol{\xi}_k$ if $k\neq i$. Recall that $S=S_{\boldsymbol{\xi}}\cup S_{\boldsymbol{\zeta}}$ and $S^{i,\boldsymbol{\xi}}=S'_{\boldsymbol{\xi}}\cup S_{\boldsymbol{\zeta}}$ only differ one element such that ξ_i in $S_{\boldsymbol{\xi}}$ is changed to ξ'_i.

$$\begin{aligned}&\left\|\mathbf{x}_{t+1}-\mathbf{x}_{t+1}^{i,\boldsymbol{\xi}}\right\|\\ \leq&\left\|\mathbf{x}_t-\eta_t\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})-\eta_t\nu_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}+\eta_t\nabla G(\mathbf{x}_t^{i,\boldsymbol{\xi}};\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1}^{i,\boldsymbol{\xi}};\boldsymbol{\xi}'_{i_t})+\eta_t\nu_t\right\|\\ =&\left\|\mathbf{x}_t-\eta_t\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})-\mathbf{x}_t^{i,\boldsymbol{\xi}}+\eta_t\nabla G(\mathbf{x}_t^{i,\boldsymbol{\xi}};\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1}^{i,\boldsymbol{\xi}};\boldsymbol{\xi}'_{i_t})\right\|.\end{aligned}$$

Note that the difference of $\left\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_{t+1}^{i,\boldsymbol{\xi}}\right\|$ has a similar recurrence.

$$\begin{aligned}&\left\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_{t+1}^{i,\boldsymbol{\xi}}\right\|\\ \leq&\left\|\tilde{\mathbf{x}}_t-\eta_t\nabla G(\tilde{\mathbf{x}}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\tilde{\mathbf{y}}_{t+1};\boldsymbol{\xi}_{i_t})-\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}}+\eta_t\nabla G(\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}};\boldsymbol{\zeta}_{j_t})\nabla F(\tilde{\mathbf{y}}_{t+1}^{i,\boldsymbol{\xi}};\boldsymbol{\xi}'_{i_t})\right\|.\end{aligned}$$

Consequently, we can follow a similar stability analysis of $\left\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_{t+1}^{i,\boldsymbol{\xi}}\right\|$ in Lemma B.3, which implies that

$$\mathbb{E}_A\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right]\leq 8C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\eta\frac{L\mu}{L+\mu}\right)^{t-k-1}\left(\mathbb{E}_A[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2]\right)^{\frac{1}{2}}$$
$$+4L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}+\frac{4L_fL_g(L+\mu)}{nL\mu}.$$

Similar to the stability analysis of $\left\|\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}}\right\|$, we obtain the following bound for $\left\|\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_t^{j,\boldsymbol{\zeta}}\right\|$.

$$\mathbb{E}_A\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \leq 8C_f L_g \eta \sup_S \sum_{k=0}^{t-1}\left(1-\eta\frac{L\mu}{L+\mu}\right)^{t-k-1}\left(\mathbb{E}_A[\|\mathbf{y}_{k+1} - g_S(\mathbf{x}_k)\|^2]\right)^{\frac{1}{2}}$$
$$+ 4L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}} + 2L_f L_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}} + \frac{12L_f L_g(L+\mu)}{mL\mu}.$$

□

Now we consider the optimization error of the NSCGD and NSCSC methods. Firstly, we present the following recurrence of $\mathbb{E}_A[\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \mid \mathcal{F}_t]$.

Lemma B.5. *Let $\Phi_S(\cdot)$ be μ-strongly convex. By running Algorithm 1, for any $\mathbf{x} \in \mathcal{X}$, it holds that*

$$\begin{aligned}
&\mathbb{E}_A[\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \mid \mathcal{F}_t] \\
\leq &\left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}_A[\|\mathbf{x}_t - \mathbf{x}\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2 \eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x})) \\
&+ \frac{2C_f^2 L_g^2 \eta_t}{\mu}\mathbb{E}_{\mathcal{A}}\left[\|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2 \mid \mathcal{F}_t\right].
\end{aligned}$$

where $\mathbb{E}_{\mathcal{A}}$ denotes the expectation with respect to the randomness of the algorithm, and $\mathcal{F}_t$ is the σ-field generated by $\{\boldsymbol{\zeta}_{j_0}, \ldots, \boldsymbol{\zeta}_{j_{t-1}}, \boldsymbol{\xi}_{i_0}, \ldots, \boldsymbol{\xi}_{i_{t-1}}\}$.

Proof. From the update of $\mathbf{x}_{t+1}$ in Algorithm 1, we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \\
\leq &\|\mathbf{x}_t - \eta_t\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t}) - \eta_t\nu_t - \mathbf{x}\|^2 \\
= &\|\mathbf{x}_t - \mathbf{x}\|^2 + \eta_t^2\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t}) + \nu_t\|^2 - 2\eta_t\langle\mathbf{x}_t - \mathbf{x}, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t}) + \nu_t\rangle \\
= &\|\mathbf{x}_t - \mathbf{x}\|^2 + \eta_t^2\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})\|^2 + 2\eta_t^2\langle\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t}), \nu_t\rangle + \eta_t^2\|\nu_t\|^2 \\
&- 2\eta_t\langle\mathbf{x}_t - \mathbf{x}, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) + \nu_t\rangle + \mathbf{u}_t,
\end{aligned}$$

where

$$\mathbf{u}_t \coloneqq 2\eta_t\langle\mathbf{x}_t - \mathbf{x}, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) - \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})\rangle.$$

Let $\mathcal{F}_t$ be the σ-field generated by $\{\boldsymbol{\zeta}_{j_0}, \ldots, \boldsymbol{\zeta}_{j_{t-1}}, \boldsymbol{\xi}_{i_0}, \ldots, \boldsymbol{\xi}_{i_{t-1}}\}$. Taking the expectation with respect to the internal randomness of the algorithm and using Assumption 3.1, we have

$$\begin{aligned}
&\mathbb{E}_A[\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \mid \mathcal{F}_t] \\
\leq &\mathbb{E}_A[\|\mathbf{x}_t - \mathbf{x}\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\mathbb{E}_{\mathcal{A}}[\langle\mathbf{x}_t - \mathbf{x}, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t})\rangle \mid \mathcal{F}_t] \\
&+ \mathbb{E}_A[\mathbf{u}_t \mid \mathcal{F}_t] \\
= &\mathbb{E}_A[\|\mathbf{x}_t - \mathbf{x}\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\langle\mathbf{x}_t - \mathbf{x}, \nabla\Phi_S(\mathbf{x}_t)\rangle + \mathbb{E}_A[\mathbf{u}_t \mid \mathcal{F}_t] \\
\leq &\mathbb{E}_A[\|\mathbf{x}_t - \mathbf{x}\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\left(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}\|^2\right) + \mathbb{E}_A[\mathbf{u}_t \mid \mathcal{F}_t],
\end{aligned}$$

where the last inequality comes from the μ-strong convexity of Φ_S. For all $\gamma_t > 0$, we have

$$\begin{aligned}
&2\eta_t\langle\mathbf{x}_t - \mathbf{x}, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) - \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})\rangle \\
\leq &2\eta_t\|\mathbf{x}_t - \mathbf{x}\|\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\|\|\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) - \nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})\| \\
\leq &2C_f\eta_t\|\mathbf{x}_t - \mathbf{x}\|\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\|\|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\| \\
\leq &2C_f\eta_t\left(\frac{\|\mathbf{x}_t - \mathbf{x}\|^2\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\|^2}{2\gamma_t} + \frac{\gamma_t}{2}\|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2\right)
\end{aligned}$$

$$\leq \frac{C_f L_g^2 \eta_t}{\gamma_t} \|\mathbf{x}_t - \mathbf{x}\|^2 + \gamma_t C_f \eta_t \|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2,$$

The first inequality follows from Cauchy–Schwartz inequality. The second inequality is due to Assumption 3.2. The third inequality follows from Young's inequality. Combining the above two inequalities, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \mid \mathcal{F}_t] \\
\leq &\left(1 + \frac{C_f L_g^2 \eta_t}{\gamma_t} - \mu\eta_t\right) \mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2 \eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x})) \\
&+ \gamma_t C_f \eta_t \mathbb{E}_A[\|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2 \mid \mathcal{F}_t].
\end{aligned}$$

By setting $\gamma_t = \gamma := \frac{2C_f L_g^2}{\mu}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \mid \mathcal{F}_t] \\
\leq &\left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2 \eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x})) \\
&+ \frac{2C_f^2 L_g^2 \eta_t}{\mu} \mathbb{E}_{\mathcal{A}}[\|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2 \mid \mathcal{F}_t].
\end{aligned}$$

□

Then we can provide the convergence analysis of the NSCGD and NSCSC methods for the strongly convex case.

Theorem B.6. *Let $\Phi_S(\cdot)$ be μ-strongly convex and $\tilde{\mathbf{x}} = (\sum_{t=1}^T (1-\mu\eta_t/2)^{T-t}\mathbf{x}_t)/(\sum_{t=1}^T (1-\mu\eta_t/2)^{T-t})$ be the output of Algorithm 1. Suppose $D_Y := \|\mathbf{y}_1 - g_S(\mathbf{x}_0)\|$, running Algorithm 1 with $\eta_t = \eta$, $\beta_t = \beta$, and $(\eta T)^{-1} \leq \mu/2$,*

- *For the SCGD update, it holds that*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*)] \\
= &\mathcal{O}\Big(\mu^{1-c}\eta^{-c}T^{-c}D_X + L_f^2 L_g^2 \eta + d\sigma^2\eta \\
&+ \frac{C_f^2 L_g^2 D_Y}{\mu}\beta^{-c}T^{-c} + \frac{C_f^2 L_f^2 L_g^6 \eta^2}{\mu\beta^2} + \frac{C_f^2 L_g^4 d\sigma^2\eta^2}{\mu\beta^2} + \frac{C_f^2 L_g^2 V_g \beta}{\mu}\Big);
\end{aligned} \tag{9}$$

- *For the SCSC update, it holds that*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*)] \\
= &\mathcal{O}\Big(\mu^{1-c}\eta^{-c}T^{-c}D_X + L_f^2 L_g^2 \eta + d\sigma^2\eta \\
&+ \frac{C_f^2 L_g^2 D_Y}{\mu}\beta^{-c}T^{-c} + \frac{C_f^2 L_f^2 L_g^6 \eta^2}{\mu\beta} + \frac{C_f^2 L_g^4 d\sigma^2\eta^2}{\mu\beta} + \frac{C_f^2 L_g^2 V_g \beta}{\mu}\Big).
\end{aligned} \tag{10}$$

Proof. We first present the proof for the SCGD update. Using Lemma B.5 and setting $\mathbf{x} = \mathbf{x}_S^*$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_{t+1} - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] \\
\leq &\left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] + L_f^2 L_g^2 \eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \\
&+ \frac{2C_f^2 L_g^2 \eta_t}{\mu} \mathbb{E}_A[\|g_S(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2 \mid \mathcal{F}_t].
\end{aligned} \tag{11}$$

Plugging Lemma B.2 into the above inequality and fix $\eta_t = \eta$, we get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_{t+1} - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] \\
\leq &\left(1 - \frac{\mu\eta}{2}\right) \mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] + L_f^2 L_g^2 \eta^2 + d\eta^2\sigma^2 - 2\eta(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*))
\end{aligned}$$

$$+\frac{2C_f^2L_g^2\eta}{\mu}\left(\left(\frac{c}{e}\right)^c(t\beta)^{-c}D_Y+\frac{L_f^2L_g^4\eta^2}{\beta^2}+\frac{dL_g^2\sigma^2\eta^2}{\beta^2}+2V_g\beta\right).$$

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

$$\begin{aligned}
&2\eta\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)]\\
\le&\left(1-\frac{\mu\eta}{2}\right)^{T}\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_1-\mathbf{x}_S^*\|^2\mid\mathcal{F}_t\right]+L_f^2L_g^2\eta^2\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}+d\eta^2\sigma^2\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\\
&+\frac{2C_f^2L_g^2D_Y\eta}{\mu}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}t^{-c}+\frac{2C_f^2L_f^2L_g^6\eta^3}{\mu\beta^2}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\\
&+\frac{2C_f^2L_g^4d\sigma^2\eta^3}{\mu\beta^2}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}+\frac{4C_f^2L_g^2V_g\eta\beta}{\mu}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_1-\mathbf{x}_S^*\|^2\right]\le&\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_0-\eta\nabla G(\mathbf{x}_0;\boldsymbol{\zeta}_{j_0})\nabla F(\mathbf{y}_1;\boldsymbol{\xi}_{i_0})-\eta\nu_0-\mathbf{x}_S^*\|^2\right]\\
\le&2\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_0-\mathbf{x}_S^*\|^2\right]+2L_f^2L_g^2\eta^2+2d\eta^2\sigma^2.
\end{aligned}$$

The second inequality follows from $\|\mathbf{a}+\mathbf{b}\|^2\le 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$ for $\forall\mathbf{a},\mathbf{b}\in\mathbb{R}^d$. Combining the above two inequalities, we obtain

$$\begin{aligned}
&2\eta\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)]\\
\le&2\left(1-\frac{\mu\eta}{2}\right)^{T}\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_0-\mathbf{x}_S^*\|^2\mid\mathcal{F}_t\right]+2L_f^2L_g^2\eta^2\sum_{t=0}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}+2d\eta^2\sigma^2\sum_{t=0}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\\
&+\frac{2C_f^2L_g^2D_Y\eta}{\mu}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}t^{-c}+\frac{2C_f^2L_f^2L_g^6\eta^3}{\mu\beta^2}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\\
&+\frac{2C_f^2L_g^4d\sigma^2\eta^3}{\mu\beta^2}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}+\frac{4C_f^2L_g^2V_g\eta\beta}{\mu}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}.
\end{aligned}$$

From Lemma A.4, we know

$$\left(1-\frac{\mu\eta}{2}\right)^{T}\le\exp(-\frac{\mu\eta T}{2})\le\left(\frac{2c}{e\mu}\right)^c(\eta T)^{-c}.$$

Divide both sides by 2η, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)\right]\\
\le&\left(\frac{2c}{e\mu}\right)^c\eta^{-c-1}T^{-c}D_X+\frac{2L_f^2L_g^2}{\mu}+\frac{2d\sigma^2}{\mu}+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^c\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}t^{-c}\\
&+\frac{2C_f^2L_f^2L_g^6\eta}{\mu^2\beta^2}+\frac{2C_f^2L_g^4d\sigma^2\eta}{\mu^2\beta^2}+\frac{4C_f^2L_g^2V_g\beta}{\mu^2\eta}.
\end{aligned}\tag{12}$$

Since $(\eta T)^{-1}\le\mu/2$, we have

$$\left(1-\frac{\mu\eta}{2}\right)^{T}\le\exp\left(-\frac{\mu\eta T}{2}\right)\le\frac{1}{2}.$$

Consequently, one can show that

$$\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}=\frac{1-\left(1-\frac{\mu\eta}{2}\right)^{T}}{\frac{\mu\eta}{2}}\geq\frac{1}{\mu\eta}. \tag{13}$$

From the choice of $\tilde{\mathbf{x}}$ and convexity of Φ_S, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}})-\Phi_S(\mathbf{x}_S^*)\right]\\
\leq&\left(\frac{2c}{e\mu}\right)^{c-1}\eta^{-c}T^{-c}D_X+2L_f^2L_g^2\eta+2d\sigma^2\eta\\
&+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^{c}\beta^{-c}\frac{\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}t^{-c}}{\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}}\\
&+\frac{2C_f^2L_f^2L_g^6\eta^2}{\mu\beta^2}+\frac{2C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta^2}+\frac{4C_f^2L_g^2V_g\beta}{\mu}
\end{aligned}$$

Note that $\left(1-\frac{\mu\eta}{2}\right)^{T-t}$ is non-decreasing with respect to t and for $c>0$, t^{-c} is non-increasing with respect to t. By Lemma A.5, we have

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

Consequently, we can show that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}})-\Phi_S(\mathbf{x}_S^*)\right]\\
\leq&\left(\frac{2c}{e\mu}\right)^{c-1}\eta^{-c}T^{-c}D_X+2L_f^2L_g^2\eta+2d\sigma^2\eta\\
&+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^{c}\beta^{-c}T^{-1}\sum_{t=1}^{T}t^{-c}\\
&+\frac{2C_f^2L_f^2L_g^6\eta^2}{\mu\beta^2}+\frac{2C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta^2}+\frac{4C_f^2L_g^2V_g\beta}{\mu}\\
=&\mathcal{O}\left(\mu^{1-c}\eta^{-c}T^{-c}D_X+L_f^2L_g^2\eta+d\sigma^2\eta\right.\\
&\left.+\frac{C_f^2L_g^2D_Y}{\mu}\beta^{-c}T^{-c}+\frac{C_f^2L_f^2L_g^6\eta^2}{\mu\beta^2}+\frac{C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta^2}+\frac{C_f^2L_g^2V_g\beta}{\mu}\right).
\end{aligned}$$

The last inequality holds as long as $c\neq 1$. Next, we present the proof for the SCSC update. Plugging Lemma B.2 into Eq. (11) and setting $\eta_t=\eta$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}_S^*\right\|^2\mid\mathcal{F}_t\right]\\
\leq&\left(1-\frac{\mu\eta}{2}\right)\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_S^*\right\|^2\mid\mathcal{F}_t\right]+L_f^2L_g^2\eta^2+d\eta^2\sigma^2-2\eta(\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*))\\
&+\frac{2C_f^2L_g^2\eta}{\mu}\left(\left(\frac{c}{e}\right)^{c}(t\beta)^{-c}D_Y+\frac{L_f^2L_g^4\eta^2}{\beta}+\frac{dL_g^2\sigma^2\eta^2}{\beta}+2V_g\beta\right).
\end{aligned}$$

Similar to the derivation of the SCGD update, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}})-\Phi_S(\mathbf{x}_S^*)\right]\\
\leq&\left(\frac{2c}{e\mu}\right)^{c-1}\eta^{-c}T^{-c}D_X+2L_f^2L_g^2\eta+2d\sigma^2\eta\\
&+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^{c}\beta^{-c}T^{-1}\sum_{t=1}^{T}t^{-c}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{2C_f^2L_f^2L_g^6\eta^2}{\mu\beta}+\frac{2C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta}+\frac{4C_f^2L_g^2V_g\beta}{\mu}\\
=&\mathcal{O}\left(\mu^{1-c}\eta^{-c}T^{-c}D_X+L_f^2L_g^2\eta+d\sigma^2\eta\right.\\
&\left.+\frac{C_f^2L_g^2D_Y}{\mu}\beta^{-c}T^{-c}+\frac{C_f^2L_f^2L_g^6\eta^2}{\mu\beta}+\frac{C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta}+\frac{C_f^2L_g^2V_g\beta}{\mu}\right).
\end{aligned}
$$

The last inequality holds as long as $c\neq 1$. □

B.1.1 Proof of Theorem 4.2 and 4.3

Proof. From Lemma A.1, we can deduce the smoothness constant of $\Phi_S(\cdot)$ is $L\coloneqq L_fC_g+L_g^2C_f$.

We first present the proof of the SCGD update. From the proof of Lemma B.4, we obtain the following result

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right]+4\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right]\\
\leq&40C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\eta\frac{L\mu}{L+\mu}\right)^{t-k-1}(\mathbb{E}_A[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2])^{\frac{1}{2}}+20L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}\\
&+2L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}+\frac{4L_fL_g(L+\mu)}{nL\mu}+8L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_fL_g(L+\mu)}{mL\mu}.
\end{aligned}
$$

Plugging the SCGD result of Lemma B.2, one has

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right]+4\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right]\\
\leq&40C_fL_g\eta\sup_S\sum_{k=0}^{t-1}\left(1-\eta\frac{L\mu}{L+\mu}\right)^{t-k-1}\left(\left(\frac{c}{e}\right)^c(k\beta)^{-c}D_Y^2+\frac{L_f^2L_g^4\eta^2}{\beta^2}+\frac{dL_g^2\sigma^2\eta^2}{\beta^2}+2V_g\beta\right)^{\frac{1}{2}}\\
&+20L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}+\frac{4L_fL_g(L+\mu)}{nL\mu}+8L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_fL_g(L+\mu)}{mL\mu}\\
\leq&40C_fL_g\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}(t\beta)^{-\frac{c}{2}}D_Y+40C_fL_fL_g^3\frac{L+\mu}{L\mu}\frac{\eta}{\beta}+40C_fL_g^2\frac{L+\mu}{L\mu}\frac{\sqrt{d}\sigma\eta}{\beta}\\
&+40\sqrt{2V_g}C_fL_g\frac{L+\mu}{L\mu}\sqrt{\beta}+20L_f\sqrt{C_g\eta\frac{L+\mu}{\mu}}\\
&+2L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}+\frac{4L_fL_g(L+\mu)}{nL\mu}+8L_fL_g\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_fL_g(L+\mu)}{mL\mu},
\end{aligned}
$$

where we use $\sqrt{a_1+\cdots+a_n}\leq\sqrt{a_1}+\cdots+\sqrt{a_n}$ for non-negative $a_1,\ldots,a_n$ and Lemma A.5 for the last inequality. Applying Theorem A.3, one has

$$
\begin{aligned}
&\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t)-\Phi_S(\mathbf{x}_t)]\\
\leq&40C_fL_fL_g^2\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}(t\beta)^{-\frac{c}{2}}D_Y+40C_fL_f^2L_g^4\frac{L+\mu}{L\mu}\frac{\eta}{\beta}+40C_fL_fL_g^3\frac{L+\mu}{L\mu}\frac{\sqrt{d}\sigma\eta}{\beta}\\
&+40\sqrt{2V_g}C_fL_fL_g^2\frac{L+\mu}{L\mu}\sqrt{\beta}+20L_f^2L_g\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}\\
&+\frac{4L_f^2L_g^2(L+\mu)}{nL\mu}+8L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_f^2L_g^2(L+\mu)}{mL\mu}+\frac{L_fV_g}{\sqrt{m}}.
\end{aligned}
$$

Multiplying both sides of the above inequality with $(1-\frac{\mu\eta}{2})^{T-t}$, telescoping from $t=1,\ldots,T$, then divide both

sides by $\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}$, we have

$$\begin{aligned}
&\frac{\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}[\Phi(\mathbf{x}_t)-\Phi_S(\mathbf{x}_t)]}{\left(\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}\right)}\\
\leq& 40C_fL_fL_g^2\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}\beta^{-\frac{c}{2}}D_Y\frac{\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}t^{-\frac{c}{2}}}{\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}}+40C_fL_f^2L_g^4\frac{L+\mu}{L\mu}\frac{\eta}{\beta}+40C_fL_fL_g^3\frac{L+\mu}{L\mu}\frac{\sqrt{d}\sigma\eta}{\beta}\\
&+40\sqrt{2V_g}C_fL_fL_g^2\frac{L+\mu}{L\mu}\sqrt{\beta}+20L_f^2L_g\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}\\
&+\frac{4L_f^2L_g^2(L+\mu)}{nL\mu}+8L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_f^2L_g^2(L+\mu)}{mL\mu}+\frac{L_fV_g}{\sqrt{m}}.
\end{aligned}$$

Lemma A.5 implies that

$$40C_fL_fL_g^2\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}\beta^{-\frac{c}{2}}D_Y\frac{\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}t^{-\frac{c}{2}}}{\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}}\leq 40C_fL_fL_g^2\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}\beta^{-\frac{c}{2}}D_Y\frac{\sum_{t=1}^{T}t^{-\frac{c}{2}}}{T}.$$

Consequently, we can show that

$$\begin{aligned}
&\frac{\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}[\Phi(\mathbf{x}_t)-\Phi_S(\mathbf{x}_t)]}{\left(\sum_{t=1}^{T}(1-\frac{\mu\eta}{2})^{T-t}\right)}\\
\leq& 40C_fL_fL_g^2\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}\beta^{-\frac{c}{2}}D_Y\frac{\sum_{t=1}^{T}t^{-\frac{c}{2}}}{T}+40C_fL_f^2L_g^4\frac{L+\mu}{L\mu}\frac{\eta}{\beta}+40C_fL_fL_g^3\frac{L+\mu}{L\mu}\frac{\sqrt{d}\sigma\eta}{\beta}\\
&+40\sqrt{2V_g}C_fL_fL_g^2\frac{L+\mu}{L\mu}\sqrt{\beta}+20L_f^2L_g\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}\\
&+\frac{4L_f^2L_g^2(L+\mu)}{nL\mu}+8L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_f^2L_g^2(L+\mu)}{mL\mu}+\frac{L_fV_g}{\sqrt{m}}.
\end{aligned}$$

Recall that the population loss can be divided in the form of Eq. (8).

$$\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t)-\Phi(\mathbf{x}^*)]\leq\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t)-\Phi_S(\mathbf{x}_t)]+\mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)]+\mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}^*)-\Phi(\mathbf{x}^*)],$$

For the second term on the right side, it is the optimization error. We can bound the second term using (12) such that

$$\begin{aligned}
&\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)]\\
\leq&\left(\frac{2c}{e\mu}\right)^{c}\eta^{-c-1}T^{-c}D_X+\frac{2L_f^2L_g^2}{\mu}+\frac{2d\sigma^2}{\mu}+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^{c}\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}t^{-c}\\
&+\frac{2C_f^2L_f^2L_g^6\eta}{\mu^2\beta^2}+\frac{2C_f^2L_g^4d\sigma^2\eta}{\mu^2\beta^2}+\frac{4C_f^2L_g^2V_g\beta}{\mu^2\eta}.
\end{aligned}$$

Divide both sides by $\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}$, and applying the lower bound (13), we have

$$\begin{aligned}
&\frac{\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)]}{\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}}\\
\leq&\left(\frac{2c}{e}\right)^{c}\mu^{1-c}\eta^{-c}T^{-c}D_X+2L_f^2L_g^2\eta+2d\sigma^2\eta+C_f^2L_g^2D_Y\eta\left(\frac{c}{e}\right)^{c}\beta^{-c}\sum_{t=1}^{T}\left(1-\frac{\mu\eta}{2}\right)^{T-t}t^{-c}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{2C_f^2L_f^2L_g^6\eta^2}{\mu\beta^2}+\frac{2C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta^2}+\frac{4C_f^2L_g^2V_g\beta}{\mu}\\
\le&\left(\frac{2c}{e}\right)^c\mu^{1-c}\eta^{-c}T^{-c}D_X+2L_f^2L_g^2\eta+2d\sigma^2\eta+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^c\beta^{-c}T^{-c}\\
&+\frac{2C_f^2L_f^2L_g^6\eta^2}{\mu\beta^2}+\frac{2C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta^2}+\frac{4C_f^2L_g^2V_g\beta}{\mu}.
\end{aligned}
$$

For the third term of (8), we can bound it as follows,

$$
\mathbb{E}[\Phi_S(\mathbf{x}^*)-\Phi(\mathbf{x}^*)]\le L_f\mathbb{E}[\|G_S(\mathbf{x}^*)-g(\mathbf{x}^*)\|]\le\frac{L_fV_g}{\sqrt{m}}.
$$

where the last two inequalities are due to Assumption 3.1 and 3.2. Combining the above three inequalities, we can obtain the following bound of excess population error using the convexity of $\Phi(\cdot)$,

$$
\begin{aligned}
&\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}})-\Phi(\mathbf{x}^*)]\\
=&\mathcal{O}(\mu^{-1}\beta^{-\frac{c}{2}}T^{-\frac{c}{2}}+\mu^{-1}\eta\beta^{-1}+\mu^{-1}\sqrt{d}\sigma\eta\beta^{-1}+\mu^{-1}\sqrt{\beta}+\mu^{-\frac{1}{2}}\sqrt{\eta}+\mu^{-\frac{1}{2}}\sqrt{\eta/n}\\
&+\mu^{-1}n^{-1}+\mu^{-\frac{1}{2}}\sqrt{\eta/m}+\mu^{-1}m^{-1}+m^{-\frac{1}{2}}+\mu^{1-c}\eta^{-c}T^{-c}+\eta+d\sigma^2\eta\\
&+\mu^{-1}\beta^{-c}T^{-c}+\mu^{-1}\eta^2\beta^{-2}+\mu^{-1}d\sigma^2\eta^2\beta^{-2}+\mu^{-1}\beta)\\
=&\mathcal{O}(\mu^{-1}\beta^{-\frac{c}{2}}T^{-\frac{c}{2}}+\mu^{-1}\sqrt{d}\sigma\eta\beta^{-1}+\mu^{-1}\sqrt{\beta}+\mu^{-\frac{1}{2}}\sqrt{\eta}\\
&+\mu^{-1}n^{-1}+\mu^{-1}m^{-1}+m^{-\frac{1}{2}}+\mu^{1-c}\eta^{-c}T^{-c}+d\sigma^2\eta),
\end{aligned}
$$

where the last equality follows from omitting non-dominant terms. Recall that we chose

$$
\sigma=\frac{c_1L_fL_g\sqrt{T\log(1/\delta)}}{n\epsilon}+\frac{c_2L_fL_g\sqrt{T\log(1/\delta)}}{m\epsilon}
$$

in Theorem 4.1.

By setting $c>3$, $T=\mathcal{O}(\min\{n^{\frac{14}{3}},m^{\frac{7}{3}}\})$, $\eta=1/\max(T^{\frac{13}{14}},T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1}+n^{-1}\epsilon^{-1}))$ and $\beta=T^{-\frac{3}{7}}$, we obtain the following utility bound

$$
\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}})-\Phi(x^*)]=\mathcal{O}\left(\frac{1}{\mu\sqrt{m}}+\frac{1}{\mu n}+\frac{\sqrt{d\log(1/\delta)}}{\mu\epsilon}\left(\frac{1}{m}+\frac{1}{n}\right)\right).
$$

Next, we present the proof for the SCSC update. Following a similar derivation as the SCGD update, we can show that

$$
\begin{aligned}
&\frac{\sum_{t=1}^T\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}[\Phi(\mathbf{x}_t)-\Phi_S(\mathbf{x}_t)]}{\left(\sum_{t=1}^T(1-\frac{\mu\eta}{2})^{T-t}\right)}\\
\le&40C_fL_fL_g^2\frac{L+\mu}{L\mu}\left(\frac{c}{e}\right)^{\frac{c}{2}}\beta^{-\frac{c}{2}}D_Y\frac{\sum_{t=1}^Tt^{-\frac{c}{2}}}{T}+40C_fL_f^2L_g^4\frac{L+\mu}{L\mu}\frac{\eta}{\sqrt{\beta}}+40C_fL_fL_g^3\frac{L+\mu}{L\mu}\frac{\sqrt{d}\sigma\eta}{\sqrt{\beta}}\\
&+40\sqrt{2V_g}C_fL_fL_g^2\frac{L+\mu}{L\mu}\sqrt{\beta}+20L_f^2L_g\sqrt{C_g\eta\frac{L+\mu}{\mu}}+2L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{n}}\\
&+\frac{4L_f^2L_g^2(L+\mu)}{nL\mu}+8L_f^2L_g^2\sqrt{\frac{L+\mu}{L\mu}}\sqrt{\frac{\eta}{m}}+\frac{48L_f^2L_g^2(L+\mu)}{mL\mu}+\frac{L_fV_g}{\sqrt{m}}.
\end{aligned}
$$

Combining this result with the optimization error,

$$
\begin{aligned}
&\frac{\sum_{t=1}^T\left(1-\frac{\mu\eta}{2}\right)^{T-t}\mathbb{E}_{\mathcal{A}}[\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)]}{\sum_{t=1}^T\left(1-\frac{\mu\eta}{2}\right)^{T-t}}\\
\le&\left(\frac{2c}{e}\right)^c\mu^{1-c}\eta^{-c}T^{-c}D_X+2L_f^2L_g^2\eta+2d\sigma^2\eta+\frac{C_f^2L_g^2D_Y}{\mu}\left(\frac{c}{e}\right)^c\beta^{-c}T^{-c}
\end{aligned}
$$

$$+\frac{2C_f^2L_f^2L_g^6\eta^2}{\mu\beta}+\frac{2C_f^2L_g^4d\sigma^2\eta^2}{\mu\beta}+\frac{4C_f^2L_g^2V_g\beta}{\mu}.$$

Consequently, we can obtain the following generalization bound

$$\begin{aligned}
&\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}})-\Phi(\mathbf{x}^*)]\\
=&\mathcal{O}(\mu^{-1}\beta^{-\frac{c}{2}}T^{-\frac{c}{2}}+\mu^{-1}\eta\beta^{-\frac{1}{2}}+\mu^{-1}\sqrt{d}\sigma\eta\beta^{-\frac{1}{2}}+\mu^{-1}\sqrt{\beta}+\mu^{-\frac{1}{2}}\sqrt{\eta}+\mu^{-\frac{1}{2}}\sqrt{\eta/n}\\
&+\mu^{-1}n^{-1}+\mu^{-\frac{1}{2}}\sqrt{\eta/m}+\mu^{-1}m^{-1}+m^{-\frac{1}{2}}+\mu^{1-c}\eta^{-c}T^{-c}+\eta+d\sigma^2\eta\\
&+\mu^{-1}\beta^{-c}T^{-c}+\mu^{-1}\eta^2\beta^{-1}+\mu^{-1}d\sigma^2\eta^2\beta^{-1}+\mu^{-1}\beta)\\
=&\mathcal{O}(\mu^{-1}\beta^{-\frac{c}{2}}T^{-\frac{c}{2}}+\mu^{-1}\sqrt{d}\sigma\eta\beta^{-\frac{1}{2}}+\mu^{-1}\sqrt{\beta}+\mu^{-\frac{1}{2}}\sqrt{\eta}\\
&+\mu^{-1}n^{-1}+\mu^{-1}m^{-1}+m^{-\frac{1}{2}}+\mu^{1-c}\eta^{-c}T^{-c}+d\sigma^2\eta),
\end{aligned}$$

where the last equality follows from omitting non-dominant terms. Recall that we chose

$$\sigma=\frac{c_1L_fL_g\sqrt{T\log(1/\delta)}}{n\epsilon}+\frac{c_2L_fL_g\sqrt{T\log(1/\delta)}}{m\epsilon}$$

in Theorem 4.1.

Setting $c>6$, $T=\mathcal{O}(\min\{n^{\frac{7}{3}},m^{\frac{7}{6}}\})$, $\eta=1/\max(T^{\frac{13}{14}},T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1}+n^{-1}\epsilon^{-1}))$ and $\beta=T^{-\frac{6}{7}}$, we obtain the following utility bound

$$\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}})-\Phi(\mathbf{x}^*)]=\mathcal{O}\left(\frac{1}{\mu\sqrt{m}}+\frac{1}{\mu n}+\frac{\sqrt{d\log(1/\delta)}}{\mu\epsilon}\left(\frac{1}{m}+\frac{1}{n}\right)\right).$$

□

B.2 Convex Case

In this subsection, we present the utility analysis of the NSCGD and NSCSC methods for the convex case. Firstly, we review the uniform stability results of the non-private SCGD and SCSC methods for the convex case as follows.

Lemma B.7 (Yang et al. (2024)). *Let $F(g_S(\cdot);\boldsymbol{\xi})$ be a convex function for any $\boldsymbol{\xi}$. Let $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$ be the sequence of iterates produced by non-private SCGD or SCSC method on dataset S, $\{\mathbf{x}_t^{i,\boldsymbol{\xi}}\}$ and $\{\mathbf{y}_t^{i,\boldsymbol{\xi}}\}$ be the sequence of iterated based on the dataset $S^{i,\boldsymbol{\xi}}$, $\{\mathbf{x}_t^{j,\boldsymbol{\zeta}}\}$ and $\{\mathbf{y}_t^{j,\boldsymbol{\zeta}}\}$ be the sequence of iterates based on the dataset $S^{j,\boldsymbol{\zeta}}$. Running the non-private SCGD or SCSC method with $\eta_t=\eta$ and $\beta_t=\beta\in(0,1)$ for any $t\in[T-1]$, then it holds that*

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right]\le 8C_fL_g\sup_S\eta\sum_{k=0}^{t-1}\left(\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2]\right)^{\frac{1}{2}}+4L_f\sqrt{C_g}\eta\sqrt{t}+\frac{6L_fL_g}{n}\eta t,$$

and

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right]\le 8C_fL_g\sup_S\eta\sum_{k=0}^{t-1}\left(\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2]\right)^{\frac{1}{2}}+4L_f\sqrt{C_g}\eta\sqrt{t}+\frac{14L_fL_g}{m}\eta t.$$

Based on the above result, we can derive the uniform stability of the NSCGD and NSCSC methods for the convex case.

Lemma B.8. *Suppose that the function $F(g_S(\cdot);\boldsymbol{\xi})$ is convex. Let $r_k=\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|$, $\{\mathbf{x}_t\}$, $\{\mathbf{x}_t^{i,\boldsymbol{\xi}}\}$ and $\{\mathbf{x}_t^{j,\boldsymbol{\zeta}}\}$ be the sequences generated by Algorithm 1 w.r.t. S, $S^{i,\boldsymbol{\xi}}$, and $S^{j,\boldsymbol{\zeta}}$, respectively. Fix $\eta_t=\eta$, then it holds that*

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right]\le&8C_fL_g\sup_S\eta\sum_{k=0}^{t-1}\left(\mathbb{E}_{\mathcal{A}}\left[r_k^2\right]\right)^{\frac{1}{2}}+\\
&4L_f\sqrt{C_g}\eta\sqrt{t}+\frac{6L_fL_g}{n}\eta t,
\end{aligned}$$

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \leq 8C_fL_g\sup_S \eta\sum_{k=0}^{t-1}\left(\mathbb{E}_{\mathcal{A}}\left[r_k^2\right]\right)^{\frac{1}{2}}+ 4L_f\sqrt{C_g}\eta\sqrt{t}+\frac{14L_fL_g}{m}\eta t.$$

Proof. Let $\{\tilde{\mathbf{x}}_t\}$, $\{\tilde{\mathbf{y}}_t\}$ be the sequence produced by the non-private NSCGD or NSCSC method based on the dataset S, $\{\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}}\}$, $\{\tilde{\mathbf{y}}_t^{i,\boldsymbol{\xi}}\}$ be the sequence based on $S^{i,\boldsymbol{\xi}}$, $\{\tilde{\mathbf{x}}_t^{j,\boldsymbol{\zeta}}\}$, $\{\tilde{\mathbf{y}}_t^{j,\boldsymbol{\zeta}}\}$ be the sequence based on $S^{j,\boldsymbol{\zeta}}$.

We begin with the estimation of $\mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}_{t+1}^{i,\boldsymbol{\xi}}\right\|\right]$. We denote the set $S'_{\boldsymbol{\xi}}=\{\boldsymbol{\xi}'_k : k=1,\ldots,n\}$ where $\boldsymbol{\xi}'_k=\boldsymbol{\xi}_k$ if $k\neq i$. Recall that $S=S_{\boldsymbol{\xi}}\cup S_{\boldsymbol{\zeta}}$ and $S^{i,\boldsymbol{\xi}}=S'_{\boldsymbol{\xi}}\cup S_{\boldsymbol{\zeta}}$ only differ one element such that ξ_i in $S_{\boldsymbol{\xi}}$ is changed to ξ'_i.

$$\begin{aligned}
&\left\|\mathbf{x}_{t+1}-\mathbf{x}_{t+1}^{i,\boldsymbol{\xi}}\right\| \\
\leq&\left\|\mathbf{x}_t-\eta_t\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})-\eta_t\nu_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}+\eta_t\nabla G(\mathbf{x}_t^{i,\boldsymbol{\xi}};\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1}^{i,\boldsymbol{\xi}};\boldsymbol{\xi}'_{i_t})+\eta_t\nu_t\right\| \\
=&\left\|\mathbf{x}_t-\eta_t\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i_t})-\mathbf{x}_t^{i,\boldsymbol{\xi}}+\eta_t\nabla G(\mathbf{x}_t^{i,\boldsymbol{\xi}};\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_{t+1}^{i,\boldsymbol{\xi}};\boldsymbol{\xi}'_{i_t})\right\|.
\end{aligned}$$

Note that the difference of $\left\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_{t+1}^{i,\boldsymbol{\xi}}\right\|$ has a similar recurrence such that

$$\begin{aligned}
&\left\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_{t+1}^{i,\boldsymbol{\xi}}\right\| \\
\leq&\left\|\tilde{\mathbf{x}}_t-\eta_t\nabla G(\tilde{\mathbf{x}}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\tilde{\mathbf{y}}_{t+1};\boldsymbol{\xi}_{i_t})-\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}}+\eta_t\nabla G(\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}};\boldsymbol{\zeta}_{j_t})\nabla F(\tilde{\mathbf{y}}_{t+1}^{i,\boldsymbol{\xi}};\boldsymbol{\xi}'_{i_t})\right\|.
\end{aligned}$$

Consequently, we can follow a similar stability analysis of $\left\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_{t+1}^{i,\boldsymbol{\xi}}\right\|$ in Lemma B.7, which implies that

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right] \leq 8C_fL_g\sup_S \eta\sum_{k=0}^{t-1}\left(\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2\right]\right)^{\frac{1}{2}}+4L_f\sqrt{C_g}\eta\sqrt{t}+\frac{6L_fL_g}{n}\eta t.$$

Similar to the stability analysis of $\left\|\tilde{\mathbf{x}}_t-\tilde{\mathbf{x}}_t^{i,\boldsymbol{\xi}}\right\|$, we can derive the following bound for $\left\|\tilde{\mathbf{x}}_t-\tilde{\mathbf{x}}_t^{j,\boldsymbol{\zeta}}\right\|$,

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t-\mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \leq 8C_fL_g\sup_S \eta\sum_{k=0}^{t-1}\left(\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{y}_{k+1}-g_S(\mathbf{x}_k)\|^2\right]\right)^{\frac{1}{2}}+4L_f\sqrt{C_g}\eta\sqrt{t}+\frac{14L_fL_g}{m}\eta t.$$

□

Next, we consider the optimization error of the NSCGD and NSCSC methods for the convex case. Firstly, we present the following recurrence of $\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right]$.

Lemma B.9. *Let $\Phi_S(\cdot)$ be a convex function. By running Algorithm 1, it holds that*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] \\
\leq&\left(1+\frac{C_fL_g^2\eta_t}{\gamma_t}\right)\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_t-\mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right]+L_f^2L_g^2\eta_t^2+d\eta_t^2\sigma^2-2\eta_t(\Phi_S(\mathbf{x}_t)-\Phi_S(\mathbf{x}_S^*)) \\
&+\gamma_tC_f\eta_t\left\|g_S(\mathbf{x}_t)-\mathbf{y}_{t+1}\right\|^2.
\end{aligned}$$

where $\mathbb{E}_{\mathcal{A}}$ denotes the expectation with respect to the randomness of the algorithm, and $\mathcal{F}_t$ is the σ-field generated by $\{\boldsymbol{\zeta}_{j_0},\ldots,\boldsymbol{\zeta}_{j_{t-1}},\boldsymbol{\xi}_{i_0},\ldots,\boldsymbol{\xi}_{i_{t-1}}\}$.

Proof. Let $\mathbf{x}_S^*=\arg\min_{\mathbf{x}\in\mathcal{X}}\Phi_S(\mathbf{x})$. From Algorithm 1, we have

$$\|\mathbf{x}_{t+1}-\mathbf{x}_S^*\|^2$$

$$
\begin{aligned}
&\leq \|\mathbf{x}_t - \eta_t \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t}) - \eta_t\nu_t - \mathbf{x}_S^*\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}_S^*\|^2 + \eta_t^2 \|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t}) + \nu_t\|^2 - 2\eta_t\langle \mathbf{x}_t - \mathbf{x}_S^*, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t}) + \nu_t\rangle \\
&= \|\mathbf{x}_t - \mathbf{x}_S^*\|^2 + \eta_t^2 \|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t})\|^2 + 2\eta_t^2\langle \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t}), \nu_t\rangle + \eta_t^2\|\nu_t\|^2 \\
&\quad - 2\eta_t\langle \mathbf{x}_t - \mathbf{x}_S^*, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) + \nu_t\rangle + \mathbf{u}_t,
\end{aligned}
$$

where

$$
\mathbf{u}_t \coloneqq 2\eta_t\langle \mathbf{x}_t - \mathbf{x}_S^*, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) - \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t})\rangle.
$$

Let $\mathcal{F}_t$ be the σ-field generated by $\{\boldsymbol{\zeta}_{j_0},\ldots,\boldsymbol{\zeta}_{j_{t-1}},\boldsymbol{\xi}_{i_0},\ldots,\boldsymbol{\xi}_{i_{t-1}}\}$. Taking the expectation with respect to the internal randomness of the algorithm and using Assumption 3.1, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_{t+1} - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] \\
\leq&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] + L_f^2L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\mathbb{E}_{\mathcal{A}}[\langle \mathbf{x}_t - \mathbf{x}_S^*, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t})\rangle \mid \mathcal{F}_t] \\
&+ \mathbb{E}_{\mathcal{A}}[\mathbf{u}_t \mid \mathcal{F}_t] \\
=&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] + L_f^2L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\langle \mathbf{x}_t - \mathbf{x}_S^*, \nabla\Phi_S(\mathbf{x}_t)\rangle + \mathbb{E}_{\mathcal{A}}[\mathbf{u}_t \mid \mathcal{F}_t] \\
\leq&\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] + L_f^2L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\left(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)\right) + \mathbb{E}_{\mathcal{A}}[\mathbf{u}_t \mid \mathcal{F}_t],
\end{aligned}
$$

where the last inequality comes from the convexity of Φ_S. For all $\gamma_t > 0$, we have

$$
\begin{aligned}
&2\eta_t\langle \mathbf{x}_t - \mathbf{x}_S^*, \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) - \nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t})\rangle \\
\leq&2\eta_t\|\mathbf{x}_t - \mathbf{x}_S^*\|\,\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\|\,\|\nabla F(g_S(\mathbf{x}_t);\boldsymbol{\xi}_{i_t}) - \nabla F(\mathbf{y}_t;\boldsymbol{\xi}_{i_t})\| \\
\leq&2C_f\eta_t\|\mathbf{x}_t - \mathbf{x}_S^*\|\,\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\|\,\|g_S(\mathbf{x}_t) - \mathbf{y}_t\| \\
\leq&2C_f\eta_t\left(\frac{\|\mathbf{x}_t - \mathbf{x}_S^*\|^2\|\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j_t})\|^2}{2\gamma_t} + \frac{\gamma_t}{2}\|g_S(\mathbf{x}_t) - \mathbf{y}_t\|^2\right) \\
\leq&\frac{C_fL_g^2\eta_t}{\gamma_t}\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 + \gamma_tC_f\eta_t\|g_S(\mathbf{x}_t) - \mathbf{y}_t\|^2,
\end{aligned}
$$

The first inequality follows from Cauchy–Schwartz inequality. The second inequality is due to Assumption 3.2. The third inequality follows from Young's inequality. Combining the above two inequalities, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] \\
\leq&\left(1 + \frac{C_fL_g^2\eta_t}{\gamma_t}\right)\mathbb{E}_{\mathcal{A}}[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t] + L_f^2L_g^2\eta_t^2 + d\eta_t^2\sigma^2 - 2\eta_t\left(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)\right) \\
&+ \gamma_tC_f\eta_t\|g_S(\mathbf{x}_t) - \mathbf{y}_t\|^2.
\end{aligned}
$$

□

Then we can provide the convergence analysis of NSCGD and NSCSC methods for the convex case.

Lemma B.10. *Let $\Phi_S(\cdot)$ be a convex function and $\tilde{\mathbf{x}} = \sum_{t=1}^T \mathbf{x}_t/T$ be the solution of Algorithm 1 w.r.t. S. Suppose $D_Y \coloneqq \|\mathbf{y}_1 - g_S(\mathbf{x}_0)\|$, running Algorithm 1 with $\eta_t = \eta$ and $\beta_t = \beta$,*

- *For SCGD update, it holds that*

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*)\right] \\
=&\mathcal{O}(D_X\eta^{-1}T^{-1} + C_fL_g^2D_X\gamma^{-1} + L_f^2L_g^2\eta + d\eta\sigma^2 \\
&+ C_f\gamma\beta^{-c}T^{-c}D_Y + C_fL_f^2L_g^4\gamma\eta^2\beta^{-2} + dC_fL_g^2\sigma^2\gamma\eta^2\beta^{-2} + C_fV_g\gamma\beta).
\end{aligned} \tag{14}
$$

- *For SCSC update, it holds that*

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*)\right] \\
=&\mathcal{O}(D_X\eta^{-1}T^{-1} + C_fL_g^2D_X\gamma^{-1} + L_f^2L_g^2\eta + d\eta\sigma^2 \\
&+ C_f\gamma\beta^{-c}T^{-c}D_Y + C_fL_f^2L_g^4\gamma\eta^2\beta^{-1} + dC_fL_g^2\sigma^2\gamma\eta^2\beta^{-1} + C_fV_g\gamma\beta).
\end{aligned} \tag{15}
$$

Proof. We first present the proof for the SCGD update. Setting $\eta_t = \eta$, $\beta_t = \beta$ and $\gamma_t = \gamma$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] \\
\leq &\left(1 + \frac{C_f L_g^2 \eta}{\gamma}\right) \mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] + L_f^2 L_g^2 \eta^2 + d\eta^2\sigma^2 - 2\eta(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \\
&+ \gamma C_f \eta \|g_S(\mathbf{x}_t) - \mathbf{y}_t\|^2 \\
\leq &\left(1 + \frac{C_f L_g^2 \eta}{\gamma}\right) \mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] + L_f^2 L_g^2 \eta^2 + d\eta^2\sigma^2 - 2\eta(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \\
&+ \gamma C_f \eta \left(\left(\frac{c}{e}\right)^c (t\beta)^{-c} D_Y + \frac{L_f^2 L_g^4 \eta^2}{\beta^2} + \frac{d L_g^2 \sigma^2 \eta^2}{\beta^2} + 2V_g\beta\right).
\end{aligned}$$

The last inequality is due to Lemma B.2. Telescope the above inequality from $t = 1$ to T, we have

$$\begin{aligned}
2\eta \sum_{t=1}^{T} (\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \leq &\left(1 + \frac{C_f L_g^2 \eta T}{\gamma}\right) D_X + L_f^2 L_g^2 \eta^2 T + d\eta^2\sigma^2 T \\
&+ \gamma C_f \eta \left(\left(\frac{c}{e}\right)^c \sum_{t=1}^{T} (t\beta)^{-c} D_Y + \frac{L_f^2 L_g^4 \eta^2 T}{\beta^2} + \frac{d L_g^2 \sigma^2 \eta^2 T}{\beta^2} + 2V_g\beta T\right).
\end{aligned}$$

By the convexity of Φ_S, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*)\right] = &\mathcal{O}(D_X \eta^{-1} T^{-1} + C_f L_g^2 D_X \gamma^{-1} + L_f^2 L_g^2 \eta + d\eta\sigma^2 \\
&+ C_f \gamma \beta^{-c} T^{-c} D_Y + C_f L_f^2 L_g^4 \gamma \eta^2 \beta^{-2} + d C_f L_g^2 \sigma^2 \gamma \eta^2 \beta^{-2} + C_f V_g \gamma \beta).
\end{aligned}$$

Next, we present the proof for the SCSC update. Setting $\eta_t = \eta$, $\beta_t = \beta$ and $\gamma_t = \gamma$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] \\
\leq &\left(1 + \frac{C_f L_g^2 \eta}{\gamma}\right) \mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] + L_f^2 L_g^2 \eta^2 + d\eta^2\sigma^2 - 2\eta(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \\
&+ \gamma C_f \eta \|g_S(\mathbf{x}_t) - \mathbf{y}_t\|^2 \\
\leq &\left(1 + \frac{C_f L_g^2 \eta}{\gamma}\right) \mathbb{E}_{\mathcal{A}}\left[\|\mathbf{x}_t - \mathbf{x}_S^*\|^2 \mid \mathcal{F}_t\right] + L_f^2 L_g^2 \eta^2 + d\eta^2\sigma^2 - 2\eta(\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \\
&+ \gamma C_f \eta \left(\left(\frac{c}{e}\right)^c (t\beta)^{-c} D_Y + \frac{L_f^2 L_g^4 \eta^2}{\beta} + \frac{d L_g^2 \sigma^2 \eta^2}{\beta} + 2V_g\beta\right).
\end{aligned}$$

The last inequality is due to Lemma B.2. Telescope the above inequality from $t = 1$ to T, we have

$$\begin{aligned}
2\eta \sum_{t=1}^{T} (\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)) \leq &\left(1 + \frac{C_f L_g^2 \eta T}{\gamma}\right) D_X + L_f^2 L_g^2 \eta^2 T + d\eta^2\sigma^2 T \\
&+ \gamma C_f \eta \left(\left(\frac{c}{e}\right)^c \sum_{t=1}^{T} (t\beta)^{-c} D_Y + \frac{L_f^2 L_g^4 \eta^2 T}{\beta} + \frac{d L_g^2 \sigma^2 \eta^2 T}{\beta} + 2V_g\beta T\right).
\end{aligned}$$

The convexity of Φ_S implies that

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}\left[\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*)\right] = &\mathcal{O}(D_X \eta^{-1} T^{-1} + C_f L_g^2 D_X \gamma^{-1} + L_f^2 L_g^2 \eta + d\eta\sigma^2 \\
&+ C_f \gamma \beta^{-c} T^{-c} D_Y + C_f L_f^2 L_g^4 \gamma \eta^2 \beta^{-1} + d C_f L_g^2 \sigma^2 \gamma \eta^2 \beta^{-1} + C_f V_g \gamma \beta).
\end{aligned}$$

□

B.2.1 Proof of Theorem 4.4 and 4.5

Proof. We first present the proof for the SCGD update. Setting $\eta_t = \eta$, $\beta_t = \beta$ for $\eta, \beta > 0$, applying Theorem B.8, for any t we have

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right] \leq 8C_f L_g \sup_S \eta \sum_{k=0}^{t-1} (\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{k+1} - g_S(\mathbf{x}_k)\|^2])^{\frac{1}{2}} + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t,$$

and

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \leq 8C_f L_g \sup_S \eta \sum_{k=0}^{t-1} (\mathbb{E}_{\mathcal{A}}[\|\mathbf{y}_{k+1} - g_S(\mathbf{x}_k)\|^2])^{\frac{1}{2}} + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{14L_f L_g}{m}\eta t.$$

Plugging the result of Lemma B.2 into the above two inequalities, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right] \leq & 8C_f L_g \eta \sum_{k=0}^{t-1} \sqrt{\left(\frac{c}{e}\right)^c (k\beta)^{-c} D_Y^2 + \frac{L_f^2 L_g^4 \eta^2}{\beta^2} + \frac{d L_g^2 \sigma^2 \eta^2}{\beta^2} + 2V_g\beta} \\ & + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + 8C_f L_g \eta D_Y,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \leq & 8C_f L_g \eta \sum_{k=0}^{t-1} \sqrt{\left(\frac{c}{e}\right)^c (k\beta)^{-c} D_Y^2 + \frac{L_f^2 L_g^4 \eta^2}{\beta^2} + \frac{d L_g^2 \sigma^2 \eta^2}{\beta^2} + 2V_g\beta} \\ & + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{14L_f L_g}{m}\eta t + 8C_f L_g \eta D_Y.\end{aligned}$$

Using $\sqrt{a_1 + \cdots + a_n} \leq \sqrt{a_1} + \cdots + \sqrt{a_n}$, we get

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right] \leq & 8C_f L_g \eta \left(\frac{c}{e\beta}\right)^{c/2} D_Y \sum_{k=0}^{t-1} k^{-\frac{c}{2}} + \frac{8C_f L_f L_g^3 \eta^2 t}{\beta} + \frac{8\sqrt{d} C_f L_g^2 \sigma \eta^2 t}{\beta} \\ & + 8\sqrt{2} C_f L_g \sqrt{V_g}\eta\sqrt{\beta} t + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + 8C_f L_g \eta D_Y,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \leq & 8C_f L_g \eta \left(\frac{c}{e\beta}\right)^{c/2} D_Y \sum_{k=0}^{t-1} k^{-\frac{c}{2}} + \frac{8C_f L_f L_g^3 \eta^2 t}{\beta} + \frac{8\sqrt{d} C_f L_g^2 \sigma \eta^2 t}{\beta} \\ & + 8\sqrt{2} C_f L_g \sqrt{V_g}\eta\sqrt{\beta} t + 4L_f\sqrt{C_g}\eta\sqrt{t} + \frac{14L_f L_g}{m}\eta t + 8C_f L_g \eta D_Y.\end{aligned}$$

Thus we get

$$\begin{aligned}& \mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{i,\boldsymbol{\xi}}\right\|\right] + 4\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{x}_t - \mathbf{x}_t^{j,\boldsymbol{\zeta}}\right\|\right] \\ \leq & 40C_f L_g \eta \left(\frac{c}{e\beta}\right)^{c/2} D_Y \sum_{k=0}^{t-1} k^{-\frac{c}{2}} + \frac{40C_f L_f L_g^3 \eta^2 t}{\beta} + \frac{40\sqrt{d} C_f L_g^2 \sigma \eta^2 t}{\beta} \\ & + 40\sqrt{2} C_f L_g \sqrt{V_g}\eta\sqrt{\beta} t + 20L_f\sqrt{C_g}\eta\sqrt{t} + \frac{6L_f L_g}{n}\eta t + \frac{56L_f L_g}{m}\eta t + 40C_f L_g \eta D_Y.\end{aligned}$$

Using Theorem A.3, we have

$$\begin{aligned}& \mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)] \\ \leq & 40C_f L_f L_g^2 \eta \left(\frac{c}{e\beta}\right)^{c/2} D_Y \sum_{k=0}^{t-1} k^{-\frac{c}{2}} + \frac{40C_f L_f^2 L_g^4 \eta^2 t}{\beta} + \frac{40\sqrt{d} C_f L_f L_g^3 \sigma \eta^2 t}{\beta}\end{aligned}$$

$$+ 40\sqrt{2}C_f L_f L_g^2 \sqrt{V_g}\eta\sqrt{\beta}t + 20L_f^2 L_g \sqrt{C_g}\eta\sqrt{t} + \frac{6L_f^2 L_g^2}{n}\eta t + \frac{56L_f^2 L_g^2}{m}\eta t$$
$$+ 40C_f L_f L_g^2 \eta D_Y + L_f \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}(\mathrm{G}(\mathrm{x_t};\boldsymbol{\zeta}))]}{m}},$$

Telescope the above inequality from $t = 1$ to T, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)] \\
\leq& 40C_f L_f L_g^2 \eta \left(\frac{c}{e\beta}\right)^{c/2} D_Y T^{2-\frac{c}{2}} + \frac{20C_f L_f^2 L_g^4 \eta^2 T^2}{\beta} + \frac{20\sqrt{d}C_f L_f L_g^3 \sigma\eta^2 T^2}{\beta} \\
&+ 20\sqrt{2}C_f L_f L_g^2 \sqrt{V_g}\eta\sqrt{\beta}T^2 + 20L_f^2 L_g \sqrt{C_g}\eta T^{\frac{3}{2}} + \frac{3L_f^2 L_g^2}{n}\eta T^2 + \frac{28L_f^2 L_g^2}{m}\eta T^2 \\
&+ 40C_f L_f L_g^2 \eta D_Y T + L_f T \frac{V_g}{\sqrt{m}}.
\end{aligned}$$

The above inequality holds as long as we choose $c \neq 2$ or 4. Recall that the population loss can be divided in the form of Eq. (8).

$$\mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}^*)] \leq \mathbb{E}_{S,\mathcal{A}}[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)] + \mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}_S^*)] + \mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}^*) - \Phi(\mathbf{x}^*)],$$

For the second term on the right side, it is the optimization error. We can bound the second term using (14) such that

$$\begin{aligned}
\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*) =& \mathcal{O}(D_X \eta^{-1} T^{-1} + C_f L_g^2 D_X \gamma^{-1} + L_f^2 L_g^2 \eta + d\eta\sigma^2 \\
&+ C_f \gamma\beta^{-c} T^{-c} D_Y + C_f L_f^2 L_g^4 \gamma\eta^2\beta^{-2} + dC_f L_g^2 \sigma^2 \gamma\eta^2\beta^{-2} + C_f V_g \gamma\beta).
\end{aligned}$$

For the third term on the right side, Assumption 3.2 implies that

$$\mathbb{E}_{S,\mathcal{A}}[\Phi_S(\mathbf{x}_S^*) - \Phi(\mathbf{x}^*)] \leq \mathbb{E}_{S,\mathcal{A}}[f_S(g_S(\mathbf{x}^*)) - f_S(g(\mathbf{x}^*))] \leq \frac{L_f V_g}{\sqrt{m}}. \tag{16}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] \\
=& \mathcal{O}\Bigg(\eta\beta^{-\frac{c}{2}} T^{1-\frac{c}{2}} + \frac{\eta^2 T}{\beta} + \frac{\sqrt{d}\sigma\eta^2 T}{\beta} + \eta\sqrt{\beta}T + \eta T^{\frac{1}{2}} + \frac{\eta T}{n} + \frac{\eta T}{m} \\
&+ \eta + \frac{1}{\sqrt{m}} + \gamma^{-1} + \gamma\beta^{-c}T^{-c} + \eta^{-1}T^{-1} + d\eta\sigma^2 + \gamma\eta^2\beta^{-2} + d\sigma^2\gamma\eta^2\beta^{-2} + \gamma\beta\Bigg) \\
=& \mathcal{O}\Bigg(\eta\beta^{-\frac{c}{2}} T^{1-\frac{c}{2}} + \frac{\sqrt{d}\sigma\eta^2 T}{\beta} + \eta\sqrt{\beta}T + \eta T^{\frac{1}{2}} + \frac{\eta T}{n} + \frac{\eta T}{m} \\
&+ \frac{1}{\sqrt{m}} + \gamma^{-1} + \gamma\beta^{-c}T^{-c} + \eta^{-1}T^{-1} + d\eta\sigma^2 + d\sigma^2\gamma\eta^2\beta^{-2} + \gamma\beta\Bigg),
\end{aligned}$$

where the last equality follows from omitting non-dominant terms. Recall that we chose

$$\sigma = \frac{c_1 L_f L_g \sqrt{T\log(1/\delta)}}{n\epsilon} + \frac{c_2 L_f L_g \sqrt{T\log(1/\delta)}}{m\epsilon}$$

in Theorem 4.1. By setting $c > 4$, $\eta = 1/\max\{T^{\frac{11}{12}}, T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1} + n^{-1}\epsilon^{-1})\}$, $\beta = T^{-\frac{1}{3}}$,$\gamma = T^{\frac{1}{6}}$ and $T = \max\{m^6, n^6\}$, we have

$$\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] = \mathcal{O}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{m} + \frac{1}{n}\right)\right).$$

Next, we present the proof for the SCSC update. Similar to the proof of the SCGD method, we have the following generalization bound

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{S,\mathcal{A}}\left[\Phi(\mathbf{x}_t) - \Phi_S(\mathbf{x}_t)\right] \\
\leq& 40 C_f L_f L_g^2 \eta \left(\frac{c}{e\beta}\right)^{c/2} D_Y T^{2-\frac{c}{2}} + \frac{20 C_f L_f^2 L_g^4 \eta^2 T^2}{\sqrt{\beta}} + \frac{20\sqrt{d} C_f L_f L_g^3 \sigma \eta^2 T^2}{\sqrt{\beta}} \\
&+ 20\sqrt{2} C_f L_f L_g^2 \sqrt{V_g} \eta \sqrt{\beta} T^2 + 20 L_f^2 L_g \sqrt{C_g} \eta T^{\frac{3}{2}} + \frac{3 L_f^2 L_g^2}{n} \eta T^2 + \frac{28 L_f^2 L_g^2}{m} \eta T^2 \\
&+ 40 C_f L_f L_g^2 \eta D_Y T + L_f T \frac{V_g}{\sqrt{m}}.
\end{aligned}$$

Similarly, we can bound the optimization error with the formula (15) such that

$$\begin{aligned}
&\Phi_S(\tilde{\mathbf{x}}) - \Phi_S(\mathbf{x}_S^*) \\
=& \mathcal{O}(D_X \eta^{-1} T^{-1} + C_f L_g^2 D_X \gamma^{-1} + L_f^2 L_g^2 \eta + d\eta\sigma^2 \\
&+ C_f \gamma \beta^{-c} T^{-c} D_Y + C_f L_f^2 L_g^4 \gamma \eta^2 \beta^{-1} + d C_f L_g^2 \sigma^2 \gamma \eta^2 \beta^{-1} + C_f V_g \gamma \beta).
\end{aligned}$$

Consequently, we have the following bound of excess population risk,

$$\begin{aligned}
&\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] \\
=& \mathcal{O}\Bigg(\eta \beta^{-\frac{c}{2}} T^{1-\frac{c}{2}} + \frac{\eta^2 T}{\sqrt{\beta}} + \frac{\sqrt{d}\sigma\eta^2 T}{\sqrt{\beta}} + \eta\sqrt{\beta} T + \eta T^{\frac{1}{2}} + \frac{\eta T}{n} + \frac{\eta T}{m} \\
&+ \eta + \frac{1}{\sqrt{m}} + \gamma^{-1} + \gamma\beta^{-c} T^{-c} + \eta^{-1} T^{-1} + d\eta\sigma^2 + \gamma\eta^2\beta^{-1} + d\sigma^2\gamma\eta^2\beta^{-1} + \gamma\beta\Bigg) \\
=& \mathcal{O}\Bigg(\eta \beta^{-\frac{c}{2}} T^{1-\frac{c}{2}} + \frac{\sqrt{d}\sigma\eta^2 T}{\sqrt{\beta}} + \eta\sqrt{\beta} T + \eta T^{\frac{1}{2}} + \frac{\eta T}{n} + \frac{\eta T}{m} \\
&+ \frac{1}{\sqrt{m}} + \gamma^{-1} + \gamma\beta^{-c} T^{-c} + \eta^{-1} T^{-1} + d\eta\sigma^2 + d\sigma^2\gamma\eta^2\beta^{-1} + \gamma\beta\Bigg),
\end{aligned}$$

where the last equality follows from omitting non-dominant terms. Recall that we chose

$$\sigma = \frac{c_1 L_f L_g \sqrt{T \log(1/\delta)}}{n\epsilon} + \frac{c_2 L_f L_g \sqrt{T \log(1/\delta)}}{m\epsilon}$$

in Theorem 4.1. By setting $c > 4$, $T = \min(m^4, n^4)$, $\eta = 1/\max(T^{\frac{7}{8}}, T\sqrt{d\log(1/\delta)}(m^{-1}\epsilon^{-1} + n^{-1}\epsilon^{-1}))$, $\beta = T^{-\frac{1}{2}}$ and $\gamma = T^{\frac{1}{4}}$, we obtain

$$\mathbb{E}_{S,\mathcal{A}}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] = \mathcal{O}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\left(\frac{1}{m} + \frac{1}{n}\right)\right).$$

□

C Utility Analysis of the Output Perturbation Method

In this section, we present the privacy and utility analysis of the output perturbation method proposed in Section 5 for solving the DP-SCO problem.

C.1 Key Lemmas

Firstly, we briefly review the concept of v-net. For any $v \in (0,1)$, $\{\mathbf{x}_l\}_{l=1}^{Q}$ is said to be a v-net of a domain $\mathcal{X}$, if $\mathbf{x}_l \in \mathcal{X}$ for any $l = 1, \ldots, Q$, and it holds that for $\forall \mathbf{x} \in \mathcal{X}$, there exists some $l(\mathbf{x}) \in [Q]$ such that $\left\|\mathbf{x} - \mathbf{x}_{l(\mathbf{x})}\right\| \leq v$.

If $\mathcal{X}$ has a finite diameter D_X, for any $v \in (0,1)$, there exists a v-net of $\mathcal{X}$, and the size of the v-net is bounded by $\mathcal{O}\big((D_X/v)^d\big)$ (Shapiro et al., 2021).

Before establishing the utility bound of the excess population risk of the output perturbation algorithm, we first show the stability and generalization of the ERM solution, which is stated in Lemma 5.1 and 5.3. Here we present the analysis of the more general Lemma 5.3, the proof of Lemma 5.1 follows by replacing the term $\Phi_S(\cdot) + \Psi(\cdot)$ with $\Phi_S(\cdot)$.

C.1.1 Proof of Lemma 5.3

Proof. We first compare the solutions obtained from the empirical risk minimization problem w.r.t. S and $S' = S^{i,\boldsymbol{\xi}}$. Since $\Phi_S(\mathbf{x}) + \Psi(\mathbf{x})$ is μ-strongly convex and $\hat{\mathbf{x}}_S^*$ is its optimal solution, we have

$$\Phi_S(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*) + \Psi(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*) \geq \Phi_S(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*) + \frac{\mu}{2}\left\|\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^* - \hat{\mathbf{x}}_S^*\right\|^2.$$

Similarly, using the μ-strong convexity of $\Phi_{S^{i,\boldsymbol{\xi}}}(\mathbf{x}) + \Psi(\mathbf{x})$ and the optimality of $\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*$, we have

$$\Phi_{S^{i,\boldsymbol{\xi}}}(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*) \geq \Phi_{S^{i,\boldsymbol{\xi}}}(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*) + \Psi(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*) + \frac{\mu}{2}\left\|\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^* - \hat{\mathbf{x}}_S^*\right\|^2.$$

Summing up the above two inequalities, we can obtain

$$\begin{aligned}
\mu\left\|\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^* - \hat{\mathbf{x}}_S^*\right\|^2 \leq& \Phi_S(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*) - \Phi_S(\hat{\mathbf{x}}_S^*) + \Phi_{S^{i,\boldsymbol{\xi}}}(\hat{\mathbf{x}}_S^*) - \Phi_{S^{i,\boldsymbol{\xi}}}(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*)\\
=&\frac{1}{n}[F(g_S(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*);\boldsymbol{\xi}_i) - F(g_S(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*);\boldsymbol{\xi}_i')] + \frac{1}{n}[F(g_S(\hat{\mathbf{x}}_S^*);\boldsymbol{\xi}_i') - F(g_S(\hat{\mathbf{x}}_S^*);\boldsymbol{\xi}_i)]\\
=&\frac{1}{n}[F(g_S(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*);\boldsymbol{\xi}_i) - F(g_S(\hat{\mathbf{x}}_S^*);\boldsymbol{\xi}_i)] + \frac{1}{n}[F(g_S(\hat{\mathbf{x}}_S^*);\boldsymbol{\xi}_i') - F(g_S(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*);\boldsymbol{\xi}_i')]\\
\leq&\frac{2L_fL_g}{n}\left\|\hat{\mathbf{x}}_S^* - \hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*\right\|.
\end{aligned}$$

The last inequality follows from Assumption 3.1. Divide both sides by $\mu\left\|\hat{\mathbf{x}}_S^* - \hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*\right\|$, we obtain

$$\left\|\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^* - \hat{\mathbf{x}}_S^*\right\| \leq \frac{2L_fL_g}{\mu n}. \tag{17}$$

Next, we compare the solutions obtained from S and $S' = S^{j,\boldsymbol{\zeta}}$. Since $\Phi_S(\mathbf{x}) + \Psi(\mathbf{x})$ is μ-strongly convex, we have

$$\Phi_S(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*) + \Psi(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*) \geq \Phi_S(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*) + \frac{\mu}{2}\left\|\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^* - \hat{\mathbf{x}}_S^*\right\|^2.$$

Similarly, using the strong convexity of $\Phi_{S^{j,\boldsymbol{\zeta}}}(\mathbf{x}) + \Psi(\mathbf{x})$ and the optimality of $\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*$, we have

$$\Phi_{S^{j,\boldsymbol{\zeta}}}(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*) \geq \Phi_{S^{j,\boldsymbol{\zeta}}}(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*) + \Psi(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*) + \frac{\mu}{2}\left\|\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^* - \hat{\mathbf{x}}_S^*\right\|^2.$$

Summing up the above two inequalities, we can obtain

$$\begin{aligned}
\mu\left\|\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^* - \hat{\mathbf{x}}_S^*\right\|^2 \leq& \Phi_S(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*) - \Phi_S(\hat{\mathbf{x}}_S^*) + \Phi_{S^{j,\boldsymbol{\zeta}}}(\hat{\mathbf{x}}_S^*) - \Phi_{S^{j,\boldsymbol{\zeta}}}(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*)\\
\leq& L_f\left\|g_S(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*) - g_{S^{j,\boldsymbol{\zeta}}}(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*)\right\| + L_f\left\|g_S(\hat{\mathbf{x}}_S^*) - g_{S^{j,\boldsymbol{\zeta}}}(\hat{\mathbf{x}}_S^*)\right\|\\
=&\frac{L_f}{m}\left\|G(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*;\boldsymbol{\zeta}_j) - G(\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^*;\boldsymbol{\zeta}_j')\right\| + \frac{L_f}{m}\left\|G(\hat{\mathbf{x}}_S^*;\boldsymbol{\zeta}_j) - G(\hat{\mathbf{x}}_S^*;\boldsymbol{\zeta}_j')\right\|\\
\leq&\frac{4L_fM_g}{m}.
\end{aligned}$$

The second inequality is due to Assumption 3.1 and the last inequality follows from Assumption 3.2. Consequently, one can show that

$$\left\|\hat{\mathbf{x}}_{S^{j,\boldsymbol{\zeta}}}^* - \hat{\mathbf{x}}_S^*\right\| \leq 2\sqrt{\frac{L_fM_g}{\mu m}}.$$

For any neighboring datasets S and S', we have

$$\|\hat{\mathbf{x}}_{S'}^* - \hat{\mathbf{x}}_S^*\| \leq \frac{2L_f L_g}{\mu n} + 2\sqrt{\frac{L_f M_g}{\mu m}}.$$

Next, we characterize the connection between the solution of the ERM problem and the solution of the SCO problem.

Recall that $\hat{\mathbf{x}}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\Phi_S(\mathbf{x}) + \Psi(\mathbf{x})\}$ and $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\Phi(\mathbf{x}) + \Psi(\mathbf{x})\}$, then

$$\begin{aligned}
&\mathbb{E}[\Phi(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*)] - \mathbb{E}[\Phi(\mathbf{x}^*) + \Psi(\mathbf{x}^*)] \\
=&\mathbb{E}[\Phi(\hat{\mathbf{x}}_S^*) - \Phi_S(\hat{\mathbf{x}}_S^*)] + \mathbb{E}[\Phi_S(\hat{\mathbf{x}}_S^*) + \Psi(\hat{\mathbf{x}}_S^*) - \Phi_S(\mathbf{x}^*) - \Psi(\mathbf{x}^*)] \\
&+ \mathbb{E}[\Phi_S(\mathbf{x}^*) - \Phi(\mathbf{x}^*)].
\end{aligned} \tag{18}$$

For the first term on the R.H.S. of (18), we can further decompose it as follows,

$$\begin{aligned}
&\mathbb{E}\left[\Phi(\hat{\mathbf{x}}_S^*) - \Phi_S(\hat{\mathbf{x}}_S^*)\right] \\
=&\mathbb{E}\left[\Phi(\hat{\mathbf{x}}_S^*) - \frac{1}{n}\sum_{i=1}^n F(g(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*); \boldsymbol{\xi}_i)\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n F(g(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*); \boldsymbol{\xi}_i) - \frac{1}{n}\sum_{i=1}^n F\left(\frac{1}{m}\sum_{j=1}^m G\left(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*; \boldsymbol{\zeta}_j\right); \boldsymbol{\xi}_i\right)\right] \\
&+ \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n F\left(\frac{1}{m}\sum_{j=1}^m G(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*; \boldsymbol{\zeta}_j); \boldsymbol{\xi}_i\right) - \Phi_S(\hat{\mathbf{x}}_S^*)\right].
\end{aligned} \tag{19}$$

Since $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_i'$ are i.i.d., $\hat{\mathbf{x}}_S^*$ and $\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*$ follow an identical distribution. In addition, $\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*$ is independent of $\boldsymbol{\xi}_i$, we can deduce the first term of Eq. (19) is 0.

To analyze the second term, denote

$$H_i(\mathbf{x}) \coloneqq F(g(\mathbf{x}); \boldsymbol{\xi}_i) - F\left(\frac{1}{m}\sum_{j=1}^m G(\mathbf{x}; \boldsymbol{\zeta}_j); \boldsymbol{\xi}_i\right).$$

We pick a v-net $\{\mathbf{x}_l\}_{l=1}^Q$ for the decision set $\mathcal{X}$, such that for any $\mathbf{x} \in \mathcal{X}$, there exists $l_0 \in \{1, \ldots, Q\}$, $\|\mathbf{x} - \mathbf{x}_{l_0}\| \leq v$. Then it holds for any $s > 0$,

$$\begin{aligned}
\exp(s\mathbb{E}[H_i(\hat{\mathbf{x}}_{S^{i,\boldsymbol{\xi}}}^*)]) \leq& \exp(s\mathbb{E}\max_{l=1,\ldots,Q} H_i(\mathbf{x}_l) + 2svL_fL_g) \\
\leq&\mathbb{E}\left[\max_{l=1,\ldots,Q}\exp(sH_i(\mathbf{x}_l) + 2svL_fL_g)\right] \\
\leq&\mathbb{E}\left[\sum_{l=1}^Q \exp(sH_i(\mathbf{x}_l) + 2svL_fL_g)\right] \\
=&\sum_{l=1}^Q \mathbb{E}\left[\exp(sH_i(\mathbf{x}_l) + 2svL_fL_g)\right].
\end{aligned} \tag{20}$$

The first inequality is due to Assumption 3.1. The second inequality follows from Jensen's inequality. Next we show that $H_i(\mathbf{x}_l) - \mathbb{E}[H_i(\mathbf{x}_l)]$ is a sub-Gaussian random variable for any given $\boldsymbol{\xi}_i$. Since $H_i(\mathbf{x}_l)$ is a function of $\{\boldsymbol{\zeta}_j\}_{j=1}^m$, denote $H_i(\mathbf{x}_l) \coloneqq \tilde{H}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_m)$. Then for any $p \in [m]$, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\zeta}_p'}\tilde{H}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{p-1}, \boldsymbol{\zeta}_p', \boldsymbol{\zeta}_{p+1}, \ldots, \boldsymbol{\zeta}_m) - \inf_{\boldsymbol{\zeta}_p''}\tilde{H}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{p-1}, \boldsymbol{\zeta}_p'', \boldsymbol{\zeta}_{p+1}, \ldots, \boldsymbol{\zeta}_m) \\
=&\sup_{\boldsymbol{\zeta}_p', \boldsymbol{\zeta}_p''}\left\{F\left(\frac{1}{m}\sum_{j\neq p} G(\mathbf{x}; \boldsymbol{\zeta}_j) + \frac{1}{m}G(\mathbf{x}; \boldsymbol{\zeta}_p''); \boldsymbol{\xi}_i\right) - F\left(\frac{1}{m}\sum_{j\neq p} G(\mathbf{x}; \boldsymbol{\zeta}_j) + \frac{1}{m}G(\mathbf{x}; \boldsymbol{\zeta}_p'); \boldsymbol{\xi}_i\right)\right\}
\end{aligned}$$

$$\leq \sup_{\boldsymbol{\zeta}_p', \boldsymbol{\zeta}_p''} \frac{L_f}{m} \left\| G(\mathbf{x}; \boldsymbol{\zeta}_p'') - G(\mathbf{x}; \boldsymbol{\zeta}_p') \right\|$$
$$\leq \frac{2M_g L_f}{m},$$

where the first inequality is due to Assumption 3.1 and the last inequality follows from Assumption 3.2. It implies that $H_i(\mathbf{x}_l)$ has bounded difference $\frac{2M_g L_f}{m}$. By McDiarmid's inequality, for any $r > 0$,

$$\mathbb{P}(H_i(\mathbf{x}_l) - \mathbb{E}[H_i(\mathbf{x}_l)] > r) \leq 2 \exp\left(-\frac{r^2 m}{2M_g^2 L_f^2}\right).$$

It implies that $H_i(\mathbf{x}_l) - \mathbb{E}[H_i(\mathbf{x}_l)]$ is a sub-Gaussian random variable with zero mean and variance proxy of $2M_g^2 L_f^2/m$ for any given $\boldsymbol{\xi}_i$. By definition, it yields

$$\mathbb{E}\left[\exp(s[H_i(\mathbf{x}_l) - \mathbb{E}[H_i(\mathbf{x}_l)]])\right] \leq \exp\left(\frac{2M_g^2 L_f^2 s^2}{m}\right).$$

Since $\mathbf{x}_l$ is independent of random variables $\{\boldsymbol{\zeta}_j\}_{j=1}^m$, we can deduce that $\mathbb{E}[H_i(\mathbf{x}_l)] \leq \frac{L_f V_g}{\sqrt{m}}$ by Assumption 3.1 and 3.2. It implies that

$$\mathbb{E}\left[\exp(s[H_i(\mathbf{x}_l)])\right] \leq \exp\left(\frac{2M_g^2 L_f^2 s^2}{m} + \frac{s L_f V_g}{\sqrt{m}}\right).$$

Substitute the above inequality into (20), we obtain

$$\exp\left(s\mathbb{E}\left[H_i(\hat{\mathbf{x}}^*_{S^i,\boldsymbol{\xi}})\right]\right) \leq Q \exp\left(\frac{2M_g^2 L_f^2 s^2}{m} + \frac{s L_f V_g}{\sqrt{m}} + 2svL_f L_g\right).$$

Taking the logarithm, dividing s on each side, and minimizing over s yields

$$\mathbb{E}\left[H_i(\hat{\mathbf{x}}^*_{S^i,\boldsymbol{\xi}})\right] \leq 2\sqrt{\frac{2\log(Q) L_f^2 M_g^2}{m}} + \frac{L_f V_g}{\sqrt{m}} + 2vL_f L_g.$$

Since $Q = \mathcal{O}((D_X/v)^d)$, we have

$$\mathbb{E}\left[H_i(\hat{\mathbf{x}}^*_{S^i,\boldsymbol{\xi}})\right] \leq \frac{cL_f M_g}{\sqrt{m}} \sqrt{d \log\left(\frac{D_X}{v}\right)} + \frac{L_f V_g}{\sqrt{m}} + 2vL_f L_g,$$

where $c > 0$ is some constant. For the third term on the right-hand side of Eq.(19), the bound (17) implies that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n F\left(\frac{1}{m}\sum_{j=1}^m G(\hat{\mathbf{x}}^*_{S^i,\boldsymbol{\xi}}; \boldsymbol{\zeta}_j); \boldsymbol{\xi}_i\right) - \Phi_S(\hat{\mathbf{x}}^*_S)\right] \leq L_f L_g \left\|\hat{\mathbf{x}}^*_{S^i,\boldsymbol{\xi}} - \hat{\mathbf{x}}^*_S\right\| \leq \frac{2L_f^2 L_g^2}{\mu n}.$$

For the second term on the right-hand side of Eq. (18), it is smaller than 0 due to the optimality of $\hat{\mathbf{x}}^*_S$. Assumption 3.1 and 3.2 implies the following bound for the third term on the right-hand side of Eq. (18) such that

$$\mathbb{E}[\Phi_S(\mathbf{x}^*) - \Phi(\mathbf{x}^*)] \leq \frac{L_f V_g}{\sqrt{m}}.$$

Putting everything together, one has

$$\mathbb{E}[\Phi(\hat{\mathbf{x}}^*_S) + \Psi(\hat{\mathbf{x}}^*_S)] - \mathbb{E}[\Phi(\mathbf{x}^*) + \Psi(\mathbf{x}^*)]$$
$$\leq \frac{2L_f^2 L_g^2}{\mu n} + \frac{cL_f M_g}{\sqrt{m}} \sqrt{d \log\left(\frac{D_X}{v}\right)} + \frac{2L_f V_g}{\sqrt{m}} + 2vL_f L_g.$$

□

The proof of Lemma 5.1 follows from that of Lemma 5.3 by replacing the occurrences of $\Phi_S(\cdot) + \Psi(\cdot)$ with $\Phi_S(\cdot)$.

C.2 Strongly Convex Case

In this subsection, we consider the privacy and utility analysis of the output perturbation method for the strongly convex case.

C.2.1 Proof of Theorem 5.2

Proof. We first prove the privacy guarantee. Given neighboring datasets $S \sim S'$, the sensitivity of $\mathcal{A}$ in Algorithm 2 is bounded as

$$\begin{aligned}&\|\mathcal{A}(S) - \mathcal{A}(S')\| \leq \|\mathcal{A}(S) - \mathbf{x}_S^*\| + \|\mathbf{x}_S^* - \mathbf{x}_{S'}^*\| + \|\mathcal{A}(S') - \mathbf{x}_{S'}^*\| \\ \leq &\frac{4L_f L_g}{\mu n} + 4\sqrt{\frac{L_f M_g}{\mu m}}.\end{aligned}$$

with probability $1 - \frac{\delta}{2}$ by the union bound, where the last inequality follows from Lemma 5.3 and the guarantee of $\mathcal{A}$. Then by Gaussian mechanism in Definition 3.5, Algorithm 2 is (ϵ, δ)-DP when setting

$$\sigma = \left(\frac{4L_f L_g}{\mu n} + 4\sqrt{\frac{L_f M_g}{\mu m}}\right)\frac{\sqrt{2\log(2.5/\delta)}}{\epsilon}.$$

We then give the guarantee for the output $\tilde{\mathbf{x}}$.

$$\begin{aligned}\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] \leq &\frac{C_f C_g}{2}\mathbb{E}\left[\|\tilde{\mathbf{x}} - \mathbf{x}^*\|^2\right] \\ \leq &\frac{3C_f C_g}{2}\left(\mathbb{E}\left[\|\tilde{\mathbf{x}} - \mathcal{A}(S)\|^2\right] + \mathbb{E}\left[\|\mathcal{A}(S) - \mathbf{x}_S^*\|^2\right] + \mathbb{E}\left[\|\mathbf{x}_S^* - \mathbf{x}^*\|^2\right]\right) \\ \leq &\frac{3C_f C_g}{2}\left(\left(\frac{L_f^2 L_g^2}{\mu^2 n^2} + \frac{L_f M_g}{\mu m}\right)\frac{64d\log(2.5/\delta)}{\epsilon^2} + \frac{\delta^2 L_f^2 L_g^2}{8\mu^2 n^2} + \frac{\delta^2 L_f M_g}{8\mu m}\right. \\ &\left. + \frac{4L_f^2 L_g^2}{\mu^2 n} + \frac{cL_f M_g}{\mu\sqrt{m}}\sqrt{d\log\left(\frac{D_X}{v}\right)} + \frac{4L_f V_g}{\mu\sqrt{m}} + 4vL_f L_g \mu^{-1}\right).\end{aligned}$$

The first inequality is due to Assumption 3.2. The last inequality follows from the definition of $\tilde{\mathbf{x}}$, convergence guarantee of $\mathcal{A}$ and Lemma 5.1. By setting $v = 1/\sqrt{m}$, we have

$$\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(x^*)] = \tilde{\mathcal{O}}\left(\frac{1}{\mu^2 n} + \frac{\sqrt{d}}{\mu\sqrt{m}} + \frac{d\log(1/\delta)}{\epsilon^2}\left(\frac{1}{\mu^2 n^2} + \frac{1}{\mu m}\right)\right).$$

□

C.3 Convex Setting

In this subsection, we consider the utility and privacy analysis of the output perturbation method for the general convex case. Firstly, we establish the following connection between the optimal solutions of the successive empirical risk minimization problems $\Phi_{S,k}(\mathbf{x})$ in Algorithm 3.

Lemma C.1. *For $1 \leq k \leq K$, running Algorithm 3 with any $\mu_k > 0$, we have*

$$\begin{aligned}&\mathbb{E}\left[\Phi(\hat{\mathbf{x}}_k^*) - \Phi(\hat{\mathbf{x}}_{k-1}^*)\right] \\ \leq &\frac{\mu_k}{2}\left\|\hat{\mathbf{x}}_{k-1}^* - \tilde{\mathbf{x}}_{k-1}\right\|^2 + \frac{2L_f^2 L_g^2}{\mu_k \bar{n}} + \frac{cL_f M_g}{\sqrt{\bar{m}}}\sqrt{d\log(D_X\sqrt{\bar{m}})} + \frac{2L_f V_g + 2L_f L_g}{\sqrt{\bar{m}}},\end{aligned}$$

where $\hat{\mathbf{x}}_k^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \hat{\Phi}_{S,k}(\mathbf{x})$.

Proof. Applying Lemma 5.3 with $v = 1/\sqrt{\bar{m}}$, we have

$$\mathbb{E}\left[\Phi(\hat{\mathbf{x}}_k^*) + \frac{\mu_k}{2}\|\hat{\mathbf{x}}_k^* - \tilde{\mathbf{x}}_{k-1}\|^2\right]$$

Algorithm 4: Output Perturbation for Regularized Objective Function

1 Inputs: data $S = S_{\boldsymbol{\xi}} \cup S_{\boldsymbol{\zeta}}$, privacy budget ϵ, δ, number of iterations T, strong-convexity parameter μ, and noise scaling parameter σ, a randomized algorithm $\mathcal{A}$.

2 Initialize $\mathbf{x}_0 \in \mathcal{X}$.

3 Run the given algorithm $\mathcal{A}$ on the smooth strongly-convex finite-sum problem $\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n F(\frac{1}{m}\sum_{j=1}^m G(\mathbf{x};\boldsymbol{\zeta}_j);\boldsymbol{\xi}_i) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{x}_0\|^2$ to obtain output $\mathcal{A}(S)$ such that $\|\mathcal{A}(S) - \mathbf{x}_S^*\| \le \frac{L_f L_g}{\mu n} + \sqrt{\frac{L_f M_g}{\mu m}}$ with probability at least $1-\delta/4$, where $\mathbf{x}_S^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_S(\mathbf{x})$ is the empirical optimal solution.

4 Return: $\tilde{\mathbf{x}} = \Pi_{\mathcal{X}}(\mathcal{A}(S) + \mathcal{N}(0, \sigma^2 I_d))$.

$$
\begin{aligned}
&\le \min_{\mathbf{x}'\in\mathcal{X}} \mathbb{E}\left[\Phi(\mathbf{x}') + \frac{\mu_k}{2}\|\mathbf{x}' - \tilde{\mathbf{x}}_{k-1}\|^2\right] + \frac{2L_f^2 L_g^2}{\mu_k \bar{n}} \\
&\quad + \frac{cL_f M_g}{\sqrt{\bar{m}}}\sqrt{d\log(D_X\sqrt{\bar{m}})} + \frac{2L_f V_g + 2L_f L_g}{\sqrt{\bar{m}}} \\
&\le \mathbb{E}\left[\Phi(\hat{\mathbf{x}}_{k-1}^*) + \frac{\mu_k}{2}\left\|\hat{\mathbf{x}}_{k-1}^* - \tilde{\mathbf{x}}_{k-1}\right\|^2\right] + \frac{2L_f^2 L_g^2}{\mu_k \bar{n}} \\
&\quad + \frac{cL_f M_g}{\sqrt{\bar{m}}}\sqrt{d\log(D_X\sqrt{\bar{m}})} + \frac{2L_f V_g + 2L_f L_g}{\sqrt{\bar{m}}}.
\end{aligned}
$$

Rearrange the terms, we have

$$
\begin{aligned}
&\mathbb{E}\left[\Phi(\hat{\mathbf{x}}_k^*) - \Phi(\hat{\mathbf{x}}_{k-1}^*)\right] \\
&\le \frac{\mu_k}{2}\left\|\hat{\mathbf{x}}_{k-1}^* - \tilde{\mathbf{x}}_{k-1}\right\|^2 + \frac{2L_f^2 L_g^2}{\mu_k \bar{n}} + \frac{cL_f M_g}{\sqrt{\bar{m}}}\sqrt{d\log(D_X\sqrt{\bar{m}})} + \frac{2L_f V_g + 2L_f L_g}{\sqrt{\bar{m}}}.
\end{aligned}
$$

□

C.3.1 A Single Regularizer Approach

A straightforward approach to solving the problem is transforming the convex problem to a strongly convex one by adding an ℓ_2 regularizer $\mu\|\mathbf{x}\|^2/2$ to the objective function. Then we apply the output perturbation method to solve the regularized problem with a careful choice of parameter μ.

Theorem C.2. *Let the function $\Phi(\cdot)$ and $F(G_S(\cdot);\boldsymbol{\xi})$ be convex for any $\boldsymbol{\xi}$ and S. By choosing the parameter*

$$
\sigma = \left(\frac{4L_f L_g}{\mu n} + 4\sqrt{\frac{L_f M_g}{\mu m}}\right)\frac{\sqrt{2\log(2.5/\delta)}}{\epsilon}, \quad \mu = \max\left\{L_f L_g/(D_X\sqrt{n}), L_f L_g^{\frac{2}{3}} M_g^{\frac{1}{3}} m^{-\frac{1}{3}}\right\},
$$

Algorithm 4 is (ϵ,δ)-DP and its output $\tilde{\mathbf{x}}_K$ satisfies that

$$
\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] = \mathcal{O}\left(\left(\frac{1}{\sqrt{n}} + \frac{1}{m^{\frac{1}{3}}}\right)\frac{\sqrt{d\log(1/\delta)}}{\epsilon} + \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\sqrt{m}} + \frac{1}{m^{\frac{1}{3}}}\right).
$$

Proof. We first present the privacy guarantees of Algorithm 4. By the stability of $\mathbf{x}_S^*$ in Lemma 5.3 and the requirement of output $A(S)$ in Algorithm 4, we know $A(S)$ has sensitivity bounded by $\frac{4L_f L_g}{\mu n} + 4\sqrt{\frac{L_f M_g}{\mu m}}$ with a probability of $1-\frac{\delta}{2}$ following a similar proof of Theorem 5.2. By setting

$$
\sigma = \left(\frac{4L_f L_g}{\mu n} + 4\sqrt{\frac{L_f M_g}{\mu m}}\right)\frac{\sqrt{2\log(2.5/\delta)}}{\epsilon}
$$

guarantees (ϵ, δ)-DP. As a result, we can obtain

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\mathbf{x}} - \mathbf{x}_S^*\|^2\right] \leq & 2\mathbb{E}\left[\|\tilde{\mathbf{x}} - A(S)\|^2\right] + 2\mathbb{E}\left[\|A(S) - \mathbf{x}_S^*\|^2\right] \\ \leq & \left(\frac{64L_f^2L_g^2}{\mu^2n^2} + \frac{64L_fM_g}{\mu m}\right)\frac{d\log(2.5/\delta)}{\epsilon^2} + \left(\frac{L_f^2L_g^2}{\mu^2n^2} + \frac{L_fM_g}{\mu m}\right)\frac{\delta^2}{8} \\ \leq & \left(\frac{65L_f^2L_g^2}{\mu^2n^2} + \frac{65L_fM_g}{\mu m}\right)\frac{d\log(2.5/\delta)}{\epsilon^2}.\end{aligned} \tag{21}$$

We decompose the generalization error as follows,

$$\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] = \mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}_S^*)] + \mathbb{E}[\Phi(\mathbf{x}_S^*) - \Phi(\mathbf{x}^*)],$$

where $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$. For the first term on the r.h.s. of the above formula, we have

$$\begin{aligned}\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}_S^*)] \leq & L_fL_g\sqrt{\mathbb{E}\left[\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}_S^*\|^2\right]} \\ \leq & L_fL_g\left(\frac{9L_fL_g}{\mu n} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu m}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon},\end{aligned}$$

where the first inequality holds by Assumption 3.1 and Cauchy–Schwarz inequality. The second inequality is due to Eq. (21). Using a similar proof of Lemma C.1, we can show that

$$\begin{aligned}&\mathbb{E}[\Phi(\mathbf{x}_S^*) - \Phi(\mathbf{x}^*)] \\ \leq & \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_0\|^2 + \frac{2L_f^2L_g^2}{\mu n} + \frac{cL_fM_g}{\sqrt{m}}\sqrt{d\log(D_X\sqrt{m})} + \frac{2L_fV_g + 2L_fL_g}{\sqrt{m}}.\end{aligned}$$

Combining the above two inequalities together, we have

$$\begin{aligned}&\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] \\ \leq & L_fL_g\left(\frac{9L_fL_g}{\mu n} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu m}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon} + \frac{\mu}{2}D_X^2 \\ & + \frac{2L_f^2L_g^2}{\mu n} + \frac{cL_fM_g}{\sqrt{m}}\sqrt{d\log(D_X\sqrt{m})} + \frac{2L_fV_g + 2L_fL_g}{\sqrt{m}}.\end{aligned}$$

By setting $\mu = \max\{L_fL_g/(D_X\sqrt{n}), L_fL_g^{\frac{2}{3}}M_g^{\frac{1}{3}}m^{-\frac{1}{3}}\}$, we have

$$\begin{aligned}&\mathbb{E}[\Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x}^*)] \\ = & \mathcal{O}\left(\left(\frac{1}{\sqrt{n}} + \frac{1}{m^{\frac{1}{3}}}\right)\frac{\sqrt{d\log(1/\delta)}}{\epsilon} + \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\sqrt{m}} + \frac{1}{m^{\frac{1}{3}}}\right).\end{aligned}$$

□

C.3.2 Phased Output Perturbation Approach

Now we show that applying the phased output perturbation approach (Algorithm 3) to solve the general convex function which leads to an improved bound of excess population risk. In particular, we present the proof of Theorem 5.4 as follows.

Proof of Theorem 5.4

Proof. We first give the guarantees of each phase in Algorithm 3. For phase $1 \leq k \leq K$, by the stability of $\hat{\mathbf{x}}_k^*$ in Lemma 5.3 and the requirement of output $\mathbf{x}_k$ in Algorithm 3, we know $\mathbf{x}_k$ has sensitivity bounded by

$\frac{4L_fL_g}{\mu_k\bar{n}} + 4\sqrt{\frac{L_fM_g}{\mu_k\bar{m}}}$ with a probability of $1-\frac{\delta}{2}$ following a similar proof of Theorem 5.2. By setting

$$\sigma_k = \left(\frac{4L_fL_g}{\mu_k\bar{n}} + 4\sqrt{\frac{L_fM_g}{\mu_k\bar{m}}}\right)\frac{\sqrt{2\log(2.5/\delta)}}{\epsilon},$$

it guarantees the output of Algorithm 3 at phase k is (ϵ,δ)-DP. Then we can give the privacy analysis of the full algorithm. By Lemma A.8, Algorithm 3 is (ϵ,δ)-DP since we use disjoint datasets for different phases and each phase is (ϵ,δ)-DP. For the utility bound, we decompose the error as

$$\mathbb{E}[\Phi(\tilde{\mathbf{x}}_K) - \Phi(\mathbf{x}^*)] = \mathbb{E}[\Phi(\tilde{\mathbf{x}}_K) - \Phi(\hat{\mathbf{x}}_K^*)] + \sum_{k=1}^{K}\{\mathbb{E}[\Phi(\hat{\mathbf{x}}_k^*) - \Phi(\hat{\mathbf{x}}_{k-1}^*)]\}, \tag{22}$$

where $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$. We let $\hat{\mathbf{x}}_0^* = \mathbf{x}^*$ for simplicity of the analysis. In addition, we can show that

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\mathbf{x}}_k - \hat{\mathbf{x}}_k^*\|^2\right] \leq& 2\mathbb{E}\left[\|\tilde{\mathbf{x}}_k - \mathbf{x}_k\|^2\right] + 2\mathbb{E}\left[\|\mathbf{x}_k - \hat{\mathbf{x}}_k^*\|^2\right] \\ \leq& \left(\frac{64L_f^2L_g^2}{\mu_k^2\bar{n}^2} + \frac{64L_fM_g}{\mu_k\bar{m}}\right)\frac{d\log(2.5/\delta)}{\epsilon^2} + \left(\frac{L_f^2L_g^2}{\mu_k^2\bar{n}^2} + \frac{L_fM_g}{\mu_k\bar{m}}\right)\frac{\delta^2}{8} \\ \leq& \left(\frac{65L_f^2L_g^2}{\mu_k^2\bar{n}^2} + \frac{65L_fM_g}{\mu_k\bar{m}}\right)\frac{d\log(2.5/\delta)}{\epsilon^2}.\end{aligned} \tag{23}$$

For the first term on the r.h.s. of the Eq. (22), we have

$$\begin{aligned}\mathbb{E}[\Phi(\tilde{\mathbf{x}}_K) - \Phi(\hat{\mathbf{x}}_K^*)] \leq& L_fL_g\sqrt{\mathbb{E}\left[\|\tilde{\mathbf{x}}_K - \hat{\mathbf{x}}_K^*\|^2\right]} \\ \leq& L_fL_g\left(\frac{9L_fL_g}{\mu_K\bar{n}} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu_K\bar{m}}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon} \\ =& L_fL_g\left(\frac{9L_fL_g}{\mu n\bar{n}} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu n\bar{m}}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon},\end{aligned}$$

where the first inequality holds by Assumption 3.1 and Cauchy–Schwarz inequality. The second inequality is due to Eq. (23). Applying Lemma C.1, we can deduce that

$$\begin{aligned}&\mathbb{E}[\Phi(\tilde{\mathbf{x}}_K) - \Phi(\mathbf{x}^*)] \\ \leq& L_fL_g\left(\frac{9L_fL_g}{\mu n\bar{n}} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu n\bar{m}}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon} \\ &+ \sum_{k=1}^{K}\left(\frac{\mu_k}{2}\left\|\hat{\mathbf{x}}_{k-1}^* - \tilde{\mathbf{x}}_{k-1}\right\|^2 + \frac{2L_f^2L_g^2}{\mu_k\bar{n}} + \frac{cL_fM_g}{\sqrt{\bar{m}}}\sqrt{d\log(D_X\sqrt{\bar{m}})} + \frac{2L_fV_g + 2L_fL_g}{\sqrt{\bar{m}}}\right) \\ \leq& L_fL_g\left(\frac{9L_fL_g}{\mu n\bar{n}} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu n\bar{m}}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon} \\ &+ \sum_{k=2}^{K}\frac{\mu_k}{2\mu_{k-1}}\left(\frac{65L_f^2L_g^2}{\mu_{k-1}\bar{n}^2} + \frac{65L_fM_g}{\bar{m}}\right)\frac{d\log(2.5/\delta)}{\epsilon^2} + \frac{2L_f^2L_g^2}{\mu\bar{n}} \\ &+ \frac{cL_fM_gK}{\sqrt{\bar{m}}}\sqrt{d\log(D_X\sqrt{\bar{m}})} + \frac{2KL_f(V_g + L_g)}{\sqrt{\bar{m}}} + \mu\left\|\mathbf{x}^* - \mathbf{x}_0\right\|^2 \\ \leq& L_fL_g\log(n)\left(\frac{9L_fL_g}{\mu n^2} + \frac{9\sqrt{L_fM_g}}{\sqrt{\mu nm}}\right)\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon} \\ &+ \left(\frac{65L_f^2L_g^2}{\mu\bar{n}^2} + \frac{65L_fM_g\log(n)}{m}\right)\frac{d\log(2.5/\delta)}{\epsilon^2} + \frac{2L_f^2L_g^2}{\mu\bar{n}}\end{aligned}$$

Table 3: We compare the convergence rate of stochastic first-order methods for solving the non-private SCO problem in the strongly convex case. T is the number of iterations. $\hat{\varkappa}$ is the generalized condition number.

Methods	Convergence Rate	Reference
SCGD	$\mathcal{O}(T^{-\frac{2}{3}})$	Wang et al. (2017a)
Accelerated SCGD	$\mathcal{O}(T^{-\frac{4}{5}})$	Wang et al. (2017b)
SCSC	$\mathcal{O}(T^{-1})$	Theorem D.2
Compositional SVRG	$\mathcal{O}\left(\rho^{\frac{T}{m+n+\hat{\varkappa}^3}}\right)$	Lian et al. (2017)

$$+\frac{cL_f M_g \log(n)^{\frac{3}{2}}}{\sqrt{m}}\sqrt{d\log(D_X\sqrt{\bar{m}})}+\frac{2\log(n)^{\frac{3}{2}}L_f(V_g+L_g)}{\sqrt{m}}+\mu\left\|\mathbf{x}^*-\mathbf{x}_0\right\|^2,$$

where we use (23) for the second inequality. By setting

$$\mu=\max\left\{\frac{L_fL_g}{D_X\sqrt{n}},\frac{L_fL_g}{D_Xn}\frac{\sqrt{d\log(2.5/\delta)}}{\epsilon}\right\},$$

we obtain

$$\mathbb{E}\left[\Phi(\tilde{\mathbf{x}}_K)-\Phi(\mathbf{x}^*)\right]=\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d}}{\sqrt{m}}+\left(\frac{1}{n}+\frac{1}{\sqrt{m}}\right)\frac{\sqrt{d\log(1/\delta)}}{\epsilon}\right).$$

□

D SCSC for Strongly Convex Functions

In this section, we first present the convergence rate of stochastic first-order methods for solving the stochastic compositional optimization problem when the objective function is strongly convex. The theoretical results of the methods are summarized in Table 3. We observe the compositional SVRG method Lian et al. (2017) achieves a linear rate of convergence, which outperforms the other methods.

Note that Chen et al. (2021) proposed the stochastically corrected stochastic compositional gradient (SCSC) method and provided the convergence analysis of the method for the nonconvex smooth optimization problem. However, the convergence analysis of the SCSC method for the strongly convex problem is still unclear. To this end, we present the convergence analysis of the non-private SCSC algorithm Chen et al. (2021) for the strongly convex function. Our proof closely follows the techniques used by Wang et al. (2017a) except we use Lemma A.7 instead of A.6. In particular, Wang et al. (2017a) introduced the following key lemma to characterize the convergence of the SCGD method, which also applies to the SCSC method.

Lemma D.1 (Wang et al. (2017a))**.** *Suppose that $\Phi(\cdot)$ is a convex function, then running Algorithm 5 guarantees that*

$$\mathbb{E}[\|\mathbf{x}_{t+1}-\mathbf{x}^*\|^2\mid\mathcal{F}_t]\le\|\mathbf{x}_t-\mathbf{x}^*\|^2+L_fL_g\eta_t^2-2\eta_t(\Phi(\mathbf{x}_t)-\Phi(\mathbf{x}^*))+\mathbb{E}[\mathbf{u}_t\mid\mathcal{F}_t],$$

where

$$\mathbf{u}_t=2\eta_t\langle\mathbf{x}_t-\mathbf{x}^*,\nabla G(\mathbf{x}_t;\boldsymbol{\zeta}_{j,t})(\nabla F(g(\mathbf{x}_t);\boldsymbol{\xi}_{i,t})-\nabla F(\mathbf{y}_{t+1};\boldsymbol{\xi}_{i,t}))\rangle.$$

We can show the convergence analysis of the SCSC method (Algorithm 5) as follows.

Theorem D.2. *Suppose $\Phi(\cdot)$ is μ-strongly convex. Running Algorithm 5 with $\eta_t=1/(\mu t)$, $\beta_t=\eta_t$, it holds that*

$$\mathbb{E}[\|\tilde{\mathbf{x}}-\mathbf{x}^*\|^2]=\mathcal{O}\left(\frac{L_fL_g}{T\mu^2}\log T+\frac{C_f^2L_g(L_fL_g^2+V_g)}{T\mu^2}\log T\right).$$

Algorithm 5: SCSC Method

Inputs: Dataset S, number of iterations T, parameters $\{\eta_t\}$, $\{\beta_t\}$
Initialize $\mathbf{x}_0 = \mathbf{x}_{-1} \in \mathcal{X}$ and $\mathbf{y}_0 \in \mathbb{R}^p$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $j_t \in [m]$
  Let $\mathbf{z}_t = (1-\beta_t)(G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t}) - G(\mathbf{x}_{t-1}; \boldsymbol{\zeta}_{j_t}))$
  $\mathbf{y}_{t+1} = (1-\beta_t)\mathbf{y}_t + \beta_t G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t}) + \mathbf{z}_t$
  Sample $i_t \in [n]$
  Let $\tilde{\mathbf{g}}_t = \nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j_t}) \nabla F(\mathbf{y}_{t+1}; \boldsymbol{\xi}_{i_t})$
  Update $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta_t \tilde{\mathbf{g}}_t)$
end
Return: $\tilde{\mathbf{x}} = \mathbf{x}_T$.

Proof. Define $\mathbf{u}_t$ to be

$$\mathbf{u}_t = 2\eta_t \langle \mathbf{x}_t - \mathbf{x}^*, \nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j,t})(\nabla F(g(\mathbf{x}_t); \boldsymbol{\xi}_{i,t}) - \nabla F(\mathbf{y}_{t+1}; \boldsymbol{\xi}_{i,t}))\rangle.$$

By Cauchy–Schwarz inequality, we can show that

$$\begin{aligned}
\mathbf{u}_t \leq & 2\eta_t \|\mathbf{x}_t - \mathbf{x}^*\| \|\nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j,t})\| \|\nabla F(g(\mathbf{x}_t); \boldsymbol{\xi}_{i,t}) - \nabla F(\mathbf{y}_{t+1}; \boldsymbol{\xi}_{i,t})\| \\
\leq & 2\eta_t C_f \|\mathbf{x}_t - \mathbf{x}^*\| \|\nabla G(\mathbf{x}_t; \boldsymbol{\zeta}_{j,t})\| \|g(\mathbf{x}_t) - \mathbf{y}_{t+1}\| \\
\leq & \eta_t \mu \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \frac{\eta_t C_f^2 L_g}{\mu} \|g(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2 .
\end{aligned}$$

The last inequality is due to Assumption 3.1 and 3.2. Combining the above inequality with Lemma D.1, we have

$$\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_t] \leq (1-\mu\eta_t)\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] + L_f L_g \eta_t^2 + \frac{\eta_t C_f^2 L_g}{\mu} \mathbb{E}[\|g(\mathbf{x}_t) - \mathbf{y}_{t+1}\|^2]. \tag{24}$$

Next, we define the variable

$$J_t = \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \Lambda(\eta_t, \beta_t) \|g(\mathbf{x}_{t-1}) - \mathbf{y}_t\|^2 ,$$

where $\Lambda(\eta_t, \beta_t) = \max\left\{\frac{C_f^2 L_g \eta_t (\beta_t - 1)/\mu}{\mu\eta_t - \beta_t}, 0\right\} = \Theta\left(\frac{C_f^2 L_g \eta_t}{\beta_t \mu}\right)$. Multiplying the formula in Lemma A.7 by $\Lambda(\eta_t, \beta_t)$ and taking its sum with Eq. (24), we obtain

$$\mathbb{E}[J_{t+1}] \leq (1-\mu\eta_t)\mathbb{E}[J_t] + L_f L_g \eta_t^2 + \mathcal{O}\left(\frac{C_f^2 L_g}{\mu}\right)\left(L_f L_g^2 \frac{\eta_t^3}{\beta_t} + V_g \eta_t \beta_t\right).$$

Taking $\eta_t = 1/(\mu t)$, $\beta_t = \eta_t$ and multiplying the preceding inequality with t on both sides, we have

$$\begin{aligned}
t\mathbb{E}[J_{t+1}] \leq & (t-1)\mathbb{E}[J_t] + \frac{L_f L_g}{t\mu^2} + \mathcal{O}\left(\frac{C_f^2 L_g (L_f L_g^2 + V_g)}{t\mu^2}\right) \\
= & \mathcal{O}\left(\frac{L_f L_g}{\mu^2} \log t + \frac{C_f^2 L_g (L_f L_g^2 + V_g)}{\mu^2} \log t\right).
\end{aligned}$$

Finally, we can show that

$$\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\right] \leq \mathbb{E}[J_{t+1}] = \mathcal{O}\left(\frac{L_f L_g}{t\mu^2} \log t + \frac{C_f^2 L_g (L_f L_g^2 + V_g)}{t\mu^2} \log t\right).$$

By noting $\tilde{\mathbf{x}} = \mathbf{x}_T$, we get the desired results. □