Stochastic Composition Optimization of Functions Without Lipschitz Continuous Gradient

Yin Liu · Sam Davanloo Tajbakhsh

Received: 14 June 2022 / Accepted: 2 February 2023 / Published online: 10 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
In this paper, we study stochastic optimization of two-level composition of functions without Lipschitz continuous gradient. The smoothness property is generalized by the notion of relative smoothness which provokes the Bregman gradient method. We propose three stochastic composition Bregman gradient algorithms for the three possible relatively smooth compositional scenarios and provide their sample complexities to achieve an $\epsilon$-approximate stationary point. For the smooth of relatively smooth composition, the first algorithm requires $O(\epsilon^{-2})$ calls to the stochastic oracles of the inner function value and gradient as well as the outer function gradient. When both functions are relatively smooth, the second algorithm requires $O(\epsilon^{-3})$ calls to the inner function value stochastic oracle and $O(\epsilon^{-2})$ calls to the inner and outer functions gradients stochastic oracles. We further improve the second algorithm by variance reduction for the setting where just the inner function is smooth. The resulting algorithm requires $O(\epsilon^{-5/2})$ calls to the inner function value stochastic oracle, $O(\epsilon^{-3/2})$ calls to the inner function gradient, and $O(\epsilon^{-2})$ calls to the outer function gradient stochastic oracles. Finally, we numerically evaluate the performance of these three algorithms over two different examples.

Keywords Composition optimization · Stochastic optimization algorithm · Bregman subproblem
1 Introduction

This paper considers the two-level stochastic composition problem

$$\min_{x \in \mathcal{X}} F(x) \triangleq f(g(x)) \quad \text{with} \quad f\left( u \right) \triangleq \mathbb{E}[f_\varphi \left( u \right)], \; g(x) \triangleq \mathbb{E}[g_\xi \left( x \right)],$$

(1)

where $\mathbb{E}[f_\varphi \left( u \right)] \triangleq \int_{\Omega_f} f_\varphi(\omega_f) dP_f(\omega_f)$, $\mathbb{E}[g_\xi \left( x \right)] \triangleq \int_{\Omega_g} g_\xi(\omega_g) dP_g(\omega_g)$, and $\mathcal{X}$ is a closed convex set. Here $P_f$ (similarly $P_g$) is a probability distribution on the sample space $\Omega_f$ (similarly $\Omega_g$), the random vector $\varphi$ (similarly $\xi$) is a mapping from $\Omega_f$ (similarly $\Omega_g$) to a measurable space $\mathbb{W}_f$ (similarly $\mathbb{W}_g$), and $f_\varphi : \mathbb{R}^d \to \mathbb{R}$ is a smooth function (similarly $g_\xi : \mathbb{R}^n \to \mathbb{R}^d$ is a smooth map).

Furthermore, we assume that the functions $f$, $g$, or both do not have Lipschitz continuous gradient. A continuously differentiable function $h$ has Lipschitz continuous gradient if for some constant $L$, we have

$$\|\nabla h(x) - \nabla h(\bar{x})\| \leq L \|x - \bar{x}\|, \quad \forall x, \; \bar{x} \in \text{dom} \; h.$$  

(2)

In the absence of Lipschitz continuity of the inner, outer, or both functions’ gradients, problem (1) appears in many applications such as policy evaluation for Markov decision processes [20], risk-averse optimization [52], low-rank nonnegative matrix factorization [10, 31], and model-agnostic meta-learning (MAML) [30]. Below, we discuss some of these applications in more detail.

Example 1.1 (Policy evaluation for Markov decision processes (MDP)) Consider a Markov chain with states $\{Y_0, Y_1, \ldots\} \subset \mathcal{Y}$, an unknown transition operator $P$, a reward function $r : \mathcal{Y} \to \mathbb{R}$ and a discount factor $\gamma \in (0, 1)$. The goal is to estimate the value function $V : \mathcal{Y} \to \mathbb{R}$, defined as $V(y) \triangleq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(Y_t) | Y_0 = y]$, for a fixed control policy $\pi$. For a finite state space $\mathcal{Y}$, the value functions of all possible initial states can be represented as a vector $v \in \mathbb{R}^{\mathcal{Y}}$ that satisfies the Bellman equation $v = r + \gamma Pv$, where $r \in \mathbb{R}^{\mathcal{Y}}$ such that $r_i = r(Y_i)$. When $|\mathcal{Y}|$ is large, one can approximate $v$ by $v \approx \Phi x$, where $\Phi$ is some matrix of basis functions and $x \in \mathbb{R}^n$ contains the coefficients with $n \ll |\mathcal{Y}|$. Since one only has access to random samples of $P$ and $r$, i.e., $\hat{P}$ and $\hat{r}$, the residual minimization problem to approximate the solution of the above system is

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^n} \text{dist}(\mathbb{E}[\hat{r}], (I - \gamma \mathbb{E}[\hat{P}]) \Phi x),$$

(3)

where $\text{dist} : \mathbb{R}^{\mathcal{Y}} \times \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}_+$ is some distance function [53]. Let $A_\xi \triangleq (I - \gamma \hat{P}) \Phi$ and $r_\varphi \triangleq \hat{r}$, the problem can then be written as $\min_{x \in \mathcal{X}} \text{dist}(\mathbb{E}[r_\varphi], \mathbb{E}[A_\xi^+ x])$ which is a special case of (1). Note that if the reward is in form of count data (e.g., number of news clicks suggested by a recommender system—see, e.g., [45] and references therein), a closely related problem to (3) is

$$\min_{x \in \mathbb{R}^n_+} D_{\text{KL}}(\mathbb{E}[r_\varphi], \mathbb{E}[A_\xi^+ x]),$$

(4)
where $A_\xi^+ \triangleq \max[(I - \gamma \hat{P})\Phi, 0\mathcal{Y}|\mathcal{X}_n]$ and $D_{KL}(\cdot, \cdot)$ is the Kullback–Leibler (KL) divergence [18] (see (10) for the definition). This problem is known as the stochastic Poisson linear inverse problem (see, e.g., [7]—Section 5.1). The KL distance is a natural measure that corresponds to noise of Poisson type and does not admit globally Lipschitz continuous gradient [7].

**Example 1.2** (Risk-averse optimization) Consider the mean-variance risk-averse optimization problem [52]

$$
\min_{x \in \mathcal{X} \subseteq \mathbb{R}^n} -\mathbb{E}[r_\xi(x)] + \lambda \text{Var}[r_\xi(x)],
$$

where $r_\xi(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ is a random function given by the decision variable $x$, $\text{Var}[-]$ denotes the variance with respect to $\xi$, and $\lambda > 0$ is the risk-aversion parameter. In the context of risk-averse optimal control of systems governed by partial differential equations (PDEs) under uncertainty, there is a need to solve (5) where $r_\xi(x)$ is the control objective whose evaluation requires the solution of a system of PDEs (see, e.g., [13] and references therein). To render the computation of the control objective and its gradient tractable, it is common to employ quadratic approximations of the objective function in the context of risk-neutral optimal control [2]. However, the variance component of the risk-averse mean-variance formulation (5) results in polynomial objective function of degree 4 which does not have Lipschitz continuous gradient (see Sect. 4.1 for more details).

**Example 1.3** (Low-rank nonnegative matrix factorization) Let $D \in \mathbb{R}^{p \times q}_+$ be an unknown nonnegative matrix. We aim to find a low-rank approximate nonnegative matrix factorization (NMF) of $D$ based on random samples $\hat{D}$ from a stochastic oracle such that $\mathbb{E}[\hat{D}] = D$. This requires solving

$$
\min_{X \in \mathbb{R}^{p \times k}_+, Y \in \mathbb{R}^{q \times k}_+} \text{dist}(\mathbb{E}[\hat{D}], XY^\top),
$$

where $k < \min\{p, q\}$ is a prespecified rank and $\text{dist}(\cdot, \cdot)$ is some distance function [10]. Low-rank approximate NMF has a range of applications in feature extraction, text mining, and spectral data analysis [39, 50]. Due to the nonnegativeness of $D$ and $XY^\top$, there are information-theoretic arguments on taking statistical divergences such as KL divergence [39] or $\varphi$-divergence [17] for the distance function, which do not have Lipschitz continuous gradient.

It is natural to use stochastic gradient descent (SGD) to solve (1). However, obtaining an unbiased estimator of the gradient of the composition is challenging as

$$
\nabla F(x) = \mathbb{E}_{\varphi, \xi} \left[ \nabla g_\xi(x) \nabla f_\varphi(\mathbb{E}_\xi[g_\xi(x)]) \right] \neq \mathbb{E}_{\varphi, \xi} \left[ \nabla g_\xi(x) \nabla f_\varphi(g_\xi(x)) \right],
$$

i.e., the stochastic gradient of the outer function evaluated at a stochastic value of the inner function results in a biased estimator.
To overcome this challenge, different methods are proposed to approximate the inner expectation which then helps to prove convergence to an approximate stationary point. These methods are discussed in Sect. 1.1. However, similar to gradient-based methods for differentiable functions, theoretical guarantees for these methods heavily make use of the Lipschitz continuity of the gradient which narrows their applications. In the absence of Lipschitz continuity of the gradient, a sequence of recent works [7, 12, 44] introduce the notion of relative smoothness which provides a new non-quadratic upper bound to the objective function and provides a new descent lemma [7]. The definition of relative smoothness is provided in Sect. 1.2. The new upper bound requires solving a Bregman gradient subproblem and is discussed in Sect. 1.1.

1.1 Related Work

This work is on the intersection of two research areas: 1. stochastic composition optimization, 2. optimization of functions without Lipschitz continuous gradient. Below, we review the literature corresponding to these areas.

**Stochastic Composition Optimization.** The stochastic composition problem dates back to [27] and regained attention due to the broad range of applications in machine learning. Large deviation bounds for the empirical optimal value were established in [28], and the central limit theorem for the composition problem was established in [22] which proved that the empirical optimal value of (1) converges to the true optimal value. Recently, different works focus on developing efficient first-order algorithms to solve (1). Wang et al. [55] proposed a two-step algorithm to solve two-level composition problems. The first step approximates the inner function value by moving average

\[ y^{k+1} = (1 - \beta_k) y^k + \beta_k g_{\xi_k}(x^k), \]

with \( \beta_k \in (0, 1) \), and the second step performs the projected stochastic gradient update

\[ x^{k+1} = \Pi_X \{ x^k - \alpha_k \nabla g_{\xi_k}(x^k) \nabla f_{\phi}(y^{k+1}) \}, \]

where \( \Pi_X \) refers to the projection operator onto the set \( X \). Since the gradient estimator is biased, the algorithm requires the stepsize \( \alpha_k \) to be small compared to \( \beta_k \), i.e., it should satisfy \( \lim_{k \to \infty} \alpha_k / \beta_k = 0 \) which results in a two-timescale algorithm. Almost sure convergence of the iterates of the algorithm to an optimal solution in the convex and to a stationary point in the nonconvex but unconstrained setting are established under the smoothness assumption. However, due to the two-timescale requirement of the algorithm, its sample complexity is inferior to SGD for single-level problems. The paper improves its sample complexity for the convex setting by a second algorithm. Furthermore, the convergence rate of the two-timescale algorithm was improved in [56] where they further considered minimizing the sum of a two-level composition and a nonsmooth but convex regularization function. Hu et al. [37] also studied this problem under the same assumptions, but the sample complexity of their algorithm is still suboptimal compared to SGD.
Recently, [33] proposed a single-timescale algorithm (i.e., the one in which the rate of inner function and the outer function gradient updates are in the same order) to solve (1). The paper also establishes almost sure convergence of the algorithm to a stationary point in the constrained setting. Motivated by gradient flow in continuous space, [16] proposes a single-timescale algorithm for the unconstrained problem. Both algorithms in [16, 33] achieve the sample complexity of SGD. The difference between their algorithms is that [33] uses moving average on the inner function value, the \( \{x_k\} \) sequence, and the gradient estimates, while [16] improves the inner function value update as

\[
y_{k+1} = (1 - \beta_k)(y_k + \nabla g_{\xi_k}(x_k)(x_k - x_{k-1})) + \beta_k g_{\xi_k}(x_k),
\]

where \( \nabla g_{\xi_k}(x_k)(x_k - x_{k-1}) \) is a correction term derived from an ODE analysis. The algorithm in [16] is also generalized for multilevel composition problems.

Besides [16], there are other works that propose algorithms for multilevel composition problems. The work of [33] is generalized in [5] to solve multilevel composition problems with and without a need to mini-batch sampling in each iteration. [61] also considers the multilevel composition problem and proposes a variance reduction method to solve it. Total sample complexity of their method has polynomial dependence on the number of levels. Yang et al. [59] proposes a multi-timescale method and [51] generalizes the work of [33] to multilevel and, further, shows the almost sure convergence of their single-timescale algorithm for problems where the composition functions are Lipschitz continuous. The paper also provides the sample complexity of the algorithm when the composition functions are smooth.

The composition problem with specific structures is considered in a number of works. Dai et al. [19] considers the setting where the outer function \( f \) is convex and the inner function \( g \) is linear. Blanchet et al. [11] proposes an unbiased estimator of the gradient for convex problems in finite-sum form. For the same setting, [23, 41–43, 58, 60] use variance reduction techniques to accelerate the convergence. Zhang and Xiao [62] considers minimizing the sum of a composition and a nonsmooth convex function \( R(x) \) and proposed a prox-linear algorithm with subproblem

\[
x_{k+1} = \arg \min_x f(\tilde{g}^k + \tilde{J}(x - x^k)) + R(x) + \frac{M}{2} \|x - x^k\|^2,
\]

where \( \tilde{g}^k \) and \( \tilde{J} \) are (mini-batch or variance-reduced) estimates of \( g(x^k) \) and \( \nabla g(x^k) \), respectively. Their theory requires the outer function \( f \) to be deterministic, convex, and possibly nonsmooth, while \( g \) be a smooth function.

Finally, we note that the composition problems are also related to biased gradient methods which focus on establishing the stationarity using biased gradients of the objective functions, not necessarily in the composition structure [36].

**Mirror Descent for Relatively Smooth Function.** The mirror descent, also known as Bregman proximal gradient method, was originally developed by Nemirovski and Yudin (see [48]) to solve \( \min_{x \in \mathcal{X}} f(x) \) in general Banach space where the primal and dual spaces are not isometric to each other—see [14] for more information. Given a continuously differentiable and strictly convex mirror function \( h \), the mirror descent
algorithm performs the update $x^{k+1} = \nabla h^{-1}(\nabla h(x^k) - \alpha_k \nabla f(x^k))$, where $\nabla h^{-1}$ is the inverse of the gradient map. Later, [8] proved that the mirror descent update is indeed the solution to the Bregman gradient subproblem

$$x^{k+1} = \underset{x \in \mathcal{X}}{\text{argmin}} \left(\nabla f(x^k), x - x^k\right) + \frac{1}{\alpha_k} D_h(x, x^k),$$

where $D_h(x, x^k) \triangleq h(x) - h(x^k) - \langle \nabla h(x^k), x - x^k \rangle$ is the Bregman distance.

From the optimization perspective, mirror descent allows obtaining convergence rates with significantly less dependence on the dimension of the ambient space $n$ in certain cases. For instance, consider minimizing a convex and $C_f$ Lipschitz continuous function $f(x)$ over the simplex $\mathcal{X} \triangleq \{x \in \mathbb{R}^n : \sum_i^n x_i = 1, x_i \geq 0\}$. As discussed in [8], the iteration complexity of the subgradient projection method for this problem is $\min_{1 \leq i \leq k} f(x^i) - \min_{x \in \mathcal{X}} f(x) \leq O(1) C_f \sqrt{n} / \sqrt{k}$, while the complexity of the mirror descent algorithm with $h(x) = 2^{-1} \|x\|^2_p$ and $p = 1 + (\log n)^{-1}$ is $O(1) \sqrt{\log n} / \sqrt{k}$, which improves the subgradient method by $\sqrt{n} / \log n$. Note that the resulting mirror descent algorithm has closed-form solution for its subproblem. The mirror descent algorithm is used in (stochastic) convex [3, 4, 8, 9, 26, 38, 47] and nonconvex [32, 57] settings. In either case, similar to other first-order methods, the Lipschitz continuity of the objective function or its gradient is crucial for the analysis.

In many problems, however, the objective function is not smooth, i.e., it does not have Lipschitz continuous gradient. Recently, a new relative smoothness condition was introduced by [7] for convex function $f$, which generalizes the commonly used smoothness condition. The function $f$ is smooth relative to $h$ if $L h - f$ is a convex function—see Definition 1.2 for the formal definition. This condition provides an upper bound on the objective function through the Bregman distance with the generating function $h$ as $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x)$, which requires solving such problems using the mirror descent update discussed above. Later, [12] generalized this smoothness condition for nonconvex functions and named it smooth adaptable (smad) property. Lu et al. [44] studied the same property with a different requirement on the generating function $h$. This property unifies the mirror descent and the gradient descent as setting $h(x) = \frac{1}{2} \|x\|^2$ results in $D_h(y, x) = \frac{1}{2} \|y - x\|^2$. If so, the above inequality recovers the traditional smoothness inequality $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ and the corresponding descent lemma [54]. Further studies focus on nonconvex [1, 6, 12, 40, 46] and stochastic [21, 24, 34, 35, 63] scenarios. The lower bound on the sample complexity under this new smoothness assumption was established in [25] which indicates the need for additional assumptions for acceleration. Finally, we note that the notion of Lipschitz continuity was also generalized to relative continuity in [21, 54].

### 1.2 Contributions

This paper proposes three algorithms to solve the two-level stochastic composition problem (1) and investigates their iteration/sample complexities in the absence of the...
smoothness of the inner, outer, or both composition functions. More specifically, the contributions of the paper are as follows.

1. We consider three different scenarios of constrained two-level stochastic composition problems: Smooth of Relative-smooth (SoR), Relative-smooth of Smooth (RoS), and Relative-smooth of Relative-smooth (RoR). For each combination, we prove that the composition function is smooth relative to a (strictly/strongly) convex distance generating function.

2. For the SoR composition, we propose a single-timescale mini-batch stochastic composition Bregman gradient method. This algorithm achieves an \( \epsilon \)-approximate stationary solution in \( O(\epsilon^{-2}) \) calls to the inner function value and gradient and outer function gradient stochastic oracles, which matches the sample complexity of SGD in the single-level problems. This result also matches the lower bound of the Bregman method with merely relative smoothness assumption [25].

3. For the RoR (and RoS) composition, we propose a mini-batch prox-linear stochastic Bregman gradient algorithm with stochastic oracle complexity of \( O(\epsilon^{-2}) \) for the gradient of the inner and outer functions, and stochastic oracle complexity of \( O(\epsilon^{-3}) \) for the inner function value.

4. For the RoS composition, we improve the RoR algorithm by a variance reduction technique in the RoS-VR algorithm. The algorithm improves the stochastic oracle complexity of the inner function value to \( O(\epsilon^{-5/2}) \) and the inner function gradient to \( O(\epsilon^{-3/2}) \), while the complexity of the outer function gradient stochastic oracle is still \( O(\epsilon^{-2}) \).

**Preliminaries**

Given a continuously differentiable function \( f : \mathbb{R}^n \to \mathbb{R}^d \), \( \nabla f(x) \in \mathbb{R}^{n \times d} \) denotes its Jacobian (equivalently its gradient when \( d = 1 \)). For a real-valued two-time continuously differentiable function \( f \), \( \nabla^2 f \) denotes its Hessian. Given a closed convex set \( X \subseteq \mathbb{R}^n \), \( \text{int} X \) denotes its interior, \( N_X(x) \) denotes the normal cone to set \( X \) at \( x \), and \( \delta_X(\cdot) \) denotes the indicator function of \( X \). Finally, given \( x \in \mathbb{R}^n \), \( \|x\| \) denotes its \( l_2 \) norm.

Next, we introduce the Bregman distance and the notion of relative smoothness which are central to the ideas discussed in this paper.

**Definition 1.1 (Bregman distance)** Let \( h \) be a proper, differentiable, and convex function on the \( \text{int dom} \ h \). The Bregman distance of \( x, y \in \text{dom} \ h \) generated by \( h \) is defined as

\[
D_h(x, y) \triangleq h(x) - h(y) - \langle \nabla h(y), x - y \rangle.
\] (7)

Note that since \( h \) is convex, \( D_h(x, y) \geq 0 \) for any \( x, y \in \text{dom} \ h \). Furthermore, if \( h \) is \( \mu \)-strongly convex, i.e., \( h(x) - h(y) - \langle \nabla h(y), x - y \rangle \geq (\mu/2)\|x - y\|^2 \), then \( D_h(x, y) \geq (\mu/2)\|x - y\|^2 \) by the definition. Note that the Bregman distance is convex with respect to the first argument. For \( h(x) = \frac{1}{2}\|x\|^2 \), \( D_h(x, y) = \frac{1}{2}\|x - y\|^2 \), i.e., the Euclidean distance.
Definition 1.2 (Relative-smoothness [7]) Assume $\text{dom} \ h \subset \text{dom} \ f$. The function $f$ is $L$-smooth relative to the convex function $h$ if

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leq LD_h(x, y), \ \forall x, y \in \text{dom} \ h. \quad (8)$$

If $f$ is a vector-valued function, then it is $L$-smooth relative to $h$ if

$$\|f(x) - f(y) - \langle \nabla f(y), x - y \rangle\| \leq LD_h(x, y), \ \forall x, y \in \text{dom} \ h. \quad (9)$$

If $h(x) = \frac{1}{2}\|x\|^2$, the relative smoothness results into the famous upper/lower bound $|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leq L\|x - y\|^2$ provided that $f$ is smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. The left-hand-side inequality from (8), i.e., $-LD_h(x, y) \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle$ is also known as $L$-weak convexity of $f$ relative to $h$ [63], which generalizes the notion of weak convexity, i.e., $-(L/2)\|x - y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle$.

Some examples of relatively smooth functions are provided below.

Example 1.4 (Relatively smooth functions)

- Let $f(x)$ be twice differentiable and the operator norm of its Hessian with respect to the $l_2$-norm satisfies $\|\nabla^2 f(x)\| \leq p_r(\|x\|)$, where $p_r(\cdot)$ is a univariate $r$th-degree polynomial that satisfies $p_r(\alpha) \leq L(1 + \alpha^r)$ for $\alpha \geq 0$ and $L > 0$. Then, $f(x)$ is $L$-smooth relative to $h(x) = \frac{1}{r+2}\|x\|^{r+2} + \frac{1}{2}\|x\|^2$ [44].

- In the Poisson linear inverse problem, given a matrix $A \in \mathbb{R}^{m \times n}_+$, one needs to recover the signal $x \in \mathbb{R}^n_+$ from noisy observations $b$ such that $Ax \approx b$ by minimizing the Kullback–Leibler divergence

$$D_{KL}(b, Ax) \triangleq \sum_{i=1}^{n} (b_i \log(b_i/(Ax)_i) + (Ax)_i - b_i), \quad (10)$$

which is $L$-smooth relative to $h(x) = -\sum_{i=1}^{n} \log x_i$ for $L \geq \|b\|_1$—see [7].

Next, we present two lemmas which are key to the analysis of the Bregman gradient methods.

Lemma 1.1 (Three-point identity, (Lemma 3.1 in [15])) Suppose $h$ is a differentiable and convex function. For any three points $y, z \in \text{int dom} \ h$ and $x \in \text{dom} \ h$, the following identity holds,

$$D_h(x, y) + D_h(y, z) = D_h(x, z) + \langle \nabla h(z) - \nabla h(y), x - y \rangle.$$

This identity is the generalization of Pythagorean identity in Euclidean geometry. The next lemma follows from the three-point identity above.
Lemma 1.2 (Three-point inequality) Let $\varphi$ be a proper, lower semicontinuous, and convex function, $h$ be a differentiable convex function. Given $\tau > 0$ and $x \in X \subset \text{int dom } h$, let

$$x^+ \in \arg\min_{y \in X} \varphi(y) + \frac{1}{\tau} D_h(y, x),$$

which is unique if $h$ is strongly convex. Then, $x^+$ satisfies

$$\tau (\varphi(x^+) - \varphi(y)) \leq D_h(y, x) - D_h(x^+, x) - D_h(y, x^+), \quad \forall y \in X.$$

Remark 1.1 In particular, if we set $\varphi(y) = \langle w, y - x \rangle$, then

$$\tau \langle w, x^+ - y \rangle \leq D_h(y, x) - D_h(x^+, x) - D_h(y, x^+), \quad \forall y \in X.$$

2 Smooth of Relative-smooth (SoR) Composition

As we deal with the composition of two functions, under different smoothness assumptions, there are four possible composition scenarios. The smooth of smooth composition is the simplest one which results in the composition function to be also smooth, and it is well studied in the literature of stochastic composition problem listed above. In this section, we study the Smooth of Relative-smooth (SoR) composition. First, we give the definition of the stationarity measure under the Bregman distance.

Lemma 2.1 (Stationarity measure) Given a $\mu_h$-strongly convex function $h$, define

$$\hat{x}^+ \triangleq \arg\min_{y \in X} \langle \nabla F(x), y - x \rangle + \frac{1}{\tau} D_h(y, x),$$ (11)

where $\tau > 0$. Then $\hat{x}^+ = x$ if and only if $-\nabla F(x) \in \mathcal{N}_h(x)$.

Proof By the optimality condition, we have

$$0 \geq \langle \nabla F(x), y - x \rangle + \frac{1}{\tau} (\nabla h(\hat{x}^+) - \nabla h(x)) + \partial \delta_h(\hat{x}^+).$$ (12)

Hence, if $\hat{x}^+ = x$, then $-\nabla F(x) \in \mathcal{N}_h(x)$. Next, assuming $-\nabla F(x) \in \mathcal{N}_h(x)$, we have $\langle \nabla F(x), y - x \rangle \geq 0, \forall y \in X$. By the strong convexity of $h$, $D_h(y, x) \geq (\mu_h/2) \| y - x \|^2 \geq 0, \forall y \in X$, with the equality when $y = x$. Hence, the minimum of (11) is unique with the solution $\hat{x}^+ = x$. □

The above lemma shows that $\text{dist}(\hat{x}^+, x)$ is a suitable measure for the stationarity. Indeed, similar measures are commonly used in the constrained optimization literature, e.g., in [32], the distance function $\text{dist}(\hat{x}^+, x) \triangleq \frac{1}{2\tau^2} \| \hat{x}^+ - x \|^2$ is used as the stationarity measure which is equal to $\| \nabla F(x) \|^2$ when $X = \mathbb{R}^n$. Following Lemma 2.1, we use $\text{dist}(\hat{x}^{k+1}, x^k) = D_h(\hat{x}^{k+1}, x^k)/\tau^2$ to measure stationarity in the SoR composition.
2.1 Smoothness of the Composition Function

Below, we provide the main assumptions for the SoR composition. These assumptions are common in stochastic (composition) optimization with the main difference being generalization of the smoothness of the inner function to relative smoothness.

Assumption 1 (Smooth of Relative-smooth (SoR) composition) The functions $f$ and $g$ satisfy the following conditions:

(a) The function $g$ is average $L_g$-smooth relative to 1-strongly convex function $h_g$, i.e., $\forall x_1, x_2 \in \text{dom } g$,

$$\mathbb{E}_\xi \left[ \| g_\xi (x_1) - g_\xi (x_2) - \langle \nabla g_\xi (x_2), x_1 - x_2 \rangle \| \right] \leq L_g D_{h_g} (x_1, x_2).$$

(b) The function $f$ is average $L_f$-smooth, i.e., $\forall u_1, u_2 \in \text{dom } f$,

$$\mathbb{E}_\varphi \left[ \| \nabla f_\varphi (u_1) - \nabla f_\varphi (u_2) \| \right] \leq L_f^2 \| u_1 - u_2 \|^2.$$

(c) The stochastic gradients of $f$ and $g$ are bounded in expectation, i.e., $\forall u \in \text{dom } f$ and $\forall x \in \text{dom } g$,

$$\mathbb{E}_\varphi \left[ \| \nabla f_\varphi (u) \|^2 \right] \leq C_f^2, \quad \mathbb{E}_\xi \left[ \| \nabla g_\xi (x) \|^2 \right] \leq C_g^2.$$

Remark 2.1 By Jensen’s inequality, Assumption 1-(a $\sim$ c) also hold for $f$ and $g$ as well, i.e., $g$ is $L_g$-smooth relative to $h_g$, $f$ is $L_f$-smooth, and $f, g$ are $C_f, C_g$-Lipschitz continuous. Furthermore, note that bounded stochastic gradient implies bounded variance, i.e., $\forall u \in \text{dom } f$, we have

$$\mathbb{E}_\varphi [\| \nabla f_\varphi (u^k) - \nabla f(u^k) \|^2] = \mathbb{E}_\varphi [\| \nabla f_\varphi (u^k) \|^2] - \| \nabla f(u^k) \|^2 \leq \mathbb{E}_\varphi [\| \nabla f_\varphi (u^k) \|^2] \leq C_f^2.$$

Similarly, $\mathbb{E}_\xi [\| \nabla g_\xi (x^k) - \nabla g(x^k) \|^2] \leq C_g^2, \forall x \in \text{dom } g.$

Under Assumption 1, the SoR composition is proved to be smooth relative to the generating function $h$ defined in Lemma 2.2, with its proof in Appendix A.1.

Lemma 2.2 Under Assumption 1, $F(x) = f(g(x))$ is 1-smooth relative to $C_f^2 L_f$-strongly convex function $h(x) = \frac{C_g^2 L_f}{2} \| x \|^2 + C_f L_g h_g(x)$.

Remark 2.2 If $h_g(x) = \frac{1}{2} \| x \|^2$, i.e., $h_g$ is also smooth, then under Assumption 1, the composition function $F(x)$ is $(C_g^2 L_f + C_f L_g)$-smooth which matches the results derived in [33].
2.2 Proposed Algorithm for the SoR Composition

By relative smoothness of the composition function, it is natural to solve this problem using the Bregman gradient method [12]. But, as discussed in Introduction, it is impossible to get an unbiased estimate of the gradient of the objective function. Given the current point \( x^k \), one needs to estimate the inner function value and then obtain a stochastic gradient of the outer function at this estimate. This requires tracking three random sequences: the iterate \( \{ x^k \} \), inner function value estimate \( \{ u^k \} \) at the current iterate, and the gradient estimate of the composition function \( \{ w^k \} \). These random sequences are defined on the probability space \( (\Omega, \mathcal{F}, P) \), where \( \Omega \) is the sample space, \( \mathcal{F}_k \) is the \( \sigma \)-algebra generated by the algorithm up to iteration \( k \), i.e.,

\[
\{ x^0, \ldots, x^{k+1}, u^0, \ldots, u^k, w^0, \ldots, w^k \},
\]

and \( P \) is the probability measure.

Given an iterate \( x \), an estimate of the inner function value \( u \), and i.i.d. sample batches \( B_g, B_{\nabla g}, \) and \( B_{\nabla f} \), the mini-batch estimate of the inner function and the mini-batch estimate of the gradients of the inner and outer functions are calculated as

\begin{align*}
\hat{u}(x; B_g) &= \frac{1}{|B_g|} \sum_{\xi \in B_g} g_{\xi}(x), \quad (13a) \\
\hat{v}(x; B_{\nabla g}) &= \frac{1}{|B_{\nabla g}|} \sum_{\xi \in B_{\nabla g}} \nabla g_{\xi}(x), \quad (13b) \\
\hat{s}(u; B_{\nabla f}) &= \frac{1}{|B_{\nabla f}|} \sum_{\varphi \in B_{\nabla f}} \nabla f_{\varphi}(u). \quad (13c)
\end{align*}

Note that not all of the above estimators are used in all of the algorithms presented later.

We propose Algorithm 1 to solve the SoR composition, and its sample complexity is derived in the rest of this section. To do so, beside Assumption 1, we require two extra assumptions discussed below. Assumption 2-(a) is a standard lower boundedness of the objective function for nonconvex optimization and (b) is the boundedness of the variance of the stochastic inner function value which is necessary to deal with stochastic estimators.

**Assumption 2** The functions \( f_{\varphi}, g_{\xi} \), and \( F \) satisfy the following conditions:

(a) The function \( F(x) \) is lower-bounded by \( F^* \).

(b) The variance of \( g_{\xi} \) is bounded, i.e.,

\[
\mathbb{E}_{\xi} \left[ \| g_{\xi}(x) - g(x) \|^2 \right] \leq \sigma_g^2, \quad \forall x \in \text{dom } g.
\]

To prove the complexity of this algorithm, we first need to establish error bounds for the stochastic estimates of the inner function value and gradient of the composition.
Algorithm 1 SoR algorithm (SoR)

Require: $x^0 \in \mathcal{X}$, $u^0 \in \mathbb{R}^d$, $(\tau_k)_{k \in \mathbb{N}_+}$, $(\beta_k)_{k \in \mathbb{N}_+} \subset (0, 1)$, $h_g$

1: Sample $B_{V_f}^0$, $B_{V_g}^0$

and update $v^0 = v(x^0; B_{V_f}^0)$, $s^0 = s(u^0; B_{V_g}^0)$ using (13b)-(13c)

2: Update $w^0 = v^0 s^0$

3: for $k = 0, 1, \ldots, K - 1$, do

4: Given $h(x) = \frac{C_f^2 L_f}{2} \|x\|^2 + C_f L_g h_g(x)$, solve

$$x^{k+1} = \arg\min_{y \in \mathcal{X}} \left( w^k, y - x^k \right) + \frac{1}{\tau_k} D_h(y, x^k) \quad (14)$$

5: Sample $B_{g}^{k+1}$ and update

$$u^{k+1} = \frac{1}{|B_g^{k+1}|} \sum_{\xi \in B_g^{k+1}} \left( 1 - \beta_k \right) (u^k + g_\xi(x^{k+1} - g_\xi(x^k)) + \beta_k g_\xi(x^{k+1}) \right) \quad (15)$$

6: Sample $B_{V_f}^{k+1}$, $B_{V_g}^{k+1}$, update $v^{k+1} = v(x^{k+1}; B_{V_f}^{k+1})$, $s^{k+1} = s(u^{k+1}; B_{V_g}^{k+1})$ by (13b)-(13c)

7: Update $w^{k+1} = v^{k+1} s^{k+1}$

8: end for

function. The error of the inner function value estimate by $u^k$ is bounded in Lemma 2.3 which is established by [16]. For the completion of the paper, we also give the proof in Appendix A.2.

Lemma 2.3 (Lemma 1 in [16]) Under Assumptions 1 and 2, given $\beta_k \in (0, 1)$, sequences $\{x^k\}$ and $\{u^k\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}[\|g(x^{k+1}) - u^{k+1}\|^2 | \mathcal{F}_k] \leq (1 - \beta_k)^2 \|g(x^k) - u^k\|^2 + 4(1 - \beta_k)^2 C_g^2 \|x^{k+1} - x^k\|^2 + 2 \beta_k^2 \sigma^2 \|B_g^{k+1}\|. \quad (16)$$

Following the bound in Lemma 2.3, the error of the composition gradient estimate $w^k$ is bounded as shown in Lemma 2.4; the proof is provided in Appendix A.3.

Lemma 2.4 Under Assumption 1, the sequences $\{x^k\}$ and $\{w^k\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}[\|\nabla F(x^k) - w^k\|^2] \leq 2C_f^2 C_g^2 |B_{V_f}^k| + 2C_f^2 C_g^2 |B_{V_g}^k| + 2C_g^2 L_f^2 \mathbb{E}[\|g(x^k) - u^k\|^2]. \quad (17)$$

We investigate the sample complexity of the SoR algorithm by analyzing the merit function

$$V(x^k, u^k) \triangleq F(x^k) - F^* + \|g(x^k) - u^k\|^2, \quad (18)$$
which is denoted by $V^k$ in the rest of the paper. The per iteration merit function decrease is essential for the analysis and is presented in Lemma 2.5 with its proof in Appendix A.4.

**Lemma 2.5** Let $\{x^k, u^k, w^k\}$ be the sequence generated by Algorithm 1. Furthermore, let Assumptions 1 and 2 hold. Then

$$
\mathbb{E}[V^{k+1}|F_k] \leq V^k + \frac{\tau_k}{2C_k^2 L_f} \|\nabla F(x^k) - w^k\|^2 - \left(\frac{1}{\tau_k} - 2\right) D_h(\hat{x}^{k+1}, x^k) - \left(\frac{1}{\tau_k} - 1 - \frac{8}{L_f}\right) D_h(x^{k+1}, x^k) + ((1 - \beta_k)^2 - 1)\|g(x^k) - u^k\|^2 + 2\beta_k^2 \sigma^2 / |B_k^1|, \tag{19}
$$

where $\hat{x}^{k+1}$ is defined as

$$
\hat{x}^{k+1} \triangleq \arg\min_{y \in \mathcal{X}} \left\{\nabla F(x^k), y - x^k\right\} + \frac{2}{\tau_k} D_h(y, x^k).
$$

This lemma establishes the connection of the merit function $V^k$, two estimates’ errors $\|g(x^k) - u^k\|^2$ and $\|\nabla F(x^k) - w^k\|^2$, and the stationarity measure $D_h(\hat{x}^{k+1}, x^k)$—see Lemma 2.1. Hence, by properly choosing the parameters, we can derive the sample complexity of this algorithm to obtain an $\epsilon$-stationary solution, which is presented in Theorem 2.1, with the proof in Appendix A.5.

**Theorem 2.1** (Sample complexity of the SoR algorithm) Let $\{x^k\}$ be a sequence generated by Algorithm 1 with $0 < \tau_k < \min\{1/2, L_f/(L_f + 8)\}$ and $\beta_k \in (0, 1)$ satisfying $(1 - \beta_k)^2 + \tau_k L_f - 1 \leq 0$. Then, under Assumptions 1 and 2, we have

$$
\mathbb{E}\left[D_h(\hat{x}^{R+1}, x^R)/\tau_R^2\right] \leq \frac{V^0}{\sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2)} + \sum_{k=0}^{K-1} \frac{\tau_k C_k^2}{L_f |B_k^1|} + \frac{\tau_k C_k^2}{L_f |B_k^1|} + \frac{2\beta_k^2 \sigma^2}{|B_k^1|}, \tag{20}
$$

where $V^0 = F(x^0) - F^* + \|g(x^0) - u^0\|^2$, and the expectation is taken with respect to all random sequences generated by the algorithm and an independent random integer $R \in \{0, \ldots, K - 1\}$ with probability distribution

$$
P(R = k) = \frac{\tau_k - 2\tau_k^2}{\sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2)}.
$$

**Remark 2.3** When $\tau_k \in (0, 1/L_f)$, with $\beta_k \in (0, 1)$, the requirement $(1 - \beta_k)^2 + \tau_k L_f - 1 \leq 0$ results in $1 - \sqrt{1 - \tau_k L_f} \leq \beta_k < 1$. Furthermore, note that $\beta_k = \tau_k L_f$ satisfies the above bound as $\tau_k < \frac{1}{L_f}$, so $\tau_k L_f < 1$, and

$$
\beta_k - (1 - \sqrt{1 - \tau_k L_f}) = (1 + \sqrt{1 - \tau_k L_f})(1 - \sqrt{1 - \tau_k L_f}) - (1 - \sqrt{1 - \tau_k L_f}) = (1 - \sqrt{1 - \tau_k L_f})\sqrt{1 - \tau_k L_f} > 0.
$$
Remark 2.4 If we set $|B^k_{\nabla f}| = |B^k_{\nabla g}| \equiv 1$ for all $k \in \{0, \ldots, K - 1\}$, then
\[
\frac{1}{\sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2)} \sum_{k=0}^{K-1} \left( \frac{\tau_k C_f^2}{L_f |B^k_{\nabla f}|} + \frac{\tau_k C_f^2}{L_f |B^k_{\nabla g}|} \right) = \frac{2C_f^2}{L_f} \sum_{k=0}^{K-1} \tau_k \sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2) \geq \frac{2C_f^2}{L_f},
\]
and hence, without the mini-batch, the right-hand side of (20) cannot be smaller than an arbitrary $\epsilon > 0$, and the upper bound of (20) cannot be made small enough.

Fixing the step and batch sizes for all iterations, the sample complexity of the algorithm is presented in a more readable form in Corollary 2.1. The proof follows from Theorem 2.1.

Corollary 2.1 Given the assumptions of Theorem 2.1, setting $\tau_k \equiv \tau < \min\{1/2, L_f/(L_f + 8), 1/L_f\}$, $\beta_k \equiv L_f \tau$, $|B^k_{\nabla f}| = |B^k_{\nabla g}| \equiv \left[ \frac{4C_f^2}{(1-2\tau)L_f} \right]$, and $|B^k_g| = \left[ \frac{4\tau^2L_f^2\sigma^2_g}{(1-2\tau)^2} \right]$, we have
\[
\frac{1}{K} \sum_{k=0}^{K-1} E |D_h(\hat{x}^{k+1}, x^k)/\tau^2| \leq \frac{V^0}{K(\tau - 2\tau^2)} + \epsilon.
\]
Hence to achieve $\epsilon$-stationarity, we can set $K = O(\epsilon^{-1})$, which means that we need $O(\epsilon^{-2})$ calls to the $g_k$, $\nabla g_k$, and $\nabla f_\psi$ stochastic oracles.

Besides the stationarity error, we can show that the errors of the inner function estimate $u^k$ and the gradient of the composition estimate $w^k$ can also be bounded with the same rate, as shown in Corollary 2.2. The proof is in Appendix A.6.

Corollary 2.2 Under the setting of Corollary 2.1, we have
\[
\frac{1}{K} \sum_{k=0}^{K-1} E \|g(x^k) - u^k\|^2 \leq \frac{\|g(x^0) - u^0\|^2}{KL_f^2 \tau^2} + \frac{8V^0}{KL_f^2 (\tau - 8L_f \tau^2)} + \left( \frac{8(1 - 2\tau)}{L_f^2 (1 - 8L_f \tau)} + \frac{1 - 2\tau}{2L_f^2 \tau} \right) \epsilon,
\]
and
\[
\frac{1}{K} \sum_{k=0}^{K-1} E \|\nabla F(x^k) - w^k\|^2 \leq \frac{2C_f^2 \|g(x^0) - u^0\|^2}{K \tau^2} + \frac{16C_g^2 V^0}{KL_f \tau (L_f + 8\tau^2)} + \left( \frac{16C_g^2 (1 - 2\tau)}{L_f - (L_f + 8\tau)} + \frac{C_g^2 (1 - 2\tau)}{\tau} + \frac{\tau - 2\tau^2}{2} \right) \epsilon.
\]
Hence, after $O(\epsilon^{-2})$ calls to the stochastic oracles, we have $\frac{1}{K} \sum_{k=0}^{K-1} E \|u^k - \nabla g(x^k)\|^2 \leq O(\epsilon)$ and $\frac{1}{K} \sum_{k=0}^{K-1} E \|\nabla F(x^k) - w^k\|^2 \leq O(\epsilon)$. 

\(\copyright\) Springer
3 Relative-smooth of Relative-smooth (RoR) and Relative-smooth of Smooth (RoS) Compositions

This section considers the other two possible composition scenarios, namely relative-smooth of relative-smooth (RoR) and relative-smooth of smooth (RoS) compositions. First, Sect. 3.1 establishes the relative smoothness of the composition function for the RoR and RoS settings—Lemma 3.1. Next, Sect. 3.2 provides an algorithm that solves both RoR and RoS settings and analyzes its sample complexity. Finally, Sect. 3.3 provides a variance-reduced algorithm with improved sample complexity for the RoS composition.

3.1 Relative-smoothness of the Composition

We first introduce our assumptions. While Assumption 3 is for more general RoR composition, we also included the Assumption 4 for the RoS setting which is specifically used in Sect. 3.3 for the analysis of the variance-reduced algorithm.

Assumption 3 (Relative-smooth of Relative-smooth (RoR) composition) The functions $f_\psi, g_\xi, h_f$ satisfy the following conditions:

(a) The function $g_\xi$ is average $L_g$-smooth relative to $1$-strongly convex function $h_g$.
(b) The function $f_\psi$ is average $L_f$-smooth relative to convex function $h_f$.
(c) The stochastic gradients of $f_\psi, g_\xi$ are bounded in expectation.
(d) The function $h_f$ is $C_{hf}$-Lipschitz continuous, i.e., $\|\nabla h_f(u)\| \leq C_{hf}$, $\forall u \in \text{dom } h_f$.

Assumption 4 (Relative-smooth of Smooth (RoS) composition) The functions $f_\psi, g_\xi, h_f$ satisfy the following conditions:

(a) The function $g_\xi$ is average $L_g$-smooth.
(b) The function $f_\psi$ is average $L_f$-smooth relative to convex function $h_f$.
(c) The stochastic gradients of $f_\psi, g_\xi$ are bounded in expectation.
(d) The function $h_f$ is $C_{hf}$-Lipschitz continuous.

The following proposition from [63] is used in Lemma 3.1 to establish the convexity of the distance generating function of the composition $f(g(x))$ by showing its relative weak convexity (relative weak convexity is defined in the preliminary below Definition 1.2). Note that Lemma 3.1 is written for the more general RoR setting and it covers the RoS composition by setting $h_g(x) = \frac{1}{2} \|x\|^2$. The proof is provided in Appendix B.1.

Proposition 3.1 (Proposition 2.2(c) in [63]) Let $X$ be a nonempty closed convex set, $f : \mathbb{R}^d \to \mathbb{R}$ is closed convex and $C_f$-Lipschitz continuous, and $g : \mathbb{R}^n \to \mathbb{R}^d$ is $L_g$-smooth relative to $h$. Then, the composition $f \circ g : \mathbb{R}^n \to \mathbb{R}$ is $C_f L_g$-weakly convex relative to $h$.

Lemma 3.1 Under Assumption 3, $F(x)$ is $1$-smooth relative to

$$h(x) = (C_f L_g + C_{hf} L_f L_g) h_g(x) + L_f h_f(g(x)),$$
which is shown to be convex. Furthermore, if \( h_g(x) \) is 1-strongly convex, then \( h(x) \) is \( C_f L_g \)-strongly convex.

### 3.2 Proposed Algorithm for RoR (and RoS) Composition

In this section, we present an algorithm for the RoR composition which can also be used for the RoS composition, as a special case.

The distance generating function \( h(x) \) in Lemma 3.1 contains the composition \( h(g(x)) \). As \( g \) has the expectation form, one can only have access to stochastic approximation of the distance generating function \( h \) which are indeed biased. Note that this is not the case in the SoR setting, as its distance generating function is independent of the stochastic oracle—see Lemma 2.2. Therefore, in the RoR (and RoS) setting, we cannot evaluate \( D_h(y, x_k) \) exactly which is needed to solve (14).

Inspired by the idea of [62], we consider to approximate \( D_h(y, x_k) \) by linearizing the inner function \( g \). This linearization is supported by Lemma 3.2, which provides a new upper bound for \( F(y) \) when replacing \( g(y) \) by its linear approximation \( g(x) + (\nabla g(x), y - x) \) inside the composite distance generating function. The proof is available in Appendix B.2.

#### Lemma 3.2

Under Assumption 3, \( \forall x, y \in \mathcal{X} \), we have

\[
h_f(g(y)) \leq h_f(g(x) + (\nabla g(x), y - x)) + C_{h_f} L_g D_{h_g}(y, x). \tag{21}
\]

Furthermore, we have

\[
F(y) \leq F(x) + (\nabla F(x), y - x) + L_f D_{h_f}(g(x) + \nabla g(x)^T(y - x), g(x)) + \lambda D_{h_g}(y, x), \tag{22}
\]

where \( \lambda \triangleq C_f L_g + 2C_{h_f} L_f L_g \).

Lemma 3.2 provides an upper bound to \( F \) that can be iteratively minimized to obtain a stationary solution. If we have access to \( \nabla F(x^k) \), \( \nabla g(x^k) \), and \( g(x^k) \) or their approximations, then we can solve

\[
\tilde{x}^{k+1} = \arg\min_{y \in \mathcal{X}} \langle \nabla F(x^k), y - x^k \rangle + \frac{L_f}{\tau_k} D_{h_f}(g(x^k) + \nabla g(x^k)^T(y - x^k), g(x^k)) + \frac{\lambda}{\tau_k} D_{h_g}(y, x^k). \tag{23}
\]

The objective function of this subproblem is convex and always upper bounds \( F(\tilde{x}^{k+1}) \) if \( \tau_k \leq 1 \). However, given that functions \( f \) and \( g \) involve expectations, it is impossible to evaluate \( \nabla F(x^k) \), \( g(x^k) \), and \( \nabla g(x^k) \) in (23) exactly. Instead, Algorithm 2 proposes mini-batch estimates of these terms and then solves (24) iteratively, to obtain a stationary solution. The \( F_k \) for this algorithm is generated from

\[\{x^0, \ldots, x^k; u^0, \ldots, u^{k-1}; v^0, \ldots, v^{k-1}; s^0, \ldots, s^{k-1}\}\].

\[\text{Springer}\]
Lemma 3.4 upper-bounds the objective function at iteration $k + 1$ with that of iteration $k$ and some estimation errors. The result is then used in Lemma 3.5 to upper bound the running average of the distance between two consecutive iterates.

**Lemma 3.4** Let $\{x^k\}$ be the sequence generated by Algorithm 2, under Assumption 3, we have

$$f(g(x^{k+1})) \leq f(g(x^k)) + 2C_f \|g(x^k) - u_k\|^2 + \frac{C_f^2}{2} \|\nabla g(x^k) - v_k\|^2$$
\[ + \frac{1}{2} \| v_k^2 \| \nabla f(u^k) - s_k^2 - \left( \frac{\lambda}{\tau_k} - C_fL_g - 2 \right) D_{h_g}(x^{k+1}, x^k), \]

(26)

where \( \lambda = C_fL_g + 2Ch_fL_fL_g \).

The proof of the lemma is provided in Appendix B.3. Fixing the step sizes and batch sizes for all iterations and using Lemma 3.4, the sample complexity of obtaining two consecutive iterates with distance less than \( \epsilon \) on average (i.e., running average) is provided in Lemma 3.5, with the proof in Appendix B.4.

Lemma 3.5 Let Assumptions 2 and 3 (or 4) hold. Then, by setting \( \tau_k \equiv \tau < \min \left\{ 1, \frac{C_fL_g + 2Ch_fL_fL_g}{C_fL_g + 2} \right\} \), \( |B^k_s| = \left\lceil \frac{16C_f^2\sigma_g^2}{M_1^2\epsilon^2} \right\rceil \), and \( |B^k_{\nabla f}| = |B^k_{\nabla g}| = \left\lceil \frac{2C_f^2C_g}{M_1\epsilon} \right\rceil \), we have

\[
\frac{1}{K} \sum_{k=0}^{K-1} E \left[ D_{h_g}(x^{k+1}, x^k) / \tau^2 \right] \leq \frac{f(g(x^{0})) - F^*}{M_1 K} + \epsilon,
\]

where \( M_1 \triangleq (C_fL_g + 2Ch_fL_fL_g)\tau - (C_fL_g + 2)\tau^2 \).

Note that while \( h_f \) is not smooth, we assume that it is smooth on any bounded subset of \( \mathbb{R}^d \).

Assumption 5 The functions \( h_f \) and \( f \) are twice differentiable. Furthermore, \( \nabla h_f \) is \( L_{h_f} \)-Lipschitz continuous on any bounded subset of \( \mathbb{R}^d \).

Remark 3.1 The above assumption is also used in [12] to derive the convergence of their algorithm. Notice that following the setting of Lemma 3.5, for any \( \epsilon > 0 \) and finite \( K \), the Bregman distance between two consecutive iterates of Algorithm 2 is bounded in expectation; hence, the sequence \( \{x^k\}_{k=0}^K \) lies in a bounded set with probability 1.

Furthermore, since \( g \) is Lipschitz continuous, \( \{g(x^k)\} \) also lies in a bounded set with probability 1. With bounded \( \{g(x^k)\} \) and the assumption that the variance of \( g_k \) is bounded, \( \{|u^k|\} \) is also guaranteed to be bounded. Hence, the argument \( u \) in \( \nabla h_f(u) \) in Assumption 5 belongs to a bounded set for any sequence generated by Algorithm 2 with probability 1.

Lemma 3.6 Under Assumptions 3 (or 4) and 5, \( f \) is \( L_fL_{h_f} \)-smooth on any bounded subset of \( \mathbb{R}^d \).

The proof of Lemma 3.6 is provided in Appendix B.5. Lemmas 3.7 and 3.8 are needed to establish the upper bound on \( \|\tilde{x}^{k+1} - x^k\|^2 \) in Theorem 3.1. The proof of Lemma 3.7 is provided in Appendix B.6, while the proof of Lemma 3.8 is similar to that of Lemma 2.4 and is omitted.

Lemma 3.7 Let \( \tilde{x}^{k+1} \) be defined as in (23) with \( \lambda = C_fL_g + 2Ch_fL_fL_g \). Under Assumptions 2, 3 (or 4) and 5, we have

\[
\|\tilde{x}^{k+1} - x^k\|^2 \leq \frac{2\tau_k^2}{C_f^2L_g^2} \|w^k - \nabla F(x^k)\|^2 + \frac{8Ch_fL_f}{C_fL_g} \|u^k - g(x^k)\|.
\]
\[
+ \frac{2C_f^2 L_f L_h^2}{C_f C_f L_g^2} \|u^k - g(x^k)\|^2 + \frac{6C_h f L_f}{C_f L_g^2} \|v^k - \nabla g(x^k)\|^2 + \frac{4(C_f + 3C_h f L_f)}{C_f} D_{h^k} (x^{k+1}, x^k).
\]

(27)

**Lemma 3.8** Under Assumptions 3 (or 4) and 5, the sequences \(\{w^k\}\) and \(\{x^k\}\) satisfy

\[
\mathbb{E}[\|\nabla F(x^k) - w^k\|^2 | \mathcal{F}^k] \leq 2C_g^2 L_f^2 L_h^2 \sigma_g^2 / |B_g^k| + 2C_f^2 C_g^2 / |B_{\nabla F}^k| + 2C_f^2 C_g / |B_{\nabla g}^k|.
\]

Finally, we can establish the sample complexity of the RoR algorithm 2 to find an approximate stationary solution, which is presented in Theorem 3.1, with the proof in Appendix B.7.

**Theorem 3.1** (Sample complexity of the RoR algorithm) Let \(\{x^k\}\) be a sequence generated by Algorithm 2 with \(\tau_k \equiv \tau < \min \left\{ 1, \frac{C_f L_g + 2C_h f L_f}{C_f L_g + 2} \right\}\), \(|B_g^k| \equiv \left| \frac{16C_f^2 \sigma_g^2}{M_1 \epsilon^2} \right|\), and \(|B_{\nabla F}^k| = |B_{\nabla g}^k| = \left| \frac{2C_f^2 C_g^2}{M_1 \epsilon^2} \right|\), where \(M_1 = (C_f L_g + 2C_h f L_g L_f) \tau - (C_f L_g + 2) \tau^2\).

Then, under Assumptions 2, 3 (or 4), and 5, we have

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[ \|x^{k+1} - x^k\|^2 / \tau^2 \right] \leq \left( \frac{4C_f + 6C_h f L_f}{M_1 C_f} \right) f(g(x^0)) - F^* + \left( \frac{C_f^2 C_g^2}{4C_f^2 L_g^2} + \frac{C_f L_g L_h^2 M_1^2}{8C_h f C_f L_g^2 \tau^2} \right) \epsilon^2 + \left( \frac{4M_1}{C_f^2 L_g^2} + \frac{2C_h f L_f M_1}{C_f L_g \tau^2} + \frac{3C_h f L_f M_1}{C_f^2 L_g \tau^2} + 4 + \frac{12C_h f L_f}{C_f} \right) \epsilon.
\]

Hence, to obtain an \(\epsilon\)-stationarity solution, by setting \(K = \mathcal{O}(\epsilon^{-1})\), the RoR algorithm 2 requires \(\mathcal{O}(\epsilon^{-3})\) calls to \(g_{\xi}\) and \(\mathcal{O}(\epsilon^{-2})\) calls to \(\nabla f_{\phi}\) and \(\nabla g_{\xi}\) stochastic oracles.

### 3.3 Variance-Reduced Algorithm for the RoS Composition

When the inner function is Lipschitz smooth, i.e., Assumption 4, the RoR algorithm can be improved by variance reduction to estimate the inner function and its gradient. We use the stochastic path integrated differential estimator to estimate \(g\) and \(\nabla g\) [29, 49].

These updates are incorporated in Algorithm 3—see (30) and (31). The corresponding \(\mathcal{F}_j^k\) is now generated from

\[
\{x_j^0, \ldots, x_{j-1}^0, x_j^1, \ldots, x_j^k; u_j^0, \ldots, u_j^{j-2}, u_j^{j-1};
\]

\[
v_j^0, \ldots, v_j^{j-2}, v_j^1, \ldots, v_j^{j-1}; s_j^0, \ldots, s_{j-1}^0, s_j^{j-1}, s_{j-1}^1, \ldots, s_{j-1}^j\}.
\]

The error bounds for the new estimators are provided in the following lemma.
Lemma 3.9 (Lemma 1 in [29]) Let Assumptions 2 and 4 hold, and $u_j^k$ and $v_j^k$ be updated as in Algorithm 3. Then, $\{u_j^k\}$, $\{v_j^k\}$, and $\{x_j^k\}$ satisfy

\[
\mathbb{E}[\|u_j^k - g(x_j^k)\|^2] \leq \frac{\sigma^2_g}{|B_g^k|} + \frac{C_g^2}{|S_{y,x}^k,j|} \mathbb{E}[\|x_{r+1}^k - x_j^k\|^2].
\] (28)

\[
\mathbb{E}[\|v_j^k - \nabla g(x_j^k)\|^2] \leq \frac{C_g^2}{|B_g^k \nabla g|} + \frac{L_g^2}{|S_{y,x}^k,j|} \mathbb{E}[\|x_{r+1}^k - x_j^k\|^2].
\] (29)

**Algorithm 3** Variance-reduced RoS algorithm (RoS-VR)

**Require:** $x_0^k \in \mathcal{X}$, $\tau_k < 1$, $\lambda \triangleq C_f L_g + 2C_h f L_f h_f$

1: for $k = 0, 1, \ldots, K - 1$
2: for $j = 0, 1, \ldots, J - 1$
3: if $j == 0$
4: Sample $B_g^k$ and use (13a) to update $u_0^k = u(x_j^k; B_g^k)$
5: Sample $B_{g,y}^k$ and use (13b) to update $v_0^k = v(x_j^k; B_{g,y}^k)$
6: else
7: Sample $S_{g,y}^{k,j}$ and calculate

\[
u_j^k = u_j^{k-1} + \frac{1}{|S_{g,y}^{k,j}|} \sum_{\xi \in S_{g,y}^{k,j}} \left( g_\xi(x_j^k) - g_\xi(x_j^{k-1}) \right)
\] (30)

8: Sample $S_{g,y}^{k,j}$ and calculate

\[
\nu_j^k = v_j^{k-1} + \frac{1}{|S_{g,y}^{k,j}|} \sum_{\xi \in S_{g,y}^{k,j}} \left( \nabla g_\xi(x_j^k) - \nabla g_\xi(x_{j-1}^k) \right)
\] (31)

9: end if
10: Sample $B_{g,y}^{k,j}$ and use (13c) to update $s_j^k = s(u_j^k; B_{g,y}^{k,j})$
11: Update $w_j^k = v_j^k s_j^k$
12: Solve

\[
x_{j+1}^k = \arg\min_{y \in \mathcal{X}} \left\{ \langle w_j^k, y - x_j^k \rangle + \frac{L_f}{\tau_k} D_{y,x} (u_j^k + (v_j^k)^T (y - x_j^k), u_j^k) + \frac{\lambda}{2 \tau_k} \|y - x_j^k\|^2 \right\}
\] (32)

13: end for
14: Set $x_0^{k+1} = x_j^k$
15: end for
Furthermore, as $\forall \delta > 0$, we have

$$
\mathbb{E}[\|u^k_j - g(x^k_j)\|] \leq \sqrt{\frac{\sigma^2_g}{|B^k_g|}} + \sqrt{\frac{C^2_g}{|S^k_{g,r+1}|}} \mathbb{E}[\|x^k_{r+1} - x^k_r\|^2] \\
\leq \frac{\sigma_g}{\sqrt{|B^k_g|}} + \frac{\delta}{2} + \frac{1}{2\delta} \sum_{r=0}^{i-1} \frac{C^2_g}{|S^k_{g,r+1}|} \mathbb{E}[\|x^k_{r+1} - x^k_r\|^2],
$$

(33)

where the first inequality follows from Jensen’s inequality and the fact that $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$, and the second inequality is due to Young’s inequality. These inequalities are used in the following lemma to provide a new error bound for the new estimator of the gradient of the composition.

**Lemma 3.10** Under Assumptions 2 and 4, the sequences $\{w^k_j\}$ and $\{x^k_j\}$ generated by Algorithm 3 satisfy

$$
\mathbb{E}[\|\nabla F(x^k_j) - w^k_j\|^2] \leq 2C^2_g L^2_f L_{h_f}^2 \mathbb{E}[\|g(x^k_j) - u^k_j\|^2] + 2C^2_f \mathbb{E}[\|\nabla g(x^k_j) - v^k_j\|^2] + 2C^2_f C^2/g/|B^k_c/f|.
$$

The proof is similar to that of Lemma 2.4 with adjustment of the indices and is omitted here. We can now derive the sample complexity of Algorithm 3. First, we will show that the average Bregman distance of two consecutive points are bounded (Lemma 3.11, with the proof in Appendix B.8); hence, Assumption 5 is applicable. Next, under smoothness of $h_f$ on bounded subsets of $\mathbb{R}^d$, we derive the bound on $D_h(\tilde{x}^k_{j+1}, x^k_j)$ and the final sample complexity.

**Lemma 3.11** Under Assumptions 2 and 4, setting $\tau_k \equiv \tau \leq \frac{C_f L_g}{C_f L_g + 2}$, $S^k_{g,j}$ and $B^k_{g,j}$ satisfy

$$
S^k_{g,j} \equiv \begin{bmatrix} 6C^2_g C^2_f \tau \\ C_h L_f L_g M_{2\epsilon} \end{bmatrix}, \quad S_{\nabla g}^k \equiv \begin{bmatrix} 3C^2_f L_g \tau J \\ \frac{36C^2_f \sigma^2 g}{M^2_{2\epsilon}} \end{bmatrix}, \quad B^k_{g,j} \equiv \begin{bmatrix} 15C^2_f C^2_g \tau \\ \frac{15C^2_f C^2_g}{2M_{2\epsilon}} \end{bmatrix}, \quad B_{\nabla g}^k \equiv \begin{bmatrix} 15C^2_f C^2_g \tau \\ \frac{15C^2_f C^2_g}{2M_{2\epsilon}} \end{bmatrix},
$$

where $M_2 \leq \frac{C_f L_g}{2} - \left(\frac{C_f L_g}{2} + 1\right) \tau^2$, the sequence $\{x^k\}$ satisfies

$$
\frac{1}{KJ} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E}\left[\|x^k_{j+1} - x^k_j\|^2 / \tau^2\right] \leq \frac{f(g(x^k_0)) - F^*}{M_2 K J} + \epsilon.
$$

**Theorem 3.2** (Sample complexity of the RoS-VR algorithm) Under Assumptions 2, 4, and 5, following the setting of Lemma 3.11, the sequence $\{x^k\}$ generated by Algorithm 3 satisfies

$$
\frac{1}{KJ} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E}\left[\frac{\|x^k_{j+1} - x^k_j\|^2}{\tau^2}\right] \leq \frac{f(g(x^k_0)) - F^*}{M_2 K J} + \epsilon.
$$
\[ \leq \left( \frac{2A_0}{C_f L_g M_1} + \frac{A_1 C_{h_f} L_f \tau}{2C_f^2 M_1^2} + \frac{2A_1 C_{h_f} L_f \tau}{3C_f^3 M_1^3} \right) \frac{f(g(x_0^0)) - F^*}{K J} + \left( \frac{A_1 C_{h_f} L_f \tau}{2C_f^2 M_1} + \frac{2A_3 C_{h_f} L_f \tau}{3C_f^3 M_1} + \frac{2A_1 M_2}{15C_f^2 L_g M_1} + \frac{4A_3 M_2}{15C_f^3 C_g^2 L_g M_1} \right) \epsilon + \frac{A_2 C_{h_f} L_f M_2 \tau}{3C_f^3 M_1^3} \frac{(f(g(x_0^0)) - F^*) \epsilon}{K J} + \left( \frac{A_2 C_{h_f} L_f M_2 \tau}{3C_f^3 M_1} + \frac{A_2 M_2}{18C_f^2 L_g M_1} \right) \epsilon^2, \]

where the coefficients are defined in the proof.

The proof is provided in Appendix B.9. By setting \( K = J = O(\epsilon^{-1/2}) \), the RoS-VR algorithm 3 achieves its best sample complexity which is \( O(\epsilon^{-5/2}) \) for the \( g_{\xi} \) oracle, \( O(\epsilon^{-3/2}) \) for the \( \nabla g_{\xi} \) oracle, and \( O(\epsilon^{-2}) \) for the \( \nabla f_{\psi} \) oracle.

### 4 Applications and Numerical Experiments

Below, we evaluate the performance of the proposed algorithms over two different applications and compared them with two other algorithms for stochastic composition optimization. All implementations are performed in Python and are available at https://github.com/samdavanloo/BSC.

#### 4.1 Risk-Averse optimization (SoR case)

The mean-variance risk-averse optimization problem (5) can be written as

\[
\min_{x \in \mathcal{X}} -\mathbb{E} \left[ r_{\xi}(x) \right] + \lambda \left( \mathbb{E} \left[ (r_{\xi}(x))^2 \right] - \mathbb{E}^2 \left[ r_{\xi}(x) \right] \right),
\]

which is an instance of two-level stochastic composition problem (1) with

\[
g(x) : \mathbb{R}^n \to \mathbb{R}^2 \triangleq \mathbb{E} \left[ r_{\xi}(x), \ (r_{\xi}(x))^2 \right], \quad f(u_1, u_2) : \mathbb{R}^2 \to \mathbb{R} \triangleq -u_1 + \lambda u_2 - \lambda u_1^2.
\]

Following our discussion in Example 1.2, we let the random reward to be a quadratic function \( r_{\xi}(x) \triangleq \frac{1}{2} x^T A_{\xi} x \) where \( A_{\xi} \) is a symmetric matrix. Furthermore, we let the feasible set to be \( \mathcal{X} \triangleq \{ x \in \mathbb{R} \mid \| x \| \leq R \} \) for some \( R > 0 \). Note that while \( f \) is quadratic and hence smooth, \( g(x) \) does not have Lipschitz continuous Jacobian as \( g_{\xi}(x) \) involves fourth-degree polynomial \( (r_{\xi}(x))^2 \). However, the following lemma shows that \( (r_{\xi}(x))^2 \) is relatively smooth which we use in Lemma 4.2 to show that \( g_{\xi} \) is average relatively smooth. Hence, the objective of (34) is an instance of the SoR composition.

**Lemma 4.1** (Lemma 5.1 in [12]) For any \( L \geq 3\| A \|^2 \), \( f(x) = \frac{1}{4}(x^T A x)^2 \) is \( L \)-smooth relative to \( h(x) = \frac{1}{4}\| x \|^4 \).
Lemma 4.2 The vector-valued function \( g_\xi(x) = \left[ \frac{1}{4} x^T A_\xi x - \frac{1}{4} x^T A_\xi x \right] \) is average 1-smooth relative to \( h_g(x) = \frac{k_1}{2} \| x \|^2 + \frac{k_2}{4} \| x \|^4 \), where \( k_1 \geq \| E[A_\xi] \| \) and \( k_2 \geq 3 \| A_\xi \|^2 \).

Remark 4.1 Note that when the feasible set \( \mathcal{X} \) is compact, the inner function \( g \) is also smooth, but its smoothness constant depends on the diameter of \( \mathcal{X} \) since \( \| \nabla^2 g_2(x) \| \leq \sup_{x \in \mathcal{X}} 3 \| A_\xi \|^2 \| x \|^2 = O(R^2) \).

For the boundedness of the stochastic gradient of \( g_\xi \) in expectation, we have
\[
\mathbb{E} \left[ \| \nabla g_\xi(x) \|^2 \right] \leq \mathbb{E} \left[ \| \nabla g_\xi(x) \|^2 \right] = \mathbb{E} \left[ \| A_\xi x \|^2 \right] + \mathbb{E} \left[ \| x^T A_\xi x \|^2 \right] \leq \mathbb{E} \left[ \| A_\xi \|^2 \right] R^2 + \mathbb{E} \left[ \| A_\xi \|^4 \right] R^6.
\]
Hence, \( C_g^2 \geq \mathbb{E} \left[ \| A_\xi \|^2 \right] R^2 + \mathbb{E} \left[ \| A_\xi \|^4 \right] R^6 \). Furthermore, the gradient and Hessian of \( f \) are
\[
\nabla f(u_1, u_2) = \begin{bmatrix} -1 - 2\lambda u_1 \\ \lambda \end{bmatrix}, \quad \nabla^2 f(u_1, u_2) = \begin{bmatrix} -2\lambda & 0 \\ 0 & 0 \end{bmatrix}.
\]

With \( A \triangleq \mathbb{E}[A_\xi] \), we have
\[
\| \nabla f(u_1, u_2) \|^2 = (1 + 2\lambda u_1)^2 + \lambda^2 \leq 1 + \lambda^2 \| A \|^2 \| x \|^4 + 2\lambda \| A \| \| x \|^2 + \lambda^2 \\
\leq 1 + \lambda^2 R^4 \| A \|^2 + 2\lambda R^2 \| A \| + \lambda^2,
\]
and \( \| \nabla^2 f(u_1, u_2) \| = 2\lambda \). So \( C_f^2 \geq 1 + \lambda^2 R^4 \| A \|^2 + 2\lambda R^2 \| A \| + \lambda^2 \). Finally, the average smoothness of \( f_\varphi \) and average relative smoothness of \( g_\xi \) follow with \( L_f = 2\lambda \) and \( L_g = 1 \). From the discussion above and Lemma 2.2, we have \( f(g(x)) \) is 1-smooth relative to \( h(x) = C_f^2 L_f + C_f L_\xi \| A \| \| x \|^2 + \frac{3 C_f L_\xi \| A \|^2}{4} \| x \|^4 \). Note that the above discussion provides loose bounds on \( C_f, C_g \). In the experiments, we use the distance generating function for the composition as
\[
h(x) \triangleq \frac{c_1}{2} \| x \|^2 + \frac{c_2}{4} \| x \|^4,
\]
with \( c_1, c_2 \) determined by grid search. To do so, we generate grids \( C_1, C_2 \) with 6 logarithmically spaced values in \([10^{-2}, 10^2]\) for \( C_1 \) and in \([10^{-2}, 10]\) for \( C_2 \). Next, for all combinations \((c_1, c_2) \in C_1 \times C_2 \), we run the algorithm for a large enough number of iterations to find the best combination \((c_1^*, c_2^*) \), and, hence, the distance generating function.

In this experiment, we consider solving (34) in a finite-sum form, i.e., the expectations are available as finite sums divided by the number of components. To simulate the data, we first randomly generate a \( 50 \times 50 \) symmetric matrix \( A \). Then, we generate 1000 random noise matrices of size \( 50 \times 50 \) with independent elements from standard normal distribution and add them to \( A \) to construct 1000 samples of \( A_\xi \). We solve (34) with \( \lambda \) equal to 10 and the radius of the feasible set \( R \) equal to 10.
Following Corollary 2.1, the step sizes are set to $\tau_k \equiv 0.025$ and $\beta_k \equiv 0.5$. The grid search over the coefficients of (35) results in $(c_1^*, c_2^*) = (15.8, 6.3 \times 10^{-3})$.

To validate the sample complexity of our algorithm, we run the experiment with three different batch sizes $|B^g|, |B^f| \in \{1, 10, 100\}$. Note that the number of calls to the inner and outer function stochastic gradient oracles is set to be equal in Corollary 2.1 (which we named $|B_V|$). For each setting, we replicate the algorithm 20 times from a fixed initial point.

Figure 1 shows the decay of the stationarity measure and function value versus the iteration and the number of stochastic gradient calls where the solid lines are the average of the 20 replicates.

Note that based on Corollary 2.1 the theoretical sample complexity for $E[D_h(\tilde{x}_{R+1}^k, x_{R}^k)]$ is $O(K^{-1} + |B_V|^{-1})$. So, for a fixed iteration, the upper bound on stationarity measure is determined by the batch size inverse as verified by the first two plots of Fig. 1. Comparing the decay of the stationarity measure from sample complexity perspective, we see that $|B_V| = 10$ performs the best followed by $|B_V| = 100$. We should also note that with $|B_V| = 1$ the algorithm has significantly higher variance compared to the bigger two batch sizes. Finally, with $|B_V| = 10$ the algorithm is able to obtain lower values of the stationarity measure compared to the other two batch sizes.

We also compare the SoR algorithm with NASA [33] and SCSC [16] which are two of the state-of-the-art algorithms for stochastic composition optimization. The batch sizes of the SCSC algorithm are set to 100 and all batch sizes of the NASA algorithm are set to 1 which is supported by their theory. For parameter turning, we also do a grid search and determine the best setting based on the decay in the stationarity measure. Figure 2 illustrates the decay of two stationarity measures and function values by different methods. The second stationarity measure is based on $\bar{x}^{k+1}$ defined as

$$\bar{x}^{k+1} \triangleq \arg\min_{y \in X} \left\{ \nabla F(x^k), y - x^k \right\} + \frac{1}{2\tau_k} \|y - x\|^2,$$

which is the Euclidean counterpart of $\tilde{x}^{k+1}$ defined in (11). SoR and SCSC have similar performance while NASA decays slower across all three measures.

### 4.2 Policy Evaluation for MDP (RoS case)

Following the discussion in Example 1.1, in this section, we solve (4) when the reward is of count type (e.g., number of clicks). It is common in this setting to take the distance function to be the KL divergence $D_{KL}(a, b) = \sum_i a_i \log \frac{a_i}{b_i} + b_i - a_i$ with $a_i, b_i > 0$ which better captures the information when the random noise follows the Poisson distribution.

The problem is formulated in (4) which is an instance of the stochastic composition problem (1) with
Fig. 1 Decay of the stationarity measure $E[D_h(\hat{x}^{R+1}, x^R)]$ and expected function value $E[F(x^R)]$ versus iteration and stochastic gradient oracle calls by the SoR algorithm 1 for the risk-averse optimization problem.
Fig. 2 Decay of the two stationarity measures and expected function values by the SoR algorithm [1], NASA [33] and SCSC [16] for the risk-averse problem.
We first generate a random matrix $A$ with independent elements from standard normal distribution and set the negative elements to zero to create a matrix $A$.

In the RoS-VR algorithm, the maximum iteration numbers in the two nested loops, i.e., $K$ and $J$, are set equal to each other and equal to the square root of the bigger batch sizes. In this experiment, we solve (4) by both the RoR 2 and RoS-VR 3 algorithms.

Similar to the SoR experiment, we consider a finite-sum setting. To simulate data, we first generate a random matrix $A$ with independent elements $A_{ij} \sim U[0,2]$, $i = 1, \cdots, 50$, $j = 1, \cdots, 30$, random vector $r$ with independent elements $r_i \sim \text{Poisson}(0.5)$. Next, we generate 30,000 random noise matrices with independent elements from standard normal distribution and add them to $A$ and set the negative elements to zero to create $A^+$. The parameters $L_f/\tau$ and $\lambda/\tau$ in (24) and (32) are determined by grid search with a grid of 6 logarithmically spaced values in $[10^{-5}, 10^2]$. The RoR algorithm 2 is run with two different batch sizes $|B_{\psi}| = 100$, $|B_\varphi| = 10^4$ and $|B_{\psi}| = 20$, $|B_\varphi| = 400$. Similarly, the RoS-VR algorithm 3 is run with two different batch sizes $|B_{\psi}| = 100$, $|B_\varphi| = 10^4$, $|S_{\psi}| = 10$, $|S_\varphi| = 100$ and $|B_{\psi}| = 20$, $|B_\varphi| = 400$, $|S_{\psi}| = 5$, $|S_\varphi| = 25$ (the bigger batch sizes for $j = 0$ match the setting of the RoR algorithm and the smaller batch sizes for $j > 0$ are roughly the square root of the bigger batch sizes). In the RoS-VR algorithm, the maximum iteration numbers in the two nested loops, i.e., $K$ and $J$, are set equal to each other and equal to the square root of the

$$g(x) = \mathbb{E}[A^+_\xi]x, \quad f(u) = \sum_{i=1}^{s} u_i - \mathbb{E}[r_\varphi]_i \log u_i, \quad (36)$$

where $s \triangleq |\mathcal{Y}|$. Note that in (36), $g_\xi$ is average Lipschitz smooth and $f_\psi$ is average smooth relative to $h_j(u) = -\sum_{i=1}^{j} \log u_i$; hence, this is an instance of the RoS composition. As RoS is a subset of the RoR composition, in this experiment, we solve (4) by both the RoR 2 and RoS-VR 3 algorithms.

Fig. 3 Decay of the stationarity measure $\mathbb{E}[D_h(\tilde{x}^{R+1}, x^R)]$ and expected function value $\mathbb{E}[F(x^R)]$ versus the inner/outer gradient and inner function value oracle calls by the RoR 2 (circle marker) and RoS-VR 3 (square marker) algorithms for the policy evaluation problem. Bigger and smaller batch sizes are shown with solid and dashed lines, respectively.
Fig. 4 Decay of the stationarity measures and expected function values by RoR 2, RoS-VR 3, NASA [33] and SCSC [16] for the policy evaluation problem.
maximum iteration number in the RoR algorithm. Both algorithms are replicated 20 times from a fixed initial point for each scenario.

Figure 3 shows the decay of the stationarity measure and expected function value versus the inner/outer gradient and inner function value stochastic oracle calls by the RoR 2 and RoS-VR 3 algorithms over 20 replicates. Results show that the decays are faster by the variance-reduced RoS-VR algorithm compared to the RoR algorithm which supports our theoretical findings.

We also compare the two algorithms with NASA [33] and SCSC [16]. The results are shown in Fig. 4. The proposed algorithms perform generally better than the other two methods with respect to the iteration and the inner/outer gradient stochastic oracle calls. However, NASA decays faster with respect to the inner function stochastic oracle calls which is expected given its single-batch oracle call per iteration. With respect to iteration, RoR has a good performance relative to RoS-VR. But, with respect to the number of stochastic oracle calls, RoS-VR results in faster decay across different measures. For most of the settings, SCSC could not obtain low values of the stationarity measure.

5 Conclusions
In this paper, we study the two-level stochastic composition problem in the absence of Lipschitz continuity of the gradient of the inner, outer, or both functions. Given the notion of relative smoothness for the single-level deterministic problems, we consider three compositions: smooth of relative-smooth (SoR), relative-smooth of smooth (RoS), and relative-smooth of relative-smooth (RoR) compositions. We then propose one algorithm to solve the SoR and one algorithm to solve the RoR and RoS compositions. We further improve the second algorithm by variance reduction for the RoS composition. We then investigate the iteration and sample complexities of the three proposed algorithms. Finally, we evaluate the performances of these algorithms over two numerical experiments and compare them with other methods.

Data availability statement The data generated and analyzed in this paper can be entirely reproduced by running our code which is publicly available on the corresponding authors’ GitHub Web site at https://github.com/samdavanloo/BSC.

A SoR Proofs

A.1 Proof of Lemma 2.2

By $L_f$-smoothness of $f$, $\forall g(x), g(y) \in \text{dom } f$, we have

$$\frac{L_f}{2} \| g(x) - g(y) \|^2$$

$$\geq | f(g(x)) - f(g(y)) - \langle \nabla f(g(y)), g(x) - g(y) \rangle |$$

$$\geq | f(g(x)) - f(g(y)) - \langle \nabla g(y) \nabla f(g(y)), x - y \rangle |$$
where the last inequality holds because \( f \) is \( C_f \)-Lipschitz continuous and \( g \) is \( L_g \)-smooth relative to \( h_g \). Rearranging the above result and using the property that \( g \) is \( C_g \)-Lipschitz continuous, with \( h(\mathbf{x}) \triangleq \frac{C_L f^2}{2} \|\mathbf{x}\|^2 + C_f L_g h_g(\mathbf{x}) \), we have

\[
|f(\mathbf{g}(\mathbf{x})) - f(\mathbf{g}(\mathbf{y})) - \langle \nabla g(\mathbf{y}) \nabla f(\mathbf{g}(\mathbf{y})), \mathbf{x} - \mathbf{y} \rangle| \leq D_h(\mathbf{x}, \mathbf{y}).
\]

### A.2 Proof of Lemma 2.3

Define \( \bar{g}_\xi(\cdot) \triangleq \frac{1}{|B_{k+1}^g|} \sum_{\xi \in B_{k+1}^g} \bar{g}_\xi(\cdot) \), we have \( \mathbb{E}[\|\bar{g}_\xi(\mathbf{x}^{k+1}) - g(\mathbf{x}^{k+1})\|^2 |\mathcal{F}_k] \leq \sigma_g^2 |B_{k+1}^g| \). Furthermore, by (15), we have

\[
\mathbb{E}[\|g(\mathbf{x}^{k+1}) - \mathbf{u}^{k+1}\|^2 |\mathcal{F}_k] \\
\quad = \mathbb{E}[\|g(\mathbf{x}^{k+1}) - \mathbf{u}^{k+1}\|^2 + (1 - \beta_k)g(\mathbf{x}^{k+1}) - g(\mathbf{x}^{k+1})\|^2 |\mathcal{F}_k] \\
\quad \leq (1 - \beta_k)^2\|g(\mathbf{x}^{k+1}) - \mathbf{u}^{k+1}\|^2 + \mathbb{E}[\|g(\mathbf{x}^{k+1}) - g(\mathbf{x}^{k+1})\|^2 |\mathcal{F}_k] \\
\quad \leq (1 - \beta_k)^2\|g(\mathbf{x}^{k+1}) - \mathbf{u}^{k+1}\|^2 + 2(1 - \beta_k)^2\|g(\mathbf{x}^{k+1}) - g(\mathbf{x}^{k+1})\|^2 |\mathcal{F}_k] \\
\quad + 4(1 - \beta_k)\beta_k\mathbb{E}[\|\bar{g}_\xi(\mathbf{x}^{k+1}) - \bar{g}_\xi(\mathbf{x}^{k+1})\| |\mathcal{F}_k] \\
\quad + 2(1 - \beta_k)^2\mathbb{E}[\|\bar{g}_\xi(\mathbf{x}^{k+1}) - \bar{g}_\xi(\mathbf{x}^{k+1})\|^2 |\mathcal{F}_k] \\
\quad \leq (1 - \beta_k)^2\|g(\mathbf{x}^{k+1}) - \mathbf{u}^{k+1}\|^2 + 4(1 - \beta_k)^2C_g^2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + 2\beta_k^2\sigma_g^2 |B_{k+1}^g|,
\]

where (a) is based on the Lipschitz continuity of \( g \).

### A.3 Proof of Lemma 2.4

By Algorithm 1, we have

\[
\|\nabla F(\mathbf{x}^k) - \mathbf{w}^k\|^2 \\
\quad = \|\nabla g(\mathbf{x}^k) \nabla f(\mathbf{g}(\mathbf{x}^k)) - \nabla g(\mathbf{x}^k) \mathbf{s}^k + \nabla g(\mathbf{x}^k) \mathbf{s}^k - \mathbf{v}^k \mathbf{s}^k\|^2 \\
\quad \leq 2\|\nabla g(\mathbf{x}^k)\|^2 \|\nabla f(\mathbf{g}(\mathbf{x}^k)) - \mathbf{s}^k\|^2 + 2\|\nabla g(\mathbf{x}^k) - \mathbf{v}^k\|^2 \|\mathbf{s}^k\|^2 \\
\quad \leq 2C_g^2 \|\nabla f(\mathbf{g}(\mathbf{x}^k)) - \nabla f(\mathbf{u}^k) + \nabla f(\mathbf{u}^k) - \mathbf{s}^k\|^2 + 2\|\nabla g(\mathbf{x}^k) - \mathbf{v}^k\|^2 \|\mathbf{s}^k\|^2 \\
\quad \leq 2C_g^2L_f^2\|\mathbf{g}(\mathbf{x}^k) - \mathbf{u}^k\|^2 + 2C_g^2 \|\nabla f(\mathbf{u}^k) - \mathbf{s}^k\|^2 \\
\quad + 4C_g^2 \|\nabla f(\mathbf{g}(\mathbf{x}^k)) - \nabla f(\mathbf{u}^k) + \nabla f(\mathbf{u}^k) - \mathbf{s}^k\|^2 + 2\|\nabla g(\mathbf{x}^k) - \mathbf{v}^k\|^2 \|\mathbf{s}^k\|^2,
\]
where the first inequality follows from the identity \((a + b)^2 \leq 2a^2 + 2b^2\), the second inequality uses the Lipschitz continuity of \(g\), and the last inequality uses the smoothness of \(f\). Taking expectation on both sides and using the tower property in the three identities

\[
\mathbb{E}[\|\nabla f(u^k) - s_k\|^2] = \mathbb{E}_{u^k}\left[\mathbb{E}[\|\nabla f(u^k) - s_k\|^2|u^k]\right] \leq C_f^2/|B_{\nabla f}|, \\
\mathbb{E}[\langle \nabla f(g(x^k)) - \nabla f(u^k), \nabla f(u^k) - s_k\rangle] = \mathbb{E}_{u^k}\left[\mathbb{E}[\langle \nabla f(g(x^k)) - \nabla f(u^k), \nabla f(u^k) - s_k\rangle|u^k]\right] = 0,
\]
and

\[
\mathbb{E}[\|\nabla g(x^k) - v_k\|^2] = \mathbb{E}_{u^k}\left[\mathbb{E}[\|\nabla g(x^k) - v_k\|^2|u^k]\right] \\
\leq C_g C_f^2/|B_{\nabla g}|,
\]
completes the proof.

**A.4 Proof of Lemma 2.5**

By Definition 1.2 and Lemma 2.2, \(\forall x, y, z \in \mathcal{X}\), we have

\[
F(x) - F(y) \leq \langle \nabla F(x), x - y \rangle + D_h(y, x), \\
F(z) - F(x) \leq \langle \nabla F(x), z - x \rangle + D_h(z, x).
\]

Summing the above two inequalities, we get

\[
F(z) - F(y) \leq \langle \nabla F(x), z - y \rangle + D_h(y, x) + D_h(z, x). \tag{37}
\]

By Lemma 1.2, \(\forall y \in \mathcal{X}\), we have

\[
\tau_k(w^k, x^{k+1} - y) \leq D_h(y, x^k) - D_h(x^{k+1}, x^k) - D_h(y, x^{k+1}), \tag{38}
\]
and

\[
\frac{\tau_k}{2} \langle \nabla F(x^k), \hat{x}^{k+1} - y \rangle \leq D_h(y, x^k) - D_h(\hat{x}^{k+1}, x^k) - D_h(y, \hat{x}^{k+1}). \tag{39}
\]

Letting \(z = x^{k+1}, x = x^k, y = \hat{x}^{k+1}\) in (37) and (38), and summing them up, we have

\[
F(x^{k+1}) - F(\hat{x}^{k+1}) \leq \langle \nabla F(x^k) - w^k, x^{k+1} - \hat{x}^{k+1} \rangle + \left(1 + \frac{1}{\tau_k}\right)D_h(\hat{x}^{k+1}, x^k) \\
+ \left(1 - \frac{1}{\tau_k}\right)D_h(x^{k+1}, x^k) - \frac{1}{\tau_k}D_h(\hat{x}^{k+1}, x^{k+1}).
\]
Letting $z = x^{k+1}$, $x = x^k$, $y = x^k$ in (37) and (39), summing them up,

$$F(\hat{x}^{k+1}) - F(x^k) \leq \left( 1 - \frac{2}{\tau_k} \right) D_h(\hat{x}^{k+1}, x^k) - \frac{2}{\tau_k} D_h(x^k, \hat{x}^{k+1}).$$

Adding the last two inequalities, we get

$$F(x^{k+1}) - F(x^k) \leq \langle \nabla F(x^k) - w^k, x^{k+1} - \hat{x}^{k+1} \rangle + \left( 2 - \frac{1}{\tau_k} \right) D_h(\hat{x}^{k+1}, \hat{x}^k) + \left( 1 - \frac{1}{\tau_k} \right) D_h(x^{k+1}, x^k) - \frac{2}{\tau_k} D_h(x^k, \hat{x}^{k+1})$$

$$\leq \frac{\tau_k}{2\mu} \|\nabla F(x^k) - w^k\|^2 + \frac{\mu}{2\tau_k} \|x^{k+1} - \hat{x}^{k+1}\|^2 + \left( 2 - \frac{1}{\tau_k} \right) D_h(\hat{x}^{k+1}, x^k)$$

$$+ \left( 1 - \frac{1}{\tau_k} \right) D_h(x^{k+1}, x^k) - \frac{2}{\tau_k} D_h(x^k, \hat{x}^{k+1})$$

$$\leq \frac{\tau_k}{2\mu} \|\nabla F(x^k) - w^k\|^2 + \frac{1}{\tau_k} D_h(\hat{x}^{k+1}, x^k) + \left( 2 - \frac{1}{\tau_k} \right) D_h(\hat{x}^{k+1}, x^k)$$

$$+ \left( 1 - \frac{1}{\tau_k} \right) D_h(x^{k+1}, x^k) - \frac{2}{\tau_k} D_h(x^k, \hat{x}^{k+1})$$

$$\leq \frac{\tau_k}{2\mu} \|\nabla F(x^k) - w^k\|^2 + \left( 2 - \frac{1}{\tau_k} \right) D_h(\hat{x}^{k+1}, x^k) + \left( 1 - \frac{1}{\tau_k} \right) D_h(x^{k+1}, x^k),$$

(40)

where the second inequality is due to Young’s inequality and the third inequality holds because $h$ is $\mu$-strongly convex.

Adding $-F(x^k) + \|g(x^{k+1}) - u^{k+1}\|^2$ on both sides of the inequality (40), taking expectation conditioned on $F_k$, we get

$$\mathbb{E}[F(x^{k+1}) - F(x^k) + \|g(x^{k+1}) - u^{k+1}\|^2 | F_k]$$

$$\leq F(x^k) - F(x^k) + \|g(x^k) - u^k\|^2 + \frac{\tau_k}{2\mu} \|\nabla F(x^k) - w^k\|^2 - \left( \frac{1}{\tau_k} - 2 \right) D_h(\hat{x}^{k+1}, x^k)$$

$$- \left( \frac{1}{\tau_k} - 1 \right) D_h(x^{k+1}, x^k) + \mathbb{E}[\|g(x^{k+1}) - u^{k+1}\|^2 | F_k] - \|g(x^k) - u^k\|^2.$$

Using the definition of $V^k$, Lemmas 2.3 and the fact that $\mu = C^2_{g}L_f$, we have

$$\mathbb{E}[V^{k+1}|F_k]$$

$$\leq V^k + \frac{\tau_k}{2C^2_{g}L_f} \|\nabla F(x^k) - w^k\|^2 - \left( \frac{1}{\tau_k} - 2 \right) D_h(\hat{x}^{k+1}, x^k) - \left( \frac{1}{\tau_k} - 1 \right) D_h(x^{k+1}, x^k)$$

$$+ ((1 - \beta_k)^2 - 1)\|g(x^k) - u^k\|^2 + 4(1 - \beta_k)^2 C^2_g \|x^{k+1} - x^k\|^2 + 2\beta_k^2 \sigma^2_{g} |g^k_{x+1}|.$$

Հ Springer
\begin{align*}
&\leq V_k + \frac{\tau_k}{2C_g^2L_f} \| \nabla F(x^k) - w^k \|^2 - \left( \frac{1}{\tau_k} - 2 \right) D_h(\hat{x}^{k+1}, x^k) - \left( \frac{1}{\tau_k} - 1 - \frac{8}{L_f} \right) D_h(x^{k+1}, x^k) \\
+& (1 - \beta_k)^2 - 1)g(x^k) - u^k\|_2^2 + 2\beta_k^2\sigma^2 / |B_g^{k+1}|, \\
\end{align*}

where the second inequality holds because $h$ is $\mu$-strongly convex.

**A.5 Proof of Theorem 2.1**

Taking expectation of (19) with respect to the random sequences generated by the algorithm and using the tower property, we get

\[
\mathbb{E}[V^{k+1}] \leq \mathbb{E}[V^k] + \frac{\tau_k}{2C_g^2L_f} \mathbb{E}[\| \nabla F(x^k) - w^k \|^2] - \left( \frac{1}{\tau_k} - 2 \right) \mathbb{E}[D_h(\hat{x}^{k+1}, x^k)] \\
- \left( \frac{1}{\tau_k} - 1 - \frac{8}{L_f} \right) \mathbb{E}[D_h(x^{k+1}, x^k)] \\
+ (1 - \beta_k)^2 - 1) \mathbb{E}[\| g(x^k) - u^k \|^2] + 2\beta_k^2\sigma^2 / |B_g^{k+1}|.
\]

Applying Lemma 2.4 to the above inequality, under the assumption $(1 - \beta_k)^2 + \tau_k L_f - 1 \leq 0$ with $0 < \tau_k < \min\{1/2, L_f/(L_f + 8)\}$, $\beta_k \in (0, 1)$, we have

\[
\mathbb{E}[V^{k+1}] \leq \mathbb{E}[V^k] - \frac{1 - 2\tau_k}{\tau_k} \mathbb{E}[D_h(\hat{x}^{k+1}, x^k)] + \frac{\tau_k C_g^2}{L_f} \left( 1/|B_{V_f}^k| + 1/|B_{V_g}^k| \right) \\
+ (1 - \beta_k)^2 + \tau_k L_f - 1) \mathbb{E}[\| g(x^k) - u^k \|^2] + 2\beta_k^2\sigma^2 / |B_g^{k+1}| \\
\leq \mathbb{E}[V^k] - \frac{1 - 2\tau_k}{\tau_k} \mathbb{E}[D_h(\hat{x}^{k+1}, x^k)] \\
+ \frac{\tau_k C_g^2}{L_f} \left( 1/|B_{V_f}^k| + 1/|B_{V_g}^k| \right) + 2\beta_k^2\sigma^2 / |B_g^{k+1}|.
\]

Rearranging the terms in the above inequality, telescoping from $k = 0, \ldots, K - 1$ and using the fact that $V^{k+1} \geq 0$, we have

\[
\sum_{k=0}^{K-1} (\tau_k - 2\tau_k^2) \mathbb{E}\left[ D_h(\hat{x}^{k+1}, x^k)/\tau_k^2 \right] \leq V^0 + \sum_{k=0}^{K-1} \frac{\tau_k C_g^2}{L_f|B_{V_f}^k|} + \frac{\tau_k C_g^2}{L_f|B_{V_g}^k|} + \frac{2\beta_k^2\sigma^2}{|B_g^{k+1}|}.
\]

Dividing both sides by $\sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2)$ and using the fact that

\[
\mathbb{E}\left[ D_h(\hat{x}^{k+1}, x^k)/\tau_k^2 \right] = \frac{\sum_{k=0}^{K-1} (\tau_k - 2\tau_k^2) \mathbb{E}\left[ D_h(\hat{x}^{k+1}, x)/\tau_k^2 \right]}{\sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2)},
\]

 Springer
we get

\[
\mathbb{E} \left[ D_h(\tilde{x}^{R+1}, x^R) / \tau_R^2 \right] \leq \frac{V^0}{\sum_{j=0}^{K-1} (\tau_j - 2\tau_j^2)} + \frac{\sum_{k=0}^{K-1} \tau_k C_f^2}{L_f |\mathcal{B}_{f}^{\tau_f}|} + \frac{\sum_{k=0}^{K-1} \tau_k C_f^2}{L_f |\mathcal{B}_{g}^{\tau_g}|} + \frac{2\beta_f^2 \sigma_g^2}{|\mathcal{B}_{g}^{\tau_g+1}|}.
\]

A.6 Proof of Corollary 2.2

By Lemma 2.3, the \( C_g^2 L_f \)-strongly convexity of \( h \) and the tower property, denoting \( \Delta_1^k \triangleq \| g(x^k) - u^k \|^2 \), we have

\[
\mathbb{E}[\Delta_1^{k+1}] \leq (1 - L_f^2 \tau^2)\mathbb{E}[\Delta_1^k] + \frac{8}{L_f} \mathbb{E}[D_h(x^{k+1}, x^k)] + \frac{\epsilon}{2}(\tau - 2\tau^2).
\]

Rearranging the above terms,

\[
\mathbb{E}[\Delta_1^k] \leq \mathbb{E} \left[ \frac{\Delta_1^k - \Delta_1^{k+1}}{L_f^2 \tau^2} \right] + \frac{8}{L_f} \mathbb{E} \left[ D_h(x^{k+1}, x^k) / \tau^2 \right] + \frac{(1 - 2\tau)\epsilon}{2L_f^2 \tau}.
\]

Telescoping the above inequality from \( k = 0 \) to \( K - 1 \), and dividing by \( K \) on both sides, we get

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\Delta_1^k] \leq \frac{\Delta_0}{K L_f^2 \tau^2} + \frac{8}{L_f^3} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[ D_h(x^{k+1}, x^k) / \tau^2 \right] + \frac{(1 - 2\tau)\epsilon}{2L_f^2 \tau}
\]

\[
\leq \frac{\Delta_0}{K L_f^2 \tau^2} + \frac{8V^0}{K L_f^3 (\tau - L_f^2 \tau^2)} + \left( \frac{8(1 - 2\tau)}{L_f^3 (1 - L_f^2 \tau)} + \frac{1 - 2\tau}{2L_f^2 \tau} \right) \epsilon.
\]

Inserting the above result into (17) from Lemma 2.4, the error of the gradient of the composition function can be bounded.

B RoR Proofs

B.1 Proof of Lemma 3.1

\( f \) is \( L_f \)-smooth relative to \( h_f \), so \( \forall x, y \in \mathcal{X} \), we have

\[
L_f D_{h_f} \left( g(x), g(y) \right) \geq |f(g(x)) - f(g(y)) - \langle \nabla f(g(y)), g(x) - g(y) \rangle |
\]

\[
\geq |f(g(x)) - f(g(y)) - \langle \nabla g(y) \nabla f(g(y)), x - y \rangle |
\]

\[
- C_f \| g(x) - g(y) - \nabla g(y)^T (x - y) \|
\]
where the first inequality follows from the reverse triangle inequality, the second inequality follows from Cauchy–Schwarz inequality, and the last inequality is due to $C_f$-Lipschitz continuity of $f$ and $L_g$-relative smoothness of $g$. Next, we upper-bound $D_{h_f}(g(x), g(y))$ as

$$D_{h_f}(g(x), g(y)) = h_f(g(x)) - h_f(g(y)) - \langle \nabla h_f(g(y)), g(x) - g(y) \rangle$$

$$+ \langle \nabla g(y) \nabla h_f(g(y)), x - y \rangle - \langle \nabla g(y) \nabla h_f(g(y)), x - y \rangle$$

$$\leq A + \| \nabla h_f(g(y)) \| \| g(x) - g(y) - \nabla g(y)^T (x - y) \|$$

$$\leq A + C_{h_f} L_g D_{h_g}(x, y), \quad (41)$$

where $A \triangleq h_f(g(x)) - h_f(g(y)) - \langle \nabla g(y) \nabla h_f(g(y)), x - y \rangle$. Combining the above two inequalities, we have

$$|f(g(x)) - f(g(y)) - \langle \nabla g(y) \nabla f(g(y)), x - y \rangle|$$

$$\leq L_f A + L_f C_{h_f} L_g D_{h_g}(x, y) + C_f L_g D_{h_g}(x, y) = D_h(x, y),$$

where $h(x) \triangleq (C_f L_g + C_{h_f} L_g L_f) h_g(x) + L_f h_f(g(x))$. Below, we show that the function $h$ is indeed convex.

Using Proposition 3.1, $h_f(g(x))$ is $C_{h_f} L_g$-weakly convex relative to $h_g$; hence, $C_{h_f} L_g L_f h_g(x) + L_f h_f(g(x))$ is convex, and hence $h$, is convex. Furthermore, if $h_g$ is 1-strongly convex, $h$ is $C_f L_g$-strongly convex.

**B.2 Proof of Lemma 3.2**

Since $h_f$ is $C_{h_f}$-Lipschitz continuous and $g$ is $L_g$-smooth relative to $h_g$, we have

$$|h_f(g(y)) - h_f(g(x) + \langle \nabla g(x), y - x \rangle)|$$

$$\leq C_{h_f} \| g(y) - g(x) - \langle \nabla g(x), y - x \rangle \| \leq C_{h_f} L_g D_{h_g}(y, x).$$

Rearranging the above terms, the first result follows. By Lemma 3.1, we have

$$F(y)$$

$$\leq F(x) + \langle \nabla F(x), y - x \rangle + D_h(y, x)$$

$$\leq F(x) + \langle \nabla F(x), y - x \rangle + (C_f L_g + C_{h_f} L_g L_f) D_{h_g}(y, x) + C_{h_f} L_g L_f D_{h_g}(y, x)$$

$$+ L_f h_f(g(x) + \langle \nabla g(x), y - x \rangle) - L_f h_f(g(x)) - L_f \| \nabla g(x) \nabla h_f(g(x)), y - x \|$$

$$= F(x) + \langle \nabla F(x), y - x \rangle + (C_f L_g + 2C_{h_f} L_g L_f) D_{h_g}(y, x)$$

$$+ L_f D_{h_f}(g(x) + \langle \nabla g(x), y - x \rangle, g(x)),$$

where the second inequality uses the first result.
B.3 Proof of Lemma 3.4

Since $f$ is $C_f$-Lipschitz continuous, we have

$$|f(u^k) - f(g(x^k))| \leq C_f\|u^k - g(x^k)\|, \quad (42)$$

and

$$|f(g(x^{k+1})) - f(u^k + (v^k)^T(x^{k+1} - x^k))|$$
$$\leq C_f\|g(x^{k+1}) - u^k - (v^k)^T(x^{k+1} - x^k)\|$$
$$\leq C_f\|g(x^{k+1}) - g(x^k) - \nabla g(x^k)^T(x^{k+1} - x^k)\| + C_f\|g(x^k) - u^k\|$$
$$+ C_f\|\nabla g(x^k) - v^k\|^T(x^{k+1} - x^k)\|$$
$$\leq C_f L_g D_{h_g}(x^{k+1}, x^k) + C_f\|g(x^k) - u^k\| + C_f\|\nabla g(x^k) - v^k\|\|x^{k+1} - x^k\|, \quad (43)$$

where the last inequality uses the fact $g$ is $L_g$-smooth relative to $h_g$. $f$ is also $L_f$-smooth relative to $h_f$, so we have

$$|f(u^k + (v^k)^T(x^{k+1} - x^k)) - f(u^k) - (v^k s^k, x^{k+1} - x^k)|$$
$$+ (\nabla f(u^k) - s^k, (v^k)^T(x^{k+1} - x^k))|$$
$$\leq L_f D_{h_f}(u^k + (v^k)^T(x^{k+1} - x^k), u^k) + \|v^k\|\|\nabla f(u^k) - s^k\|\|x^{k+1} - x^k\|. \quad (44)$$

Combining these three inequalities, we have

$$f(g(x^{k+1}))$$
$$\leq f(u^k + (v^k)^T(x^{k+1} - x^k)) + C_f L_g D_{h_g}(x^{k+1}, x^k) + C_f\|g(x^k) - u^k\|$$
$$+ C_f\|\nabla g(x^k) - v^k\|\|x^{k+1} - x^k\|$$
$$\leq f(g(x^k)) + (v^k s^k, x^{k+1} - x^k) + \frac{L_f}{\tau_k} D_{h_f}(u^k + (v^k)^T(x^{k+1} - x^k), u^k) + \frac{\lambda}{\tau_k} D_{h_g}(x^{k+1}, x^k)$$
$$+ 2C_f\|g(x^k) - u^k\| + C_f\|\nabla g(x^k) - v^k\|\|x^{k+1} - x^k\| + \|v^k\|\|\nabla f(u^k) - s^k\|\|x^{k+1} - x^k\|$$
$$+ L_f(1 - \frac{1}{\tau_k}) D_{h_f}(u^k + (v^k)^T(x^{k+1} - x^k), u^k) + (C_f L_g - \frac{\lambda}{\tau_k}) D_{h_g}(x^{k+1}, x^k), \quad (45)$$

where $\lambda = C_f L_g + 2C_{h_f} L_g L_f$ as defined in Lemma 3.2. By the definition of $x^{k+1}$, we know that

$$\langle v^k s^k, x^{k+1} - x^k \rangle + \frac{L_f}{\tau_k} D_{h_f}(u^k + (v^k)^T(x^{k+1} - x^k), u^k) + \frac{\lambda}{\tau_k} D_{h_g}(x^{k+1}, x^k)$$
$$\leq \langle v^k s^k, x^k - x^k \rangle + \frac{L_f}{\tau_k} D_{h_f}(u^k + (v^k)^T(x^k - x^k), u^k) + \frac{\lambda}{\tau_k} D_{h_g}(x^k, x^k) = 0.$$

 Springer
Furthermore, given that $\tau_k \leq 1$, we have $L_f (1 - \frac{1}{\tau_k}) D_{h_x} (u^k + (v^k)\top (x^{k+1} - x^k), u^k) \leq 0$. Using these two inequalities in the right-hand side of (45), we have

$$f(g(x^{k+1})) \leq f(g(x^k)) + 2C_f \|g(x^k) - u^k\| + C_f \|\nabla g(x^k) - v^k\| x^{k+1} - x^k\|
+ \|v^k\| \|\nabla f(u^k) - s^k\| x^{k+1} - x^k\| + (C_f L_g - \frac{1}{\tau_k}) D_{h_x} (x^{k+1}, x^k)
\leq f(g(x^k)) + 2C_f \|g(x^k) - u^k\| + \frac{C_f^2}{2} \|\nabla g(x^k) - v^k\|^2 + \frac{1}{2} \|x^{k+1} - x^k\|^2
+ \frac{1}{2} \|v^k\|^2 \|\nabla f(u^k) - s^k\|^2 + \frac{1}{2} \|x^{k+1} - x^k\|^2 + (C_f L_g - \frac{1}{\tau_k}) D_{h_x} (x^{k+1}, x^k)
\leq f(g(x^k)) + 2C_f \|g(x^k) - u^k\| + \frac{C_f^2}{2} \|\nabla g(x^k) - v^k\|^2
+ \frac{1}{2} \|v^k\|^2 \|\nabla f(u^k) - s^k\|^2 - (\frac{1}{\tau_k} - C_f L_g - 2) D_{h_x} (x^{k+1}, x^k),$$

where the second inequality follows from the Young’s inequality and the last one uses the fact $h_g$ is 1-strongly convex.

**B.4 Proof of Lemma 3.5**

By Assumption 2, we have $\mathbb{E}[\|u^k - g(x^k)\|^2 | F_k] \leq \sigma_g^2 / |B_g^k|$ and by Jensen’s inequality, we get $\mathbb{E}[\|u^k - g(x^k)\|^2 | F_k] \leq \sqrt{\mathbb{E}[\|u^k - g(x^k)\|^2 | F_k]} \leq \sigma_g / \sqrt{|B_g^k|}$. Taking expectation on both sides of (26) conditioned on $F_k$, we have

$$\mathbb{E}[f(g(x^{k+1}) | F_k]
\leq f(g(x^k)) + 2C_f \mathbb{E}[\|g(x^k) - u^k\| | F_k] + \frac{C_f^2}{2} \mathbb{E}[\|\nabla g(x^k) - v^k\|^2 | F_k]
+ \frac{1}{2} \mathbb{E}[\|v^k\|^2 \|\nabla f(u^k) - s^k\|^2 | F_k] - \left(\frac{1}{\tau} - C_f L_g - 2\right) \mathbb{E}[D_{h_x} (x^{k+1}, x^k) | F_k]
\leq f(g(x^k)) + 2C_f \sigma_g \sqrt{|B_g^k|} + \frac{C_f^2}{2} \mathbb{E}[\|\nabla g(x^k) - v^k\|^2 | F_k]
+ \frac{C_f^2}{2} \mathbb{E}[\|\nabla f(u^k) - s^k\|^2 | F_k] - \left(\frac{1}{\tau} - C_f L_g - 2\right) \mathbb{E}[D_{h_x} (x^{k+1}, x^k) | F_k],$$

where the second inequality holds because $v^k$, $u^k$, and $s^k$ are independent conditioned on $F_k$ and Remark 2.1. Adding $-F^*$ on both sides and using the tower property, we get

$$\mathbb{E}[f(g(x^{k+1})) - F^*]
\leq \mathbb{E}[f(g(x^k)) - F^*] + \frac{2C_f \sigma_g}{\sqrt{|B_g^k|}} + \frac{C_f^2}{2} \mathbb{E}[\|\nabla g(x^k) - v^k\|^2 | F_k].$$
\[ + \frac{C_f^2 C_g^2}{2 |B_{\nabla f}^k|} - \left( \frac{\lambda}{\tau} - C_f L_g - 2 \right) \mathbb{E}[D_{h_g}(x^{k+1}, x^k)]. \]

Rearranging the above terms, we have

\[
\left( \lambda \tau - (C_f L_g + 2) \tau^2 \right) \mathbb{E} \left[ D_{h_g}(x^{k+1}, x^k)/\tau^2 \right] \\
\leq -\mathbb{E}[f(g(x^{k+1})) - F^*] + \mathbb{E}[f(g(x^k)) - F^*] + \frac{2C_f \sigma_g}{\sqrt{|B_{g}^k|}} + \frac{C_f^2 C_g^2}{2 |B_{\nabla g}^k|}.
\]

(46)

Telescoping from \( k = 0 \) to \( K - 1 \),

\[
\left( \lambda \tau - (C_f L_g + 2) \tau^2 \right) \sum_{k=0}^{K-1} \mathbb{E} \left[ D_{h_g}(x^{k+1}, x^k)/\tau^2 \right] \\
\leq f(g(x^0)) - F^* + \sum_{k=0}^{K-1} \left( \frac{2C_f \sigma_g}{\sqrt{|B_{g}^k|}} + \frac{C_f^2 C_g^2}{2 |B_{\nabla g}^k|} + \frac{C_f^2 C_g^2}{2 |B_{\nabla f}^k|} \right)
\]

Define \( M_1 \triangleq \lambda \tau - (C_f L_g + 2) \tau^2 \) and let \( \tau < \frac{C_f L_g + 2C_f L_f L_g}{C_f L_g + 2C_f L_f L_g} \) which results in \( M_1 > 0 \). Setting \( |B_{g}^k| = \left\lceil \frac{16C_f^2 \sigma_g^2}{M_1^2 \epsilon^2} \right\rceil \), \( |B_{\nabla g}^k| = |B_{\nabla f}^k| = \left\lceil \frac{2C_f^2 C_g^2}{M_1 \epsilon} \right\rceil \) and dividing both sides of the above inequality by \( M_1 K \), we get

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[ D_{h_g}(x^{k+1}, x^k)/\tau^2 \right] \leq \frac{f(g(x^0)) - F^*}{M_1 K} + \epsilon.
\]

B.5 Proof of Lemma 3.6

\( h_f \) is twice differentiable and \( L_{h_f} \)-smooth over \( \mathcal{S} \), i.e.,

\[ \nabla^2 h_f(x) \preceq L_{h_f} I, \quad \forall x \in \mathcal{S} \subset \mathbb{R}^d, \]

where \( I \) is the identity matrix and \( \mathcal{S} \) is a bounded subset of \( \mathbb{R}^d \). As \( f \) is \( L_{h_f} \)-smooth relative to \( h_f \), we have

\[ \nabla^2 f(x) \preceq L_f \nabla^2 h_f(x) \preceq L_f L_{h_f} I, \quad \forall x \in \mathcal{S} \subset \mathbb{R}^d. \]
which means \( f \) is \( L_f L_{h_f} \)-smooth over \( S \).

**B.6 Proof of Lemma 3.7**

By the three-point inequality, \( \forall y \in X \), we have

\[
\tau_k \left( w^k, x^{k+1} - y \right) + L_f h_f(u^k + (v^k)^T (x^{k+1} - x^k)) - L_f h_f(u^k + (v^k)^T (y - x^k)) + L_f \left( v^k \nabla h_f(u^k), y - x^{k+1} \right) \leq \lambda(D_{h_y}(y, x^k) - D_{h_y}(x^{k+1}, x^k) - D_{h_y}(y, \tilde{x}^{k+1})),
\]

and

\[
\tau_k \left( \nabla F(x^k), \tilde{x}^{k+1} - y \right) + L_f h_f(g(x^k) + \nabla g(x^k)^T (\tilde{x}^{k+1} - x^k)) - L_f h_f(g(x^k) + \nabla g(x^k)^T (y - x^k)) + L_f \left( \nabla g(x^k) \nabla h_f(g(x^k)), y - \tilde{x}^{k+1} \right) \leq \lambda(D_{h_y}(y, x^k) - D_{h_y}(\tilde{x}^{k+1}, x^k) - D_{h_y}(y, \tilde{x}^{k+1})).
\]

Letting \( y = \tilde{x}^{k+1} \) in the first inequality, \( y = x^{k+1} \) in the second one, summing them together, and rearranging the terms, we have

\[
\lambda(D_{h_y}(\tilde{x}^{k+1}, x^{k+1}) + D_{h_y}(x^{k+1}, \tilde{x}^{k+1})) \leq \tau_k \left( \nabla F(x^k) - w^k, x^{k+1} - \tilde{x}^{k+1} \right) + L_f \left( v^k \nabla h_f(u^k) - \nabla g(x^k) \nabla h_f(g(x^k)), x^{k+1} - \tilde{x}^{k+1} \right) + L_f \left[ h_f(u^k + (v^k)^T (\tilde{x}^{k+1} - x^k)) - h_f(g(x^k) + \nabla g(x^k)^T (\tilde{x}^{k+1} - x^k)) \right] + L_f \left[ h_f(g(x^k) + \nabla g(x^k)^T (x^{k+1} - x^k)) - h_f(u^k + (v^k)^T (x^{k+1} - x^k)) \right].
\]

We will next bound these four terms individually. By Young’s inequality,

\[
\tau_k T_1 \leq \frac{\tau_k^2}{2C_f L_g} \|w^k - \nabla F(x^k)\|^2 + \frac{C_f L_g}{2} \|x^{k+1} - \tilde{x}^{k+1}\|^2.
\]

Similarly,

\[
L_f T_2 \leq \frac{L_f}{4C_h L_g} \|v^k \nabla h_f(u^k) - \nabla g(x^k) \nabla h_f(g(x^k))\|^2 + C_h L_f L_g \|x^{k+1} - \tilde{x}^{k+1}\|^2 + C_h L_f L_g \|x^{k+1} - \tilde{x}^{k+1}\|^2.
\]
Using Lipschitz continuity of $h_f$,

\[
L_f T_3 \leq C_{h_f} L_f \|u^k - g(x^k)\| + \frac{L_f}{2C_{h_f} L_g} \|g(x^k)\|^2 + \frac{C_{h_f} L_f L_g^2}{2C_{h_f} L_g} \|v^k - \nabla g(x^k)\|^2 + C_{h_f} L_f L_g \|\hat{x}^{k+1} - \tilde{x}^{k+1}\|^2.
\]

By Lipschitz continuity of $h_f$,

\[
L_f T_4 \leq C_{h_f} L_f \|g(x^k) - u^k\| + \frac{L_f}{2C_{h_f} L_g} \|\tilde{g}^k - v^k\|^2 + \frac{C_{h_f} L_f L_g^2}{2C_{h_f} L_g} \|\hat{x}^{k+1} - \tilde{x}^{k+1}\|^2.
\]

$h_g$ is 1-strongly convex, so

\[
\lambda \|\tilde{x}^{k+1} - x^{k+1}\|^2 \leq \lambda (D_{h_g} (\tilde{x}^{k+1}, x^{k+1}) + D_{h_g} (x^{k+1}, \tilde{x}^{k+1})).
\]

Combining the upper and lower bounds we obtained for (47), we have

\[
\lambda \|\tilde{x}^{k+1} - x^{k+1}\|^2 \leq \lambda \left( D_{h_g} (\tilde{x}^{k+1}, x^{k+1}) + D_{h_g} (x^{k+1}, \tilde{x}^{k+1}) \right).
\]

Rearranging the terms and using $\lambda = \frac{C_f L_g}{2} + C_{h_f} L_f L_g$, we have

\[
\left( \frac{C_f L_g}{2} + C_{h_f} L_f L_g \right) \|\tilde{x}^{k+1} - x^{k+1}\|^2 \leq \frac{\tau_k^2}{2C_f L_g} \|w^k - \nabla F(x^k)\|^2 + \frac{3C_{h_f} L_f L_g}{2C_{h_f} L_g} \|\nabla F(x)\|^2 + \frac{C_{h_f} L_f L_g^2}{2C_{h_f} L_g} \|u^k - g(x^k)\|^2 + 2C_{h_f} L_f \|u^k - g(x^k)\| + \frac{C_{h_f} L_f L_g}{2} \left( \|\hat{x}^{k+1} - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2 \right).
\]

Using $\|x - y\|^2 \leq 2 \|x - z\|^2 + 2 \|z - y\|^2$ and above inequalities, we have

\[
\left( \frac{C_f L_g}{2} + C_{h_f} L_f L_g \right) \|\tilde{x}^{k+1} - x^k\|^2
\]
\[
\leq 2\left(\frac{C_f L_g}{2} + C_{h_f} L_f L_g\right)\|\tilde{x}^{k+1} - x^{k+1}\|^2 + 2\left(\frac{C_f L_g}{2} + C_{h_f} L_f L_g\right)\|x^{k+1} - x^k\|^2
\]
\[
\leq \frac{\tau^2}{C_f L_g} \|w^k - \nabla F(x^k)\|^2 + \frac{3C_{h_f} L_f L_g}{C_{h_f} L_g} \|\nabla g(x^k)\|^2 + \frac{C^2_f L_f L^2_{h_f}}{C_{h_f} L_g} \|u^k - g(x^k)\|^2
\]
\[
+ 4C_{h_f} L_f \|u^k - g(x^k)\| + C_{h_f} L_f L_g \left(\|x^{k+1} - x^k\|^2 + \|x^{k+1} - x^k\|^2\right)
\]
\[
+ 2\left(\frac{C_f L_g}{2} + C_{h_f} L_f L_g\right)\|x^{k+1} - x^k\|^2.
\]

Rearranging the terms,
\[
\frac{C_f L_g}{2} \|\tilde{x}^{k+1} - x^k\|^2
\]
\[
\leq \frac{\tau^2}{C_f L_g} \|w^k - \nabla F(x^k)\|^2 + \frac{3C_{h_f} L_f}{C_{h_f} L_g} \|\nabla g(x^k)\|^2 + \frac{C^2_f}{C_{h_f} L_g} \|u^k - g(x^k)\|^2
\]
\[
+ \frac{3C_{h_f} L_f}{L_g} \|\nabla g(x^k)\|^2 + \left(C_f L_g + 3C_{h_f} L_f L_g\right)\|x^{k+1} - x^k\|^2.
\]

Multiplying the inequality above by \(2/(C_f L_g)\) and using 1-strong convexity of \(h_g\), the proof is complete.

**B.7 Proof of Theorem 3.1**

Multiplying (26) by \(2(C_f L_g + 3C_{h_f} L_f L_g)\), multiplying (27) by \(C_f L_g M_1/2\tau^2\), and adding the two inequalities, we get
\[
\frac{C_f L_g M_1}{2} \|\tilde{x}^{k+1} - x^k\|^2
\]
\[
\leq (2C_f L_g + 6C_{h_f} L_f L_g)(f(g(x^k)) - f(g(x^{k+1}))) + \frac{M_1}{C_f L_g} \|w^k - \nabla F(x^k)\|^2
\]
\[
+ \left(\frac{4C_{h_f} L_f M_1}{\tau^2} + 4C^2_f L_g + 12C_{h_f} C_f L_f L_g\right)\|u^k - g(x^k)\|^2
\]
\[
+ \frac{C^2_f L_f L^2_{h_f} M_1}{C_{h_f} L_g \tau^2} \|u^k - g(x^k)\|^2
\]
\[
+ \left(\frac{3C_{h_f} L_f M_1}{L_g \tau^2} + C^3_f L_g + 3C_{h_f} C^2_f L_f L_g\right)\|v^k - \nabla g(x^k)\|^2
\]
\[
+ \left(3C_{h_f} L_f L_g\right)\|v^k\|^2 \|\nabla f(u^k) - s^k\|^2.
\]

Taking expectation of both sides conditioned on \(F^k\), we have
\[
\frac{C_f L_g M_1}{2} \|\tilde{x}^{k+1} - x^k\|^2
\]
\[
\begin{align*}
&\leq (2C_f L_g + 6C_h f L_f L_g)(f(g(x^k)) - \mathbb{E}[f(g(x^{k+1}))]) \\
&+ \frac{M_1}{C_f L_g} \left( \frac{2C_f^2 L_f^2 L_h^2 g^2}{|B^{k_f}_g|} + \frac{2C_f^2 C_f^2}{|B^{k_f}_{\nabla f}|} + \frac{2C_f^2 C_f^2}{|B^{k_f}_g|} \right) \\
&+ \left( \frac{4C_h f L_f M_1}{\tau^2} + 4C_f^2 L_g + 12C_h f C_f L_f L_g \right) \frac{\sigma_g}{|B^{k}_g|} + \frac{C_h f L_f^2 M_1}{C_h f L_g \tau^2} \frac{\sigma_g}{|B^{k}_g|} \\
&+ \left( \frac{3C_h f L_f M_1}{L_g \tau^2} + C_f^2 L_g + 3C_h f C_f^2 L_f L_g \right) \frac{C_f^2 C_f^2}{|B^{k}_{\nabla g}|} + (C_f L_g + 3C_h f L_f L_g) \frac{C_f^2 C_f^2}{|B^{k}_{\nabla f}|}.
\end{align*}
\]

Taking expectation over the algorithm, applying the tower property, we get

\[
\begin{align*}
&\frac{C_f L_g M_1}{2} \mathbb{E} \left[ \|\hat{x}^{k+1} - x^k\|^2 \right] \\
&\leq (2C_f L_g + 6C_h f L_f L_g) \mathbb{E} \left[ f(g(x^k)) - f(g(x^{k+1})) \right] \\
&+ \left( \frac{C_f^2 L_f^2 L_h^2 M_1^3}{8C_f^2 L_g} + \frac{C_f^2 L_f^2 M_1^3}{16C_h f C_f^2 L_g \tau^2} \right) \epsilon^2 \\
&+ \left( \frac{2M_1^2}{C_f L_g} + \frac{C_h f L_f M_1^2}{C_f \tau^2} + \frac{3C_h f L_f M_1^2}{2C_f^2 L_g \tau^2} + 2C_f L_g M_1 + 6C_h f L_f L_g M_1 \right) \epsilon.
\end{align*}
\]

Telescoping from \( k = 0 \) to \( K - 1 \), dividing both sides by \( \frac{C_f L_g M_1 K}{2} \), the result will follow.

**B.8 Proof of Lemma 3.11**

By triangle inequality, we have

\[
\mathbb{E} [\|v^k_j\|^2 | F^k_j] \leq 2\mathbb{E} [\|v^k_j - \nabla g(x^k_j)\|^2 | F^k_j] + 2\mathbb{E} [\|\nabla g(x^k_j)\|^2 | F^k_j] \\
\leq 2\mathbb{E} [\|v^k_j - \nabla g(x^k_j)\|^2 | F^k_j] + 2C^2_g.
\] (49)

Adjusting (26) based on Algorithm 3, we have

\[
\begin{align*}
f(g(x^k_{j+1})) &\leq f(g(x^k_j)) + 2C_f g(x^k_j) - u^k_j + \frac{C_f^2}{2} \|\nabla g(x^k_j) - v^k_j\|^2 \\
&+ \frac{1}{2} \|v^k_j\|^2 + \|\nabla f(u^k_j) - s^k_j\|^2 - \frac{1}{2} \frac{\lambda}{\tau} - C_f L_g - 2) \|x^k_{j+1} - x^k_j\|^2.
\end{align*}
\]
Taking expectation of both sides, using the tower property and (49), we have

\[
\frac{1}{2} \mathbb{E} [\| \mathbf{v}_j^k \|^2] \leq \mathbb{E} [\| \mathbf{v}_j^k \|^2] + \mathbb{E} [\| \nabla f (\mathbf{u}_j^k) - \mathbf{s}_j^k \|^2] - \frac{1}{2} \mathbb{E} \left[ \sum_{r=0}^{j-1} \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2] \right] + \frac{C_f^2 C_g^2}{2 |B_{\mathbf{v}_j^k}^k|} \mathbb{E} [\| \nabla g (\mathbf{x}_j^k) \|^2] + \frac{C_f^2 C_g^2}{2 |B_{\mathbf{v}_j^k}^k|} \mathbb{E} [\| \nabla f (\mathbf{u}_j^k) - \mathbf{s}_j^k \|^2]
\]

Hence, we have

\[
\mathbb{E} [f(\mathbf{x}_{j+1}^k)] \leq \mathbb{E} [f(\mathbf{x}_j^k)] + 2C_f \mathbb{E} [\| g(\mathbf{x}_j^k) - \mathbf{u}_j^k \|] + \frac{C_f^2 C_g^2}{2 |B_{\mathbf{v}_j^k}^k|} \mathbb{E} [\| \nabla g(\mathbf{x}_j^k) - \mathbf{v}_j^k \|^2] + \frac{C_f^2 C_g^2}{2 |B_{\mathbf{v}_j^k}^k|} - \frac{1}{2} \left( \lambda - C_f L_g - 2 \right) \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2].
\]

Applying Lemma 3.9 and inequality (33), we have

\[
\mathbb{E} [f(\mathbf{x}_{j+1}^k)] \leq \mathbb{E} [f(\mathbf{x}_j^k)] + 2C_f \sigma_g \frac{C_f^2}{|B_{\mathbf{v}_j^k}^k|} \mathbb{E} [\| g^{rk+1} \|] + \frac{C_f^2 C_g^2}{2 |B_{\mathbf{v}_j^k}^k|} - \frac{1}{2} \left( \lambda - C_f L_g - 2 \right) \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2].
\]

Let \( S_{g}^{k,r} \equiv S_{g} \), \( S_{v_{g}}^{k,r} \equiv S_{v_{g}} \), \( B_{g}^{k} \equiv B_{g} \), \( B_{v_{g}}^{k} \equiv B_{v_{g}} \), \( B_{v_{g}}^{k,j} \equiv B_{v_{g}} \), and note that

\[
\sum_{r=0}^{j-1} \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2] \leq \sum_{r=0}^{j-1} \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2].
\]

Also, we have

\[
\mathbb{E} [f(\mathbf{x}_{j+1}^k)] \leq \mathbb{E} [f(\mathbf{x}_j^k)] + \left( \frac{C_f C_g^2}{\delta S_{g}} + \frac{C_f^2 L_g^2}{2 S_{v_{g}}} + \frac{C_f^2 L_g^2}{B_{v_{g}} S_{v_{g}}} \right) \sum_{j=0}^{J-1} \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2] - \frac{1}{2} \left( \lambda - C_f L_g - 2 \right) \mathbb{E} [\| \mathbf{x}_{j+1}^k - \mathbf{x}_j^k \|^2].
\]
Telescoping the above inequality from $j = 0$ to $J - 1$ and rearranging the terms,

$$
\frac{\lambda}{2\tau} - \frac{C_f L_g}{2} - 1 - J \left( \frac{C_f C_g^2}{\delta S_g} + \frac{C_f^2 L_g^2}{2S_g V_g} \right) \sum_{j=0}^{J-1} \mathbb{E}[\|x_{j+1}^k - x_j^k\|^2] 
\leq \mathbb{E}[f(g(x_0^k))] - f(g(x_j^k))] + J \left( \frac{2C_f \sigma_g}{\sqrt{B_g}} + C_f \delta + \frac{C_f^2 C_g^2}{2B_g V_g} + \frac{C_f^2 C_g^2}{2B_v f B_g} + \frac{C_f^2 C_g^2}{2B_v f} \right).
$$

Telescoping the above inequality from $k = 0$ to $K - 1$, and noting that by Algorithm 3, $x_0^{k+1} = x_f^k$, we have

$$
\frac{\lambda}{2\tau} - \frac{C_f L_g}{2} - 1 - J \left( \frac{C_f C_g^2}{\delta S_g} + \frac{C_f^2 L_g^2}{2S_g V_g} \right) \sum_{j=0}^{K-1} \sum_{k=0}^{J-1} \mathbb{E}[\|x_{j+1}^k - x_j^k\|^2] 
\leq f(g(x_0^0)) - F^* + JK \left( \frac{2C_f \sigma_g}{\sqrt{B_g}} + C_f \delta + \frac{C_f^2 C_g^2}{2B_g V_g} + \frac{C_f^2 C_g^2}{2B_v f B_g} + \frac{C_f^2 C_g^2}{2B_v f} \right).
$$

Since $B_v f \geq 1$, the term $-J \left( \frac{C_f C_g^2}{\delta S_g} + \frac{C_f^2 L_g^2}{2S_g V_g} + \frac{C_f^2 L_g^2}{2B_v f S_g} \right)$ in the left-hand side of the above inequality can be lower-bounded as

$$
-J \left( \frac{C_f C_g^2}{\delta S_g} + \frac{C_f^2 L_g^2}{2S_g V_g} + \frac{C_f^2 L_g^2}{2B_v f S_g} \right) \geq -J \left( \frac{C_f C_g^2}{\delta S_g} + \frac{3C_f^2 L_g^2}{2S_g V_g} \right).
$$

Define $M_2 \triangleq \left( \frac{C_f L_g \tau}{2} - \left( \frac{C_f L_g}{2} + 1 \right) \tau^2 \right)$, set $\delta = \frac{M_2 \epsilon}{3C_f\tau}$, $S_g = \left[ \frac{6C_f^2 C_g^2 \tau}{C_h f L_f L_g M_2 \epsilon^2} J \right]$, $S_{\sqrt{g}} = \left[ \frac{3C_f^2 L_g \tau}{C_h f L_f} \right]$, $B_g = \left[ \frac{36C_f^2 \sigma_g^2}{M_2 \epsilon^2} \right]$, $B_{\sqrt{f}} = B_{\sqrt{f}} = \left[ \frac{15C_f^2 C_g^2}{2M_2 \epsilon^2} \right]$ and note that $\frac{C_f^2 C_g^2}{B_v f B_g} \leq \frac{C_f^2 C_g^2}{B_v f}$. Hence, we have

$$
M_2 \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E}[\|x_{j+1}^k - x_j^k\|^2] \leq f(g(x_0^0)) - F^* + M_2 JK \epsilon.
$$

With $\tau < \frac{C_f L_g}{C_f L_g + 1}$, we have $M_2 > 0$. Dividing both sides by $M_2 JK \epsilon$, the proof is complete.

### B.9 Proof of Theorem 3.2

By Lemma 3.11, for any given $K$, $J$ and $\epsilon$, the sequence $\{x^k\}$ generated by Algorithm 3 lies in a bounded set with probability 1. Hence Assumption 5 is applicable. Adjusting (48) based on Algorithm 3, we have

$$
\frac{C_f L_g M_1}{2} \frac{\|\tilde{x}_{j+1}^k - x_j^k\|^2}{\tau^2}
$$
By Lemma 3.9, similar to the proof of Lemma 3.11, after telescoping twice, we get
\[
1 \leq (2C_f L_g + 6C_{h_f} C_f L_f L_g) (f(g(x_j^k)) - f(g(x_{j+1}^k))) + \frac{M_1}{C_f L_g} \| \omega^k_j - \nabla F(x_j^k) \|^2 \\
+ \left( \frac{4C_{h_f} C_f M_1}{\tau^2} + 4C_f^2 L_g + 12C_{h_f} C_f L_f L_g \right) \| u^k_j - g(x_j^k) \|^2 + \frac{C_f^2 L_f^2 L_g^2 M_1}{C_{h_f} L_g \tau^2} \| u^k_j - g(x_j^k) \|^2 \\
+ \left( \frac{3C_{h_f} C_f M_1}{L_g \tau^2} + C_f^2 L_g + 3C_{h_f} C_f^2 L_f L_g \right) \| v^k_j - \nabla g(x_j^k) \|^2 \\
+ \left( C_f L_g + 3C_{h_f} C_f L_f L_g \right) \| v^k_j \|^2 \| \nabla f(u_j^k) - s_j^k \|^2.
\]

Taking expectation of both sides, by Lemma 3.10, inequality (50), and the fact \( |B^k_j \|_f \geq 1 \),
\[
\frac{C_f L_g M_1}{2} \mathbb{E} \left[ \| \tilde{x}_{j+1}^k - x_j^k \|^2 / \tau^2 \right] \\
\leq A_0 \mathbb{E}[f(g(x_j^k)) - f(g(x_{j+1}^k))] + A_1 \mathbb{E}[\| u_j^k - g(x_j^k) \|] \\
+ A_2 \mathbb{E}[\| u_j^k - g(x_j^k) \|^2] + A_3 \mathbb{E}[\| v_j^k - \nabla g(x_j^k) \|^2] + A_4 / |B^k_j |_f,
\]

where
\[
A_0 \triangleq 2C_f L_g + 6C_{h_f} C_f L_f L_g, \quad A_1 \triangleq \frac{4C_{h_f} C_f M_1}{\tau^2} + 4C_f^2 L_g + 12C_{h_f} C_f L_f L_g,
\]
\[
A_2 \triangleq \frac{2C_g L_f^2 L_{h_f} M_1}{C_f L_g} + \frac{C_f C_g L_f^2 L_{h_f} M_1}{C_{h_f} C_f L_f L_g \tau^2}, \\
A_3 \triangleq \frac{2C_f M_1}{L_g} + \frac{3C_{h_f} C_f M_1}{L_g \tau^2} + \frac{3C_f^3 L_g}{L_g} + 9C_{h_f} C_f^2 L_f L_g,
\]
\[
A_4 \triangleq \frac{2C_f C_g^2 M_1}{L_g} + \frac{2C_f^3 C_g L_f L_g}{L_g} + 6C_{h_f} C_f^2 C_g^2 L_f L_g.
\]

By Lemma 3.9, similar to the proof of Lemma 3.11, after telescoping twice, we get
\[
\frac{1}{KJ} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E} \left[ \| \tilde{x}_{j+1}^k - x_j^k \|^2 / \tau^2 \right] \\
\leq 2A_0 \left( f(g(x_0^0)) - F^* \right) \frac{2\tau^2}{C_f L_g M_1 K J} + \frac{2\tau^2}{C_f L_g M_1 K J} \left( \frac{A_1 C_g^2}{2\delta S_g} + \frac{A_2 C_g^2}{S_g} + \frac{A_3 C_g^2}{S_{B'}} + \frac{A_4}{B_{\nabla f}} \right) \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E} \left[ \| \tilde{x}_{j+1}^k - x_j^k \|^2 / \tau^2 \right] \\
+ \frac{2}{C_f L_g M_1} \left( \frac{A_1 \sigma_g}{\sqrt{B_g}} + \frac{A_1 \delta}{2} + \frac{A_2 \sigma_g}{B_g} + \frac{A_3 C_g^2}{B_{\nabla g}} + \frac{A_4}{B_{\nabla f}} \right).
\]

By Lemma 3.11, the right-hand side of the above inequality can be further bounded as
\[
\frac{1}{KJ} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E} \left[ \| \tilde{x}_{j+1}^k - x_j^k \|^2 / \tau^2 \right]
\]
\[
\frac{1}{KJ} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E} \left[ \| \tilde{x}_{j+1}^k - x_j^k \|^2 / \tau^2 \right] \\
\leq \left( \frac{2A_0}{C_f L_g M_1} + \frac{A_1 C_g^2 \tau^2 J}{C_f L_g \delta S_g M_1} + \frac{2A_2 C_g^2 \tau^2 J}{C_f L_g S_g M_1} + \frac{2A_3 L_g \tau^2 J}{C_f S_{\nabla g} M_1} \right) \frac{f(g(\mathbf{x}_0^0)) - F^*}{M_1 K J} \\
+ \left( \frac{A_1 C_g^2 \tau^2 J}{C_f L_g \delta S_g M_1} + \frac{2A_2 C_g^2 \tau^2 J}{C_f L_g S_g M_1} + \frac{2A_3 L_g \tau^2 J}{C_f S_{\nabla g} M_1} \right) \epsilon \\
+ \frac{2}{C_f L_g M_1} \left( \frac{A_1 \sigma_g}{\sqrt{B_g}} + \frac{A_1 \delta}{B_g} + \frac{A_2 \sigma_g^2}{B_{\nabla g}} + \frac{A_3 C_g^2}{B_{\nabla g}} + \frac{A_4}{B_{\nabla g}} \right).
\]
Replacing coefficients with the values set in Lemma 3.11, the final result is

\[
\frac{1}{KJ} \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} \mathbb{E} \left[ \| \tilde{x}_{j+1}^k - x_j^k \|^2 / \tau^2 \right] \\
\leq \left( \frac{2A_0}{C_f L_g M_1} + \frac{A_1 C_h f L_f \tau}{2C_f^2 M_1^2} + \frac{2A_2 C_h f L_f \tau}{3C_f^3 M_1^2} \right) \frac{f(g(\mathbf{x}_0^0)) - F^*}{K J} \\
+ \left( \frac{A_1 C_h f L_f \tau}{2C_f^2 M_1^2} + \frac{2A_2 C_h f L_f \tau}{3C_f^3 M_1^2} + \frac{2A_1 M_2}{3C_f^2 L_g M_1} + \frac{4A_3 M_2}{15C_f^3 L_g M_1} + \frac{4A_4 M_2}{15C_f^3 C_g L_g M_1} \right) \epsilon \\
+ \frac{A_2 C_h f L_f M_2 \tau}{3C_f^3 M_1^2} \frac{f(g(\mathbf{x}_0^0)) - F^* \epsilon}{K J} + \left( \frac{A_2 C_h f L_f M_2 \tau}{3C_f^3 M_1} + \frac{A_2 M_2^2}{18C_f^3 L_g M_1} \right) \epsilon^2.
\]

To achieve \( \epsilon \)-stationarity, we need to set \( KJ = \mathcal{O}(\epsilon^{-1}) \), then the number of calls to the \( g_\xi \) oracle is

\[
KB_g + KJS_g = \frac{\mathcal{O}(\epsilon^{-1})}{J} \mathcal{O}(\epsilon^{-2}) + \mathcal{O}(\epsilon^{-2}) J,
\]
as \( S_g = J \times \mathcal{O}(\epsilon^{-1}) \). Minimizing the right-hand side of above equality w.r.t. \( J \) results in \( J = \mathcal{O}(\epsilon^{-1/2}) \) and the sample complexity for \( g_\xi \) equal to \( \mathcal{O}(\epsilon^{-5/2}) \). Similarly, the total number of calls to the \( \nabla g_\xi \) oracle is

\[
KB_{\nabla g} + KJS_{\nabla g} = \frac{\mathcal{O}(\epsilon^{-1})}{J} \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-1}) J,
\]
which achieves its minimum with \( J = \mathcal{O}(\epsilon^{-1/2}) \) and the sample complexity of \( \nabla g_\xi \) is \( \mathcal{O}(\epsilon^{-3/2}) \).

C Experiment Proofs

C.1 Proof of Lemma 4.1

\( f(\mathbf{x}) \) is \( L \)-relative-smooth to \( h(\mathbf{x}) \) is equivalent to \( L h(\mathbf{x}) \pm f(\mathbf{x}) \) is convex, i.e., \( L \nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0 \) since \( h, f \) are twice differentiable. It is sufficient to show
\( \lambda_{\text{max}}(\pm \nabla^2 f(x)) \leq L \lambda_{\text{min}}(\nabla^2 h(x)) \) where \( \lambda_{\text{max}}(A), \lambda_{\text{min}}(A) \) are the maximal and minimum eigenvalues of matrix \( A \). We have \( \nabla^2 f(x) = 2Ax^T A + x^TAxA \) and \( \nabla^2 h(x) = \|x\|^2_I + 2xx^T \), where \( I \) is the identity matrix. Note that

\[
\lambda_{\text{max}}(\pm \nabla^2 f(x)) \leq \| \pm \nabla^2 f(x) \|_2 = \| \nabla^2 f(x) \|_2 \leq 3\|A\|_2^2\|x\|_2^2.
\]

Let \( L \geq 3\|A\|_2^2 \), we have

\[
\lambda_{\text{max}}(\pm \nabla^2 f(x)) \leq L\|x\|_2^2 \leq L \lambda_{\text{min}}(\nabla^2 h(x)).
\]

which finishes the proof.

### C.2 Proof of Lemma 4.2

Let \( g_1(x) \triangleq \mathbb{E}\left[\frac{1}{2}x^TA_\xi x\right] \) and \( g_2(x) \triangleq \mathbb{E}\left[\frac{1}{4}(x^TA_\xi x)^2\right] \). Following the proof of Lemma 4.1, we have

\[
\nabla^2 g_2(x) = \mathbb{E}[2A_\xi xx^TA_\xi + x^TA_\xi xA_\xi].
\]

To derive the sufficient condition \( \lambda_{\text{max}}(\pm \nabla^2 g_2(x)) \leq Lg_{\lambda_{\text{min}}}(\nabla^2 h(x)) \), we have

\[
\lambda_{\text{max}}(\pm \nabla^2 g_2(x)) \leq \| \nabla^2 g_2(x) \|_2 = \| \mathbb{E}[2A_\xi xx^TA_\xi + x^TA_\xi xA_\xi] \|_2 \leq 3\mathbb{E}[\|A_\xi\|_2^2]\|x\|_2^2.
\]

With \( L \geq \mathbb{E}[\|A_\xi\|_2^2] \), \( h_2(x) = \frac{1}{4}\|x\|^4 \), we can get

\[
\lambda_{\text{max}}(\pm \nabla^2 g_2(x)) \leq L\|x\|_2^2 \leq L \lambda_{\text{min}}(\nabla^2 h_2(x)).
\]

We have

\[
\|g(x) - g(y) - \nabla g(y)^T(x - y)\|_2 \\
\leq \|g(x) - g(y) - \nabla g(y)^T(x - y)\|_1 \\
= |g_1(x) - g_1(y) - \langle \nabla g_1(y), x - y \rangle| + |g_2(x) - g_2(y) - \langle \nabla g_2(y), x - y \rangle| \\
\leq \frac{k_1}{2} \|x - y\|_2^2 + k_2 D_{h_2}(x, y) = D_h(x, y),
\]

where \( h(x) = \frac{k_1}{2} \|x\|_2^2 + \frac{k_2}{4} \|x\|_4^4 \), \( k_1 \geq \|\mathbb{E}[A_\xi]\|_2 \), \( k_2 \geq 3\mathbb{E}[\|A_\xi\|^2] \), and the second inequality is due to the smoothness of the quadratic function and the relative smoothness of \( g_2 \).
References

1. Ahookhosh, M., Themelis, A., Patrinos, P.: A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. SIAM J. Optim. 31(1), 653–685 (2021). https://doi.org/10.1137/19M1264783
2. Alexanderian, A., Petra, N., Stadler, G., Ghattività, O.: Mean-variance risk-averse optimal control of systems governed by PDEs with random parameter fields using quadratic approximations. SIAM/ASA J. Uncertain. Quant. 5(1), 1166–1192 (2017). https://doi.org/10.1137/16M106306X
3. Asi, H., Duchi, J.C.: Modeling simple structures and geometry for better stochastic optimization algorithms. In: Chaudhuri, K., Sugiyama, M. (eds) The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan, Proceedings of Machine Learning Research, vol. 89, pp. 2425–2434. PMLR (2019)
4. Auslender, A., Teboulle, M.: Projected subgradient methods for non-Euclidean distances with non-differentiable convex minimization and variational inequalities. Math. Program. 120(1), 27–48 (2009). https://doi.org/10.1007/s10107-007-0147-z
5. Balasubramanian, K., Ghadimi, S., Nguyen, A.: Stochastic multilevel composition optimization algorithms with level-independent convergence rates. SIAM J. Optim. 32(2), 519–544 (2022). https://doi.org/10.1137/21M1406222
6. Bauschke, H.H., Bolte, J., Chen, J., Teboulle, M., Wang, X.: On linear convergence of non-Euclidean gradient method without strong convexity and Lipschitz gradient continuity. J. Optim. Theory Appl. 182(3), 1068–1087 (2019). https://doi.org/10.1007/s10957-019-01516-9
7. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. 42(2), 330–348 (2017). https://doi.org/10.1287/moor.2016.0817
8. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. 31(3), 167–175 (2003). https://doi.org/10.1016/S0167-6377(02)00231-6
9. Ben-Tal, A., Margalit, T., Nemirovski, A.: The ordered subsets mirror descent optimization with applications to tomography. SIAM J. Optim. 12(1), 79–108 (2001). https://doi.org/10.1137/S1052623499354564
10. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal 52(1), 155–173 (2007). https://doi.org/10.1016/j.csda.2006.11.006
11. Blanchet, J., Goldfarb, D., Iyengar, G., Li, F., Zhou, C.: Unbiased simulation for optimizing stochastic function compositions. arXiv:1711.07564 (2017)
12. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. 28(3), 2131–2151 (2018). https://doi.org/10.1137/17M1138558
13. Börzl, A., von Winckel, G.: Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients. SIAM J. Sci. Comput. 31(3), 2172–2192 (2009). https://doi.org/10.1137/070711311
14. Bubeck, S.: Convex optimization: algorithms and complexity. Found® Trends Mach Learn 8(3–4), 231–357 (2015). https://doi.org/10.1561/2200000050
15. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. 3(3), 538–543 (1993). https://doi.org/10.1137/0803026
16. Chen, T., Sun, Y., Yin, W.: Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. IEEE Trans. Signal Process. 69, 4937–4948 (2021). https://doi.org/10.1109/TSP.2021.3092377
17. Cichocki, A., Zdunek, R., Amari, S.I.: Csiszar’s divergences for non-negative matrix factorization: family of new algorithms. In: Rosca, J., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) Independent Component Analysis and Blind Signal Separation, pp. 32–39. Springer, Berlin (2006)
18. Csiszar, I.: Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. 19(4), 2032–2066 (1991). https://doi.org/10.1214/aos/1176348385
19. Dai, B., He, N., Pan, Y., Boots, B., Song, L.: Learning from conditional distributions via dual embeddings. In: Singh, A., Zhu, X.J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA, Proceedings of Machine Learning Research, vol. 54, pp. 1458–1467. PMLR (2017)
20. Dann, C., Neumann, G., Peters, J.: Policy evaluation with temporal differences: A survey and comparison. J. Mach. Learn. Res. 15(24), 809–883 (2014)

21. Davis, D., Drusvyatskiy, D., MacPhee, K.J.: Stochastic model-based minimization under high-order growth. arXiv:1807.00255 (2018)

22. Dentcheva, D., Penev, S., Ruszczyński, A.: Statistical estimation of composite risk functionals and risk optimization problems. Ann. Inst. Stat. Math. 69(4), 737–760 (2017). https://doi.org/10.1007/s10463-016-0559-8

23. Devraj, A.M., Chen, J.: Stochastic variance reduced primal dual algorithms for empirical composition optimization. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp. 9878–9888 (2019)

24. Dragomir, R.A., Even, M., Hendrix, H.: Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 139, pp. 2815–2825. PMLR (2021)

25. Dragomir, R.A., Taylor, A.B., d’Aspremont, A., Bolte, J.: Optimal complexity and certification of Bregman first-order methods. Math. Program. 194(1), 41–83 (2022). https://doi.org/10.1007/s10107-021-01618-1

26. Duchi, J.C., Shalev-Shwartz, S., Singer, Y., Tewari, A.: Composite objective mirror descent. In: Kalai, A.T., Mohri, M. (eds.) COLT 2010—The 23rd Conference on Learning Theory, Haifa, Israel, June 27–29, 2010, pp. 14–26. Omnipress (2010)

27. Ermoliev, Y.: Stochastic Programming Methods. Nauka (1976)

28. Ermoliev, Y.M., Norkin, V.I.: Sample average approximation method for compound stochastic optimization problems. SIAM J. Optim. 23(4), 2231–2263 (2013). https://doi.org/10.1137/120863277

29. Fang, C., Li, C.J., Lin, Z., Zhang, T.: Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018)

30. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135. PMLR (2017)

31. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points - online stochastic gradient for tensor decomposition. In: Grünwald, P., Hazan, E., Kale, S. (eds.) Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3–6, 2015, JMLR Workshop and Conference Proceedings, vol. 40, pp. 797–842. JMLR.org (2015)

32. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program. 155(1), 267–305 (2016). https://doi.org/10.1007/s10107-014-0846-1

33. Ghadimi, S., Ruszczyński, A., Wang, M.: A single timescale stochastic approximation method for nested stochastic optimization. SIAM J. Optim. 30(1), 960–979 (2020). https://doi.org/10.1137/18M1230542

34. Hanzely, F., Richtárik, P.: Fastest rates for stochastic mirror descent methods. Comput. Optim. Appl. 79(3), 717–766 (2021). https://doi.org/10.1007/s10589-021-00284-5

35. Hanzely, F., Richtárik, P., Xiao, L.: Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. Comput. Optim. Appl. 79(2), 405–440 (2021). https://doi.org/10.1007/s10589-021-00273-8

36. Hu, B., Seiler, P., Lessard, L.: Analysis of biased stochastic gradient descent using sequential semidefinite programs. Math. Program. 187(1), 383–408 (2021). https://doi.org/10.1007/s10107-020-01486-1

37. Hu, Y., Zhang, S., Chen, X., He, N.: Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 2759–2770. Curran Associates, Inc. (2020)

38. Juditsky, A., Nemirovski, A., et al.: First order methods for nonsmooth convex large-scale optimization, I: general purpose methods. Optim. Mach. Learn. 30(9), 121–148 (2011)
39. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999). https://doi.org/10.1038/44565
40. Li, Q., Zhu, Z., Tang, G., Wakin, M.B.: Provable Bregman-divergence based methods for nonconvex and non-Lipschitz problems. arXiv:1904.09712 (2019)
41. Lian, X., Wang, M., Liu, J.: Finite-sum composition optimization via variance reduced gradient descent. In: Singh, A., Zhu, X.J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA, Proceedings of Machine Learning Research, vol. 54, pp. 1159–1167. PMLR (2017)
42. Lin, T., Fan, C., Wang, M., Jordon, M.I.: Improved sample complexity for stochastic compositional variance reduced gradient. In: 2020 American Control Conference (ACC), pp. 126–131 (2020). https://doi.org/10.23919/ACC45564.2020.9147515
43. Liu, L., Liu, J., Tao, D.: Dualityfree methods for stochastic composition optimization. IEEE Trans. Neural Netw. Learn. Syst. 30(4), 1205–1217 (2019). https://doi.org/10.1109/TNNLS.2018.2866699
44. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. 28(1), 333–354 (2018). https://doi.org/10.1137/16M1099546
45. Luo, X., Liu, Z., Xiao, S., Xie, X., Li, D.: Mindsim: user simulator for news recommenders. In: Proceedings of the ACM Web Conference 2022, WWW ’22, pp. 2067–2077. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3485447.3512080
46. Mukkamala, M.C., Ochs, P., Pock, T., Sabach, S.: Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. SIAM J. Math. Data Sci. 2(3), 658–682 (2020). https://doi.org/10.1137/19M1298007
47. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. 19(4), 1574–1609 (2009). https://doi.org/10.1137/070704277
48. Nemirovskij, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
49. Nguyen, L.M., Liu, J., Scheinberg, K., Takác, M.: SARAH: A novel method for machine learning problems using stochastic recursive gradient. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, Proceedings of Machine Learning Research, vol. 70, pp. 2613–2621. PMLR (2017)
50. Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5(2), 111–126 (1994). https://doi.org/10.1002/env.3170050203
51. Ruszczyński, A.: A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. SIAM J. Control. Optim. 59(3), 2301–2320 (2021). https://doi.org/10.1137/20M1312952
52. Ruszczyński, A., Shapiro, A.: Chapter 6: risk averse optimization. In: SIAM, pp. 223–305 (2021). https://doi.org/10.1137/1.9781611976595.ch6
53. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
54. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. 170(1), 67–96 (2018). https://doi.org/10.1007/s10107-017-1284-2
55. Wang, M., Fang, E.X., Liu, H.: Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. Math. Program. 161(1), 419–449 (2017). https://doi.org/10.1007/s10107-016-1017-3
56. Wang, M., Liu, J., Fang, E.X.: Accelerating stochastic composition optimization. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, pp. 1714–1722 (2016)
57. Wang, Z., Ji, K., Zhou, Y., Liang, Y., Tarokh, V.: Spiderboost and momentum: Faster variance reduction algorithms. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp. 2403–2413 (2019)
58. Xu, Y., Xu, Y.: Katyusha acceleration for convex finite-sum compositional optimization. INFORMS J. Optim. 3(4), 418–443 (2021). https://doi.org/10.1287/joio.2021.0055
59. Yang, S., Wang, M., Fang, E.X.: Multilevel stochastic gradient methods for nested composition optimization. SIAM J. Optim. 29(1), 616–659 (2019). https://doi.org/10.1137/18M1164846
60. Yu, Y., Huang, L.: Fast stochastic variance reduced ADMM for stochastic composition optimization. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pp. 3364–3370. ijcai.org (2017). https://doi.org/10.24963/ijcai.2017/470

61. Zhang, J., Xiao, L.: Multilevel composite stochastic optimization via nested variance reduction. SIAM J. Optim. 31(2), 1131–1157 (2021). https://doi.org/10.1137/19M1285457

62. Zhang, J., Xiao, L.: Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. Math. Program. 195(1), 649–691 (2022). https://doi.org/10.1007/s10107-021-01709-z

63. Zhang, S., He, N.: On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. arXiv:1806.04781 (2018)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.