Non-smooth stochastic gradient descent using smoothing functions

T. Giovannelli* J. Tan† L. N. Vicente‡

July 15, 2025

Abstract

In this paper, we address stochastic optimization problems involving a composition of a non-smooth outer function and a smooth inner function, a formulation frequently encountered in machine learning and operations research. To deal with the non-differentiability of the outer function, we approximate the original non-smooth function using smoothing functions, which are continuously differentiable and approach the original function as a smoothing parameter goes to zero (at the price of increasingly higher Lipschitz constants). The proposed smoothing stochastic gradient method iteratively drives the smoothing parameter to zero at a designated rate. We establish convergence guarantees under strongly convex, convex, and non-convex settings, proving convergence rates that match known results for non-smooth stochastic compositional optimization. In particular, for convex objectives, smoothing stochastic gradient achieves a $1/T^{1/4}$ rate in terms of the number of stochastic gradient evaluations. We further show how general compositional and finite-sum compositional problems (widely used frameworks in large-scale machine learning and risk-averse optimization) fit the assumptions needed for the rates (unbiased gradient estimates, bounded second moments, and accurate smoothing errors). We present preliminary numerical results indicating that smoothing stochastic gradient descent can be competitive for certain classes of problems.

1 Introduction

Stochastic compositional optimization problems, in which the objective function involves a compositional structure of an inner and an outer function, have found increasing applications in machine learning and risk-averse optimization (see [5, 40]). In this paper, we will consider the following two stochastic compositional problems

$$\min_{x\in\mathbb{R}^n} \quad \phi(\mathbb{E}[\psi(x,\xi)]), \tag{1.1}$$

*Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, OH 45221, USA (giovanto@ucmail.uc.edu).

†Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (jit423@lehigh.edu).

‡Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (lnv@lehigh.edu).

and

$$\min_{x\in\mathbb{R}^n} \ \mathbb{E}[\phi(\psi(x,\xi))], \tag{1.2}$$

where x is the vector of decision variables, ξ is a random vector, $\mathbb{E}$ is taken with respect to the probability distribution of ξ, ψ is a differentiable function, and ϕ is a function which is not necessarily smooth (meaning that the function ϕ may be non-differentiable).

Both types of problems naturally occur in fields like reinforcement learning [8, 22], risk management [1, 26], meta-learning [17], and distributionally robust optimization [23]. An example of $\phi(\mathbb{E}[\psi(x,\xi)])$ involves super and sub-quantiles and functions that can be written as

$$\max\ \{0, \mathbb{E}[\psi(x,\xi)])\}.$$

An instance of $\min_{x\in\mathbb{R}^n} \ \mathbb{E}[\phi(\psi(x,\xi))]$ is expected risk minimization, which is widely adopted in machine learning applications. A specific example is given as follows

$$\min_{x\in\mathbb{R}^n} \ \mathbb{E}[|x^\top u + b - v|],$$

where the outer function is the $L1$ loss, the inner function is linear, and $\xi = (u, v)$. We will discuss the formulations (1.1) and (1.2) in more detail in Sections 4 and 5.

Due to the existence of non-differentiable points of the non-smooth outer function ϕ, traditional gradient-based methods cannot be directly applied to problems (1.1) and (1.2). To deal with the non-smoothness of ϕ, we will use a smoothing function (see Definition 1 below). Smoothing functions have been introduced in [4, 51].

Definition 1 *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function. We call $\tilde{g} : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}$ a smoothing function of g if, for any $\mu \in (0, +\infty)$, $\tilde{g}(\cdot, \mu)$ is continuously differentiable in $\mathbb{R}^n$ and, for any $x \in \mathbb{R}^n$,*

$$\lim_{z\to x, \mu\downarrow 0} \ \tilde{g}(z,\mu) \ = \ g(x). \tag{1.3}$$

In Section 2, we will provide an introduction to the essential properties of smoothing functions. We will also see in Sections 4 and 5 what conditions $\tilde{\phi}$ and ψ need to satisfy to ensure that $\mathbb{E}[\tilde{\phi}(\psi(\cdot,\xi),\mu)]$ and $\tilde{\phi}(\mathbb{E}[\psi(\cdot,\xi)],\mu)$ are smoothing functions of $\mathbb{E}[\phi(\psi(\cdot,\xi))]$ and $\phi(\mathbb{E}[\psi(\cdot,\xi)])$, respectively (satisfying appropriate accuracy and convexity requirements). We will assume that we can compute stochastic smoothing gradients for both $\mathbb{E}[\tilde{\phi}(\psi(\cdot,\xi),\mu)]$ and $\tilde{\phi}(\mathbb{E}[\psi(x,\xi)],\mu)$ (the explicit forms will be shown in Sections 4 and 5). As the smoothing parameter approaches zero, the smoothing function will approach the original function but at the price of larger Lipschitz constants.

1.1 Short review of convergence rates for gradient descent

Gradient descent (GD) and stochastic gradient descent (SGD) are widely used in optimization and machine learning problems due to their simplicity, efficiency, and effectiveness, particularly in terms of low iteration cost and memory storage requirements. The first development of GD can be traced back to 1847. Cauchy introduced a method of iterative improvement, laying out the fundamental idea of moving toward function minima by repeatedly taking steps proportional to

the negative gradient. SGD arose in the context of the statistical approximation method, which was first introduced by Robbins and Monro [38] in 1951. Over the subsequent decades, SGD gained widespread popularity in large-scale problems [3], owing to its efficiency and scalability in handling large datasets and high-dimensional parameter spaces.

Modern convergence theory traces back to Polyak's seminal work on GD for L-smooth objectives in the 1960s [35, 36]. For a convex and L-smooth function f, taking a fixed stepsize $1/L$ yields the classical sub-linear rate $\mathcal{O}(1/T)$ [33]. When f is additionally μ-strongly convex, the same scheme converges at the linear rate $\mathcal{O}((1-\mu/L)^T)$ [36]. In the non-smooth convex setting, replacing gradients with subgradients and using the diminishing stepsizes $\gamma_t = \Theta(1/\sqrt{t})$ attains the optimal rate $\mathcal{O}(1/\sqrt{T})$ [31, 41]. Furthermore, for non-smooth and μ-strongly convex objectives, choosing $\gamma_t = \Theta(1/t)$ tightens the bound to $\mathcal{O}(1/T)$ [37].

For SGD, analogous complexity results were developed. On a convex and L-smooth objective with unbiased stochastic gradients of bounded second moments, SGD with diminishing stepsizes $\gamma_t = \Theta(1/\sqrt{t})$ achieves the optimal sub-linear rate $\mathcal{O}(1/\sqrt{T})$ [30]. If the objective is additionally μ-strongly convex, choosing $\gamma_t = \Theta(1/t)$ tightens the rate to $\mathcal{O}(1/T)$ [29]. If smoothness is absent, in the convex setting, replacing gradients with subgradients while retaining $\gamma_t = \Theta(1/\sqrt{t})$ yields the classical $\mathcal{O}(1/\sqrt{T})$ bound [31, 41]. Finally, for non-smooth and μ-strongly convex objectives, a subgradient scheme with $\gamma_t = \Theta(1/t)$ attains the optimal $\mathcal{O}(1/T)$ convergence rate [29, 30].

We identify two main categories of optimization problems: the additive form $f(x) = g(x) + h(x)$, which leads to a composite optimization problem, and the inner–outer (nested) form $f(x) = \phi(\psi(x))$, which leads to a compositional optimization problem, each bringing its own algorithmic challenges and solution methods. Considering the additive composite formulation $f(x) = g(x) + h(x)$, suppose that g is smooth and L-Lipschitz differentiable, and h is proximable and potentially non-smooth. Proximal-gradient methods have been extensively studied and widely used for solving this type of problem. Notable methods include ISTA [9] and its accelerated variant FISTA [2]. Stochastic variants, such as FOBOS [12] and RDA [47], have also been explored. Additionally, accelerated stochastic proximal methods like Prox-SVRG and SAGA [10, 48] have gained popularity due to their improved convergence rates. In contrast to the well-studied additive composite form, research on stochastic compositional optimization, defined by its nested inner-outer structure, has begun to attract attention in recent years. An early contribution is due to Ermoliev [14], who proposed a two-timescale stochastic approximation method for optimizing the composition of two expectation-valued functions; see Section 6.7 of [15] for a brief related discussion. A more recent advance came from Wang et al. [43], who formally introduced the stochastic compositional gradient descent (SCGD) method to solve problems of the form $f(x) = \mathbb{E}[\phi(\mathbb{E}[\psi(x, \xi_1)], \xi_2)]$, where the outer expectation is taken with respect to the probability distribution of ξ_2, the inner one is taken with respect to the probability distribution of ξ_1, and the inner function may be non-smooth. The SCGD algorithm operates on two intertwined time-scales, employing an auxiliary sequence to estimate the inner expectation via an exponential moving average, while the outer iterate descends along a stochastic quasi-gradient formed from this estimation. For non-smooth objectives, SCGD achieves sample complexities of $\mathcal{O}(T^{-1/4})$ and $\mathcal{O}(T^{-2/3})$ in the convex and strongly convex cases, respectively. Compared to our method, SCGD primarily focuses on handling the non-smoothness of the inner function, whereas our approach specifically targets scenarios where the outer function is potentially non-smooth. Additionally, our method employs an iterative smoothing parameter scheme, ensuring controlled smoothing accuracy throughout the optimization process.

Building on SCGD, subsequent work further developed the nested type stochastic compositional optimization problems (and their variants). Wang et al. [24] integrated SVRG into SCGD for finite-sum compositional objectives $f(x) = \frac{1}{N}\sum_{i=1}^{N} \phi_i \left(\frac{1}{M} \sum_{j=1}^{M} \psi_j(x) \right)$ (where ϕ_i and ψ_j denote the values of ϕ and ψ for specific realizations of their respective random vectors) and proved a linear convergence rate under strong convexity. For smooth but non-convex stochastic compositional objectives $f(x) = \mathbb{E}[\phi(\mathbb{E}[\psi(x, \xi_1)], \xi_2)]$, advances have been made by [27, 50]. In addition, several works extend proximal-gradient methods and their variants [20, 44], employ ADMM [49], use SARAH [53], or adopt SAGA [52] to address stochastic compositional problems of the form $f(x) = \mathbb{E}[\phi(\mathbb{E}[\psi(x, \xi_1)], \xi_2)] + h(x)$, where both the inner and outer functions are smooth and a non-smooth regularization term h is incorporated.

1.2 Gradient descent using smoothing techniques

Besides the methods discussed in Section 1.1, another prominent approach for handling non-smooth objectives is the use of smoothing functions, which approximate a non-smooth function by a smooth surrogate that converges to the original function as the smoothing parameter diminishes. Nesterov proposes a smoothing framework [32] (infimal convolution smoothing) and demonstrates that structured non-smooth convex problems admit an $\mathcal{O}(1/\varepsilon)$ first-order complexity bound for obtaining an ε-accurate solution, improving upon the classical subgradient rate of $\mathcal{O}(1/\varepsilon^2)$. Alternative smoothing techniques, such as randomized smoothing [13] and Moreau envelope smoothing [28], have been extensively explored in the literature. In derivative-free settings, notable smoothing approaches include Gaussian smoothing [34] and direct search using smoothing functions [18].

For additive composite optimization problems of the form $f(x) = g(x) + h(x)$, where $h(x)$ may be non-smooth, various smoothing techniques have been explored in the literature. Randomized smoothing [13] was introduced to address this type of problem. Accelerated proximal gradient methods enhanced by Nesterov's smoothing technique have also been explored [25, 45]. Additionally, an approach using the Moreau envelope along with a variable sample-size accelerated proximal scheme to manage non-smoothness has been proposed in [21]. By contrast, to the best of our knowledge, existing studies on stochastic compositional problems with an inner–outer structure do not employ a smoothing gradient method that iteratively updates the smoothing parameter, as proposed in this paper.

1.3 Compositional case using smoothing functions—our contribution

The main goal of this paper is to develop a smoothing stochastic gradient (SSG) method using the gradients of smoothing functions, where the smoothing parameter is decreased at a controlled rate. Our motivation is drawn from compositional problems (1.1) and (1.2), but we will abstract from these two settings and develop the SSG method and its convergence theory for a general scenario (see formulation (3.1) in Section 3) for which there is an accurate smoothing function, and for which there is the possibility of drawing unbiased smoothing stochastic gradients satisfying a bounded second order moment dependent on the Lipschitz constant of the smoothing function. We will investigate the convergence of the SSG method for this general scenario under strongly convex, convex, and non-convex regimes (see Sections 3.2–3.3 and Appendix B). Notably, for convex objectives, the SSG method attains a $1/T^{1/4}$ convergence rate in terms of stochastic gradient evaluations. We also show in Sections 4 and 5 that widely used

frameworks such as compositional and finite-sum compositional problems (which are instances of problems (1.1) and (1.2), respectively) meet the required smoothing assumptions, including convexity, unbiased gradient estimates, bounded second moments, and controlled errors. The illustrative numerical experiments reported in Section 6 show that the SSG method can perform competitively for certain problem classes.

2 Introduction to smoothing functions

In this section, we will present two main examples and their relevant properties of smoothing functions for the context of our paper. We specifically address widely used non-smooth functions in optimization, namely the $L1$ and the Hinge losses. At the end of the section, we mention one way of constructing smoothing functions through convolution with mollifiers, yielding smoothed functions that satisfy the key properties we outlined.

The $L1$ norm is widely used in optimization, for instance in sparse optimization and compressed sensing (see [11, 42]). Due to its non-smooth nature, it is advantageous to replace it with a smooth approximation. One way to smooth $|\cdot|$ is introduced in [7]

$$\tilde{s}(t,\mu) \;=\; \int_{-\infty}^{+\infty} \big|t-\mu\tau\big|\,\rho(\tau)\,d\tau, \tag{2.1}$$

where $\rho:\mathbb{R}^n\to[0,+\infty)$ is a symmetric and piecewise continuous density function with a finite number of pieces, satisfying $\int_{-\infty}^{+\infty}|\tau|\,\rho(\tau)\,d\tau<+\infty$.

The function (2.1) holds several important properties. Firstly, $\tilde{s}(\cdot,\mu)$ is continuously differentiable on $\mathbb{R}$, with its derivative expressed as $\tilde{s}'(t,\mu)=2\int_0^{t/\mu}\rho(\tau)d\tau$. As μ approaches zero, the function converges uniformly to $|t|$ on $\mathbb{R}$, with the approximation error bounded by a multiple of the smoothing parameter μ. Moreover, this smoothing function satisfies the so-called gradient consistency, meaning that the limit points of $\tilde{s}'(t,\mu)$ approach the Clarke subdifferential of $|t|$, more specifically, it holds that $\{\lim_{t\to 0,\mu\downarrow 0}\tilde{s}'(t,\mu)\}=[-1,1]=\partial_{|\cdot|}(0)$. Finally, for any fixed $\mu>0$, the derivative $\tilde{s}'(t,\mu)$ is Lipschitz continuous with a constant that is proportional to $1/\mu$.

An explicit example of (2.1) with a uniform density

$$\rho(\tau) \;=\; \begin{cases} \frac{1}{2}, & \tau\in\left[-\frac{1}{2},\frac{1}{2}\right],\\ 0, & \text{otherwise,}\end{cases}$$

yields

$$\tilde{s}(t,\mu) \;=\; \begin{cases} \frac{t^2}{\mu}+\frac{\mu}{4}, & t\in\left[-\frac{\mu}{2},\frac{\mu}{2}\right],\\ |t|, & \text{otherwise.}\end{cases} \tag{2.2}$$

Besides the $L1$ loss, the Hinge loss is also broadly used in machine learning and optimization applications. Similarly to the $L1$ case, a particular smoothing function for the Hinge loss is

$$\tilde{h}(t,\mu) \;=\; \begin{cases} 1-t, & t\le 1-\mu,\\ \frac{1}{4\mu}(1-t+\mu)^2, & 1-\mu<t\le 1+\mu,\\ 0, & t>1+\mu.\end{cases} \tag{2.3}$$

This smoothing function holds the same properties as (2.2).

Generally speaking, there are various techniques to construct a smoothing function. For example, one can smooth the non-smooth function by convolution with mollifiers (see [39]). A parameterized family of measurable functions $\{\psi_\mu : \mathbb{R}^n \to [0, +\infty),\ \mu \in (0, \infty)\}$ is called a mollifier (or mollifier sequence) if it satisfies $\int_{\mathbb{R}^n} \psi_\mu(z)\, dz = 1$ and if $B_\mu = \{z : \psi_\mu(z) > 0\}$ is bounded and converges to $\{0\}$ as $\mu \downarrow 0$. A smoothing function $\tilde{f}$ for a non-smooth function f is then constructed via convolution with mollifiers as follows

$$\tilde{f}(x, \mu) = \int_{\mathbb{R}^n} f(x - z)\, \psi_\mu(z)\, dz = \int_{\mathbb{R}^n} f(z)\, \psi_\mu(x - z)\, dz. \tag{2.4}$$

This construction guarantees pointwise convergence from $\tilde{f}(x, \mu)$ to $f(x)$ as μ goes to 0 (thus satisfies (1.3) in Definition 1). If the mollifiers are continuous on $\mathbb{R}^n$, then $\tilde{f}(\cdot, \mu)$ is continuously differentiable. Note that when f is locally Lipschitz continuous at a limit point x_*, one always has $\partial f(x_*) \subseteq \operatorname{co} G_{\tilde{f}}(x_*)$ (see [6]), where $\partial f(x_*)$ is the subdifferential, co denotes the convex hull, and $G_{\tilde{f}}(x_*) = \{\lim_{x \to x_*, \mu \downarrow 0} \nabla \tilde{f}(x, \mu)\}$. Moreover, using construction (2.4), when f is Lipschitz continuous near x_*, the so-called gradient consistency property, which says that $\partial f(x_*) = \operatorname{co} G_{\tilde{f}}(x_*)$, is satisfied. Other than convolution with mollifiers, smoothing functions can also be constructed by convolution with probability density functions (see [39]).

3 The smoothing stochastic gradient method

3.1 Algorithm and assumptions

We will be stating and analyzing a smoothing stochastic gradient method to solve the expected risk problem written in the general form

$$\min_{x \in \mathbb{R}^n}\ f(x)\ \ (f \text{ involves } \xi,\ \psi \text{ and } \phi), \tag{3.1}$$

where again x is the vector of decision variable, ξ is a random vector, ψ is a differentiable function, and ϕ is a function which is not necessarily smooth. Problem (3.1) is meant to include cases (1.1) and (1.2) as two particular instances. We consider that a smoothing function $\tilde{f}(x, \mu)$ is available for $f(x)$ involving a smoothing function $\tilde{\phi}$ of ϕ and also including the randomness given by ξ. We also assume that one can draw gradient estimates $\nabla_x \tilde{f}(x, \xi, \mu)$ for the smoothing function $\tilde{f}(x, \mu)$. A smoothing stochastic gradient method can then be stated in Algorithm 1 and it only differs from the traditional stochastic gradient algorithm in the update of the smoothing parameter μ_k.

Algorithm 1 Smoothing stochastic gradient (**SSG**) method

Input: x_1, $\{\alpha_k\}$, and $\{\mu_k\}$.

For $k = 1, 2, \ldots$ **do**
  Step 1. Generate a realization ξ_k of the random variable ξ. Then compute $\nabla \tilde{f}_x(x_k, \xi_k, \mu_k)$.
  Step 2. Update iterate $x_{k+1} = x_k - \alpha_k \nabla \tilde{f}_x(x_k, \xi_k, \mu_k)$.
  Step 3. Update smoothing parameter $\mu_{k+1} \leq \mu_k$.
End do

In this section, we will analyze the convergence of the SSG algorithm in the convex and non-convex regimes. The strongly convex case has less applicability and is left to the appendix of this paper. We first introduce the general assumptions required for the convergence of the SSG algorithm in all three regimes.

Assumption 3.1 states the unbiasedness of the smoothing gradient $\nabla_x \tilde{f}(x, \xi, \mu)$, meaning that its expectation over random vector ξ coincides with $\nabla_x \tilde{f}(x, \mu)$. This requirement is typically mild and will be shown to hold under reasonable conditions in Sections 4 and 5.

Assumption 3.1 (Unbiasedness of the smoothing gradient) *For all $x \in \mathbb{R}^n$ and for all $\mu > 0$, the smoothing gradient $\nabla_x \tilde{f}(x, \xi, \mu)$ serves as an unbiased stochastic estimator of $\nabla_x \tilde{f}(x, \mu)$, meaning that*

$$\mathbb{E}_\xi[\nabla_x \tilde{f}(x, \xi, \mu)] \;=\; \nabla_x \tilde{f}(x, \mu).$$

Assumption 3.2 requires that the second moment of the smoothing gradient is bounded by a constant of the same order as $1/\mu^2$. This is again a mild requirement, and it will be shown to hold under suitable conditions for most smoothing functions in Sections 4 and 5.

Assumption 3.2 (Bounded second moment of the smoothing gradient) *For all $x \in \mathbb{R}^n$ and for all $\mu > 0$, the smoothing gradient $\nabla_x \tilde{f}(x, \xi, \mu)$ is bounded as follows*

$$\mathbb{E}_\xi[\|\nabla_x \tilde{f}(x, \xi, \mu)\|^2] \;\leq\; \frac{1}{\mu^2} G^2,$$

where $G > 0$ is a positive constant.

Lastly, Assumption 3.3 ensures that the smoothing function $\tilde{f}(x, \mu)$ closely approximates the true function $f(x)$, with their difference (accuracy) controlled by the smoothing parameter μ. Similarly as for Assumption 3.1 and Assumption 3.2, such a requirement is mild and will be shown to hold under appropriate conditions for most smoothing functions in Sections 4 and 5.

Assumption 3.3 (Accuracy of the smoothing function) *For all $x \in \mathbb{R}^n$ and for all $\mu > 0$, the difference between the true function and the smoothing function is bounded by some constant times the smoothing parameter in the following way*

$$|f(x) - \tilde{f}(x, \mu)| \;\leq\; C\mu, \tag{3.2}$$

for some $C > 0$.

By now, we have outlined the general assumptions required for the SSG algorithm. Next, we will present additional specific assumptions for the convex and non-convex cases, and then develop the corresponding convergence results.

3.2 Rate in the convex case

In this subsection, we examine the convex case. While this generally leads to milder convergence guarantees compared to the strongly convex setting, it applies to a wider range of problems. We begin by outlining the assumption specific to the convex case and then present the convergence analysis for the SSG algorithm under such a condition. Assumption 3.4 states the convexity of the smoothing function.

Assumption 3.4 (Convexity of the smoothing function) *For all $x \in \mathbb{R}^n$ and for all $\mu > 0$, the smoothing function $\tilde{f}(x, \mu)$ is convex.*

To establish the convergence results, we also need to consider an additional assumption regarding the boundness of the sequence of iterates.

Assumption 3.5 (Boundness of the sequence of iterates) *There exists a positive constant $M > 0$ such that $\|x_k - x_*\| \leq M$ for all k, where x_* is any minimizer of f.*

Building upon the convexity assumption and considering specific rates for both the smoothing parameter and the stepsize in the SSG method, we now present the convergence results in the convex setting, which provides a weaker convergence rate compared to the strongly convex case given in the appendix.

Theorem 3.1 (Convergence rate in the convex case) *Under Assumptions 3.1–3.5, suppose that the SSG algorithm runs with the smoothing parameter $\mu_k = k^{-a}$ and stepsize $\alpha_k = k^{-b}$, where $a > 0$ and $b > 0$. Define $f_{best}^T = \min_{k=\{0,1,\ldots,T\}} f(x_k)$. Then for any minimizer x_* of f, one has*

$$\mathbb{E}[f_{best}^T] - f(x_*) \;\leq\; \frac{2Cc_1}{T^a} + \frac{M^2}{T^{1-b}} + \frac{c_2 G^2}{2}\frac{1}{T^{-2a+b}}, \tag{3.3}$$

where c_1 and c_2 are some positive constants.

Moreover, one can show that the best convergence rate of (3.3) is of the order of $1/T^{1/4}$.

Proof. First, analogous to the proof for the strongly convex case, we apply the update rule of the SSG algorithm to obtain the following

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x_*\|^2 | x_k] \\
&= \mathbb{E}[\|x_k - \alpha_k \nabla_x \tilde{f}(x_k, \xi_k, \mu_k) - x_*\|^2 | x_k] \\
&= \mathbb{E}\|x_k - x_*\|^2 | x_k] - 2\alpha_k \mathbb{E}[\langle \nabla_x \tilde{f}(x_k, \xi_k, \mu_k), x_k - x_* \rangle | x_k] + \alpha_k^2 \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2 | x_k] \\
&= \|x_k - x_*\|^2 - 2\alpha_k \langle \nabla_x \tilde{f}(x_k, \mu_k), x_k - x_* \rangle + \alpha_k^2 \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2 | x_k].
\end{aligned} \tag{3.4}$$

Then, by the convexity assumption of $\tilde{f}(x, \mu)$ (Assumption 3.4), we have

$$\langle \nabla_x \tilde{f}(x_k, \mu_k), x_k - x_* \rangle \geq \tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k). \tag{3.5}$$

By combining (3.4) and (3.5), and taking the total expectation on both sides, we get

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \mathbb{E}[\|x_k - x_*\|^2] - 2\alpha_k \mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] + \alpha_k^2 \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2].$$

Using the boundedness assumption of the smoothing gradient estimates (Assumption 3.2), and dividing both sides by α_k, we obtain

$$2\mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] \leq \frac{\mathbb{E}[\|x_k - x_*\|^2] - \mathbb{E}[\|x_{k+1} - x_*\|^2]}{\alpha_k} + \frac{\alpha_k}{\mu_k^2} G^2.$$

Then, by summing the inequality from $k = 1$ to T for some $T > 0$ and using the boundedness assumption of the sequence of iterates (Assumption 3.5), we obtain

$$\begin{aligned}\sum_{k=1}^{T} 2\mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] &\leq \frac{1}{\alpha_1}\mathbb{E}[\|x_1 - x_*\|^2] + \sum_{k=2}^{T}(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}})\mathbb{E}[\|x_k - x_*\|^2] + \sum_{k=1}^{T}\frac{\alpha_k}{\mu_k^2}G^2 \\ &\leq \frac{2M^2}{\alpha_T} + \sum_{k=1}^{T}\frac{\alpha_k}{\mu_k^2}G^2.\end{aligned} \tag{3.6}$$

From inequalities (3.2) (for x_n and x_*) and (3.6), we have

$$\begin{aligned}&\sum_{k=1}^{T} 2\mathbb{E}[f(x_k) - f(x_*)] \\ &\leq -\sum_{k=1}^{T} 2\mathbb{E}[\tilde{f}(x_k, \mu_k) - f(x_k)] + \sum_{k=1}^{T} 2\mathbb{E}[\tilde{f}(x_*, \mu_k) - f(x_*)] + \frac{2M^2}{\alpha_T} + \sum_{k=1}^{T}\frac{\alpha_k}{\mu_k^2}G^2 \\ &\leq 4C\sum_{k=1}^{T}\mu_k + \frac{2M^2}{\alpha_T} + \sum_{k=1}^{T}\frac{\alpha_k}{\mu_k^2}G^2.\end{aligned} \tag{3.7}$$

Now let us first consider $\sum_{k=1}^{T}\mu_k = \sum_{k=1}^{T}\frac{1}{k^a}$, when $a \in (0, 1)$, one can easily show the following inequality hold

$$\sum_{k=1}^{T}\frac{1}{k^a} \leq 1 + \int_1^T k^{-a} \leq c_1 T^{1-a}, \tag{3.8}$$

where $c_1 > 0$ is some positive constant.

Similarly, consider $\sum_{k=1}^{T}\frac{\alpha_k}{\mu_k^2} = \sum_{k=1}^{T} k^{2a-b}$, when $b - 2a < 1$, one can also prove that it satisfies the following inequality

$$\sum_{k=1}^{T} k^{2a-b} \leq c_2 T^{2a-b+1}, \tag{3.9}$$

where $c_2 > 0$ is some positive constant.

Next, let us define $f_{\text{best}}^T = \min_{k=\{0,1,\ldots,T\}} f(x_k)$. Consider the left hand side of the inequality (3.7), we have

$$\sum_{k=1}^{T} 2\mathbb{E}[f(x_k) - f(x_*)] \geq 2T\mathbb{E}[f_{\text{best}}^T - f(x_*)]. \tag{3.10}$$

Thus, combining (3.7), (3.8), (3.9), and (3.10), one arrives at (3.3).

Finally, we will establish the optimal convergence bound achievable by the algorithm. Under the constraints $b - 2a < 1$ and $0 < a < 1$, we aim to maximize

$$t = \min_{a,b\in\mathbb{R}_+}\{a, 1 - b,\ -2a + b\},$$

One can easily show using an LP formulation that the maximum value of t is $1/4$ with the optimal solution $(a, b) = (1/4, 3/4)$. This result shows that the best bound the SSG algorithm could achieve for the expected risk problem (3.1) in the convex case is of the order of $1/T^{1/4}$. □

Before moving on to the non-convex case, we note that the best convergence rate of the SSG algorithm in the convex setting is of the order of $T^{1/4}$, which matches the rate for a different non-smooth convex stochastic composition problem (see [43]), where the non-smoothness lies instead in the inner function ψ. In Section 4 and Section 5, we will demonstrate how the non-smooth convex compositional problems (1.1) and (1.2) can be fitted in the SSG algorithm as instances of the general problem (3.1), thereby allowing the SSG method to achieve the developed convergence rate of the non-smooth convex compositional setting.

3.3 Rate in the non-convex case

In this subsection, we examine the non-convex setting of the SSG method. The lack of convexity in this context introduces additional analytical challenges compared to the strongly convex and convex cases. We begin by outlining the assumptions necessary for the non-convex scenario and subsequently develop the corresponding convergence results for the SSG algorithm under these assumptions.

Assumption 3.6 requires that the gradient of the smoothing function is Lipschitz continuous with a constant inversely proportional to the smoothing parameter μ. This inverse relationship indicates that decreasing μ, thereby reducing the degree of smoothing and allowing the function to more closely approximate the original non-smooth objective, results in an increase in the Lipschitz constant. Most smoothing functions (see Section 2) satisfy this property. We also require that the stepsize should be of a diminishing type, as shown in Assumption 3.7.

Assumption 3.6 *For all $x \in \mathbb{R}^n$ and for all $\mu > 0$, the gradient of the smoothing function $\tilde{f}(x, \mu)$ is Lipschitz continuous with a Lipschitz constant C_0/μ for some $C_0 > 0$.*

Assumption 3.7 *The sequence of diminishing stepsizes $\{\alpha_k\}$ satisfies $\sum_{k=1}^{\infty} \alpha_k = \infty$, and $\sum_{k=1}^{\infty} \alpha_k^d < \infty$ for some $d > 1$.*

After having listed all the required assumptions, we now establish a first convergence result for the non-convex setting on the gradient of the smoothing function. In what follows, we present the asymptotic convergence analysis, drawing inspiration from the proof in [19].

Theorem 3.2 (Convergence in the non-convex case) *Under Assumptions 3.1–3.3 and 3.6–3.7, for $d \in (1, 2)$, by choosing $\mu_k = \alpha_k^{-\frac{d}{3}+\frac{2}{3}}$, one has*

$$\lim_{T\to\infty} \mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|\nabla_x \tilde{f}(x_k, \mu_k)\|^2\right] < \infty. \tag{3.11}$$

Moreover, with $A_T \coloneqq \sum_{k=1}^{T} \alpha_k$, one has

$$\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{A_T}\sum_{k=1}^{T} \alpha_k \|\nabla_x \tilde{f}(x_k, \mu_k)\|^2\right] = 0. \tag{3.12}$$

Proof. From Assumption 3.6, one has

$$\tilde{f}(x_{k+1}) - \tilde{f}(x_k) \leq \langle \nabla_x \tilde{f}(x_k, \mu_k), x_{k+1} - x_k \rangle + \frac{C_0}{2\mu_k}\|x_{k+1} - x_k\|^2.$$

Recalling the update rule of the SSG algorithm $x_{k+1} = x_k - \alpha_k \nabla_x \tilde{f}(x_k, \xi_k, \mu_k)$, we have

$$\tilde{f}(x_{k+1}) - \tilde{f}(x_k) \leq -\alpha_k \langle \nabla_x \tilde{f}(x_k, \mu_k), \nabla_x \tilde{f}(x_k, \xi_k, \mu_k) \rangle + \frac{\alpha_k^2 C_0}{2\mu_k} \|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2. \tag{3.13}$$

Then, by taking conditional expectations with respect to the distribution of ξ_k of (3.13), we obtain

$$\mathbb{E}[\tilde{f}(x_{k+1})|\xi_k] - \tilde{f}(x_k) \leq -\alpha_k \|\nabla_x \tilde{f}(x_k, \mu_k)\|^2 + \frac{\alpha_k^2 C_0}{2\mu_k} \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2 | \xi_k]. \tag{3.14}$$

Now we take the total expectation of (3.14) and obtain

$$\mathbb{E}[\tilde{f}(x_{k+1})] - \mathbb{E}[\tilde{f}(x_k)] \leq -\alpha_k \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \mu_k)\|^2] + \frac{\alpha_k^2 C_0}{2\mu_k} \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2]. \tag{3.15}$$

Next, by summing the inequality (3.15) from $k = 1$ to T for some $T > 0$, we have

$$\mathbb{E}[\tilde{f}(x_{T+1})] - \mathbb{E}[\tilde{f}(x_1)] \leq -\sum_{k=1}^{T} \alpha_k \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \mu_k)\|^2] + \frac{C_0}{2} \sum_{k=1}^{T} \frac{\alpha_k^2}{\mu_k} \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2]. \tag{3.16}$$

By defining $\tilde{f}_{\text{best}}^T = \min_{k=\{0,1,\ldots,T\}} \tilde{f}(x_k)$ and applying Assumption 3.2, we obtain

$$\sum_{k=1}^{T} \alpha_k \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \mu_k)\|^2] \leq \mathbb{E}[\tilde{f}(x_1)] - \tilde{f}_{\text{best}}^T + \frac{C_0 G^2}{2} \sum_{k=1}^{T} \frac{\alpha_k^2}{\mu_k^3}. \tag{3.17}$$

By choosing $\mu_k = \alpha_k^{-\frac{d}{3}+\frac{2}{3}}$ for some $d \in (1, 2)$, the diminishing stepsize assumption (Assumption 3.7) ensures that the right-hand side of (3.17) converges to a finite limit as $T \to \infty$. Since the left-hand side is a nondecreasing sum (all terms are nonnegative), we prove that the theorem's first statement (3.11) holds.

Next, we turn to prove the second statement of the theorem. We divide both sides of (3.17) by $A_T = \sum_{k=1}^{T} \alpha_k$ and get

$$\frac{\sum_{k=1}^{T} \alpha_k \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \mu_k)\|^2]}{A_T} \leq \frac{\mathbb{E}[\tilde{f}(x_1)] - \tilde{f}_{\text{best}}^T}{A_T} + \frac{\frac{C_0 G^2}{2} \sum_{k=1}^{T} \frac{\alpha_k^2}{\mu_k^3}}{A_T}.$$

Taking the limit as $T \to \infty$, using (3.11), recalling $\mu_k = \alpha_k^{-\frac{d}{3}+\frac{2}{3}}$, and using the assumption of the stepsize (Assumption 3.7), we prove that the theorem's second statement (3.12) holds. □

Theorem 3.2 provides the convergence of the SSG method for gradients of the smoothing function $\tilde{f}$. We want to further investigate the convergence for subgradients of the true nonsmooth function f. For this purpose, we impose Assumption 3.8 which guarantees the existence of a nearby true subgradient within a distance of the order of μ to any smoothing gradient.

Assumption 3.8 *For all $x \in \mathbb{R}^n$, and for all $\mu > 0$, there exists a $y \in \mathcal{B}(x, E_1\mu)$ and a subgradient $g(y) \in \partial f(y)$ such that*

$$\|\nabla_x \tilde{f}(x, \mu) - g(y)\| \ \leq \ E_2 \mu, \tag{3.18}$$

where E_1 and E_2 are some positive constants.

Note that when f is locally Lipschitz continuous at a limit point x_*, Assumption 3.8 implies the gradient consistency property (see details in Section 2). In fact, consider any sequence $\{(x_k, \mu_k)\}$ satisfying $x_k \to x_*$ and $\mu_k \downarrow 0$, and let y_k be the corresponding points that satisfy Assumption 3.8. It is easy to show that $y_k \to x_*$ as $\mu_k \downarrow 0$. Since f is locally Lipschitz continuous at x_*, its Clarke subdifferential is outer semicontinuous at x_* (see [39, Theorem 5.19]), which leads to $g(y_k) \to g(x_*) \in \partial f(x_*)$. Thus, taking limits in $\|\nabla_x f(x_k, \mu_k) - g(y_k)\| \leq E_2 \mu_k$ yields

$$\lim_{x \to x_*,\, \mu \downarrow 0} \nabla_x \tilde{f}(x, \mu) = g(x_*) \in \partial f(x_*). \tag{3.19}$$

Using (3.19) and the fact that the Clarke subdifferential $\partial f(x_*)$ is always a convex set, we have $\operatorname{co} G_{\tilde{f}}(x_*) = \operatorname{co}\{\lim_{x \to x_*,\, \mu \downarrow 0} \nabla_x \tilde{f}(x, \mu)\} \subseteq \partial f(x_*)$. Since one always has $\partial f(x_*) \subseteq \operatorname{co} G_{\tilde{f}}(x_*)$ for a locally Lipschitz continuous f (see in Section 2), we have thus showed that Assumption 3.8 implies the gradient consistency property.

Corollary 3.1 (Convergence in the non-convex case) *Under Assumptions 3.1–3.3 and 3.6–3.8, for $d \in (1, 2)$, by choosing $\mu_k = \alpha_k^{-\frac{d}{3}+\frac{2}{3}}$, one has*

$$\lim_{T \to \infty} \mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|g(y_k)\|^2\right] \;<\; \infty. \tag{3.20}$$

Moreover, with $A_T \coloneqq \sum_{k=1}^{T} \alpha_k$, one has

$$\lim_{T \to \infty} \mathbb{E}\left[\frac{1}{A_T} \sum_{k=1}^{T} \alpha_k \|g(y_k)\|^2\right] \;=\; 0, \tag{3.21}$$

where $g(y_k)$ is a subgradient satisfying (3.18).

Proof. For any $y_k \in \mathcal{B}(x_k, E_1 \mu)$ such that the subgradient $g(y_k)$ satisfies Assumption 3.8, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|g(y_k)\|^2\right] &= \mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|g(y_k) - \nabla_x \tilde{f}(x_k, \mu_k) + \nabla_x \tilde{f}(x_k, \mu_k)\|^2\right] \\
&\leq 2\mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|g(y_k)) - \nabla_x \tilde{f}(x_k, \mu_k)\|^2\right] + 2\mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|\nabla_x \tilde{f}(x_k, \mu_k)\|^2\right] \\
&\leq 2E_2^2 \sum_{k=1}^{T} \alpha_k \mu_k^2 + 2\mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|\nabla_x \tilde{f}(x_k, \mu_k)\|^2\right].
\end{aligned} \tag{3.22}$$

Given our choice of $\mu_k = \alpha_k^{-\frac{d}{3}+\frac{2}{3}}$ with $d \in (1, 2)$, we observe that $\lim_{T \to \infty} E_2^2 \sum_{k=1}^{T} \alpha_k \mu_k^2 < \infty$. Combining $\lim_{T \to \infty} E_2^2 \sum_{k=1}^{T} \alpha_k \mu_k^2 < \infty$ with (3.11), we know that the right-hand side of (3.22) is upper bounded when $T \to \infty$. Since $\sum_{k=1}^{T} \alpha_k \|g(y_k)\|^2$ is a nonnegative-term sum, one can conclude that (3.20) holds.

Next, we prove the second statement of the theorem. Notice that (3.12) from Theorem 3.2 also holds here. Consider (3.22) again and divide its both sides by $A_T = \sum_{k=1}^{T} \alpha_k$, obtaining

$$\frac{\mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|g(y_k)\|^2\right]}{A_T} \leq \frac{2E_2^2 \sum_{k=1}^{T} \alpha_k \mu_k^2}{A_T} + \frac{2\mathbb{E}\left[\sum_{k=1}^{T} \alpha_k \|\nabla_x \tilde{f}(x_k, \mu_k)\|^2\right]}{A_T}. \tag{3.23}$$

When $d \in (1,2)$, we use the finiteness of $\lim_{T\to\infty} E_2^2 \sum_{k=1}^{T} \alpha_k \mu_k^2$ together with (3.12) to establish the second statement (3.21) of the theorem.

□

4 Case $f(x) = \phi(\mathbb{E}[\psi(x,\xi)])$

An objective function of the type (3.1) takes a compositional form in many real-world applications, such as risk-averse optimization. An outer function $\phi(\cdot)$ is composed with an inner function $\mathbb{E}[\psi(\cdot,\xi)]$, leading to the problem formulation (1.1). In this case, the smoothing function of f (with $\psi(x) = \mathbb{E}[\psi(x,\xi)]$) is

$$\tilde{f}(x,\mu) \;=\; \tilde{\phi}(\psi(x),\mu).$$

And the gradient of $\nabla_x \tilde{f}(x,\mu)$ is

$$\nabla_x \tilde{f}(x,\mu) \;=\; \tilde{\phi}'(\psi(x),\mu)\nabla\psi(x).$$

To compute the smoothing gradient as an unbiased estimator of $\nabla_x \tilde{f}(x,\mu)$ (to satisfy Assumption 3.1), we independently sample two batches ξ^1 and ξ^2 from the random vector ξ and then compute $\psi(x_k,\xi_k^1)$ and $\nabla_x\psi(x_k,\xi_k^2)$ separately. By doing so, the smoothing gradient estimator is written as follows

$$\nabla_x \tilde{f}(x,\xi^1,\xi^2,\mu) \;=\; \tilde{\phi}'(\psi(x,\xi^1),\mu)\nabla_x\psi(x,\xi^2), \tag{4.1}$$

where $\tilde{\phi}'(\cdot,\mu)$ denotes the derivative of the smoothing function $\tilde{\phi}$ evaluated at $\psi(x,\xi^1)$. We will show in Proposition 4.2 that $\nabla_x \tilde{f}(x,\xi^1,\xi^2,\mu)$ is indeed an unbiased estimator of $\nabla_x \tilde{f}(x,\mu)$.

To deal with this compositional setting, we present a particular SSG method in Algorithm 2. We assume that the sequences $\{\alpha_k\}$ and $\{\mu_k\}$ are both decreasing and converging to zero. An initial point x_1, along with a sequence of stepsize $\{\alpha_k\}$, and a sequence of smoothing parameter $\{\mu_k\}$ are required as input. At each iteration, the algorithm samples realizations ξ_k^1 and ξ_k^2 of the random variable ξ independently and then computes $\psi(x_k,\xi_k^1)$ and $\nabla_x\psi(x_k,\xi_k^2)$ separately. Next, the algorithm compute smoothing gradient $\nabla_x \tilde{f}(x,\xi^1,\xi^2,\mu)$ as in (4.1). This smoothing gradient $\nabla_x \tilde{f}(x,\xi^1,\xi^2,\mu)$ is used in a standard stochastic gradient descent update procedure, scaled by the stepsize α_k. Lastly, the SSG algorithm updates the smoothing parameter μ at the end of each iteration.

Algorithm 2 SSG method for case $f(x) = \phi(\mathbb{E}[\psi(x,\xi)])$

Input: x_1, $\{\alpha_k\}$, and $\{\mu_k\}$.

For $k = 1, 2, \ldots$ **do**

Step 1. Generate realizations ξ_k^1 and ξ_k^2 of the random variable ξ. Then compute $\psi(x_k,\xi_k^1)$ and $\nabla_x\psi(x_k,\xi_k^2)$, respectively.

Step 2. Compute $\nabla_x \tilde{f}(x_k,\xi_k^1,\xi_k^2,\mu_k) = \tilde{\phi}'(\psi(x_k,\xi_k^1),\mu_k)\,\nabla_x\psi(x_k,\xi_k^2)$.

Step 3. Update iterate $x_{k+1} = x_k - \alpha_k \nabla_x \tilde{f}(x_k,\xi_k^1,\xi_k^2,\mu_k)$.

Step 4. Update smoothing parameter $\mu_{k+1} \le \mu_k$.

End do

4.1 About convexity

In this section, we will investigate under what properties of ϕ and ψ the assumptions needed in Section 3 are met.

In many real-world optimization problems that are under uncertainty, the non-smoothness of the outer function ϕ arises from max-operations and absolute values. For example, problems that we are looking at could be like $\max\{0, \mathbb{E}[\psi(x,\xi)]\}$ or $|\mathbb{E}[\psi(x,\xi)]|$. Consider a finance scenario as an instance, the inner function ψ computes the profit for a given market realization and the expectation $\mathbb{E}[\psi(\cdot,\xi)]$ averages this profit over all possible market conditions.

To smooth the outer function ϕ with a hinge type function $\max\{0,\cdot\}$, we can use smoothing functions provided in Section 2. In such a case, $\tilde{\phi}(t,\mu)$ is non-decreasing and convex with respect to t. Then if one has that $\psi(x) = \mathbb{E}[\psi(x,\xi)]$ is convex, the convexity assumption of the compositional smoothing function $\tilde{f}(x,\mu) = \tilde{\phi}(\psi(x),\mu)$ is satisfied. A similar property holds for the smoothing of $|\cdot|$ if $\psi(x)$ is always evaluated with nonnegative values.

4.2 About stochasticity

Next, we will show that the bounded second moment assumption given in Section 3 (Assumption 3.2) can be met in this compositional case. For this purpose, we first require that the second moment of the gradient of the inner function be bounded in expectation. Such an assumption is typically met in machine learning models (for example, linear regression and logistic regression) with bounded feasible regions $\mathcal{X} \subset \mathbb{R}^n$.

Assumption 4.1 *For all $x \in \mathcal{X}$, the gradient of the inner function $\psi(x,\xi)$ has a bounded second moment*

$$\mathbb{E}[\|\nabla\psi(x,\xi)\|^2] \;\leq\; G_1^2,$$

for some $G_1 > 0$.

We also need to assume that the expected squared derivative of $\tilde{\phi}$ does not exceed a positive quantity that inversely depends on the smoothing parameter μ.

Assumption 4.2 *For all $x \in \mathcal{X}$ and for all $\mu > 0$, the gradient of $\tilde{\phi}(\psi(x,\xi),\mu)$ satisfies*

$$\mathbb{E}[\tilde{\phi}'(\psi(x,\xi),\mu)^2] \;\leq\; \left(\frac{G_2}{\mu}\right)^2,$$

for some $G_2 > 0$.

We note that Assumption 4.2 is satisfied when using smoothing functions ϕ of the types shown in Section 2 (see (2.2) and (2.3)), which correspond to functions of type $|\cdot|$ and $\max\{0,\cdot\}$. In such cases, $\phi'(\psi(x,\xi),\mu)$ is a multiple of $1/\mu$ of $\psi(x,\xi)$, and Assumption 4.2 is met as long as $\psi(x,\xi)$ is bounded. Hence, Assumption 4.2 is a combination of the boundness of $\psi(x,\xi)$ and the properties of the smoothing function $\tilde{\phi}(\cdot,\mu)$.

Using these two assumptions, we can ensure the bounded second moment of the smoothing gradient and thus confirm that Assumption 3.2 is valid in this compositional setting.

Proposition 4.1 *Suppose Assumptions 4.1 and 4.2 hold. Additionally, assume that $\psi(x,\xi^1)$ is sampled independently from $\nabla_x\psi(x,\xi^2)$. Then, the smoothing gradient satisfies*

$$\mathbb{E}[\|\tilde{\phi}'(\psi(x,\xi^1),\mu)\nabla_x\psi(x,\xi^2)\|^2] \;\leq\; \left(\frac{G_1G_2}{\mu}\right)^2.$$

Proof. By the definition of the smoothing gradient estimate (4.1) and the independence of ξ^1 and ξ^2, we have the following equalities

$$\begin{aligned}\mathbb{E}[\|\tilde{\phi}'(\psi(x,\xi^1),\mu)\nabla_x\psi(x,\xi^2)\|^2] &= \mathbb{E}[|\tilde{\phi}'(\psi(x,\xi^1),\mu)|^2\|\nabla_x\psi(x,\xi^2)\|^2] \\ &= \mathbb{E}_{\xi^1}[|\tilde{\phi}'(\psi(x,\xi^1),\mu)|^2]\mathbb{E}_{\xi^2}[\|\nabla_x\psi(x,\xi^2)\|^2].\end{aligned}$$

The final bound can be determined from Assumptions 4.1 and 4.2. □

Having established the bounded second moment of the smoothing gradient, we now proceed to demonstrate the unbiasedness of the smoothing gradient estimator using the dominated convergence theorem [16, Theorem 1.19] (to confirm that Assumption 3.1 can also be satisfied in this setting).

Proposition 4.2 *Suppose Assumptions 4.1 and 4.2 hold. Assuming the unbiasedness of $\psi(x,\xi^1)$ and $\nabla\psi(x,\xi^2)$, the smoothing gradient $\nabla_x\tilde{f}(x,\xi^1,\xi^2,\mu) = \tilde{\phi}'(\psi(x,\xi^1),\mu)\nabla_x\psi(x,\xi^2)$ serves as an unbiased stochastic estimator of $\nabla_x\tilde{f}(x,\mu)$, meaning that*

$$\nabla_x\tilde{f}(x,\mu) \;=\; \mathbb{E}[\tilde{\phi}'(\psi(x,\xi^1),\mu)\nabla_x\psi(x,\xi^2)].$$

Proof. Again, since we sample ξ^1 and ξ^2 independently, we have

$$\mathbb{E}[\tilde{\phi}'(\psi(x,\xi^1),\mu)\nabla_x\psi(x,\xi^2)] = \mathbb{E}_{\xi^1}[\tilde{\phi}'(\psi(x,\xi^1),\mu)]\mathbb{E}_{\xi^2}[\nabla_x\psi(x,\xi^2)] \tag{4.2}$$

Then using the dominated convergence theorem and the unbiasedness of $\psi(x,\xi^1)$ and $\nabla\psi(x,\xi^2)$ on (4.2) completes the proof. □

4.3 About smoothing

Lastly, to fit the convergence theory that we discussed in Section 3.2 into this compositional case, we also need to make sure that $\tilde{f}$ is a smoothing function with an accuracy of the order of μ. This is guaranteed in Proposition 4.3 as long as the smoothing function $\tilde{\phi}$ satisfies the same property (Assumption 4.3).

Assumption 4.3 *For all $y \in \mathbb{R}$ and for all $\mu > 0$, the difference between $\phi(y)$ and $\tilde{\phi}(y,\mu)$ is bounded by some constant times the smoothing parameter in the following way*

$$|\phi(y) - \tilde{\phi}(y,\mu)| \;\leq\; C\mu,$$

where $C > 0$ is some positive constant.

Proposition 4.3 *Under Assumption 4.3, for all $\mu > 0$, the smoothing function $\tilde{f}(x,\mu)$ satisfies Assumption 3.3.*

Proof. By Assumption 4.3, the smoothing accuracy assumption of $\tilde{f}(x,\mu)$ is naturally satisfied as follows

$$|f(x) - \tilde{f}(x,\mu)| = |\phi(\psi(x)) - \tilde{\phi}(\psi(x),\mu)| \leq C\mu.$$

□

Notice that as a consequence of Proposition 4.3, one can infer that $\tilde{f}(x,\mu)$ is indeed a smoothing function for $f(x)$ (see Definition 1).

Building upon established assumptions and propositions of this section, we have demonstrated that the SSG method and its convergence results hold in a convex compositional setting of type $f(x) = \phi(\mathbb{E}[\psi(x, \xi)])$. Specifically, Proposition 4.1 and 4.2 ensure the bounded second moment and unbiasedness of the smoothing gradient estimates, and Proposition 4.3 ensures the accuracy of the smoothing function. Consequently, our analysis paves the way for practical application in diverse optimization problems characterized by such a convex compositional structure.

5 Finite-sum compositional case

In this section, we will focus on case $f(x) = \mathbb{E}[\phi(\psi(x, \xi))]$. In particular, we are interested in a finite-sum compositional structure because of its wide applicability. In many real-world applications, particularly in machine learning and large-scale optimization, the objective function naturally takes a finite-sum compositional structure. The finite-sum compositional objective function is defined in the following way

$$f(x) \;=\; \frac{1}{N} \sum_{i=1}^{N} \phi(\psi(x, \xi_i)), \tag{5.1}$$

where $N \in \mathbb{N}_+$ denotes the number of components and ξ_i represents the i-th data sample. Note that this finite-sum compositional function (5.1) approximate function $f(x) = \mathbb{E}[\phi(\psi(x, \xi))]$ using finite samples.

We then define the smoothing function $\tilde{f}$ of the finite-sum compositional function f as follows

$$\tilde{f}(x, \mu) \;=\; \frac{1}{N} \sum_{i=1}^{N} \tilde{\phi}(\psi(x, \xi_i), \mu),$$

where $\tilde{\phi}(\psi(x, \xi_i), \mu)$ serves as a smoothing function of $\phi(\psi(x, \xi_i))$ for each $i \in \{1, 2, \ldots, N\}$. To compute the stochastic gradient in this finite-sum setting, we use a stochastic batch gradient approach

$$\tilde{g}(x_k, \mu_k) \;=\; \frac{1}{|B_k|} \sum_{i \in B_k} \tilde{\phi}'(\psi(x_k, \xi_i), \mu_k) \nabla_x \psi(x_k, \xi_i), \tag{5.2}$$

where B_k denotes the sample batches. As a result, the update rule of the SSG method for the finite-sum compositional scenario is formulated as $x_{k+1} = x_k - \alpha_k \tilde{g}(x_k, \mu_k)$.

After pointing out the modifications of the SSG method for the finite-sum compositional setting, we now present the corresponding particular method in Algorithm 3.

Algorithm 3 SSG method for the finite-sum compositional case

Input: x_1, $\{\alpha_k\}$, and $\{\mu_k\}$.

For $k = 1, 2, \ldots$ **do**

Step 1. Select a batch B_k from $\{1, 2, \ldots, N\}$. For all $i \in B_k$, compute $\psi(x_k, \xi_i)$ and $\nabla_x \psi(x_k, \xi_i)$.

Step 2. Use (5.2) to compute the stochastic batch smoothing gradient $\tilde{g}(x_k, \mu_k)$.

Step 3. Update iterate $x_{k+1} = x_k - \alpha_k \tilde{g}(x_k, \mu_k)$.

Step 4. Update smoothing parameter $\mu_{k+1} \leq \mu_k$.

End do

5.1 About convexity

In the context of machine learning, the outer function $\phi(\cdot)$ often represents a loss function that measures the discrepancies between predictions and observed data. Many commonly used loss functions, such as $L1$ loss, hinge loss, and quantile loss, are convex but non-smooth. In these cases, by selecting an appropriate smoothing function $\tilde{\phi}(\cdot, \mu)$, it is possible to reformulate the compositional problem to fit the convex framework discussed in Section 3.

For example, consider the setting where the outer function is the hinge loss. Let $\xi = (u, v)$ represent a pair of features and labels. The inner function is a linear function defined as $\psi(x, \xi) = x^\top u + b - v$. Suppose we have N realizations of $\xi = (u, v)$ denoted by $\xi_i = (u_i, v_i)$, where $i \in \{1, \ldots, N\}$. In such a setting, the problem can be expressed as

$$f(x) \;=\; \frac{1}{N}\sum_{i=1}^{N} \phi(\psi(x, \xi_i)) \;=\; \frac{1}{N}\sum_{i=1}^{N} \max\{0, x^\top u_i + b - v_i\}.$$

A popular smoothing function for $\max\{0, \cdot\}$ was described in Section 2 (see (2.3)). In such a case, $\tilde{\phi}(t, \mu)$ is nondecreasing with respect to nonnegative t. If the inner function ψ is convex, then for any realization ξ_i of ξ, $\tilde{\phi}(\psi(x, \xi_i), \mu)$ is convex. Consequently, the sum of convex functions in $\tilde{f}(x, \mu) = \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(\psi(x, \xi_i), \mu)$ is convex and Assumption 3.4 is guaranteed.

Another example is the $L1$ loss, where we have

$$f(x) \;=\; \frac{1}{N}\sum_{i=1}^{N} \phi(\psi(x, \xi_i)) \;=\; \frac{1}{N}\sum_{i=1}^{N} |x^\top u_i + b - v_i|.$$

By using the smoothing function described in Section 2 (see (2.2)), if $\psi(x, \xi_i)$ is always evaluated with nonnegative values, we can also obtain the convexity of $\tilde{\phi}(\psi(x, \xi_i), \mu)$, therefore also meet the convexity of $\tilde{f}(x, \mu)$ required in Assumption 3.4.

5.2 About stochasticity

In this subsection, we introduce the specific assumptions and propositions relevant to the finite-sum compositional case, thereby showing that this case can be fitted into the convergence result established in Section 3.2. Assumption 5.1 requires that the mini-batch is sampled uniformly at random.

Assumption 5.1 *The mini-batch B is sampled uniformly at random from $\{1, 2, \ldots, N\}$, with each index i equally likely to be selected without replacement.*

This standard uniform sampling mechanism is crucial for maintaining the unbiasedness of the stochastic gradient estimates derived from the mini-batch. Thus, under Assumption 5.1, we have the following proposition.

Proposition 5.1 *Under Assumption 5.1, the smoothing gradient $\tilde{g}(x, \mu)$ given in* (5.2) *serves as an unbiased stochastic estimator of $\nabla_x \tilde{f}(x, \mu)$, satisfying*

$$\nabla_x \tilde{f}(x, \mu) \;=\; \mathbb{E}[\tilde{g}(x, \mu)].$$

Proof. By the chain rule, we can express $\nabla_x \tilde{f}(x, \mu)$ as

$$\nabla_x \tilde{f}(x, \mu) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\phi}'(\psi(x, \xi_i), \mu) \nabla_x \psi(x, \xi_i).$$

The stochastic batch gradient estimator is given by (5.2). By taking the expectation over the random sampling of B_k and utilizing Assumption 5.1, which guarantees a uniform selection probability for each index i, the unbiasedness directly follows from standard results in stochastic gradient methods (see, e.g., [3]). □

Assumption 5.2 ensures that the gradient of every inner function $\psi(x, \xi_i)$ has a bounded second moment.

Assumption 5.2 *For all $x \in \mathcal{X}$, the gradient of $\psi(x, \xi_i)$ is bounded for each $i \in \{1, 2, \ldots, N\}$ by*

$$\mathbb{E}[\|\nabla \psi(x, \xi_i)\|^2] \;\leq\; G_1^2,$$

for some $G_1 > 0$.

Similar to Section 4, Assumption 5.3 guarantees that, given the boundedness of the decision variables, the expected squared derivative of the smoothing function $\tilde{\phi}$ is controlled by a number that depends inversely on the smoothing parameter μ.

Assumption 5.3 *For all $\mu > 0$, the gradient of $\tilde{\phi}(\psi(x, \xi_i), \mu)$ satisfies*

$$E[\tilde{\phi}'(\psi(x, \xi_i), \mu)^2] \;\leq\; \left(\frac{G_2}{\mu}\right)^2,$$

for some $G_2 > 0$.

Proposition 5.2 states that the second moment of the smoothing stochastic batch gradient is bounded under the assumptions listed above.

Proposition 5.2 *Suppose Assumption 5.1–5.3 hold. Additionally, assume that $\psi(x, \xi_i)$ is sampled independently from $\nabla_x \psi(x, \xi_i)$ for each i. Then, the smoothing gradient estimate satisfies the following condition*

$$E[\|\tilde{g}(x, \mu)\|^2] \;\leq\; \left(\frac{G_1 G_2}{\mu}\right)^2. \tag{5.3}$$

Proof. The stochastic gradient estimate is given by

$$\tilde{g}(x,\mu) = \frac{1}{|B|}\sum_{i\in B} g_i, \quad \text{where} \quad g_i = \widetilde{\phi}'(\psi(x,\xi_i),\mu)\nabla_x\psi(x,\xi_i),$$

from which we have

$$\mathbb{E}\left[\|\tilde{g}(x,\mu)\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{|B|}\sum_{i\in B} g_i\right\|^2\right] \leq \frac{1}{|B|^2}|B|\sum_{i\in B}\mathbb{E}\left[\|g_i\|^2\right]. \tag{5.4}$$

From Assumption 5.2 and 5.3, each g_i satisfies

$$\mathbb{E}\left[\|g_i\|^2\right] \leq \mathbb{E}\left[\left(\widetilde{\phi}'(\psi(x,\xi_i),\mu)\right)^2\|\nabla_x\psi(x,\xi_i)\|^2\right] \leq \left(\frac{G_1G_2}{\mu}\right)^2.$$

Substituting the above bound into (5.4), we conclude (5.3). □

5.3 About smoothing

Lastly, to fit the convergence theory that we discussed in the previous section into this case, we need to consider Assumption 4.3 on the bound of the difference between each function $\phi(\cdot)$ and each smoothing function $\tilde{\phi}(\cdot,\mu)$. The final result concerns the approximation accuracy of the smoothing function.

Proposition 5.3 *Under Assumption 4.3, we have*

$$|f(x) - \tilde{f}(x,\mu)| \leq C\mu.$$

Proof. Using Assumption 4.3, it is easy to show that

$$\begin{aligned}
|f(x) - \tilde{f}(x,\mu)| &= \left|\frac{1}{N}\sum_{i=1}^{N}\left(\phi(\psi(x,\xi_i)) - \tilde{\phi}(\psi(x,\xi_i),\mu)\right)\right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left|\phi(\psi(x,\xi_i)) - \tilde{\phi}(\psi(x,\xi_i),\mu)\right| \\
&\leq C\mu.
\end{aligned}$$

□

Using the assumptions and propositions outlined in this section, we can fit the finite-sum compositional case into our convergence results presented in Section 3.2.

6 Numerical results

In this section, we are going to report some preliminary numerical results on the performance of the SSG method for the finite-sum compositional case using Algorithm 3. We will use a hinge loss as the outer function and a linear model as the inner function (see Section 5.1 for the detailed problem formulation). The SSG is run using the smoothing version of hinge given in (2.3), and

it is compared against a standard smooth SGD method ignoring the non-differentiability of the hinge-loss at the non-smooth point.

We test a dataset called Breast Cancer Wisconsin [46], which is widely used for benchmarking binary classification in medical diagnostics. It comprises $N = 569$ fine-needle aspirate samples of breast masses, each described by 30 real-valued features derived from digitized images of cell nuclei. The target label indicates the clinical diagnosis—malignant (0) or benignant (1). We map the original labels $\{0, 1\}$ to $\{-1, +1\}$ for hinge–loss training. The dataset is randomly split into an 80% train set and a 20% test set. For both methods, we are using decaying stepsizes and, for the SSG method, we are also using a decaying smoothing parameter. The decaying rules are $\alpha_k = \alpha_0 k^{-3/4}$ and $\mu_k = \mu_0 k^{-1/4}$ (which are the choices for the SSG method to achieve the best convergence rate in the convex case as discussed in Theorem 3.1). We performed a grid search (α_0 and $\mu_0 \in \{1, 5, 10, 15, 20, 30, 50, 75, 100, 150, 200\}$) to determine the best initial $\alpha_0 = 50$ (for both methods) and $\mu_0 = 15$. We run $T = 50{,}000$ epochs with mini-batches of size 128.

Figure 1 plots the training and test loss curves for the standard SGD and the SSG method on the Breast Cancer Wisconsin dataset. In the first few hundred epochs, the SSG curves descend more steeply than the SGD curves (both on training and test). This accelerated descent can be attributed to the smoothing of the hinge loss which yields μ-related Lipschitz-continuous derivatives instead of discontinuous gradients of the standard hinge loss. The smoothing function eliminates erratic jumps of gradient updates near the non-smooth point, reduces variance in mini-batch gradients, and allows for more consistent descent direction, thus leading to faster convergence. Moreover, by the end of training, the SSG method achieves lower losses on both the training and the test sets.

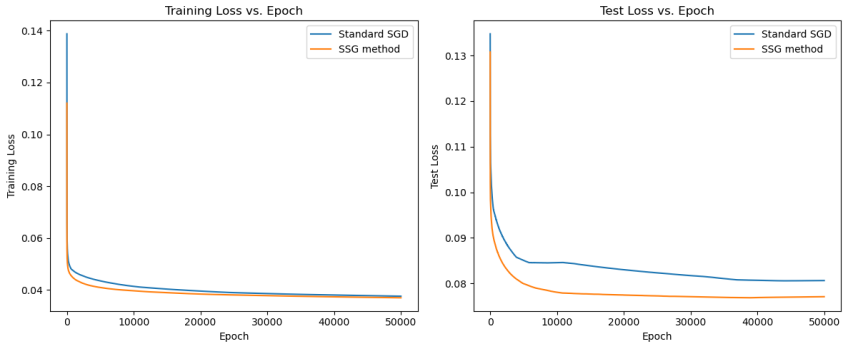

Figure 1: Training and test loss curves for standard SGD vs. SSG method

7 Conclusions and future works

This paper presented a smoothing-based stochastic gradient method addressing compositional optimization problems with a non-smooth outer function and a smooth inner function. The SSG method smooths the non-smooth component and progressively decreases the smoothing parameter. We highlight the role of the smoothing parameter, noting that iterative decreases in this parameter allow the smoothing function to more closely approximate the original non-smooth function at each iteration. Additionally, this controlled smoothing scheme prevents erratic jumps

in gradient updates near non-smooth points, reduces variance in gradient estimates, and provides more consistent descent directions. Consequently, the SSG method can converge fast without significantly compromising accuracy in approximating the non-smooth problem. We established convergence guarantees for SSG under convex, non-convex, and strongly convex settings. In the convex case, SSG matches the best-known convergence rate $\mathcal{O}(1/T^{1/4})$ for comparable compositional problems in which the inner function may be non-smooth. In the strongly convex case, the rate of SSG is arbitrarily close to $\mathcal{O}(1/T^{1/2})$. To achieve optimal convergence rates, the smoothing parameter decreases more rapidly in the strongly convex setting with $\mu_k = k^{-1/2}$, than in the convex setting with $\mu_k = k^{-1/4}$. This illustrates how stronger curvature conditions permit more aggressive reductions in the smoothing parameter, enabling faster smooth approximations. Similarly, the stepsize α_k must also decay more quickly ($\alpha_k = k^{-3/4}$ in the convex case and approximately $\alpha_k = k^{-3/2}$ in the strongly convex case), revealing a dependency in SSG methods, where a faster decrease in the smoothing parameter necessarily demands a correspondingly faster decay in the stepsize to preserve stability and achieve optimal convergence.

We further demonstrated how the general SSG algorithm can be specialized to two particular compositional settings of interest by satisfying the required assumptions. These two settings frequently arise in risk-averse optimization, machine learning, and large-scale optimization. We did not consider the double-expectation formulation examined in previous works, as the applications we target involve only a single expectation.

For future work, a promising direction is the integration of variance-reduction techniques—such as SVRG and SAGA—into this smoothing stochastic gradient framework, which may significantly accelerate convergence, especially in large-scale machine learning and optimization settings. Another direction is the exploration of practical applications in machine learning and risk-sensitive optimization, where compositional non-smooth structures naturally arise, such as in CVaR minimization. Evaluating the empirical performance of the SSG method in such domains remains an important task for further study.

Acknowledgments

This work is partially supported by the U.S. Air Force Office of Scientific Research (AFOSR) award FA9550-23-1-0217 and the U.S. Office of Naval Research (ONR) award N000142412656.

References

[1] S. Ahmed, U. Çakmak, and A. Shapiro. Coherent risk measures in inventory problems. *European J. Oper. Res.*, 182:226–238, 2007.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIIMS*, 2:183–202, 2009.

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60:223–311, 2018.

[4] C. Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Comput. Optim. Appl.*, 5:97–138, 1996.

[5] T. Chen, Y. Sun, and W. Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Trans. Signal Process.*, 69:4937–4948, 2021.

[6] X. Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.*, 134:71–99, 2012.

[7] X. Chen and W. Zhou. Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.*, 3:765–790, 2010.

[8] C. Dann, G. Neumann, and J. Peters. Policy evaluation with temporal differences: A survey and comparison. *J. Mach. Learn. Res.*, 15:809–883, 2014.

[9] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1413–1457, 2004.

[10] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *NeurIPS*, 27, 2014.

[11] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[12] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

[13] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIOPT*, 22:674–701, 2012.

[14] Y. Ermoliev. Stochastic programming methods, 1976.

[15] Y. Ermoliev, M. Yu, and R. B. Wets. *Numerical Techniques for Stochastic Optimization.* Springer-Verlag, Heidelberg, 1988.

[16] L. Evans. *Measure Theory and Fine Properties of Functions.* Routledge, Milton Park, Oxfordshire, UK, 2018.

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

[18] R. Garmanjani and L. N. Vicente. Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization. *IMA J. Numer. Anal.*, 33:1008–1028, 2013.

[19] T. Giovannelli, G. Kent, and L. N. Vicente. Inexact bilevel stochastic gradient methods for constrained and unconstrained lower-level problems. *ISE Technical Report 21T-025, Lehigh University*, December 2022.

[20] Z. Huo, B. Gu, J. Liu, and H. Huang. Accelerated method for stochastic composition optimization with nonsmooth regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] A. Jalilzadeh, U. Shanbhag, J. Blanchet, and P. W. Glynn. Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs. *Stoch. Syst.*, 12: 373–410, 2022.

[22] Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued Var. Anal.*, 29: 967–996, 2021.

[23] T. Li, A. Beirami, M. Sanjabi, and V. Smith. On tilted losses in machine learning: Theory and applications. *J. Mach. Learn. Res.*, 24:1–79, 2023.

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

[25] Q. Lin, X. Chen, and J. Pena. A smoothing stochastic gradient method for composite optimization. *Optim. Methods Softw.*, 29:1281–1301, 2014.

[26] J. Liu, Y. Cui, and J. Pang. Solving nonsmooth and nonconvex compound stochastic programs with applications to risk measure minimization. *Math. Oper. Res.*, 47:3051–3083, 2022.

[27] L. Liu, J. Liu, and D. Tao. Variance reduced methods for non-convex composition optimization. *IEEE TPAMI*, 44:5813–5825, 2021.

[28] J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[29] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *NeurIPS*, 24, 2011.

[30] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19:1574–1609, 2009.

[31] A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization. *SIAM Rev.*, 1983.

[32] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103:127–152, 2005.

[33] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, Berlin, 2013.

[34] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *FoCM*, 17:527–566, 2017.

[35] B. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Math. Math. Phys.*, 3:864–878, 1963.

[36] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Math. Math. Phys.*, 4:1–17, 1964.

[37] B.T. Polyak. *Introduction to Optimization.* Optimization Software, New York, 1987.

[38] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, pages 400–407, 1951.

[39] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, Berlin, 1998.

[40] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, PA, USA, 2021.

[41] N. Z. Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, Berlin, 2012.

[42] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58:267–288, 1996.

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

[44] M. Wang, J. Liu, and E. X. Fang. Accelerating stochastic composition optimization. *J. Mach. Learn. Res.*, 18:1–23, 2017.

[45] R. Wang, C. Zhang, L. Wang, and Y. Shao. A stochastic nesterov's smoothing accelerated method for general nonsmooth constrained stochastic composite convex optimization. *J. Sci. Comput.*, 93:52, 2022.

[46] W. Wolberg, O. Mangasarian, N. Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

[47] L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. *NeurIPS*, 22, 2009.

[48] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim*, 24:2057–2075, 2014.

[49] Y. Yu and L. Huang. Fast stochastic variance reduced admm for stochastic composition optimization. *arXiv preprint arXiv:1705.04138*, 2017.

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

[51] C. Zhang and X. Chen. Smoothing projected gradient method and its application to stochastic linear complementarity problems. *SIOPT*, 20:627–649, 2009.

[52] J. Zhang and L. Xiao. A composite randomized incremental gradient method. In *ICML*, pages 7454–7462. PMLR, 2019.

[53] J. Zhang and L. Xiao. A stochastic composite gradient method with incremental variance reduction. *NeurIPS*, 32, 2019.

A Assumption 3.8 for the case $f(x) = \phi(\psi(x))$

In this part of the appendix, we want to show strong evidence, for the case $f(x) = \phi(\psi(x))$, that there exists $y \in \mathcal{B}(x, E_1\mu)$ for some $E_1 > 0$, such that $\|\nabla_x \tilde{f}(x,\mu) - g(y)\| \leq E_2\mu$ for some $E_2 > 0$, where $g(y) \in \partial f(y)$. In this case, we have $\partial f(x) = \partial\phi(x)\nabla\psi(x)$, $\tilde{f}(x,\mu) = \tilde{\phi}(\psi(x),\mu)$, and $\nabla_x \tilde{f}(x,\mu) = \tilde{\phi}'(x,\mu)\nabla\psi(x)$.

If $\psi(x) > \frac{\mu}{2}$, then $\tilde{f}$ and f coincide, and the result is trivially satisfied for $y = x$. If $\psi(x) \leq \frac{\mu}{2}$, assume the existence of $y \in \mathcal{B}(x, E_1\mu)$ such that $\psi(y) = 0$. For most ϕ, such as in the $L1$ and Hinge cases (see Section 2), one has $|\tilde{\phi}'(t,\mu)| \leq E_3|t|/\mu$ for some $E_3 > 0$ and $\tilde{\phi}'(t,\mu) \in \partial\phi(0)$. Hence, $g(y) = \tilde{\phi}'(\psi(x),\mu)\nabla\psi(x) \in \partial f(y)$, and we can derive (assuming that $\nabla\psi$ is $L_{\nabla\psi}$-Lipschitz continuous for some $L_{\nabla\psi} > 0$)

$$\begin{aligned}\|\nabla_x \tilde{f}(x,\mu) - g(y)\| &= |\tilde{\phi}'(\psi(x),\mu)|\|\nabla\psi(x) - \nabla\psi(y)\| \\ &\leq \frac{E_3|\psi(x)|}{\mu} L_{\nabla\psi}\|x - y\| \\ &\leq \frac{E_1 E_3}{2} L_{\nabla\psi} \equiv E_2\mu.\end{aligned}$$

Note that for most ψ, it is reasonable to assume that such a y exists. If that was not the case, then for all $E_1 > 0$ and $\psi(y) = 0$, one would have $\|y - x\| > E_1\mu$. This would imply that $\|y - x\| > 2E_1\psi(x)$, and one would obtain $(\psi(x) - \psi(y))/\|x - y\| \leq 1/(2E_2)$ for all $E_2 > 0$, which would mean that ψ would be arbitrarily flat.

B Rate in the strongly convex case

In this part of the appendix, we analyze the rate of convergence of the SSG algorithm under the assumption that the smoothing function is strongly convex. We begin by stating the strong convexity setting and subsequently present the corresponding convergence theorem. Assumption B.1 states that the smoothing function $\tilde{f}$ is strongly convex, with the strong convexity constant depending on the smoothing parameter μ.

Assumption B.1 *For all $x \in \mathbb{R}^n$, the smoothing function $\tilde{f}(x,\mu)$ is strongly convex with a strong convexity constant $c(\mu)$ given by $c(\mu) = c/\mu$, where $c > 0$ is a constant.*

Assumption B.1 is satisfied for the smoothing functions (2.2) and (2.3) when x is relatively close to the kink point. After having stated the assumption we need, we now establish the rate of convergence for the strongly convex case.

Theorem B.1 (Convergence rate in the strongly convex case) *Under Assumptions 3.1–3.3 and B.1, suppose that the SSG algorithm runs with the smoothing parameter $\mu_k = k^{-a}$, where $a > 0$, and stepsize $\alpha_k = \frac{1}{(k+1)c(\mu_k)}$. Define $f_{best}^T = \min_{k=\{0,1,\ldots,T\}} f(x_k)$. Assume that x_* is a minimizer of the function f and x_1 is the initial iterate. Then for all $s > 0$ and $a > \frac{1}{2}$, we have*

$$\mathbb{E}[f_{best}^T] - f(x_*) \;\leq\; \left(\frac{4aC}{2a-1} + \frac{1}{2}\|x_1 - x_*\|^2 + \frac{1}{2}G^2\right)\frac{1}{T^{1-a}} + \frac{G^2}{2s}\frac{1}{T^{1-s-a}}. \tag{B.1}$$

Proof. We first recall the update rule of the SSG algorithm. For each iteration, we set the new iterate as $x_{k+1} = x_k - \alpha_k \nabla_x \tilde{f}(x_k, \xi_k, \mu_k)$. We use the update rule to get the following

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x_*\|^2 | x_k] \\
&= \mathbb{E}[\|x_k - \alpha_k \nabla_x \tilde{f}(x_k, \xi_k, \mu_k) - x_*\|^2 | x_k] \\
&= \mathbb{E}\|x_k - x_*\|^2 | x_k] - 2\alpha_k \mathbb{E}[\langle \nabla_x \tilde{f}(x_k, \xi_k, \mu_k), x_k - x_* \rangle | x_k] + \alpha_k^2 \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2 | x_k] \\
&= \|x_k - x_*\|^2 - 2\alpha_k \langle \nabla_x \tilde{f}(x_k, \mu_k), x_k - x_* \rangle + \alpha_k^2 \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2 | x_k].
\end{aligned} \tag{B.2}$$

By the strongly convexity assumption of $\tilde{f}(x, \mu)$, we have

$$\langle \nabla_x \tilde{f}(x_k, \mu_k), x_k - x_* \rangle \geq \frac{c(\mu_k)}{2} \|x_k - x_*\|^2 + \tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k). \tag{B.3}$$

Combining (B.2) and (B.3), and taking the unconditional expectation $\mathbb{E}[\cdot]$ of both sides, the following inequality is obtained:

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \mathbb{E}[\|x_k - x_*\|^2] - \alpha_k c(\mu_k) \mathbb{E}[\|x_k - x_*\|^2] \\
- 2\alpha_k \mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] + \alpha_k^2 \mathbb{E}[\|\nabla_x \tilde{f}(x_k, \xi_k, \mu_k)\|^2].
\end{aligned}$$

Then, by applying the bounded second moment assumption of the smoothing gradient as stated in Assumption 3.2, we have

$$2\alpha_k \mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] \leq (1 - \alpha_k c(\mu_k)) \mathbb{E}[\|x_k - x_*\|^2] - \mathbb{E}[\|x_{k+1} - x_*\|^2] + \frac{\alpha_k^2}{\mu_k^2} G^2. \tag{B.4}$$

Recall that we have $\alpha_k = \frac{1}{k^a(k+1)}$, $c(\mu_k) = k^a$, and $\mu_k = k^{-a}$. Consequently, inequality (B.4) can be written as:

$$\frac{2}{k^a(k+1)} \mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] \leq \frac{k}{k+1} \mathbb{E}[\|x_k - x_*\|^2] - \mathbb{E}[\|x_{k+1} - x_*\|^2] + \frac{1}{(k+1)^2} G^2.$$

Multiplying both sides by $k+1$, it follows that

$$\frac{2}{k^a} \mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] \leq k \mathbb{E}[\|x_k - x_*\|^2] - (k+1) \mathbb{E}[\|x_{k+1} - x_*\|^2] + \frac{1}{(k+1)} G^2. \tag{B.5}$$

Next, summing the inequality (B.5) from $k = 1$ to T for some $T > 0$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{T} \frac{2}{k^a} \mathbb{E}[\tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] = \sum_{k=1}^{T} \frac{2}{k^a} \mathbb{E}[f(x_k) - f(x_k) + f(x_*) - f(x_*) + \tilde{f}(x_k, \mu_k) - \tilde{f}(x_*, \mu_k)] \\
&\leq \mathbb{E}[\|x_1 - x_*\|^2] - (T+1) \mathbb{E}[\|x_{T+1} - x_*\|^2] + G^2 \sum_{k=1}^{T} \frac{1}{(k+1)} \\
&\leq \|x_1 - x_*\|^2 + G^2 \sum_{k=1}^{T} \frac{1}{(k+1)}.
\end{aligned} \tag{B.6}$$

From inequality (B.6), together with (3.2) (for x_n and x_*), we have

$$\begin{aligned}
&2\sum_{k=1}^{T}\frac{1}{k^a}\mathbb{E}[f(x_k)-f(x_*)]\\
&\leq -\sum_{k=1}^{T}\frac{2}{k^a}\mathbb{E}[\tilde{f}(x_k,\mu_k)-f(x_k)]+\sum_{k=1}^{T}\frac{2}{k^a}\mathbb{E}[\tilde{f}(x_*,\mu_k)-f(x_*)]+\|x_1-x_*\|^2+G^2\sum_{k=1}^{T}\frac{1}{(k+1)}\\
&\leq 4C\sum_{k=1}^{T}\frac{1}{k^{2a}}+\|x_1-x_*\|^2+G^2\sum_{k=1}^{T}\frac{1}{(k+1)}.
\end{aligned} \tag{B.7}$$

Now, our goal is to demonstrate that the right-hand side of inequality (B.7) can be bounded by a constant. To achieve this, let us first consider the term $\sum_{k=1}^{T}\frac{1}{k^{2a}}$, it holds that

$$\sum_{k=1}^{T}\frac{1}{k^{2a}}\leq 1+\int_1^T\frac{1}{x^{2a}}dx=1+\frac{x^{1-2a}}{1-2a}\bigg|_1^T=\frac{T^{1-2a}-2a}{1-2a}.$$

Particularly, when $a>\frac{1}{2}$, we have

$$\sum_{k=1}^{T}\frac{1}{k^{2a}}\leq\frac{2a-T^{1-2a}}{2a-1}\leq\frac{2a}{2a-1}. \tag{B.8}$$

Also, consider $\sum_{k=1}^{T}\frac{1}{(k+1)}$, it holds that

$$\sum_{k=1}^{T}\frac{1}{(k+1)}\leq\ln T+1.$$

Notice that for all $T\geq 0$ and $s>0$, the inequality $\ln T\leq\frac{T^s}{s}$ holds. Therefore, we have

$$\sum_{k=1}^{T}\frac{1}{(k+1)}\leq\ln T+1\leq\frac{T^s}{s}+1. \tag{B.9}$$

Next, by defining $f_{\text{best}}^T=\min_{k=\{0,1,\ldots,T\}}f(x_k)$ and considering again inequality (B.7), we have

$$2\sum_{k=1}^{T}\frac{1}{k^a}\mathbb{E}[f(x_k)-f(x_*)]\geq 2\mathbb{E}[f_{\text{best}}^T-f(x_*)]\sum_{k=1}^{T}\frac{1}{k^a}\geq 2\mathbb{E}[f_{\text{best}}^T-f(x_*)]\frac{1}{T^{a-1}}. \tag{B.10}$$

Thus, when $a>\frac{1}{2}$ and $s>0$, combining (B.7), (B.8), (B.9), and (B.10) gives the following:

$$\mathbb{E}[f_{\text{best}}^T-f(x_*)]\leq\left(\frac{4aC}{2a-1}+\frac{1}{2}\|x_1-x_*\|^2+G^2\frac{T^s+s}{2s}\right)\frac{1}{T^{1-a}},$$

from which we arrive at (B.1). □

The rate in the strongly convex case is arbitrarily close to $1/T^{1/2}$ in the sense of being worse than $1/T^{1/2}$ but better than $1/T^{1/p}$ for any $p>2$. This rate can be translated into a worst-case complexity bound arbitrarily close to ε^{-2}, in the sense of being slower than ε^{-2} but faster than any bound of rate ε^{-p} with $p>2$. Note that the rate in [43] is $1/T^{2/3}$, but in their approach, the authors used a similar but different problem formulation and also a form of momentum.

C Strongly convex case rate for $f(x) = \phi(\mathbb{E}[\psi(x, \xi)])$

In this part of the appendix, we consider the case $f(x) = \phi(\mathbb{E}[\psi(x, \xi)])$ (which follows the same structure as in Section 4) when f is strongly convex. In this setting, we aim to improve the convergence rate established for the general case of f in Appendix B. The specific algorithm we will use is shown in Algorithm 4, which follows the lines of the SCGD algorithm in [43]. The main modification in Algorithm 4 compared to Algorithm 2 in Section 4 lies in Step 2, which updates a moving average of the inner function ψ and computes the outer derivative at that averaged value. By doing so, the modified algorithm reduces the variance of $\psi(x_k, \xi_k)$ and provides more consistent outer derivative estimates. Besides, also notice that here we will be requiring $\{\alpha_k\}$ decaying faster than $\{\beta_k\}$, meaning $\alpha_k/\beta_k \to 0$.

Algorithm 4 SSG method for strongly convex $f(x) = \phi(\mathbb{E}[\psi(x, \xi)])$

Input: x_1, y_1, $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\mu_k\}$.

For $k = 1, 2, \ldots$ **do**

Step 1. Generate realizations ξ_k^1 and ξ_k^2 of the random variable ξ. Then compute $\psi(x_k, \xi_k^1)$ and $\nabla_x \psi(x_k, \xi_k^2)$, respectively.

Step 2. Update $y_{k+1} = (1 - \beta_k) y_k + \beta_k \psi(x_k, \xi_k^1)$ and compute $\tilde{\phi}'(y_{k+1}, \mu_k)$.

Step 3. Update iterate $x_{k+1} = x_k - \alpha_k \tilde{\phi}'(y_{k+1}, \mu_k) \nabla_x \psi(x_k, \xi_k^2)$.

Step 4. Update smoothing parameter $\mu_{k+1} \leq \mu_k$.

End do

Since the problem in this appendix has the same structure as in Section 4, we retain Assumptions 4.1 and 4.2. Following [43], we also introduce the additional assumptions stated below to establish the convergence result. Assumption C.1 states that the function f is strongly convex.

Assumption C.1 *For all $x \in \mathbb{R}^n$, the true function f is strongly convex with a strong convexity constant 2σ, where σ is some positive scalar.*

In the proof of Theorem C.2 below, we will use the following inequality known for strongly convex functions with minimizer x_*

$$f(x) - f(x_*) \;\geq\; \sigma \|x - x_*\|^2. \tag{C.1}$$

Assumption C.2 states the Lipschitz continuity of the derivative of the smoothing outer function $\tilde{\phi}$, where the Lipschitz constant is inversely proportional to the smoothing parameter μ.

Assumption C.2 *The smoothing outer function $\tilde{\phi}$ has a Lipschitz continuous derivative with constant $L_{\tilde{\phi}}/\mu$, meaning that for all $y, \bar{y} \in \mathbb{R}$ and all $\mu > 0$,*

$$\|\tilde{\phi}'(y, \mu) - \tilde{\phi}'(\bar{y}, \mu)\| \;\leq\; \frac{1}{\mu} L_{\tilde{\phi}} \|y - \bar{y}\|,$$

where $L_{\tilde{\phi}}$ is some positive scalar.

Assumption C.3 requires that the inner function ψ is Lipschitz continuous, and its stochastic sample $\psi(\cdot, \xi)$ is an unbiased estimator with a bounded variance.

Assumption C.3 *For all $x \in \mathbb{R}^n$, the inner function $\psi(x)$ is Lipschitz continuous with Lipschitz constant G_1 (and here we are using the same scalar as in Assumption 4.1), and its stochastic estimator $\psi(x, \xi)$ satisfies*

(i) Unbiasedness: $\mathbb{E}[\psi(x, \xi)] = \psi(x)$.

(ii) Bounded variance: $\mathbb{E}\big[\|\psi(x, \xi) - \psi(x)\|^2\big] \leq V_\psi$, where $V_\psi > 0$ is some positive scalar.

Assumption C.4 states that the term $\mathbb{E}[\tilde{\phi}'(\psi(x), \mu)\nabla_x \psi(x, \xi)]$ always lies in the subdifferential of the function f.

Assumption C.4 *For all $x \in \mathbb{R}^n$ and for all $\mu > 0$, it holds that*

$$\mathbb{E}[\tilde{\phi}'(\psi(x), \mu)\nabla_x \psi(x, \xi)] \ \in \ \partial f(x).$$

Before showing the convergence analysis of Algorithm 4, we first present a lemma ([43, Lemma 2(a)]) bounding the difference between y_{k+1} and $\psi(x_k)$ at each iteration k.

Lemma C.1 *Let Assumptions 4.1 and C.3 hold. Consider the sequences $\{(x_k, y_k)\}$ generated by Algorithm 4, it holds that*

$$\mathbb{E}[\|y_{k+1} - \psi(x_k)\|^2] \ \leq \ (1 - \beta_k)\mathbb{E}[\|y_k - \psi(x_{k-1})\|^2] + \frac{G_1^2}{\beta_k}\mathbb{E}[\|x_k - x_{k-1}\|^2] + 2V_\psi \beta_k^2. \tag{C.2}$$

Proof. See [43, Supplementary Materials Section G.1].

□

With the necessary assumptions and lemma listed, we now present the following convergence results. The proof of the theorem is inspired by [43].

Theorem C.2 *Under Assumptions 4.1, 4.2, and C.1–C.4, suppose that Algorithm 4 runs with stepsizes $\alpha_k = \frac{1}{k\sigma}$, $\beta_k = \frac{1}{k^{2/3}}$, and smoothing parameter $\mu_k = k^{-a}$, where $a > 0$. Assume that x_* is the minimizer of the function f. For all $a \in (0, \frac{1}{6})$ and for sufficiently large T, we have*

$$\begin{aligned}\mathbb{E}[\|x_{T+1} - x_*\|^2] \ &\leq \ \mathcal{O}\left(\frac{G_1^2 G_2^2}{\sigma^2}\frac{1}{T^{1-2a}} + \frac{L_{\tilde{\phi}}^2 G_1^6 G_2^2}{\sigma^4}\frac{1}{T^{2/3-4a}} + \frac{L_{\tilde{\phi}}^2 G_1^2 V_\psi}{\sigma^2}\frac{1}{T^{2/3-2a}}\right) \\ &= \ \mathcal{O}\left(\frac{1}{T^{2/3-4a}}\right).\end{aligned}$$

Proof. First, from the update rule of Algorithm 4, we have

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &= \|x_k - \alpha_k \tilde{\phi}'(y_{k+1}, \mu_k)\nabla_x \psi(x_k, \xi_k^2) - x_*\|^2 \\ &= \|x_k - x_*\|^2 - 2\alpha_k (x_k - x_*)^\top \tilde{\phi}'(y_{k+1}, \mu_k)\nabla_x \psi(x_k, \xi_k^2) \\ &\quad + \alpha_k^2 \|\tilde{\phi}'(y_{k+1}, \mu_k)\nabla_x \psi(x_k, \xi_k^2)\|^2 \\ &= \|x_k - x_*\|^2 - 2\alpha_k (x_k - x_*)^\top \tilde{\phi}'(\psi(x_k), \mu_k)\nabla_x \psi(x_k, \xi_k^2) + u_k \\ &\quad + \alpha_k^2 \|\tilde{\phi}'(y_{k+1}, \mu_k)\nabla_x \psi(x_k, \xi_k^2)\|^2,\end{aligned} \tag{C.3}$$

where u_k is defined as

$$u_k = 2\alpha_k(x_k - x_*)^\top \nabla_x \psi(x_k, \xi_k^2)(\tilde{\phi}'(\psi(x_k), \mu_k) - \tilde{\phi}'(y_{k+1}, \mu_k)).$$

To bound u_k, we use Assumption C.2 and Young's inequality for products

$$\begin{aligned} u_k &\leq 2\alpha_k \|x_k - x_*\| \|\nabla_x \psi(x_k, \xi_k^2)\| \|\tilde{\phi}'(\psi(x_k), \mu_k) - \tilde{\phi}'(y_{k+1}, \mu_k)\| \\ &\leq \frac{2\alpha_k L_{\tilde{\phi}}}{\mu_k} \|x_k - x_*\| \|\nabla_x \psi(x_k, \xi_k^2)\| \|\psi(x_k) - y_{k+1}\| \\ &\leq \alpha_k \sigma \|x_k - x_*\|^2 + \frac{\alpha_k L_{\tilde{\phi}}^2}{\sigma \mu_k^2} \|\nabla_x \psi(x_k, \xi_k^2)\|^2 \|\psi(x_k) - y_{k+1}\|^2 \end{aligned} \tag{C.4}$$

Taking the conditional expectation with respect to x_k on both sides of (C.3) and applying the convexity of f together with Assumption C.4, we obtain

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x_*\|^2 | x_k] &= \|x_k - x_*\|^2 - 2\alpha_k (x_k - x_*)^\top \mathbb{E}[\tilde{\phi}'(\psi(x_k), \mu_k) \nabla_x \psi(x_k, \xi_k^2) | x_k] \\ &\quad + \alpha_k^2 \mathbb{E}[\|\tilde{\phi}'(y_{k+1}, \mu_k) \nabla_x \psi(x_k, \xi_k^2)\|^2 | x_k] + \mathbb{E}[u_k | x_k] \\ &\leq \|x_k - x_*\|^2 - 2\alpha_k (f(x_k) - f^*) + \alpha_k^2 \mathbb{E}[\|\tilde{\phi}'(y_{k+1}, \mu_k) \nabla_x \psi(x_k, \xi_k^2)\|^2 | x_k] \\ &\quad + \mathbb{E}[u_k | x_k]. \end{aligned} \tag{C.5}$$

Next, by taking conditional expectation in (C.4), plugging it into (C.5), and taking total expectation on both sides, we have

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x_*\|^2] \leq (1 + \sigma\alpha_k) \|x_k - x_*\|^2 - 2\alpha_k (f(x_k) - f^*) + \alpha_k^2 \mathbb{E}[\|\tilde{\phi}'(y_{k+1}, \mu_k) \nabla_x \psi(x_k, \xi_k^2)\|^2] \\ + \frac{\alpha_k L_{\tilde{\phi}}^2}{\sigma \mu_k^2} \mathbb{E}[\|\nabla_x \psi(x_k, \xi_k^2)\|^2 \|\psi(x_k) - y_{k+1}\|^2]. \end{aligned}$$

We now use Assumptions 4.1–4.2 and (C.1), and obtain

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq (1 - \sigma\alpha_k) \|x_k - x_*\|^2 + \alpha_k^2 \left(\frac{G_1 G_2}{\mu_k}\right)^2 + \frac{\alpha_k L_{\tilde{\phi}}^2 G_1^2}{\sigma \mu_k^2} \mathbb{E}[\|\psi(x_k) - y_{k+1}\|^2]. \tag{C.6}$$

Now we denote $a_k = \mathbb{E}[\|x_k - x_*\|^2]$ and $b_k = \mathbb{E}[\|y_k - \psi(x_{k-1})\|^2]$. Let $\Lambda_{k+1} = \max\{\frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{(\beta_k - \sigma\alpha_k)\sigma\mu_k^2} - \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2}, 0\}$. Given the choice of α_k and β_k ($\alpha_k/\beta_k \to 0$), one has $\Lambda_{k+1} = \Theta(\frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\beta_k \sigma \mu_k^2})$. Multiplying (C.2) by $\Lambda_{k+1} + \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2}$ and summing it with (C.6), we obtain

$$\begin{aligned} a_{k+1} + (\Lambda_{k+1} + \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2}) b_{k+1} &\leq (1 - \sigma\alpha_k) a_k + (1 - \beta_k)(\Lambda_{k+1} + \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2}) b_k \\ &\quad + w_k + \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2} b_{k+1}, \end{aligned} \tag{C.7}$$

where

$$\begin{aligned} w_k &= \alpha_k^2 \left(\frac{G_1 G_2}{\mu_k}\right)^2 + \left(\Lambda_{k+1} + \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2}\right) \left(\frac{G_1^2}{\beta_k} \mathbb{E}[\|x_k - x_{k-1}\|^2] + 2V_\psi \beta_k^2\right) \\ &= \alpha_k^2 \left(\frac{G_1 G_2}{\mu_k}\right)^2 + (\frac{L_{\tilde{\phi}}^2 G_1^2}{\sigma \mu_k^2}) \mathcal{O}(\frac{\alpha_k^3 G_1^4 G_2^2}{\beta_k^2 \mu_k^2} + V_\psi \alpha_k \beta_k). \end{aligned}$$

From (C.7) and the expression for Λ_{k+1}, we can infer that

$$\begin{aligned} a_{k+1} + \Lambda_{k+1} b_{k+1} &\leq (1 - \sigma\alpha_k) a_k + (1 - \beta_k)(\Lambda_{k+1} + \frac{L_{\tilde{\phi}}^2 G_1^2 \alpha_k}{\sigma \mu_k^2}) b_k + w_k \\ &\leq (1 - \sigma\alpha_k)(a_k + \Lambda_{k+1} b_k) + w_k, \end{aligned}$$

where the second inequality comes from the expression for Λ_{k+1} and $\sigma\alpha_k \leq \beta_k$. Similarly to [43], it can be shown that $0 < \Lambda_{k+1} \leq \Lambda_k$ for sufficiently large k. Thus, if k is large enough, we have

$$a_{k+1} + \Lambda_{k+1} b_{k+1} \leq (1 - \sigma\alpha_k)(a_k + \Lambda_k b_k) + w_k. \tag{C.8}$$

Now, letting $J_k = a_k + \Lambda_k b_k = \mathbb{E}[\|x_k - x_*\|^2] + \Lambda_k \mathbb{E}[\|y_k - \psi(x_{k-1})\|^2]$, (C.8) is rewritten as $\mathbb{E}[J_{k+1}] \leq (1 - \sigma\alpha_k)\mathbb{E}[J_k] + w_k$. By applying $\alpha_k = 1/(\sigma k)$, $\beta_k = k^{-2/3}$, $\mu_k = k^{-a}$, and multiplying both sides of this inequality by k, we obtain

$$k\mathbb{E}[J_{k+1}] \leq (k-1)\mathbb{E}[J_k] + \frac{G_1^2 G_2^2}{\sigma^2 k^{1-2a}} + \mathcal{O}\left(\frac{L_{\tilde{\phi}}^2 G_1^6 G_2^2}{\sigma^4 k^{2/3-4a}} + \frac{L_{\tilde{\phi}}^2 G_1^2 V_\psi}{\sigma^2 k^{2/3-2a}} \right). \tag{C.9}$$

Next, summing the inequality (C.9) from $k = 1$ to T for some large $T > 0$, we obtain

$$T\mathbb{E}[J_{T+1}] \leq \mathcal{O}\left(\frac{G_1^2 G_2^2}{\sigma^2} T^{2a} + \frac{L_{\tilde{\phi}}^2 G_1^6 G_2^2}{\sigma^4} T^{4a+1/3} + \frac{L_{\tilde{\phi}}^2 G_1^2 V_\psi}{\sigma^2} T^{2a+1/3} \right).$$

Thus, it follows that

$$\begin{aligned} \mathbb{E}[\|x_{T+1} - x_*\|^2] \leq E[J_{T+1}] &\leq \mathcal{O}\left(\frac{G_1^2 G_2^2}{\sigma^2} \frac{1}{T^{1-2a}} + \frac{L_{\tilde{\phi}}^2 G_1^6 G_2^2}{\sigma^4} \frac{1}{T^{2/3-4a}} + \frac{L_{\tilde{\phi}}^2 G_1^2 V_\psi}{\sigma^2} \frac{1}{T^{2/3-2a}} \right) \\ &= \mathcal{O}\left(\frac{1}{T^{2/3-4a}} \right). \end{aligned}$$

□

Notice that the rate in this specific strongly convex case using Algorithm 4 is arbitrarily close to $1/T^{2/3}$ when $a \to 0$, which matches the rate in [43] under a similar setting.