Stochastic nested primal-dual method for nonconvex constrained composition optimization*

Lingzi Jin † Xiao Wang ‡

August 15, 2022

Abstract

In this paper we study the nonconvex constrained composition optimization, in which the objective contains a composition of two expected-value functions whose accurate information is normally expensive to calculate. We propose a STochastic nEsted Primal-dual (STEP) method for such problems. In each iteration, with an auxiliary variable introduced to track inner layer function values we compute stochastic gradients of the nested function based on the subsampling strategy. To alleviate difficulties caused by possibly nonconvex constraints, we construct a stochastic approximation to the linearized augmented Lagrangian function to update the primal variable, which further motivates to update the dual variable in a weighted-average way. Moreover, to better understand the asymptotic dynamics of the update schemes we consider a deterministic continuous-time system from the perspective of ODE. We analyze the KKT measure at the output by the STEP method and establish its iteration and sample complexities to find an ϵ-stationary point, with expected stationarity, feasibility as well as complementary slackness below accuracy ϵ. Numerical results on a risk-averse portfolio optimization problem reveal the effectiveness of the proposed algorithm.

Keywords: Composition, nonconvex constraints, augmented Lagrangian function, stationarity, feasibility, complementary slackness, iteration complexity, sample complexity

Mathematics Subject Classification 2020: 65K05, 90C30, 90C46, 90C60

1 Introduction

In this paper, we consider the nonconvex constrained composition optimization

$$\begin{aligned} \min_{x\in X} \quad & \{\Gamma(x) \equiv (f\circ h)(x)\} + \chi(x) \\ \text{s.t.} \quad & g_i(x) \le 0, \quad i = 1,\dots,m, \end{aligned} \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set, $f : \mathbb{R}^{\bar{n}} \to \mathbb{R}$ with $f(\cdot) = \mathbb{E}_\xi[F(\cdot;\xi)]$, $h : \mathbb{R}^n \to \mathbb{R}^{\bar{n}}$ with $h(\cdot) = \mathbb{E}_\phi[H(\cdot;\phi)]$ and $g_i, i = 1,\dots,m$ are continuously differentiable but possibly nonconvex, and $\chi : \mathbb{R}^n \to \mathbb{R}$ is a proper convex lower semicontinuous function. Here, ϕ, ξ are random variables independent of $x \in X$ and each other in the probability space Ξ_l, Ξ_u respectively, and mappings $F(\cdot;\xi)$ and $H(\cdot;\phi)$ are differentiable almost surely for $\xi \in \Xi_u$ and $\phi \in \Xi_l$. We assume that the feasible set $\mathcal{X} := \{x \in X \mid g_i(x) \le 0, i = 1,\dots,m\}$ is nonempty and the objective function value of (1.1) over X is lower bounded by C^*. Problem (1.1), embracing the single-layer stochastic optimization (corresponding to identity mapping f and $\bar{n} = 1$), covers a wide range of applications, such as risk-averse optimization [7, 29, 31], reinforcement learning [31] and meta-learning [7], where the explicit constraints are normally introduced to characterize some prior knowledge or field knowledge [21].

Challenges often arise when solving problem (1.1). Due to the nest structure of $(f\circ h)$, it is generally impossible to obtain an unbiased estimate of its gradient based on stochastic information of f and h. Moreover, the existence

*Part of this research work was supported by the National Natural Science Foundation of China (No. 11871453) and the Major Key Project of PCL (No. PCL2022A05)

†jinlingzi19@mails.ucas.ac.cn, Peng Cheng Laboratory, Shenzhen, 518066, China; School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China.

‡wangx07@pcl.ac.cn, Peng Cheng Laboratory, Shenzhen, 518066, China.

of nonconvex constraints makes it impractical to keep feasibility through simple projection. Study on stochastic composition optimization has attracted much interest. In [29] a stochastic compositional gradient descent (SCGD) algorithm is proposed for minimizing a composition of two expected-value functions, denoted as $\Gamma(x)$, over a convex set which is assumed easy to project onto. SCGD employs an extra variable y to track the expected value of inner function values. By updating y in a moving average way and x through a stochastic quasi-gradient iteration, the basic SCGD converges at the rate of $\mathcal{O}(k^{-1/4})$ in terms of $\mathbb{E}[\Gamma(\hat{x}^k)-\Gamma(x^*)]$ when Γ is convex, and $\mathcal{O}(k^{-2/3})$ in terms of $\mathbb{E}[\|\hat{x}^k-x^*\|^2]$ in strongly convex case, where $\hat{x}^k$ is a moving-average of $\{x^k\}$. Based on extrapolation/momentum acceleration, accelerated SCGD enjoys an improved convergence rate of $\mathcal{O}(k^{-2/7})$ when Γ is convex, and $\mathcal{O}(k^{-4/5})$ in strongly convex case. Later, [31] extends the accelerated SCGD to an accelerated stochastic compositional proximal gradient (ASC-PG) method to handle the nonsmooth regularization penalty. ASC-PG enjoys the same convergence rate and sample complexity as the accelerated SCGD when the objective is (strongly) convex, while $\mathcal{O}(\epsilon^{-4.5})$ complexity to achieve the expected gradient norm less than ϵ for smooth (possibly nonconvex) unconstrained case. [40] further generalizes SCGD type methods for two-layer optimization to algorithms for multi-layer problems with complexity theories analyzed. Meanwhile, in [28] an adaptive solver is designed for unconstrained (possibily nonconvex) smooth problems by integrating Adam [16] and mini-batch technique into the accelerated SCGD. All the aforementioned algorithms require step sizes along variables in different time scales in order to derive desired properties. Recently, by lifting the problem into a higher dimensional space, [12] proposes a single time-scale algorithm for two-layer nested stochastic smooth nonconvex optimization with convex constraints, which is further generalized into multi-level cases by [2]. Later, Ruszczyński [26] adds a linear correction on path-averaged inner function estimates on the basis of [12] and proposes a stochastic subgradient algorithm for nonconvex nonsmooth multi-level composition optimization with convex constraints, but no convergence in terms of the gradient norm is analyzed until being provided in [2] later. Based on the estimations of the inner function in [26] and quasi-gradient updates of x in basic SCGD [29], [7] further discusses the validity of the linear correction in [26] from the view of ODE and establishes the convergence rate regarding gradient norm for unconstrained smooth stochastic compositional optimization. Adam-type and multi-level variants are also proposed and studied in [7].

Apart from works aforementioned, algorithms for several special cases of stochastic composition optimization have also been developed recently. For instance, two-layer problems with a linear inner function are considered in both [9] and [31], while the latter one shows the convergence rate of $\mathcal{O}\left(K^{-1}\right)$ (resp. $\mathcal{O}\left(K^{-1/2}\right)$) in strongly convex (resp. general convex) case, which matches the optimal rate for single-layer stochastic optimization. [43] studies the unconstrained multi-level compositional optimization with functions in finite-sum form and proposes a proximal algorithm with nested variance reduced gradient. More related work can also be found in [10, 14, 18, 41, 42]. Prox-linear type algorithms are studied in [27,44] for nonsmooth stochastic compositional optimization with deterministic outer function. There is also some work focusing on the cases where samples are corrupted with Markov noise rather than the common zero-mean noise, such as [30]. In a different way, [45] studies the convex nested stochastic composite optimization and transforms each layer function, which is assumed convex and monotonously non-decreasing, into a minimax problem, then proposes stochastic sequential dual (SSD) methods for two-layer and multi-layer problems. Similar idea is used in [9] for kernel estimation. Except above methods directly aiming for stochastic composition optimization, algorithms proposed for stochastic bilevel problems, such as STABLE [5] and ALSET [6], can also be applied to minimizing $\Gamma(x)+\chi(x)$ over $\mathbb{R}^n$ as mentioned in [6]. Note that all the literature above assumes that the feasible region is easy to project. In [41] an ADMM-type method is studied for convex stochastic composition optimization with linear equality constraints. To the best of our knowledge, however, study on stochastic composition optimization with general constraints, such as (1.1), is still limited.

When it comes to general constrained optimization, for which we assume the feasibility of iterates can not be realized through simple projection, without considering the stochastic nested structure in the objective of (1.1), there has been a surge of works in past decades [22]. And motivated by recent progress in convex constrained stochastic optimization [4, 17, 35], some research effort has been devoted to more general constrained stochastic optimization, such as penalty methods for nonconvex equality constrained stochastic optimization [32], inexact constrained proximal point algorithm for nonconvex inequality constrained stochastic problem [4], stochastic descent methods for the Lagrangian minimax form of nonconvex constrained stochastic problem [20, 23, 39], SQP type methods for nonconvex smooth constrained optimiation with deterministic equality constraints [3, 8] and so on. We studied a stochastic primal-dual method (SPD) for nonconvex optimization with many nonconvex constraints in recent paper [15]. At each iteration, SPD [15] minimizes a linearized augmented Lagrangian function constructed based on the unbiased stochastic gradient of the objective and information of a randomly subsampled set of constraints, to cope with the difficulties caused by the possibly nonconvex feasible set. Nevertheless, these algorithms for constrained optimization only apply to problems with single-layer expectation in the objective and are not suitable for (1.1) with nest structure due to the absence of unbiased stochastic gradients.

1.1 Contributions

When solving nonconvex constrained composition optimization problem (1.1), challenges arise due to the nest structure involved in the objective and the possibly nonconvex constraints. To cope with these issues we propose a STochastic nEsted Primal-dual (STEP) method, aiming for an ϵ-stationary point, whose expected norm of KKT measure in terms of stationarity, primal feasibility as well as complementary slackness in Euclidean norm is less than accuracy ϵ. Introducing an auxiliary variable to track inner layer function values and applying subsampling strategies to calculate stochastic gradients, we construct a stochastic approximation to the linearized augmented Lagrangian function to update primal variable, based on which to further update the dual variable in a weighted-average way. In addition, to understand the asymptotic dynamics of the stochastic update schemes, we consider a deterministic continuous-time system. Under mild conditions, we establish that the iteration and sample complexities of the proposed algorithm to find an ϵ-stationary point are bounded by $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(\epsilon^{-6})$. When initial guess of the primal variable is (nearly) feasible, previous orders can be reduced to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-5})$, respectively. Finally we test our algorithm on solving a risk-averse portfolio optimization problem and its numerical performances look promising. To the best of our knowledge, this is the first work on algorithms for stochastic composition optimization with nonconvex constraints.

In Table 1, we summarize our STEP method and other related algorithms for nonconvex stochastic composition optimization, with total sample complexities and problem types as well as associated assumptions on the problems listed. Here we add a few comments on the differences of our problem settings from those listed in the table. The problem (1.1) we consider here owns a more general (possibly nonconvex) feasible set $\mathcal{X} = \{x \in X \mid g_i(x) \leq 0, i = 1, \ldots, m\}$. Compared with the convex set assumed in NASA [12] and STABLE [5] and the full space $\mathbb{R}^n$ in other literature, the feasibility of iterates with respect to the general set $\mathcal{X}$ could be much more hard to maintain, which raises challenges when solving (1.1). Different from existing work assuming that $\chi \equiv 0$ or an indicator function, we allow a more general nonsmooth regularizer, which broadens the scope of problems our algorithm can solve. In addition, the uniform Lipschitz smoothness assumed in other work is more stringent than the Lipschitz smoothness required in this paper. Besides, in our paper it only requires standard unbiasedness and variance-boundedness of $\mathcal{SFO}$ (Stochastic First-order Oracle to access approximate gradients) and $\mathcal{SZO}$ (Stochastic Zeroth-order Oracle to access approximation function values) which is weaker than the boundedness of fourth (central) moment of $\mathcal{SFO}$ and uniform boundedness of $\mathcal{SFO}$ in other works.

Algorithm	Complexity	$\mathcal{X}$	χ	Smoothness	Ass. ($\mathcal{SFO}$)
a-SCGD [29]	$\mathcal{O}\left(\epsilon^{-7}\right)$	$\mathbb{R}^n$	$\times$	$\oplus$	fourth moment
ASC-PG [31]	$\mathcal{O}\left(\epsilon^{-4.5}\right)$	$\mathbb{R}^n$	$\times$	$\oplus$	uniform
SCSC [7]	$\mathcal{O}\left(\epsilon^{-4}\right)$	$\mathbb{R}^n$	$\times$	$\oplus$	standard
ALSET [6]	$\mathcal{O}\left(\epsilon^{-4}\right)$	$\mathbb{R}^n$	$\times$	$\bigcirc$	standard
NPAG [43]	$\mathcal{O}\left(\epsilon^{-3}\right)$	$\mathbb{R}^n$	$\checkmark$	$\oplus$	standard
STABLE [5]	$\mathcal{O}\left(\epsilon^{-4}\right)$	convex	$\times$	$\oplus$	fourth moment
NASA [12]	$\mathcal{O}\left(\epsilon^{-4}\right)$	convex	$\times$	$\bigcirc$	standard
STEP (this paper)	$\mathcal{O}\left(\epsilon^{-4}\right)$	convex	$\checkmark$	$\bigcirc$	standard
STEP (this paper)	$\mathcal{O}\left(\epsilon^{-5}\right)$	general	$\checkmark$	$\bigcirc$	standard

Table 1: Complexities of different algorithms for minimizing the nonconvex objective function $\Gamma(x) + \chi(x)$ within a feasible region $\mathcal{X} \subseteq \mathbb{R}^n$. Here, "Complexity" refers to the total number of $\mathcal{SFO}$-calls and $\mathcal{SZO}$-calls to achieve some optimality measure, i.e. the KKT measure (for STEP, see Definition 2.2), (generalized)gradient of (nonsmooth) objective for other works below ϵ. In the "Smoothness" column, we list the smoothness assumption on Γ, where "$\oplus$" represents the uniform Lipschitz smoothness, i.e. for all ξ, $\|\nabla F(x;\xi) - \nabla F(y;\xi)\| \leq L_{f,1}\|x - y\|$ with $L_{f,1} > 0$, while "$\bigcirc$" represents Lipschitz smoothness, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L_{f,1}\|x - y\|$. In the last column, "standard" boundedness of $\mathcal{SFO}$ refers to the assumption that the variances or second moment of $\nabla F(y;\xi)$ and $\nabla H(x;\phi)$ are bounded for any $y \in \mathbb{R}^{\bar{n}}, x \in \mathbb{R}^n$, while "uniform" refers to the boundedness of $\nabla F(y;\xi)$ and $\nabla H(x;\phi)$ for any $y \in \mathbb{R}^{\bar{n}}, x \in \mathbb{R}^n$ and any $\phi \in \Xi_l, \xi \in \Xi_u$.

1.2 Notations and organization

We reserve some space for notations used throughout the remainder of the paper. We define $[k] := \{1, \ldots, k\}$ for any positive integer k and $\mathbb{R}^m_+ := \{v \in \mathbb{R}^m : v \geq \mathbf{0}\}$. For any $u \in \mathbb{R}$, its positive and negative parts are denoted by $[u]_+ := \max(0, u)$ and $[u]_- := \max(0, -u)$, respectively, while for any $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$, we define

$[v]_+ := ([v_1]_+, \ldots, [v_n]_+)^T$ and $[v]_- := ([v_1]_-, \ldots, [v_n]_-)^T$, respectively. For a differentiable function F, its gradient at x is denoted by $\nabla F(x)$. For a function $F(x, y)$ differentiable in y, we denote its partial derivative with respect to y at (x, y) as $\nabla_y F(x, y)$. For simplicity, we denote $g = (g_1, \ldots, g_m)^T$ and $J_g = \nabla g = (\nabla g_1, \ldots, \nabla g_m)^T$. Without specification we use $\|\cdot\|$ to denote the Euclidean norm. Given $x, y \in \mathbb{R}^n$, define the inner product $\langle x, y\rangle := x^T y$ and Hadamard product $x \odot y := (x_1 y_1, \ldots, x_n y_n)^T$. Given nonempty sets $X, Y \subseteq \mathbb{R}^n$, $\mathbf{d}(X, Y) := \min_{x\in X, y\in Y} \|x - y\|$ refers to the distance between these two sets and $\mathbf{1}_X$ refers to the indicator function of X, i.e., $\mathbf{1}_X(x)$ equals 0 for $x \in X$ and $+\infty$, otherwise. For random variables ξ and ζ, $\mathbb{E}[\xi]$ represents the expectation of ξ and $\mathbb{E}[\xi \mid \zeta]$ represents the expectation of ξ conditioned on ζ. Besides we use superscript ^{k} to represent the iteration number of iterates, and the subscript $_k$ otherwise.

The rest of the paper is organized as follows. In Section 2, we present the detailed description of a stochastic nested primal-dual method for (1.1). Theoretical properties of the proposed algorithm are established in Section 3. We first bound the Lagrange multipliers through a proper parameter setting scheme, then analyze the theoretical bounds on KKT measure in terms of stationarity, feasibility as well as complementary slackness, and then deduce the iteration and the sample complexities accordingly. In Section 4, numerical experiments on a risk-adverse portfolio optimization problem are implemented. Finally, we give some conclusional remarks.

2 Algorithm description

In this section we will present details of a stochastic nested primal-dual method for (1.1). In the following we assume that only stochastic oracles to $\nabla f(y), \nabla h(x), h(x)$ can be obtained, while $g_i(x), \nabla g_i(x)$, $i \in [m]$, can always be calculated accurately at any inquired point $x \in \mathbb{R}^n$.

In general, it is difficult to obtain a global or even a local minimizer for nonconvex constrained optimization. Attention is often paid to seeking more trackable solutions. It has been shown that under certain constraint qualifications [22], a local minimizer satisfies first-order necessary conditions, known as Karush-Kuhn-Tucker (KKT) conditions. Those satisfying KKT conditions are normally called KKT points. Without specifying the constraint qualifications, in this paper we just assume the existence of a KKT point of (1.1). Before giving its definition, we need following concepts. The normal cone [13] of a convex set X at a point $x \in X$ is denoted by

$$\mathcal{N}_X(x) := \{v \mid \langle v, y - x\rangle \le 0, \forall y \in X\}. \tag{2.1}$$

For the proper lower-semicontinuous convex function $\chi : \mathbb{R}^n \to \mathbb{R}$, its subdifferential [24] at x is defined as

$$\partial\chi(x) := \{d \in \mathbb{R}^n \mid \chi(y) \ge \chi(x) + \langle d, y - x\rangle, \forall y \in \operatorname{dom}\chi \subseteq \mathbb{R}^n\}.$$

And each element of $\partial\chi(x)$ is called as subgradient of χ at x.

Definition **2.1.** *A point $x^* \in X$ is a KKT point of* (1.1)*, if there exists $z^* = (z_1^*, \ldots, z_m^*)^T \in \mathbb{R}_+^m$ such that*

$$\mathbf{0} \in \nabla\Gamma(x^*) + \partial\chi(x^*) + \sum_{i=1}^m z_i^* \nabla g_i(x^*) + \mathcal{N}_X(x^*), \qquad g(x^*) \le \mathbf{0}, \qquad z^* \odot g(x^*) = \mathbf{0}.$$

As is known, the augmented Lagrangian (AL) function plays an important role in characterizing the optimality conditions for constrained optimization and helping design effective algorithms, thus it is widely used in optimization community. The AL function [25] associated with (1.1) is defined as

$$\mathcal{L}_\beta(x, z) = \mathcal{D}_\beta(x, z) + \chi(x), \tag{2.2}$$

where

$$\mathcal{D}_\beta(x, z) := \Gamma(x) + \Psi_\beta(x, z), \quad \Psi_\beta(x, z) := \sum_{i=1}^m \psi_\beta(g_i(x), z_i) \text{ with } \psi_\beta(u, v) = \begin{cases} uv + \frac{\beta}{2}u^2, & \text{if } \beta u + v \ge 0, \\ -\frac{v^2}{2\beta}, & \text{otherwise.} \end{cases}$$

Let $\mathcal{P}_{k,1}, \mathcal{P}_{k,2}$ be two randomly sampled sets from Ξ_l with sizes $P_{k,1}$ and $P_{k,2}$. Calculate function values and gradients of H: $\{H(x^k; \phi), \phi \in \mathcal{P}_{k,1}\}$, $\{\nabla H(x^k; \phi), \phi \in \mathcal{P}_{k,2}\}$. Similar to [29], we use an auxiliary variable y to track the inner layer function value and update it in a moving average way:

$$y^{k+1} = (1 - \eta_k) y^k + \frac{\eta_k}{P_{k,1}} \sum_{\phi\in\mathcal{P}_{k,1}} H(x^k; \phi) \tag{2.3}$$

with $\eta_k \in (0,1]$. Then we randomly generate a sample set $\mathcal{J}_k$ from Ξ_u with $J_k := |\mathcal{J}_k|$ and compute a set of gradients $\{\nabla F(y^{k+1};\xi), \xi \in \mathcal{J}_k\}$ based on which we obtain

$$\overline{\nabla}\Gamma^k := \left[\frac{1}{P_{k,2}} \sum_{\phi \in \mathcal{P}_{k,2}} \nabla H\left(x^k;\phi\right)\right]^T \left[\frac{1}{J_k} \sum_{\xi \in \mathcal{J}_k} \nabla F\left(y^{k+1};\xi\right)\right]. \tag{2.4}$$

By getting access to $\nabla g_i(x^k)$, $g_i(x^k), i \in [m]$, we can further calculate

$$\overline{\nabla}G^k := \nabla_x \Psi_\beta\left(x^k, z^k\right) = \sum_{i=1}^{m} \left[\beta g_i\left(x^k\right) + z_i^k\right]_+ \nabla g_i\left(x^k\right). \tag{2.5}$$

It is easy to see that $\overline{\nabla}\Gamma^k$ is a stochastic approximation to $\nabla h(x^k)^T \nabla f(y^{k+1})$ and also an approximation to true gradient $\nabla\Gamma(x^k)$, since $\nabla\Gamma(x^k) = \nabla h(x^k)\nabla f(h(x^k))$. Thus, if y^{k+1} is sufficiently close to $h\left(x^k\right)$, $\left(\overline{\nabla}\Gamma^k + \overline{\nabla}G^k\right)$ will be a good approximation to $\nabla_x \mathcal{D}_\beta(x^k, z^k)$. Then it is straightforward to have

$$\mathcal{L}_\beta(x, z^k) \approx \mathcal{D}_\beta(x^k, z^k) + \langle \overline{\nabla}\Gamma^k + \overline{\nabla}G^k, x - x^k\rangle + \frac{1}{2\alpha_k}\|x - x^k\|^2 + \chi(x),$$

with $\alpha_k > 0$, which leads to the proximal subproblem we are about to solve to update primal variable:

$$x^{k+1} = \operatorname*{argmin}_{x \in X} \left\{\left\langle \overline{\nabla}\Gamma^k + \overline{\nabla}G^k, x\right\rangle + \frac{1}{2\alpha_k}\left\|x - x^k\right\|^2 + \chi(x)\right\}. \tag{2.6}$$

It is noteworthy that different from classical double-loop AL methods, only an approximation to the original AL function is minimized in (2.6). This idea is actually motivated by linearized AL method, such as [33,34,36], which makes the whole algorithm become a single-loop algorithm. Recall that in classical AL methods the Lagrange multiplier is calculated as

$$z^{k+1} = \bar{z}^{k+1}, \text{ where } \bar{z}^{k+1} = z^k + \beta \max\left(-\frac{z^k}{\beta}, g\left(x^{k+1}\right)\right).$$

Differently, however, since the subproblem (2.6) is built on an approximation to the AL function, it seems unnecessary to follow above regime to update z. Instead, to control the deviation between z^k and z^{k+1} we can adopt the following moving average way:

$$\begin{aligned} z^{k+1} &= \frac{\rho_k}{\beta}\bar{z}^{k+1} + \left(1 - \frac{\rho_k}{\beta}\right) z^k \\ &= \frac{\rho_k}{\beta}\left(z^k + \beta \max\left(-\frac{z^k}{\beta}, g\left(x^{k+1}\right)\right)\right) + \left(1 - \frac{\rho_k}{\beta}\right) z^k \\ &= z^k + \rho_k \cdot \max\left(-\frac{z^k}{\beta}, g\left(x^{k+1}\right)\right), \end{aligned} \tag{2.7}$$

where $\rho_k \in (0, \beta]$. This strategy can help to derive the boundedness of z^k in next section.

In order to better understand the asymptotic dynamics of the stochastic discrete-time update schemes (2.6) and (2.7), let us consider a corresponding continuous-time system in deterministic setting. We take a special case for an example: problem (1.1) with $X = \mathbb{R}^n$ and $\chi \equiv 0$. Consider the tendency of $\mathcal{L}_\beta(x, z)$ following the trajectories $x = x(t)$ and $z = z(t)$ defined by a system of ordinary differential equations:

$$\dot{x}\left(t\right) = -\alpha\left(\nabla h\left(x\left(t\right)\right)^T \nabla f\left(y\left(t\right)\right) + \nabla_x \Psi_\beta\left(x\left(t\right), z\right)\right), \tag{2.8}$$

$$\dot{z}\left(t\right) = \rho \nabla_z \Psi_\beta\left(x, z\left(t\right)\right), \tag{2.9}$$

where $\alpha > 0, \rho > 0$ and $y(t)$ is an approximation to $h(x(t))$. By [11], (2.8) and (2.9) can be regarded as the continuous analogues of (2.6) and (2.7), respectively. From (2.8), it holds that

$$\begin{aligned} \frac{d}{dt}\mathcal{L}_\beta\left(x\left(t\right), z\right) &= \left[\nabla_x \mathcal{L}_\beta(x\left(t\right), z)\right]^T \dot{x}\left(t\right) \\ &= -\alpha\left[\nabla_x \mathcal{L}_\beta\left(x\left(t\right), z\right)\right]^T \left[\nabla_x \mathcal{L}_\beta\left(x\left(t\right), z\right) + \nabla h\left(x\left(t\right)\right)^T \nabla f\left(y\left(t\right)\right) - \nabla h\left(x\left(t\right)\right)^T \nabla f\left(h\left(x\left(t\right)\right)\right)\right] \end{aligned}$$

$$
\begin{aligned}
&\leq -\alpha \left\|\nabla_x \mathcal{L}_\beta\left(x\left(t\right), z\right)\right\|^2 + \alpha \left\|\nabla_x \mathcal{L}_\beta\left(x\left(t\right), z\right)\right\| \left\|\nabla h\left(x\left(t\right)\right)^T \nabla f\left(y\left(t\right)\right) - \nabla h\left(x\left(t\right)\right)^T \nabla f\left(h\left(x\left(t\right)\right)\right)\right\| \\
&\leq -\frac{\alpha}{2} \left\|\nabla_x \mathcal{L}_\beta\left(x\left(t\right), z\right)\right\|^2 + \frac{\alpha}{2} \left\|\nabla h\left(x\left(t\right)\right)^T \nabla f\left(y\left(t\right)\right) - \nabla h\left(x\left(t\right)\right)^T \nabla f\left(h\left(x\left(t\right)\right)\right)\right\|^2, \qquad (2.10)
\end{aligned}
$$

where the last inequality is due to the Young's inequality. Then $\frac{d}{dt}\mathcal{L}_\beta(x\left(t\right), z)$ can be nonpositive when the second term of the right hand side of (2.10) is no larger than the first one. Especially when $\nabla_x \mathcal{L}_\beta(x(t), z) \neq \mathbf{0}$, which normally implies that $(x(t), z)$ is not a saddle point for $\min_{x\in\mathbb{R}^n} \max_{z\in\mathbb{R}^m_+} \mathcal{L}_\beta(x, z)$, $\mathcal{L}_\beta(x(t), z)$ can be monotonically decreasing provided that $y(t)$ is sufficiently close to $h(x(t))$. Meanwhile, it follows from (2.9) that $\mathcal{L}_\beta(x, z(t))$ keeps monotonicity with respect to t, due to the fact that

$$
\frac{d}{d_t}\mathcal{L}_\beta(x, z(t)) = \nabla_z \mathcal{L}_\beta(x, z(t))^T \dot{z}(t) = \rho \|\nabla_z \mathcal{L}_\beta(x, z(t))\|^2 \geq 0.
$$

We are now ready to present the stochastic nested primal-dual method for (1.1) in Algorithm 2.1 below.

Algorithm 2.1 STochastic nEsted Primal-dual (STEP) method for (1.1)

Input: Initial points $x^0 \in X, y^0 \in \mathbb{R}^{\bar{n}}, z^0 = \mathbf{0} \in \mathbb{R}^m$, parameters $\beta > 0$, $\{\alpha_k\} \subseteq (0, +\infty)$, $\{\eta_k\} \subseteq (0, 1]$, $\{\rho_k\} \subseteq (0, \beta]$ and a positive integer K;

Output: x^{R+1} and $\bar{z} := [\beta g(x^{R+1}) + z^{R+1}]_+$ where R is uniformly randomly chosen from $\{1, \ldots, K\}$.

1: **for** $k = 0$ to K **do**
2: Choose independent identical-distributed samples $\mathcal{P}_{k,1} \subseteq \Xi_l, \mathcal{P}_{k,2} \subseteq \Xi_l, \mathcal{J}_k \subseteq \Xi_u$ according to the probability distribution function on the respective probability space.
3: Compute y^{k+1} through (2.3).
4: Compute $\overline{\nabla}\Gamma^k$ and $\overline{\nabla}G^k$ through (2.4) and (2.5), respectively.
5: Compute x^{k+1} through (2.6).
6: Compute z^{k+1} through (2.7).
7: **end for**

It is straightforward to obtain the nonnegativity of z^k from $z^0 = \mathbf{0}$, $\rho_k \subseteq (0, \beta]$ and (2.7). So we state the lemma omitting the proof.

Lemma **2.1.** *For any $k \geq 0$, $z^k \in \mathbb{R}^m_+$.*

Since the iteration process applying the STEP method for (1.1) is random, we aim for an ϵ-stationary point of (1.1) with the expected KKT measure below a given tolerance ϵ.

Definition **2.2.** *Given $\epsilon > 0$, a point x is called an ϵ-stationary point of* (1.1)*, if there exists $z \in \mathbb{R}^m_+$ such that*

$$
\mathbb{E}\left[\mathbf{d}\left(\nabla\Gamma\left(x\right) + \partial\chi\left(x\right) + \sum_{i=1}^m z_i \nabla g_i\left(x\right) + \mathcal{N}_X\left(x\right), \mathbf{0}\right)\right] < \epsilon, \qquad (2.11)
$$

$$
\mathbb{E}\left[\left\|\left[g\left(x\right)\right]_+\right\|\right] < \epsilon, \qquad (2.12)
$$

$$
\mathbb{E}\left[\left\|z \odot g(x)\right\|\right] < \epsilon, \qquad (2.13)
$$

where the expectation is taken with respect to all the random variables generated in the iteration process.

3 Theoretical analysis

Our goal in this section is to study theoretical properties of the STEP method. Eventually we will establish its iteration and sample complexities to find an ϵ-stationary point of (1.1).

3.1 Preliminaries

We first lay out some preliminary assumptions and lemmas preparing for later analysis on theoretical complexities of the STEP method. Throughout the remainder of this paper, we make the following assumptions.

Assumption 3.1. *$F(\cdot;\xi)$ and $H(\cdot;\phi)$ are differentiable almost surely for $\xi \in \Xi_u$ and $\phi \in \Xi_l$, respectively. Functions $f, \nabla f, h$ and ∇h are Lipschitz continuous with constants $L_{f,0}$, $L_{f,1}, L_{h,0}$ and $L_{h,1}$, respectively. Functions $g_i, i \in [m]$, are $L_{g,0}$-Lipschitz continuous and $\nabla g_i, i \in [m]$, are $L_{g,1}$-Lipschitz continuous. That is for any $x, \bar{x} \in \mathbb{R}^n$ and $y, \bar{y} \in \mathbb{R}^{\bar{n}}$,*

$$\|f(y) - f(\bar{y})\| \le L_{f,0} \|y - \bar{y}\|, \ \|\nabla f(y) - \nabla f(\bar{y})\| \le L_{f,1} \|y - \bar{y}\|, \tag{3.1}$$

$$\|h(x) - h(\bar{x})\| \le L_{h,0} \|x - \bar{x}\|, \ \|\nabla h(x) - \nabla h(\bar{x})\| \le L_{h,1} \|x - \bar{x}\|, \tag{3.2}$$

$$\|g_i(x) - g_i(\bar{x})\| \le L_{g,0} \|x - \bar{x}\|, \ \|\nabla g_i(x) - \nabla g_i(\bar{x})\| \le L_{g,1} \|x - \bar{x}\|, i \in [m]. \tag{3.3}$$

Assumption 3.2. *There exist positive constants G and G_χ such that*

$$\left[g_i(x^k)\right]_+ \le G \text{ and } \|\partial\chi(x^k)\| \le G_\chi, \quad \forall k \ge 0; \forall i \in [m]. \tag{3.4}$$

From Assumption 3.1 it holds that for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^{\bar{n}}$,

$$\|\nabla f(y)\| \le L_{f,0}, \|\nabla h(x)\| \le L_{h,0}, \|\nabla g_i(x)\| \le L_{g,0}, \quad \forall i \in [m]. \tag{3.5}$$

It is worthy to mention that those two assumptions are not more stringent compared with existing literature. For example, (3.1)-(3.2) are also required in ALSET [6] and NASA [12], while the work on constrained stochastic optimization [3] also assumes (3.3). In addition, the boundedness on $[g_i]_+$ and $\|\partial\chi\|$ is also required in [3] and ASC-PG [31] respectively. In [35] it assumes the boundedness of $|g_i|$ which is stronger than ours.

Under Assumption 3.1 we have the following lemma.

Lemma **3.1.** *Under Assumption 3.1, it holds that for any $x, \bar{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}^{\bar{n}}$,*

$$\left\|\nabla h(x)^T \nabla f(h(x)) - \nabla h(x)^T \nabla f(y)\right\| \le L_f \|h(x) - y\|, \quad \|\nabla\Gamma(x) - \nabla\Gamma(\bar{x})\| \le L_\Gamma \|x - \bar{x}\|,$$

where $L_f := L_{h,0}L_{f,1}, L_\Gamma := L_{h,0}^2 L_{f,1} + L_{f,0}L_{h,1}$.

Proof. It follows from (3.1) and (3.5) that

$$\left\|\nabla h(x)^T \nabla f(h(x)) - \nabla h(x)^T \nabla f(y)\right\| \le \|\nabla h(x)\| \|\nabla f(h(x)) - \nabla f(y)\| \le L_{h,0}L_{f,1} \|h(x) - y\|.$$

And (3.1)-(3.2) and (3.5) indicate

$$\begin{aligned}
\|\nabla\Gamma(x) - \nabla\Gamma(\bar{x})\| &= \left\|\nabla h(x)^T \nabla f(h(x)) - \nabla h(\bar{x})^T \nabla f(h(\bar{x}))\right\| \\
&\le \|\nabla h(x)\| \|\nabla f(h(x)) - \nabla f(h(\bar{x}))\| + \|\nabla h(x) - \nabla h(\bar{x})\| \|\nabla f(h(\bar{x}))\| \\
&\le L_{h,0}L_{f,1} \|h(x) - h(\bar{x})\| + L_{h,1}L_{f,0} \|x - \bar{x}\| \\
&\le \left(L_{h,0}^2 L_{f,1} + L_{h,1}L_{f,0}\right) \|x - \bar{x}\|.
\end{aligned}$$

The proof is completed. □

Besides the nonnegativity of z^k shown in Lemma 2.1, we can also guarantee its boundedness under proper setting of ρ_k. Without loss of generality, we assume

$$\rho_k \in \left(0, \frac{\rho}{K}\right] \subseteq (0, \beta], \quad k = 0, \ldots, K,$$

where $\rho > 0$ is independent of K.

Lemma **3.2.** *Under Assumption 3.2, it holds that for any $k \in [K+1]$,*

$$z_i^k \le G \sum_{t=0}^{k-1} \rho_t \le 2G\rho, \quad \forall i \in [m]. \tag{3.6}$$

Proof. We will show the result by induction. First, by the update formula of z^k we have

$$z_i^1 = z_i^0 + \rho_0 \max\left(\frac{-z_i^0}{\beta}, g_i\left(x^1\right)\right) = \rho_0\left[g_i\left(x^1\right)\right]_+ \le \rho_0 G \le G\rho, \quad \forall i \in [m].$$

Assume (3.6) holds for k. It follows from Lemma 2.1 that for any $i \in [m]$,

$$z_i^{k+1} = z_i^k + \rho_k \max\left(\frac{-z_i^k}{\beta}, g_i\left(x^{k+1}\right)\right) \le \begin{cases} z_i^k \le G\sum_{t=0}^{k-1}\rho_t \le G\sum_{t=0}^{k}\rho_t, & \text{if } g_i\left(x^{k+1}\right) \le 0; \\ z_i^k + \rho_k g_i\left(x^{k+1}\right) \le z_i^k + \rho_k G \le G\sum_{t=0}^{k}\rho_t, & \text{otherwise.} \end{cases}$$

Together with $\sum_{t=0}^{k-1}\rho_t \le \rho k/K \le 2\rho$ for any $k \in [K+1]$, it yields the conclusion. □

Applying the nonnegativity and boundedness of z^k, we can provide estimations of one-iteration increment of z and the smoothness of $\Psi_\beta(x,z)$ with respect to x for fixed z as follows.

LEMMA **3.3.** *Under assumptions 3.1 and 3.2, we have that for any $k = 0, 1, \ldots, K$,*

$$\left|z_i^{k+1} - z_i^k\right| \le \frac{\rho G}{K}\left(\frac{2\rho}{\beta} + 1\right), \quad \forall i \in [m], \tag{3.7}$$

$$\|\nabla_x \Psi_\beta\left(x, z^k\right) - \nabla_x \Psi_\beta\left(x^k, z^k\right)\| \le L_\beta \|x - x^k\|, \quad \forall x \in \mathbb{R}^n, \tag{3.8}$$

where $L_\beta = \beta L_{g,0}^2 m + \beta G L_{g,1} m + 2L_{g,1} G\rho m$.

Proof. Firstly, it follows from lemmas 2.1 and 3.2 that for any $i \in [m]$,

$$\left|z_i^{k+1} - z_i^k\right| = \rho_k \left|\max\left(\frac{-z_i^k}{\beta}, g_i\left(x^{k+1}\right)\right)\right| \le \begin{cases} \rho_k \frac{z_i^k}{\beta} \le \rho_k \frac{2G\rho}{\beta}, & \text{if } g_i(x^{k+1}) \le 0, \\ \rho_k\left[g_i\left(x^{k+1}\right)\right]_+ \le \rho_k G, & \text{otherwise.} \end{cases}$$

which yields (3.7).

Secondly, it follows from (2.5), (3.3) and (3.4) that for any $x \in \mathbb{R}^n$,

$$\begin{aligned} &\left\|\nabla_x \Psi_\beta(x,z) - \nabla_x \Psi_\beta\left(x^k, z\right)\right\| \\ \le &\sum_{i=1}^m \left\|[\beta g_i(x) + z_i]_+ \nabla g_i(x) - \left[\beta g_i\left(x^k\right) + z_i\right]_+ \nabla g_i\left(x^k\right)\right\| \\ = &\sum_{i=1}^m \left\|\left[[\beta g_i(x) + z_i]_+ - \left[\beta g_i\left(x^k\right) + z_i\right]_+\right] \nabla g_i(x) + \left[\beta g_i\left(x^k\right) + z_i\right]_+ \left[\nabla g_i(x) - \nabla g_i\left(x^k\right)\right]\right\| \\ \le &\sum_{i=1}^m \left[\beta\left|g_i(x) - g_i\left(x^k\right)\right| \|\nabla g_i(x)\| + \left[\beta g_i\left(x^k\right) + z_i\right]_+ L_{g,1}\left\|x - x^k\right\|\right] \\ \le &\sum_{i=1}^m \left[\beta L_{g,0}^2\left\|x - x^k\right\| + (\beta G + |z_i|) L_{g,1}\left\|x - x^k\right\|\right] \\ = &\left(\beta L_{g,0}^2 m + \beta G L_{g,1} m + L_{g,1}\|z\|_1\right)\left\|x - x^k\right\| \end{aligned} \tag{3.9}$$

which together with Lemma 3.2 indicates (3.8). □

To proceed our analysis we need another assumption, which is commonly used in stochastic optimization.

Assumption 3.3. *There exist positive constants $\sigma_f, \sigma_{h,0}, \sigma_{h,1}$ such that for any $x \in \mathbb{R}^n, y \in \mathbb{R}^{\bar{n}}$,*

(1) $\mathbb{E}_\xi[\nabla F(y;\xi)] = \nabla f(y), \mathbb{E}_\xi[\|\nabla F(y;\xi) - \nabla f(y)\|^2] \le \sigma_f^2$ *;*

(2) $\mathbb{E}_\phi[H(x;\phi)] = h(x), \mathbb{E}_\phi[\|H(x;\phi) - h(x)\|^2] \le \sigma_{h,0}^2$ *;*

(3) $\mathbb{E}_\phi[\nabla H(x;\phi)] = \nabla h(x), \mathbb{E}_\phi[\|\nabla H(x;\phi) - \nabla h(x)\|^2] \le \sigma_{h,1}^2$.

In the following, define the filtration

$$\mathcal{H}^k = \left\{x^0, y^0, z^0, \ldots, x^k, y^k, z^k\right\} \text{ and } \overline{\mathcal{H}}^k = \left\{x^0, y^0, z^0, \ldots, x^k, y^k, z^k, y^{k+1}\right\}, \quad k \geq 0.$$

The next lemma characterizes properties of the stochastic approximation $\overline{\nabla}\Gamma^k$.

LEMMA **3.4.** *Under assumptions 3.1 and 3.3, it holds that for any $k = 0, \ldots, K$,*

$$\mathbb{E}\left[\left\|\overline{\nabla}\Gamma^k - \nabla h\left(x^k\right)^T \nabla f\left(y^{k+1}\right)\right\|^2\right] \leq \sigma_{\Gamma_k}^2, \tag{3.10}$$

$$\mathbb{E}\left[\left\|\overline{\nabla}\Gamma^k - \nabla\Gamma\left(x^k\right)\right\|^2\right] \leq 2L_f^2\mathbb{E}\left[\left\|h\left(x^k\right) - y^{k+1}\right\|^2\right] + 2\sigma_{\Gamma_k}^2, \tag{3.11}$$

where $\sigma_{\Gamma_k}^2 = 2(L_{f,0}^2 + \sigma_f^2)\sigma_{h,1}^2/P_{k,2} + 2L_{h,0}^2\sigma_f^2/J_k$, L_f is defined in Lemma 3.1 and the expectation is taken with respect to all the random variables generated up to kth iteration.

Proof. It indicates from (2.4) that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}_k,\mathcal{P}_{k,2}}\left[\left\|\overline{\nabla}\Gamma^k - \nabla h\left(x^k\right)^T \nabla f\left(y^{k+1}\right)\right\|^2 \mid \overline{\mathcal{H}}^k\right] \\
=&\mathbb{E}_{\mathcal{J}_k,\mathcal{P}_{k,2}}\left[\left\|\left[\frac{1}{P_{k,2}}\sum_{\phi\in\mathcal{P}_{k,2}}\nabla H\left(x^k,\phi\right)\right]^T\left[\frac{1}{J_k}\sum_{\xi\in\mathcal{J}_k}\nabla F\left(y^{k+1},\xi\right)\right] - \nabla h\left(x^k\right)^T\nabla f\left(y^{k+1}\right)\right\|^2 \mid \overline{\mathcal{H}}^k\right] \\
\leq&\mathbb{E}_{\mathcal{J}_k,\mathcal{P}_{k,2}}\left[2\left\|\left[\frac{1}{P_{k,2}}\sum_{\phi\in\mathcal{P}_{k,2}}\nabla H\left(x^k,\phi\right) - \nabla h\left(x^k\right)\right]^T\left[\frac{1}{J_k}\sum_{\xi\in\mathcal{J}_k}\nabla F\left(y^{k+1},\xi\right)\right]\right\|^2 \mid \overline{\mathcal{H}}^k\right] \\
&+\mathbb{E}_{\mathcal{J}_k,\mathcal{P}_{k,2}}\left[2\left\|\nabla h\left(x^k\right)^T\left[\frac{1}{J_k}\sum_{\xi\in\mathcal{J}_k}\nabla F\left(y^{k+1},\xi\right) - \nabla f\left(y^{k+1}\right)\right]\right\|^2 \mid \overline{\mathcal{H}}^k\right] \\
=&2\mathbb{E}_{\mathcal{P}_{k,2}}\left[\left\|\frac{1}{P_{k,2}}\sum_{\phi\in\mathcal{P}_{k,2}}\nabla H\left(x^k,\phi\right) - \nabla h\left(x^k\right)\right\|^2 \mid \overline{\mathcal{H}}^k\right]\mathbb{E}_{\mathcal{J}_k}\left[\left\|\frac{1}{J_k}\sum_{\xi\in\mathcal{J}_k}\nabla F\left(y^{k+1},\xi\right)\right\|^2 \mid \overline{\mathcal{H}}^k\right] \\
&+2\left\|\nabla h\left(x^k\right)\right\|^2\mathbb{E}_{\mathcal{J}_k}\left[\left\|\frac{1}{J_k}\sum_{\xi\in\mathcal{J}_k}\nabla F\left(y^{k+1},\xi\right) - \nabla f\left(y^{k+1}\right)\right\|^2 \mid \overline{\mathcal{H}}^k\right] \\
\leq&2\frac{\sigma_{h,1}^2}{P_{k,2}}\left(L_{f,0}^2 + \frac{\sigma_f^2}{J_k}\right) + 2L_{h,0}^2\frac{\sigma_f^2}{J_k},
\end{aligned} \tag{3.12}$$

where the first inequality comes from $\|A+B\|^2 \leq 2\|A\|^2 + 2\|B\|^2$, the second equality comes from the independence of $\mathcal{P}_{k,2}$ and $\mathcal{J}_k$, the last inequality uses the independence of $(\mathcal{J}_k, \mathcal{P}_{k,2})$ and $\overline{\mathcal{H}}^k$, Assumption 3.3 and (3.5). Taking expectation of (3.12) with respect to all the samples related with $\overline{\mathcal{H}}^k$, we obtain (3.10).

In addition, by the definition of $\overline{\nabla}\Gamma^k$, Lemma 3.1 and (3.10), it holds that

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{\nabla}\Gamma^k - \nabla\Gamma\left(x^k\right)\right\|^2\right] \leq& 2\mathbb{E}\left[\left\|\nabla h\left(x^k\right)^T\nabla f\left(h\left(x^k\right)\right) - \nabla h\left(x^k\right)^T\nabla f\left(y^{k+1}\right)\right\|^2 + \left\|\nabla h\left(x^k\right)^T\nabla f\left(y^{k+1}\right) - \overline{\nabla}\Gamma^k\right\|^2\right] \\
\leq& 2L_{h,0}^2L_{f,1}^2\mathbb{E}\left[\left\|h\left(x^k\right) - y^{k+1}\right\|^2\right] + 2\sigma_{\Gamma_k}^2
\end{aligned}$$

which derives (3.11). □

3.2 Iteration and sample complexities

In this part we aim for characterizing iteration and sample complexities of the STEP method to find an ϵ-stationary point of (1.1). To achieve this goal, we need to analyze the KKT measure defined in Definition 2.2 in terms of stationarity, feasibility and complementary slackness separately.

In the following, we will first analyze the stationarity measure. Without any specification, the expectation is taken with respect to all random variables generated up to the latest iteration.

LEMMA **3.5.** *Under assumptions 3.1-3.3, it holds that for any $k = 0, \ldots, K$,*

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{k+1}\right) + \partial\chi\left(x^{k+1}\right) + \nabla_x\Psi_\beta\left(x^{k+1}, z^{k+1}\right) + \mathcal{N}_X\left(x^{k+1}\right), \mathbf{0}\right)\right] \\
\leq & 3\left(L_\Gamma + L_\beta + \frac{1}{\alpha_k}\right)^2 \mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] + 6L_f^2\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2\right] + 3L_{g,0}^2 m\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|^2\right] + 6\sigma_{\Gamma_k}^2,
\end{aligned}$$

where σ_{Γ_k} is defined in Lemma 3.4.

Proof. For any $k \geq 0$, it follows from optimality conditions for (2.6) that

$$-\overline{\nabla}\Gamma^k - \overline{\nabla}G^k - \frac{1}{\alpha_k}\left(x^{k+1} - x^k\right) \in \partial\chi\left(x^{k+1}\right) + \mathcal{N}_X\left(x^{k+1}\right).$$

Applying this relation and $\nabla_x\Psi_\beta\left(x^k, z^k\right) = \overline{\nabla}G^k$ we obtain

$$\begin{aligned}
&\mathbf{d}\left(\nabla\Gamma\left(x^{k+1}\right) + \partial\chi\left(x^{k+1}\right) + \nabla_x\Psi_\beta\left(x^{k+1}, z^{k+1}\right) + \mathcal{N}_X\left(x^{k+1}\right), \mathbf{0}\right) \\
\leq & \left\|\nabla\Gamma\left(x^{k+1}\right) + \nabla_x\Psi_\beta\left(x^{k+1}, z^{k+1}\right) - \overline{\nabla}\Gamma^k - \overline{\nabla}G^k - \frac{1}{\alpha_k}\left(x^{k+1} - x^k\right)\right\| \\
\leq & \left\|\nabla\Gamma\left(x^{k+1}\right) - \nabla\Gamma\left(x^k\right)\right\| + \left\|\nabla\Gamma\left(x^k\right) - \overline{\nabla}\Gamma^k\right\| + \left\|\nabla_x\Psi_\beta\left(x^{k+1}, z^{k+1}\right) - \nabla_x\Psi_\beta\left(x^{k+1}, z^k\right)\right\| \\
& + \left\|\nabla_x\Psi_\beta\left(x^{k+1}, z^k\right) - \nabla_x\Psi_\beta\left(x^k, z^k\right)\right\| + \frac{1}{\alpha_k}\left\|x^{k+1} - x^k\right\| \\
\leq & \left(L_\Gamma + L_\beta + \frac{1}{\alpha_k}\right)\left\|x^{k+1} - x^k\right\| + \left\|\nabla\Gamma\left(x^k\right) - \overline{\nabla}\Gamma^k\right\| + L_{g,0}\sum_{i=1}^m \left|z_i^{k+1} - z_i^k\right|,
\end{aligned}$$

where the last inequality is indicated by Lemma 3.1, Lemma 3.3 and

$$\begin{aligned}
\left\|\nabla_x\Psi_\beta\left(x^{k+1}, z^{k+1}\right) - \nabla_x\Psi_\beta\left(x^{k+1}, z^k\right)\right\| &\leq \sum_{i=1}^m \left\|\left[\left[\beta g_i\left(x^{k+1}\right) + z_i^{k+1}\right]_+ - \left[\beta g_i\left(x^{k+1}\right) + z_i^k\right]_+\right]\nabla g_i\left(x^{k+1}\right)\right\| \\
&\leq \sum_{i=1}^m \left|z_i^{k+1} - z_i^k\right| \left\|\nabla g_i(x^{k+1})\right\| \leq L_{g,0}\left\|z^{k+1} - z^k\right\|_1.
\end{aligned}$$

Together with $\left(\sum_{i=1}^m a_i\right)^2 \leq m\sum_{i=1}^m a_i^2$ for any $a_1, \ldots, a_m \in \mathbb{R}$, it implies that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{k+1}\right) + \partial\chi\left(x^{k+1}\right) + \nabla_x\Psi_\beta\left(x^{k+1}, z^{k+1}\right), \mathbf{0}\right)\right] \\
\leq & 3\left(L_\Gamma + L_\beta + \frac{1}{\alpha_k}\right)^2 \mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] + 3\mathbb{E}\left[\left\|\nabla\Gamma\left(x^k\right) - \overline{\nabla}\Gamma^k\right\|^2\right] + 3L_{g,0}^2 m\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|^2\right],
\end{aligned}$$

which together with (3.11) yields the conclusion. □

Motivated by Lemma 3.5, we can further estimate the increment of two successive iterates.

LEMMA **3.6.** *Under assumptions 3.1-3.3, it holds that for any $k = 0, \ldots, K$,*

$$\begin{aligned}
&\left(\frac{1}{2\alpha_k} - \frac{L_\Gamma + L_\beta}{2}\right)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
\leq & \mathbb{E}\left[\mathcal{L}_\beta\left(x^k, z^k\right) - \mathcal{L}_\beta\left(x^{k+1}, z^{k+1}\right)\right] + \mathbb{E}\left[\sum_{i=1}^m \max\left(G, \frac{z_i^k}{\beta}, \frac{z_i^{k+1}}{\beta}\right)\left|z_i^k - z_i^{k+1}\right|\right] \\
& + \alpha_k L_f^2\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2\right] + \alpha_k\sigma_{\Gamma_k}^2. \qquad (3.13)
\end{aligned}$$

Proof. By lemmas 3.1 and 3.3, $\mathcal{D}_\beta\left(x, z^k\right)$ is $(L_\Gamma + L_\beta)$-smooth in x for any k. Then we have

$$\mathcal{D}_\beta\left(x^{k+1}, z^k\right) \leq \mathcal{D}_\beta\left(x^k, z^k\right) + \left\langle\nabla\Gamma\left(x^k\right) + \nabla_x\Psi_\beta\left(x^k, z^k\right), x^{k+1} - x^k\right\rangle + \frac{L_\Gamma + L_\beta}{2}\left\|x^{k+1} - x^k\right\|^2. \qquad (3.14)$$

It follows from optimality conditions for (2.6) that there exists a vector $v \in \mathcal{N}_X\left(x^{k+1}\right)$ satisfying

$$-v-\overline{\nabla}\Gamma^k-\overline{\nabla}G^k-\frac{1}{\alpha_k}\left(x^{k+1}-x^k\right)\in\partial\chi\left(x^{k+1}\right),$$

which by the convexity of χ and (2.1) indicates

$$\begin{aligned}\chi\left(x^{k+1}\right)\leq&\,\chi\left(x^k\right)+\left\langle -v-\overline{\nabla}\Gamma^k-\overline{\nabla}G^k-\frac{1}{\alpha_k}\left(x^{k+1}-x^k\right),x^{k+1}-x^k\right\rangle\\ \leq&\,\chi\left(x^k\right)-\left\langle\overline{\nabla}\Gamma^k+\overline{\nabla}G^k,x^{k+1}-x^k\right\rangle-\frac{1}{\alpha_k}\left\|x^{k+1}-x^k\right\|^2.\end{aligned}\tag{3.15}$$

Summing up (3.14) and (3.15) and by $\mathcal{L}_\beta(x,z)=\mathcal{D}_\beta(x,z)+\chi(x)$ we have

$$\begin{aligned}&\mathcal{L}_\beta\left(x^{k+1},z^k\right)\\ \leq&\,\mathcal{L}_\beta\left(x^k,z^k\right)+\left\langle\nabla\Gamma\left(x^k\right)+\nabla_x\Psi_\beta\left(x^k,z^k\right)-\overline{\nabla}\Gamma^k-\overline{\nabla}G^k,x^{k+1}-x^k\right\rangle+\left(\frac{L_\Gamma+L_\beta}{2}-\frac{1}{\alpha_k}\right)\left\|x^{k+1}-x^k\right\|^2\\ \leq&\,\mathcal{L}_\beta\left(x^k,z^k\right)+\frac{\alpha_k}{2}\left\|\nabla\Gamma\left(x^k\right)-\overline{\nabla}\Gamma^k\right\|^2+\left(\frac{L_\Gamma+L_\beta}{2}-\frac{1}{2\alpha_k}\right)\left\|x^{k+1}-x^k\right\|^2,\end{aligned}\tag{3.16}$$

where the last inequality follows from $\nabla_x\Psi_\beta\left(x^k,z^k\right)=\overline{\nabla}G^k$ and $\langle u,v\rangle\leq\frac{a}{2}\|u\|^2+\frac{1}{2a}\|v\|^2$ for any $a>0$. Taking expectation with respect to all the samples generated up to kth iteration on both sides of (3.16) and applying (3.11), we obtain

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}_\beta\left(x^{k+1},z^k\right)\right]\\ \leq&\,\mathbb{E}\left[\mathcal{L}_\beta\left(x^k,z^k\right)\right]+\alpha_kL_f^2\mathbb{E}\left[\left\|h\left(x^k\right)-y^{k+1}\right\|^2\right]+\alpha_k\sigma_{\Gamma_k}^2+\left(\frac{L_\Gamma+L_\beta}{2}-\frac{1}{2\alpha_k}\right)\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right].\end{aligned}\tag{3.17}$$

We next consider the term $\mathcal{L}_\beta\left(x^{k+1},z^k\right)-\mathcal{L}_\beta\left(x^{k+1},z^{k+1}\right)$ and note

$$\mathcal{L}_\beta\left(x^{k+1},z^k\right)-\mathcal{L}_\beta\left(x^{k+1},z^{k+1}\right)=\sum_{i=1}^m\left[\psi_\beta\left(g_i\left(x^{k+1}\right),z_i^k\right)-\psi_\beta\left(g_i\left(x^{k+1}\right),z_i^{k+1}\right)\right].$$

It can be derived from the definition of $\psi_\beta(u,v)$ in (2.2) that for any $u\in[0,G]$ with $G>0$ and $(v_1,v_2)\in\mathbb{R}_+^2$,

$$\left|\psi_\beta(u,v_1)-\psi_\beta(u,v_2)\right|=\left|uv_1-uv_2\right|=u\left|v_1-v_2\right|\leq G\left|v_1-v_2\right|.$$

And for any $u<0$ and $(v_1,v_2)\in\mathbb{R}_+^2$, we have

$$\begin{aligned}&\left|\psi_\beta(u,v_1)-\psi_\beta(u,v_2)\right|\\ =&\begin{cases}\left|uv_1-uv_2\right|=-u\left|v_1-v_2\right|\leq\frac{\max(v_1,v_2)}{\beta}\left|v_1-v_2\right|, & \text{if } \beta u+v_1\geq0,\beta u+v_2\geq0,\\ -\frac{v_2^2}{2\beta}-uv_1-\frac{\beta}{2}u^2\leq uv_2+\frac{\beta}{2}u^2-uv_1-\frac{\beta}{2}u^2=-u\left|v_1-v_2\right|\leq\frac{\max(v_1,v_2)}{\beta}\left|v_1-v_2\right|, & \text{if } \beta u+v_1\geq0>\beta u+v_2,\\ -\frac{v_1^2}{2\beta}-uv_2-\frac{\beta}{2}u^2\leq uv_1+\frac{\beta}{2}u^2-uv_2-\frac{\beta}{2}u^2=-u\left|v_1-v_2\right|\leq\frac{\max(v_1,v_2)}{\beta}\left|v_1-v_2\right|, & \text{if } \beta u+v_2\geq0>\beta u+v_1,\\ \frac{\left|v_1^2-v_2^2\right|}{2\beta}=\frac{v_1+v_2}{2\beta}\left|v_1-v_2\right|\leq\frac{\max(v_1,v_2)}{\beta}\left|v_1-v_2\right|, & \text{if } \beta u+v_1<0,\beta u+v_2<0,\end{cases}\end{aligned}$$

where equalities in the second and third cases use the monotonically decreasing property of $\psi_\beta(u,v)$ in $v\geq0$ when $u<0$. Therefore, by letting $u=g_i(x^{k+1})$, $v_1=z_i^k$ and $v_2=z_i^{k+1}$ in above relations, it implies that

$$\begin{aligned}\mathcal{L}_\beta\left(x^{k+1},z^k\right)\geq&\,\mathcal{L}_\beta\left(x^{k+1},z^{k+1}\right)-\sum_{i=1}^m\left|\psi_\beta\left(g_i\left(x^{k+1}\right),z_i^k\right)-\psi_\beta\left(g_i\left(x^{k+1}\right),z_i^{k+1}\right)\right|\\ \geq&\,\mathcal{L}_\beta\left(x^{k+1},z^{k+1}\right)-\sum_{i=1}^m\max\left(G,\frac{z_i^k}{\beta},\frac{z_i^{k+1}}{\beta}\right)\left|z_i^k-z_i^{k+1}\right|,\quad k=0,\ldots,K,\end{aligned}\tag{3.18}$$

which together with (3.17) yields the conclusion. □

The following lemma provides us an estimate on the difference between y^{k+1} and $h\left(x^k\right)$.

LEMMA **3.7.** *Under assumptions 3.1-3.3, it holds that for any $\gamma_k > 0$ and $k \in [K]$,*

$$\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2\right] \leq (1+\gamma_k)(1-\eta_k)^2\,\mathbb{E}\left[\left\|y^k - h\left(x^{k-1}\right)\right\|^2\right] + \frac{\eta_k^2}{P_{k,1}}\sigma_{h,0}^2(1+\gamma_k) + \left(1+\gamma_k^{-1}\right)(1-\eta_k)^2 L_{h,0}^2\mathbb{E}\left[\left\|x^k - x^{k-1}\right\|^2\right]. \tag{3.19}$$

Proof. By Young's inequality we can obtain that for any $\gamma_k > 0$,

$$\begin{aligned}\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2 \mid \mathcal{H}^k\right] &\leq (1+\gamma_k)\,\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right) + (1-\eta_k)\left(h\left(x^k\right) - h\left(x^{k-1}\right)\right)\right\|^2 \mid \mathcal{H}^k\right] \\ &\quad + \left(1+\gamma_k^{-1}\right)\mathbb{E}\left[\left\|(1-\eta_k)\left(h\left(x^k\right) - h\left(x^{k-1}\right)\right)\right\|^2 \mid \mathcal{H}^k\right] \\ &\leq (1+\gamma_k)\,\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right) + (1-\eta_k)\left(h\left(x^k\right) - h\left(x^{k-1}\right)\right)\right\|^2 \mid \mathcal{H}^k\right] \\ &\quad + \left(1+\gamma_k^{-1}\right)(1-\eta_k)^2 L_{h,0}^2\left\|x^k - x^{k-1}\right\|^2.\end{aligned}$$

Notably, (2.3) and Assumption 3.3 imply

$$\begin{aligned}&\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right) + (1-\eta_k)\left(h\left(x^k\right) - h\left(x^{k-1}\right)\right)\right\|^2 \mid \mathcal{H}^k\right] \\ =&\,\mathbb{E}\left[\left\|(1-\eta_k)\left(y^k - h\left(x^{k-1}\right)\right) + \eta_k\left(\frac{1}{P_{k,1}}\sum_{\phi\in\mathcal{P}_{k,1}} H\left(x^k;\phi\right) - h\left(x^k\right)\right)\right\|^2 \mid \mathcal{H}^k\right] \\ \leq&\,(1-\eta_k)^2\left\|y^k - h\left(x^{k-1}\right)\right\|^2 + \eta_k^2\frac{\sigma_{h,0}^2}{P_{k,1}},\end{aligned}$$

which yields the conclusion by taking expectations with respect to all the samples generated in $\mathcal{H}^k$. □

Now we are ready to state the main theorem regarding the stationarity of x^{R+1} associated with $\bar{z}$.

THEOREM **3.1. (Stationarity)** *Under assumptions 3.1-3.3, set $\beta = K^{1/4}$ and*

$$\begin{aligned}&\rho_k \equiv \frac{\rho}{K}, P_{k,1} \equiv P_1 := \lceil K^{1/4}\rceil, P_{k,2} \equiv P_2 := \lceil K^{1/2}\rceil, J_k \equiv J := \lceil K^{1/2}\rceil, \\ &\alpha_k \equiv \bar{\alpha} := \min\left(\bar{\alpha}_1, \bar{\alpha}_2\right), \eta_k \equiv \bar{\eta} := 2\bar{\alpha}\max\left(L_f^2, 8L_{h,0}^2\right), \gamma_k \equiv \bar{\gamma} := 8L_{h,0}^2\bar{\alpha},\end{aligned} \tag{3.20}$$

for $k = 0, \ldots, K$, where $\bar{\alpha}_1 := [2\max(L_f^2, 8L_{h,0}^2)]^{-1}, \bar{\alpha}_2 = [2L_\Gamma + 2L_\beta + (12 + L_{f,1}^2)L_{h,0}^2]^{-1}$. Then it holds that

$$\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right) + \partial\chi\left(x^{R+1}\right) + \sum_{i=1}^m \bar{z}_i\nabla g_i\left(x^{R+1}\right) + \mathcal{N}_X\left(x^{R+1}\right), \mathbf{0}\right)\right] = \mathcal{O}\left(K^{-\frac{1}{2}}\right), \tag{3.21}$$

with $\bar{z} = [\beta g(x^{R+1}) + z^{R+1}]_+$.

Proof. It follows from Lemma 3.5 that

$$\begin{aligned}&\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right) + \partial\chi\left(x^{R+1}\right) + \sum_{i=1}^m\left[\beta g_i\left(x^{R+1}\right) + z_i^{R+1}\right]_+\nabla g_i\left(x^{R+1}\right) + \mathcal{N}_X\left(x^{R+1}\right), \mathbf{0}\right)\right] \\ =&\,\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{k+1}\right) + \partial\chi\left(x^{k+1}\right) + \sum_{i=1}^m\left[\beta g_i\left(x^{k+1}\right) + z_i^{k+1}\right]_+\nabla g_i\left(x^{k+1}\right) + \mathcal{N}_X\left(x^{k+1}\right), \mathbf{0}\right)\right] \\ \leq&\,\frac{3}{K}\sum_{k=1}^K\left(L_\Gamma + L_\beta + \frac{1}{\alpha_k}\right)^2\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] + \frac{6L_f^2}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2\right] \\ &+ \frac{3L_{g,0}^2 m}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|^2\right] + \frac{6}{K}\sum_{k=1}^K\sigma_{\Gamma_k}^2\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left(L_\Gamma + L_\beta + \frac{1}{\bar{\alpha}}\right)^2 \left\|x^{k+1}-x^k\right\|^2 + 2L_f^2\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right] \\
&\quad + \frac{3L_{g,0}^2 m}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|^2\right] + 6\sigma_\Gamma^2,
\end{aligned} \tag{3.22}
$$

where the last inequality follows from the constant setting of α_k and

$$
\sigma_{\Gamma_k}^2 \equiv \sigma_\Gamma^2 := 2\left(L_{f,0}^2+\sigma_f^2\right)\frac{\sigma_{h,1}^2}{P_2} + 2L_{h,0}^2\frac{\sigma_f^2}{J}, \quad k=0,1,\ldots,K. \tag{3.23}
$$

We next analyze terms of the right hand side of (3.22). Firstly, it follows from $P_2 = J \geq K^{1/2}$ and (3.23) that

$$
6\sigma_\Gamma^2 = \mathcal{O}\left(\frac{1}{P_2}+\frac{1}{J}\right) = \mathcal{O}\left(K^{-\frac{1}{2}}\right). \tag{3.24}
$$

Secondly, applying (3.7) we obtain

$$
\begin{aligned}
\frac{3L_{g,0}^2 m}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|^2\right] \leq \frac{3L_{g,0}^2 m}{K}\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\left|z_i^{k+1}-z_i^k\right|^2\right] &\leq \frac{3L_{g,0}^2 m}{K}\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\frac{G^2\rho^2}{K^2}\left(\frac{2\rho}{\beta}+1\right)^2\right] \\
&= \frac{3L_{g,0}^2 G^2\rho^2 m^2}{K^2}\left(\frac{2\rho}{\beta}+1\right)^2 \\
&= \mathcal{O}\left(K^{-2}\right).
\end{aligned} \tag{3.25}
$$

We now estimate the remaining terms of (3.22) regarding $\|x^{k+1}-x^k\|^2$ and $\|y^{k+1}-h(x^k)\|^2$. Multiplying (3.19) by $(1+\bar{\alpha}L_f^2)$ and plugging it into (3.13), by Lemma 3.2 we have the following inequality for any $k \in [K]$

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{2\bar{\alpha}} - \frac{L_\Gamma + L_\beta}{2}\right)\left\|x^{k+1}-x^k\right\|^2 + \left\|y^{k+1}-h\left(x^k\right)\right\|^2\right] \\
\leq &\mathbb{E}\left[\mathcal{L}_\beta\left(x^k,z^k\right) - \mathcal{L}_\beta\left(x^{k+1},z^{k+1}\right)\right] + G\max\left(1,\frac{2\rho}{\beta}\right)\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right] + \bar{\alpha}\sigma_\Gamma^2 \\
&+ \left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\left(1-\bar{\eta}\right)^2\mathbb{E}\left[\left\|y^k - h\left(x^{k-1}\right)\right\|^2\right] + \left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}^{-1}\right)\left(1-\bar{\eta}\right)^2 L_{h,0}^2\mathbb{E}\left[\left\|x^k-x^{k-1}\right\|^2\right] \\
&+ \left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\bar{\eta}^2\frac{\sigma_{h,0}^2}{P_1}.
\end{aligned} \tag{3.26}
$$

Under parameter settings in (3.20), we can deduce from $L_f = L_{h,0}L_{f,1}$ defined in Lemma 3.1, $\bar{\alpha} \leq \bar{\alpha}_1$ and $\bar{\alpha} \leq \bar{\alpha}_2$ that $0 < \bar{\eta} \leq 1$,

$$
\begin{aligned}
\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\left(1-\bar{\eta}\right)^2 - 1 \leq \left(\left(1+\frac{\bar{\eta}}{2}\right)\left(1-\bar{\eta}\right)\right)^2 - 1 &\leq 1 - \frac{\bar{\eta}}{2} - \frac{\bar{\eta}^2}{2} - 1 \\
&\leq -\frac{\bar{\eta}}{2} = -\bar{\alpha}\max\left(L_f^2, 8L_{h,0}^2\right)
\end{aligned} \tag{3.27}
$$

and

$$
\begin{aligned}
0 \leq \left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}^{-1}\right)\left(1-\bar{\eta}\right)^2 L_{h,0}^2 \leq \left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}^{-1}\right)L_{h,0}^2 &\leq \left(1+\frac{L_{f,1}^2}{8}+L_f^2\bar{\alpha}_1\right)L_{h,0}^2 + \frac{1}{8\bar{\alpha}} \\
&\leq \left(\frac{3}{2}+\frac{L_{f,1}^2}{8}\right)L_{h,0}^2 + \frac{1}{8\bar{\alpha}} \leq \frac{1}{4\bar{\alpha}} - \frac{L_\Gamma + L_\beta}{4}.
\end{aligned} \tag{3.28}
$$

Summing up (3.26) over $k = 1,\ldots,K$ and applying (3.27)-(3.28) to sort out the coefficients of $\left\|x^{k+1}-x^k\right\|^2$ and $\left\|y^{k+1}-h\left(x^k\right)\right\|^2$, we obtain

$$
\sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{1}{4\bar{\alpha}} - \frac{L_\Gamma + L_\beta}{4}\right)\left\|x^{k+1}-x^k\right\|^2 + \bar{\alpha}\max\left(L_f^2, 8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\mathcal{L}_\beta\left(x^1,z^1\right)-\mathcal{L}_\beta\left(x^{K+1},z^{K+1}\right)\right]+\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}^{-1}\right)\left(1-\bar{\eta}\right)^2 L_{h,0}^2\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]\\
&\quad+\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\left(1-\bar{\eta}\right)^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]+G\left(1+\frac{2\rho}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]\\
&\quad+\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\bar{\eta}^2K\frac{\sigma_{h,0}^2}{P_1}+\bar{\alpha}K\sigma_\Gamma^2. \qquad (3.29)
\end{aligned}
$$

To give a more concrete bound on the right hand side of (3.29), we come to analyze two of its terms separately:

$$
\mathcal{L}_\beta\left(x^{K+1},z^{K+1}\right)\ \text{ and }\ \left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}^{-1}\right)\left(1-\bar{\eta}\right)^2 L_{h,0}^2\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]+\mathbb{E}\left[\mathcal{L}_\beta\left(x^1,z^1\right)\right].
$$

Firstly, due to $\psi_\beta(u,v)\geq\frac{-v^2}{2\beta}$ for $u,v\in\mathbb{R}$ and $\beta=K^{1/4}$, it yields from Assumption 3.2 and Lemma 3.2 that

$$
\begin{aligned}
\mathcal{L}_\beta\left(x^{K+1},z^{K+1}\right)=\Gamma\left(x^{K+1}\right)+\chi\left(x^{K+1}\right)+\sum_{i=1}^{m}\psi_\beta\left(g_i\left(x^{K+1}\right),z_i^{K+1}\right)&\geq C^*-\sum_{i=1}^{m}\frac{\left(z_i^{K+1}\right)^2}{2\beta}\\
&\geq C^*-\frac{4mG^2\rho^2}{2\beta}\\
&\geq C^*-2G^2\rho^2m. \qquad (3.30)
\end{aligned}
$$

Secondly, lemmas 3.2 and 3.6 indicate that

$$
\begin{aligned}
\left(\frac{1}{2\bar{\alpha}}-\frac{L_\Gamma+L_\beta}{2}\right)\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]\leq&\,\mathbb{E}\left[\mathcal{L}_\beta\left(x^0,z^0\right)-\mathcal{L}_\beta\left(x^1,z^1\right)\right]+\bar{\alpha}\sigma_\Gamma^2+\bar{\alpha}L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]\\
&+G\left(1+\frac{2\rho}{\beta}\right)\mathbb{E}\left[\left\|z^1-z^0\right\|_1\right].
\end{aligned}
$$

Then it together with (3.28) yields

$$
\begin{aligned}
&\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}^{-1}\right)\left(1-\bar{\eta}\right)^2 L_{h,0}^2\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]+\mathbb{E}\left[\mathcal{L}_\beta\left(x^1,z^1\right)\right]\\
\leq&\,\mathcal{L}_\beta\left(x^0,z^0\right)+\bar{\alpha}\sigma_\Gamma^2+G\left(1+\frac{2\rho}{\beta}\right)\mathbb{E}\left[\left\|z^1-z^0\right\|_1\right]+\bar{\alpha}L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]\\
\leq&\,\mathcal{L}_\beta\left(x^0,z^0\right)+\bar{\alpha}\sigma_\Gamma^2+2G^2m\rho\left(1+\frac{2\rho}{\beta}\right)+\bar{\alpha}L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]\\
\leq&\,\Gamma\left(x^0\right)+\chi\left(x^0\right)+\frac{\beta}{2}G^2m+\bar{\alpha}\sigma_\Gamma^2+2G^2m\rho\left(1+\frac{2\rho}{\beta}\right)+\bar{\alpha}L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right], \qquad (3.31)
\end{aligned}
$$

where the second inequality is due to $z^0=\mathbf{0}$ and Lemma 3.2, and the last inequality uses

$$
\mathcal{L}_\beta\left(x^0,z^0\right)=\Gamma\left(x^0\right)+\chi\left(x^0\right)+\frac{\beta}{2}\sum_{i=1}^{m}\left(\left[g_i\left(x^0\right)\right]_+\right)^2\leq\Gamma\left(x^0\right)+\chi\left(x^0\right)+\frac{\beta}{2}G^2m. \qquad (3.32)
$$

Hence, plugging (3.30) and (3.31) into (3.29) we obtain

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{1}{4\bar{\alpha}}-\frac{L_\Gamma+L_\beta}{4}\right)\left\|x^{k+1}-x^k\right\|^2+\bar{\alpha}\max\left(L_f^2,8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\,\Gamma\left(x^0\right)+\chi\left(x^0\right)-C^*+2G^2\rho^2m+\frac{\beta}{2}G^2m+\bar{\alpha}\left(K+1\right)\sigma_\Gamma^2+2G^2m\rho\left(1+\frac{2\rho}{\beta}\right)+\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\bar{\eta}^2K\frac{\sigma_{h,0}^2}{P_1}\\
&+\left[\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\left(1-\bar{\eta}\right)^2+\bar{\alpha}L_f^2\right]\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]+G\left(1+\frac{2\rho}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]\\
\leq&\,\Gamma\left(x^0\right)+\chi\left(x^0\right)-C^*+2G^2\rho^2m+\frac{\beta}{2}G^2m+\bar{\alpha}\left(K+1\right)\sigma_\Gamma^2+2G^2m\rho\left(1+\frac{2\rho}{\beta}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\mathbb{E}\left[\left\|y^0-h\left(x^0\right)\right\|^2\right]+G\left(1+\frac{2\rho}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]+\left[\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)K+1\right]\bar{\eta}^2\frac{\sigma_{h,0}^2}{P_1}\\
\leq&\Gamma\left(x^0\right)+\chi\left(x^0\right)-C^*+2G^2\rho^2m+\frac{\beta}{2}G^2m+\bar{\alpha}\left(K+1\right)\sigma_\Gamma^2+2G^2m\rho\left(1+\frac{2\rho}{\beta}\right)\\
&+\left\|y^0-h\left(x^0\right)\right\|^2+G^2m\rho\left(1+\frac{2\rho}{\beta}\right)^2+\left(3K+1\right)\bar{\eta}^2\frac{\sigma_{h,0}^2}{P_1}\\
=&\mathcal{O}\left(1+\beta+\bar{\alpha}K\left(\frac{1}{P_2}+\frac{1}{J}\right)+\frac{\bar{\alpha}^2K}{P_1}\right),
\end{aligned}
\tag{3.33}
$$

where the second inequality holds due to $0<\bar{\eta}\leq 1$,

$$
\left(1+\bar{\alpha}L_f^2\right)\left(1+\bar{\gamma}\right)\left(1-\bar{\eta}\right)^2+\bar{\alpha}L_f^2\leq 1-\bar{\alpha}\max\left(L_f^2,8L_{h,0}^2\right)+\bar{\alpha}L_f^2\leq 1
$$

from (3.27), and

$$
\begin{aligned}
\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]=&\mathbb{E}\left[\left\|\left(1-\bar{\eta}\right)\left(y^0-h\left(x^0\right)\right)+\bar{\eta}\left(\frac{1}{P_{0,1}}\sum_{\phi\in\mathcal{P}_{0,1}}H\left(x^0;\phi\right)-h\left(x^0\right)\right)\right\|^2\right]\\
\leq&\left(1-\bar{\eta}\right)^2\left\|y^0-h\left(x^0\right)\right\|^2+\bar{\eta}^2\frac{\sigma_{h,0}^2}{P_1},
\end{aligned}
$$

the third inequality follows from

$$
\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]\leq\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\frac{\rho G}{K}\left(\frac{2\rho}{\beta}+1\right)\right]\leq Gm\rho\left(\frac{2\rho}{\beta}+1\right)
$$

by (3.7), and $(1+\bar{\alpha}L_f^2)(1+\bar{\gamma})\leq(3/2)^2<3$ from $\bar{\gamma}=8L_{h,0}^2\bar{\alpha},\bar{\alpha}\leq\bar{\alpha}_1$, and the last equality follows from (3.23) and $\bar{\eta}=2\bar{\alpha}\max\left(L_f^2,8L_{h,0}^2\right)$. Thus, (3.33) implies that

$$
\begin{aligned}
&\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left(L_\Gamma+L_\beta+\frac{1}{\bar{\alpha}}\right)^2\left\|x^{k+1}-x^k\right\|^2+2L_f^2\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\frac{3}{K}\max\left(\frac{\left(L_\Gamma+L_\beta+\frac{1}{\bar{\alpha}}\right)^2}{\frac{1}{4\bar{\alpha}}-\frac{L_\Gamma+L_\beta}{4}},\frac{2}{\bar{\alpha}}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{1}{4\bar{\alpha}}-\frac{L_\Gamma+L_\beta}{4}\right)\left\|x^{k+1}-x^k\right\|^2+\bar{\alpha}\max\left(L_f^2,8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\frac{54}{\bar{\alpha}K}\sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{1}{4\bar{\alpha}}-\frac{L_\Gamma+L_\beta}{4}\right)\left\|x^{k+1}-x^k\right\|^2+\bar{\alpha}\max\left(L_f^2,8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
=&\mathcal{O}\left(\frac{1+\beta}{K\bar{\alpha}}+\frac{1}{P_2}+\frac{1}{J}+\frac{\bar{\alpha}}{P_1}\right)\\
=&\mathcal{O}\left(\frac{1+\beta}{K\bar{\alpha}_1}+\frac{1+\beta}{K\bar{\alpha}_2}+\frac{1}{P_2}+\frac{1}{J}+\frac{\bar{\alpha}_2}{P_1}\right)=\mathcal{O}\left(K^{-\frac{1}{2}}\right),
\end{aligned}
\tag{3.34}
$$

where the second inequality follows from $L_\Gamma+L_\beta\leq(2\bar{\alpha}_2)^{-1}\leq(2\bar{\alpha})^{-1}$, the second equality comes from $\bar{\alpha}:=\min\left(\bar{\alpha}_1,\bar{\alpha}_2\right)$ defined in (3.20), the last equality holds due to $\beta=K^{1/4}$, $\bar{\alpha}_1=\mathcal{O}(1),\bar{\alpha}_2=\mathcal{O}(L_\beta^{-1})=\mathcal{O}(\beta^{-1})=\mathcal{O}(K^{-1/4})$ and $P_2=J\geq\sqrt{K},P_1\geq K^{1/4}$. Plugging (3.24), (3.25) and (3.34) into (3.22), we obtain (3.21). □

Recall that y^{k+1} was initially introduced to track the function value $h(x^k)$. As can be seen from (3.34) that

$$
\mathbb{E}[\|y^{R+1}-h(x^R)\|^2]=\mathcal{O}(K^{-\frac{1}{2}}).
$$

This actually gives a theoretical explanation that y^{k+1} can be close to $h(x^k)$ in expectation when K is sufficiently large. And it keeps in line with our expectation as discussed for (2.10). Besides, (3.32) indicates that the value $\mathcal{L}_\beta(x^0,z^0)$ has a crucial effect on orders obtained in (3.34) and (3.21). Specifically when initial point x^0 is feasible to (1.1), $\mathcal{L}_\beta(x^0,z^0)$ will be irrelevant to parameter β, thus improved results can be derived as shown in the corollary below. Proof is straightforward following similar analysis to Theorem 3.1, so we omit it here.

Corollary 3.2. *Under assumptions 3.1-3.3 and same parameter settings as Theorem 3.1 except*

$$\beta = K^{1/3},\ P_{k,1} \equiv P_1 := \lceil K^{1/3} \rceil,\ P_{k,2} \equiv P_2 := \lceil K^{2/3} \rceil,\ J_k \equiv J := \lceil K^{2/3} \rceil$$

for $k = 0, \ldots, K$, if initial point x^0 is feasible to (1.1)*, it holds that*

$$\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right) + \partial\chi\left(x^{R+1}\right) + \sum_{i=1}^{m} \bar{z}_i \nabla g_i\left(x^{R+1}\right) + \mathcal{N}_X\left(x^{R+1}\right), \mathbf{0}\right)\right] = \mathcal{O}\left(K^{-\frac{2}{3}}\right)$$

with $\bar{z} := \left[\beta g\left(x^{R+1}\right) + z^{R+1}\right]_+$, and $\mathbb{E}[\|y^{R+1} - h(x^R)\|^2] = \mathcal{O}(K^{-2/3})$.

Remark 3.1. *Let us consider the convexly constrained optimization problem*

$$\min_{x \in X} \quad \Gamma(x) + \chi(x). \tag{3.35}$$

Without the constraints $g(x) \leq \mathbf{0}$, the term in (3.22) *associated with $\|z^{k+1} - z^k\|^2$ will vanish. And of course β will not appear. In this case, if we slightly modify the parameter settings in Theorem 3.1 by assuming*

$$P_{k,1} \equiv P_1 := K,\ P_{k,2} \equiv P_2 := K,\ J_k \equiv J := K, \quad k = 0, \ldots, K,$$

we can obtain

$$\mathbb{E}[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right) + \partial\chi\left(x^{R+1}\right) + \mathcal{N}_X\left(x^{R+1}\right), \mathbf{0}\right)] = \mathcal{O}(K^{-1}).$$

Then the iteration and sample complexities to find an ϵ-stationary point of (3.35)*, i.e. satisfying* (2.11)*, are bounded by $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$, respectively, where the sample complexity is same as NASA [12].*

We now consider the feasibility measure $\|[g(x^{R+1})]_+\|$. As in general it may be impossible to find a feasible solution for a constrained optimization, to guarantee near-feasibility we give another assumption.

Assumption 3.4. *(NonSingularity Condition, NSC) There exists a parameter $\nu > 0$ such that*

$$\nu \left\| \left[g\left(x^k\right)\right]_+ \right\| \leq \mathbf{d}\left(\left(J_g\left(x^k\right)\right)^T \left[g\left(x^k\right)\right]_+ + \mathcal{N}_X\left(x^k\right), \mathbf{0}\right), \quad \forall k \in [K+1].$$

NSC can be regarded as a special case of the Kurdyka-Łojasiewicz (KL) condition[1] for minimizing $[\|[g(x)]_+\|^2 + \mathbf{1}_X(x)]$ with $\bar{x} \in \mathcal{X} := \{x \in X \mid g(x) \leq \mathbf{0}\}, c = \nu^{-1}, \theta = 1/2$, and $B_1 = B_2 = +\infty$. This is also mentioned in Assumption 4 of [19]. KL condition has been widely used in analysis for nonconvex optimization and there are types of nonconvex functions satisfying the KL inequality, such as second-order continuously differentiable Morse functions, semialgebraic functions and so on [1, 37, 38]. In linearly constrained case with $X = \mathbb{R}^n$, i.e. $g(x) = Ax + b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, NSC can be guaranteed by the linear independence of $\{\nabla g_i(x) \mid i \in [m], g_i(x) \geq 0\}$.

THEOREM **3.3. (Feasibility)** *Under the conditions of Theorem 3.1 and Assumption 3.4, it holds that*

$$\mathbb{E}\left[\left\| \left[g\left(x^{R+1}\right)\right]_+ \right\|^2\right] = \mathcal{O}\left(K^{-\frac{1}{2}}\right).$$

Proof. For any $k \geq 0$, denote

$$\left(u^k, v^k\right) := \underset{u \in \partial\chi(x^k), v \in \mathcal{N}_X(x^k)}{\operatorname{argmin}} \left\| \sum_{i=1}^{m} \left[\beta g_i\left(x^k\right)\right]_+ \nabla g_i\left(x^k\right) + u + v \right\|.$$

Then it follows from Assumption 3.4, (2.1) and (3.4) that for any $k \in [K+1]$,

$$\left\| \left[g\left(x^k\right)\right]_+ \right\| \leq \frac{1}{\nu\beta} \mathbf{d}\left(\beta\left(J_g\left(x^k\right)\right)^T \left[g\left(x^k\right)\right]_+ + \mathcal{N}_X\left(x^k\right), \mathbf{0}\right)$$

[1] (Kurdyka-Łojasiewicz condition [1, 37]) A function $\psi(x)$ satisfies the KL property at point $\bar{x} \in \operatorname{dom}(\partial\psi)$ if there exist $B_2 > 0, \mathcal{B}_{B_1}(\bar{x}) := \{x \mid \|x - \bar{x}\| < B_1\}$ and a concave function $\phi(a) = c \cdot a^{1-\theta}$ for some $c > 0, \theta \in [0, 1)$ such that

$$\nabla\phi\left(|\psi(x) - \psi(\bar{x})|\right) \mathbf{d}\left(\partial\psi(x), \mathbf{0}\right) \geq 1, \quad \text{for all } x \in \mathcal{B}_{B_1}(\bar{x}) \cap \operatorname{dom}(\partial\psi) \text{ and } \psi(\bar{x}) < \psi(x) < \psi(\bar{x}) + B_2.$$

$$
\begin{aligned}
&\leq \frac{1}{\nu\beta}\left\|\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)\right]_+\nabla g_i\left(x^k\right)+v^k\right\| \\
&\leq \frac{1}{\nu\beta}\left(\left\|\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)\right]_+\nabla g_i\left(x^k\right)+u^k+v^k\right\|+\left\|u^k\right\|\right) \\
&\leq \frac{1}{\nu\beta}\left(\mathbf{d}\left(\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)\right]_+\nabla g_i\left(x^k\right)+\partial\chi\left(x^k\right)+\mathcal{N}_X\left(x^k\right),\mathbf{0}\right)+G_\chi\right) \\
&\leq \frac{1}{\nu\beta}\mathbf{d}\left(\nabla\Gamma\left(x^k\right)+\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\nabla g_i\left(x^k\right)+\partial\chi\left(x^k\right)+\mathcal{N}_X\left(x^k\right),\mathbf{0}\right)+\frac{1}{\nu\beta}\left\|\nabla\Gamma\left(x^k\right)\right\|+\frac{G_\chi}{\nu\beta} \\
&\quad+\frac{1}{\nu\beta}\sum_{i=1}^{m}\left|\left[\beta g_i\left(x^k\right)\right]_+-\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\right|\left\|\nabla g_i\left(x^k\right)\right\| \\
&\leq \frac{1}{\nu\beta}\left(\mathbf{d}\left(\nabla\Gamma\left(x^k\right)+\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\nabla g_i\left(x^k\right)+\partial\chi\left(x^k\right)+\mathcal{N}_X\left(x^k\right),\mathbf{0}\right)+L_{g,0}\sum_{i=1}^{m}\left|z_i^k\right|+L_{f,0}L_{h,0}+G_\chi\right),
\end{aligned}
$$

where the fifth inequality follows from $\mathbf{d}\left(a+A,\mathbf{0}\right)\leq\|a\|+\mathbf{d}\left(A,\mathbf{0}\right)$ for any $a\in\mathbb{R}^n$ and $A\subseteq\mathbb{R}^n$, and the last inequality uses $\nabla\Gamma\left(x\right)=\nabla h\left(x\right)^T\nabla f\left(h\left(x\right)\right)$. Applying Lemma 3.2 we further obtain that

$$
\begin{aligned}
&\left\|\left[g\left(x^k\right)\right]_+\right\|^2 \\
&\leq \frac{4}{\beta^2\nu^2}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^k\right)+\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\nabla g_i\left(x^k\right)+\partial\chi\left(x^k\right)+\mathcal{N}_X\left(x^k\right),\mathbf{0}\right)+L_{g,0}^2\left(\sum_{i=1}^{m}\left|z_i^k\right|\right)^2+L_{f,0}^2L_{h,0}^2+G_\chi^2\right] \\
&\leq \frac{4}{\beta^2\nu^2}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^k\right)+\partial\chi\left(x^k\right)+\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\nabla g_i\left(x^k\right)+\mathcal{N}_X\left(x^k\right),\mathbf{0}\right)+4m^2L_{g,0}^2G^2\rho^2+L_{f,0}^2L_{h,0}^2+G_\chi^2\right].
\end{aligned}
$$

Hence, it indicates from (3.21) and $\beta=K^{1/4}$ that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left[g\left(x^{R+1}\right)\right]_+\right\|^2\right]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\left[g\left(x^{k+1}\right)\right]_+\right\|^2\right] \\
&\leq \frac{4}{\nu^2\beta^2K}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{k+1}\right)+\sum_{i=1}^{m}\left[\beta g_i\left(x^{k+1}\right)+z_i^{k+1}\right]_+\nabla g_i\left(x^{k+1}\right)+\partial\chi\left(x^{k+1}\right)+\mathcal{N}_X\left(x^{k+1}\right),\mathbf{0}\right)\right] \\
&\quad+\frac{4}{\nu^2\beta^2K}\sum_{k=1}^{K}\left[4m^2L_{g,0}^2G^2\rho^2+L_{f,0}^2L_{h,0}^2+G_\chi^2\right] \\
&=\frac{4}{\nu^2\beta^2}\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right)+\sum_{i=1}^{m}\left[\beta g_i\left(x^{R+1}\right)+z_i^{R+1}\right]_+\nabla g_i\left(x^{R+1}\right)+\partial\chi\left(x^{R+1}\right)+\mathcal{N}_X\left(x^{R+1}\right),\mathbf{0}\right)\right] \\
&\quad+\frac{4}{\nu^2\beta^2}\left[4m^2L_{g,0}^2G^2\rho^2+L_{f,0}^2L_{h,0}^2+G_\chi^2\right] \\
&=\mathcal{O}\left(K^{-1}+K^{-\frac{1}{2}}\right)=\mathcal{O}\left(K^{-\frac{1}{2}}\right).
\end{aligned}
$$

The proof is completed. □

Similar to Corollary 3.2, if x^0 is feasible we can obtain the following result.

Corollary 3.4. *Under assumptions 3.1-3.4 and parameter settings of Corollary 3.2, if x^0 is feasible to* (1.1)*, it holds that $\mathbb{E}[\|[g(x^{R+1})]_+\|^2]=\mathcal{O}(K^{-\frac{2}{3}})$ for any $K\geq 1$.*

We are now ready to estimate the complementary slackness measure $\|\bar{z}\odot g(x^{R+1})\|$.

THEOREM **3.5. (Complementary slackness)** *Under conditions of Theorem 3.1 and Assumption 3.4, it holds that with $\bar{z}=[\beta g(x^{R+1})+z^{R+1}]_+$,*

$$
\mathbb{E}\left[\left\|\bar{z}\odot g\left(x^{R+1}\right)\right\|\right]=\mathcal{O}\left(K^{-\frac{1}{4}}\right).
$$

Proof. On the one hand, we have

$$\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\left[g_i\left(x^k\right)\right]_- = \begin{cases} -\left(\beta g_i\left(x^k\right)+z_i^k\right) g_i\left(x^k\right), & \text{if } -\frac{z_i^k}{\beta} \leq g_i\left(x^k\right)<0\,; \\ 0, & \text{otherwise}\,. \end{cases}$$

Then it follows from $-u\left(bu+a\right) \leq \frac{a^2}{4b}$ for any $u \in \mathbb{R}$ and $b>0$ together with Lemma 3.2 that

$$\left[\beta g_i\left(x^k\right)+z_i^k\right]_+\left[g_i\left(x^k\right)\right]_- \leq \frac{\left|z_i^k\right|^2}{4\beta} \leq \frac{G^2\rho^2}{\beta},$$

which further yields

$$\mathbb{E}\left[\sum_{i=1}^m \bar{z}_i\left[g_i\left(x^{R+1}\right)\right]_-\right]=\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\sum_{i=1}^m\left[\beta g_i\left(x^{k+1}\right)+z_i^{k+1}\right]_+\left[g_i\left(x^{k+1}\right)\right]_-\right] \leq \frac{mG^2\rho^2}{\beta}=\mathcal{O}\left(K^{-\frac{1}{4}}\right). \tag{3.36}$$

On the other hand, applying $\left|z_i^k\right|=z_i^k \leq 2G\rho$ from Lemma 3.2 and $\|u\|_1 \leq \sqrt{m}\,\|u\|$ for all $u \in \mathbb{R}^m$, we have

$$\mathbb{E}\left[\sum_{i=1}^m z_i^{R+1}\left[g_i\left(x^{R+1}\right)\right]_+\right] \leq \mathbb{E}\left[2G\rho\sum_{i=1}^m\left[g_i\left(x^{R+1}\right)\right]_+\right] \leq \mathbb{E}\left[2G\rho\sqrt{m}\left\|\left[g\left(x^{R+1}\right)\right]_+\right\|\right]=\mathcal{O}\left(K^{-\frac{1}{4}}\right), \tag{3.37}$$

where the equality follows from Theorem 3.3 and $\left(\mathbb{E}\left[u\right]\right)^2 \leq \mathbb{E}\left[u^2\right]$ for any random variable $u \in \mathbb{R}$. Combining (3.37) with $\beta=K^{1/4}$ and Theorem 3.3, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^m \bar{z}_i\left[g_i\left(x^{R+1}\right)\right]_+\right] &= \mathbb{E}\left[\sum_{i=1}^m\left[\beta g_i\left(x^{R+1}\right)+z_i^{R+1}\right]_+\left[g_i\left(x^{R+1}\right)\right]_+\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^m\left(\beta\left[g_i\left(x^{R+1}\right)\right]_+^2+\left[z_i^{R+1}\right]_+\left[g_i\left(x^{R+1}\right)\right]_+\right)\right] \\
&= \mathbb{E}\left[\beta\left\|\left[g\left(x^{R+1}\right)\right]_+\right\|^2+\sum_{i=1}^m z_i^{R+1}\left[g_i\left(x^{R+1}\right)\right]_+\right]=\mathcal{O}\left(K^{-\frac{1}{4}}\right),
\end{aligned}$$

which together with (3.36) yields

$$\mathbb{E}[\|\bar{z}\odot g(x^{R+1})\|_1]=\mathbb{E}\left[\sum_{i=1}^m \bar{z}_i\left|g_i\left(x^{R+1}\right)\right|\right]=\mathbb{E}\left[\sum_{i=1}^m \bar{z}_i\left[g_i\left(x^{R+1}\right)\right]_+\right]+\mathbb{E}\left[\sum_{i=1}^m \bar{z}_i\left[g_i\left(x^{R+1}\right)\right]_-\right]=\mathcal{O}\left(K^{-\frac{1}{4}}\right).$$

Applying $\|\cdot\| \leq \|\cdot\|_1$, we complete the proof. □

Similar to Corollary 3.2, we have the following result.

Corollary 3.6. *Under assumptions 3.1- 3.4 and parameter settings in Corollary 3.2, if x^0 is feasible to* (1.1)*, it holds that $\mathbb{E}[\|\bar{z}\odot g(x^{R+1})\|]=\mathcal{O}(K^{-\frac{1}{3}})$ with $\bar{z}=[\beta g(x^{R+1})+z^{R+1}]_+$.*

We now summarize the previous analysis into the theorem below characterizing both iteration and sample complexities of the STEP method to find an ϵ-stationary point of (1.1). Here the sample complexity refers to the total number of samples that are used to compute gradients of F as well as function values and gradients of H as in (2.3) and (2.4).

THEOREM **3.7. (Iteration and sample complexities)** *Under conditions of Theorem 3.1 and Assumption 3.4, given $\epsilon>0$, the STEP method can find an ϵ-stationary point of* (1.1) *after $\mathcal{O}\left(\epsilon^{-4}\right)$ iterations with associated sample complexity bounded by $\mathcal{O}\left(\epsilon^{-6}\right)$. If initial point x^0 is feasible, under conditions as Corollary 3.2, to find an ϵ-stationary point of* (1.1)*, the total number of iterations of the STEP method is in order $\mathcal{O}\left(\epsilon^{-3}\right)$ and sample complexity is bounded by $\mathcal{O}\left(\epsilon^{-5}\right)$.*

Proof. It is straightforward to obtain the iteration complexity from theorems 3.1, 3.3 and 3.5, as well as Corollaries 3.2, 3.4 and 3.6 when x^0 is feasible to (1.1), by the fact that $\left(\mathbb{E}\left[u\right]\right)^2 \leq \mathbb{E}\left[u^2\right]$ for any random variable u. Regarding the sample complexity, as the total number of samples is $\sum_{k=1}^K\left(J_k+P_{k,1}+P_{k,2}\right)$, by (3.20) and parameter settings in Corollary 3.2 we can derive the conclusions. □

Remark 3.2. *In Theorem 3.7 we characterized the sample complexities in terms of the total number of samples used to evaluate function information including the function values and/or gradients of H and F. If we count in the calculation of function values and gradients of all m constraints at each iteration as well, $2mK$ function information evaluations will be added, which can still keep the related calculation complexity in the same order as the sample complexity in Theorem 3.7. Moreover, under KL condition of the problem minimizing $\|[g(x)]_+\|^2 + \mathbf{1}_X(x)$ (i.e. a stronger version of NSC), we can apply the algorithm in [37] as a preprocessor to obtain an initial point $x^0 \in X$ such that $\|[g(x^0)]_+\|^2 \leq \epsilon$ for which the total number of function information evaluations bounded by $\mathcal{O}\left(\epsilon^{-5}\right)$. Therefore, it can be obtained from* (3.32) *that the calculation complexity of the STEP method initialized with a near-feasible x^0 remains $\mathcal{O}\left(\epsilon^{-5}\right)$.*

4 Numerical experiments

In this section, we would like to carry out some numerical experiments for solving the following risk-averse portfolio optimization problem [7, 41, 43]:

$$\min_{x \in \Delta^n} \Gamma(x) := -\mathbb{E}_\phi\left[r_\phi(x)\right] + \lambda \operatorname{Var}\left[r_\phi(x)\right] \quad \text{s.t. } Ax \leq b, \tag{4.1}$$

where $\Delta^n = \{x \in \mathbb{R}^n_+ \mid \sum_{j=1}^n x_j = 1\}$, x is the decision variable with each component x_j representing the percentage of the total investment allocated to asset j, $j = 1, \ldots, n$, $r_\phi(x)$ is the random return under portfolio x, $\operatorname{Var}\left[r_\phi(x)\right]$ is the variance of $r_\phi(x)$, $\lambda = 0.2$ is the mean-variance trade-off parameter and $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$. We can reformulate (4.1) into the form (1.1) by defining $h(x) = \mathbb{E}_\phi\left[H(x;\phi)\right]$ with $H(x;\phi) = [r_\phi(x), r_\phi^2(x)]^T$ and $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(y) = -y_1 + \lambda y_2 - \lambda y_1^2$. All the implementations are conducted in MATLAB version R2019a on a laptop of 8GB RAM and Intel(R) Celeron(R) CPU N2940 @ 1.83GHz 1.83 GHz.

We compare our algorithm STEP with com-SVR-ADMM (Algorithm 2 in [41]) on four real world portfolio datasets: industrial-49, -48, -38 and -30 datasets from the Keneth R. French Data Library[2], which are commonly used in numerical experiments for stochastic compositional optimization [7, 43]. More specifically, we assume that the random return is defined as $r_\phi(x) = R_\phi^T x$, where R_ϕ is chosen from "Average Value Weighted Returns (Monthly)" of the aforementioned portfolio datasets with ϕ as a discrete random variable from $\{1, \ldots, P\}$. Here, P is equal to 1148 in industrial-49 dataset portfolios and 1149 in other three datasets. With these settings, we have

$$\mathbb{E}_\phi\left[r_\phi(x)\right] := \frac{1}{P}\sum_{\phi=1}^{P} R_\phi^T x, \quad \mathbb{E}_\phi\left[r_\phi^2(x)\right] := \frac{1}{P}\sum_{\phi=1}^{P}\left(R_\phi^T x\right)^2. \tag{4.2}$$

And the constraints are randomly generated by the following Matlab scripts

$$m = 100; rand('state', 4); x00 = rand(n, 1); x00 = x00/sum(x00); A = rand(m, n); b = A * x00 + rand(m, 1);$$

through which we can make sure a nonempty feasible set for (4.1). Since the test problem is convex, we also implement $fmincon$ in Matlab toolbox and take its objective function value to serve as a benchmark drawn as black horizontal lines in the following figures. For both com-SVR-ADMM and STEP, based on the theoretical parameter settings in [41] and in this paper, we have made attempts on finding the parameter settings as best as possible for each of the two algorithms when solving (4.1). Below are the specific settings for both algorithms.

- com-SVR-ADMM:

$$\tilde{x}^0 = x00, \tilde{w} = b - A\tilde{x}^0, \tilde{\lambda}^0 = \mathbf{0}, S = 20, K = 100, \rho = (KS)^{0.25},$$
$$\eta_s = 1/(200n(s+1))/, G_k = \mathbb{I}_n/(s + k/(K-1))$$

  for $s \in \{1, \ldots, S\}, k \in \{0, \ldots, K-1\}$, where $\tilde{w}$ is the slack variable transforming inequality constraints $Ax \leq b$ into $Ax + w = b, w \in \mathbb{R}^m_+$, $\tilde{x}^0$ and $\tilde{\lambda}^0$ are the initial point and initial guess of multiplier vector, S and K are the numbers of outer-loops and inner-loops respectively, and $\mathbb{I}_n$ refers to n-dimensional identity matrix.

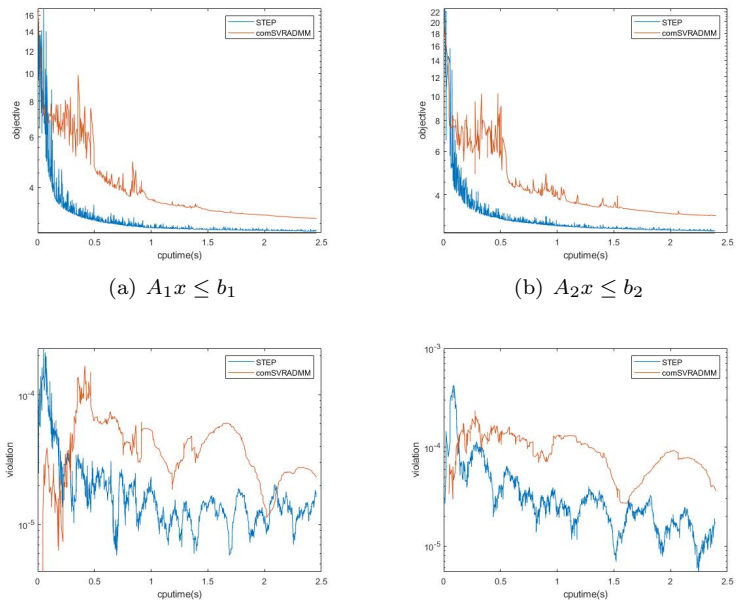
- STEP:

$$x^0 = x00, y^0 = h\left(x^0\right), z^0 = \mathbf{0}, K = 2000, \beta = K^{0.25},$$
$$\eta_k = K^{-0.25}, \alpha_k = 1/\left(50n(k+1)^{0.25}\right), \rho_k = \beta, P_{k,1} = \lceil (k+1)^{0.25} \rceil, P_{k,2} = \lceil (k+1)^{0.5} \rceil$$

  for $k = 0, \ldots, K-1$.

[2] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Figures 1-4 show the average changes of the objective function value and constraint violation $\|[Ax-b]_+\|_1/m$ by two algorithms with respect to cputime in 10 repeated tests on four industry portfolios. And for each industry portfolios we test two randomly generated groups of data (A, b). From figures 1-4, we can see that objective function values by STEP decreases faster. Although the constraint violations by both algorithm are gradually approaching similar values, com-SVR-ADMM oscillates more dramatically than STEP.



(a) $A_1x \leq b_1$ (b) $A_2x \leq b_2$

(c) $A_1x \leq b_1$ (d) $A_2x \leq b_2$

Figure 1: 49-Industrial Portfolio Dataset

5 Conclusion

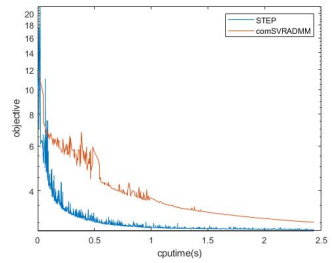
In this paper we propose a stochastic nested primal-dual (STEP) method for nonconvex constrained composition optimization, whose objective involves a nest structure on two expectation functions and feasibility determined by constraints are possibly hard to maintain at a given point. Motivated by recent progress on stochastic compositional optimization, we introduce an extra variable to track inner layer function value and update it in a moving average way. With this variable together with subsampled gradients for both layers functions, we calculate the stochastic gradient for the nested function. To cope with the general possibly nonconvex constraints, at each iteration we construct a stochastic approximation to the linearized augmented Lagrangian function to update the primal variable and then update the dual variable in a moving-average way. Under a nonsingularity condition for constraint functions, we establish the iteration and sample complexities of the proposed algorithm to find an ϵ-stationary point, and as a byproduct validate the reducing gap between the real value of inner function and its estimation. Finally, we conduct some numerical experiments on a risk-averse portfolio optimization and reveal promising numerical performances of the STEP method.

Acknowledgements

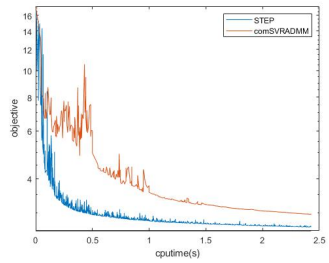
We would like to thank Prof. Xiantao Xiao for his valuable comments on an earlier version of this paper.

References

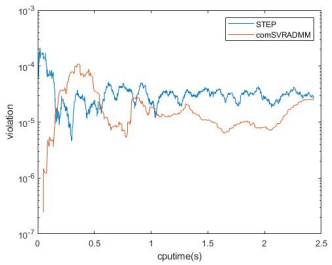
[1] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.



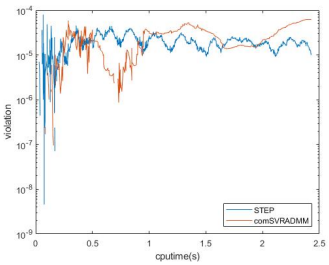
(a) $A_1x \leq b_1$



(b) $A_2x \leq b_2$

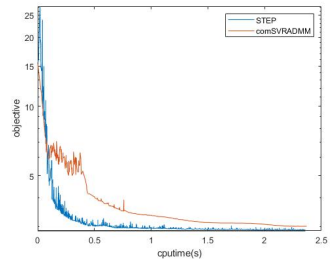


(c) $A_1x \leq b_1$

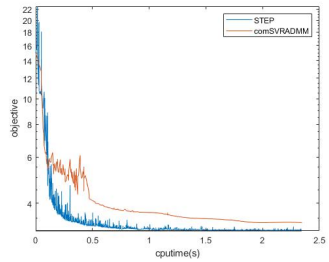


(d) $A_2x \leq b_2$

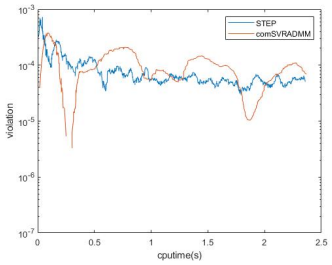
Figure 2: 40-Industrial Portfolio Dataset



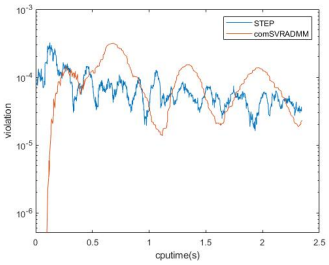
(a) $A_1x \leq b_1$



(b) $A_2x \leq b_2$

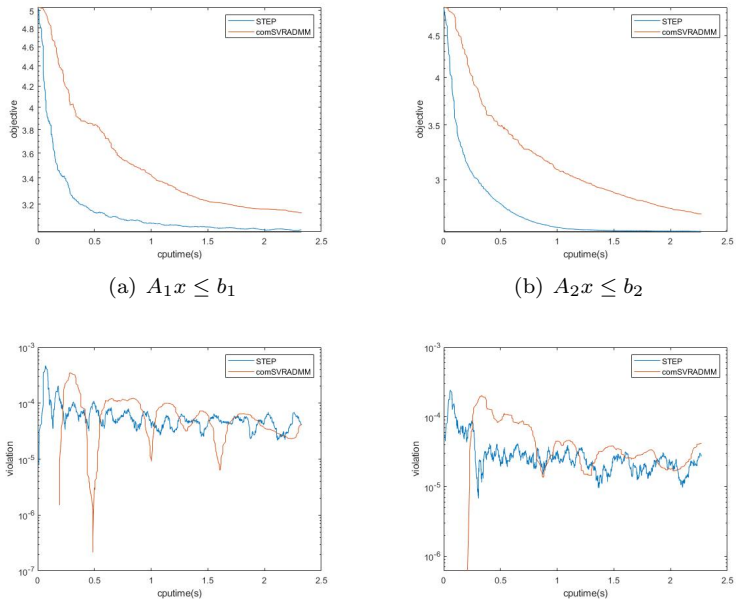


(c) $A_1x \leq b_1$



(d) $A_2x \leq b_2$

Figure 3: 38-Industrial Portfolio Dataset



(a) $A_1 x \leq b_1$ (b) $A_2 x \leq b_2$

(c) $A_1 x \leq b_1$ (d) $A_2 x \leq b_2$

Figure 4: 30-Industrial Portfolio Dataset

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

[3] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM J. Optim.*, 31(2):1352–1379, 2021.

[4] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Math. Program.*, 2022.

[5] T. Chen, Y. Sun, Q. Xiao, and W. Yin. A single-timescale method for stochastic bilevel optimization. In *25th AISTATS*, volume 151 of *PMLR*, pages 2466–2488. PMLR, 2022.

[6] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *NIPS*, volume 34, pages 25294–25307. Curran Associates, Inc., 2021.

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

[8] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *arXiv:2112.14799v2 [math.OC] 6 Jan*, 2022.

[9] B. Dai, N. He, Y. Pan, B. Boots, and L. Song. Learning from Conditional Distributions via Dual Embeddings. In *20th AISTATS*, volume 54 of *PMLR*, pages 1458–1467. PMLR, 2017.

[10] A. M. Devraj and J. Chen. Stochastic variance reduced primal dual algorithms for empirical composition optimization. In *NIPS*, volume 32. Curran Associates, Inc., 2019.

[11] Y. Evtushenko. Numerical methods for solving problems of non-linear programming. *USSR Comput. Math. Math. Phys.*, 16(2):24–42, 1976.

[12] S. Ghadimi, A. Ruszczyński, and M. Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. Optim.*, 30(1):960–979, 2020.

[13] J.-B. Hiriart-Urruty and C. Lemarchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[14] W. Hu, C. J. Li, X. Lian, J. Liu, and H. Yuan. Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. In *NIPS*, volume 32. Curran Associates, Inc., 2019.

[15] L. Jin and X. Wang. A stochastic primal-dual algorithm for a class of nonconvex constrained optimization. *Comput. Optim. Appl.*, 2022. DOI: https://doi.org/10.1007/s10589-022-00384-w.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[17] G. Lan and Z. Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Comput. Optim. Appl.*, 76(2):461–498, 2020.

[18] X. Lian, M. Wang, and J. Liu. Finite-sum Composition Optimization via Variance Reduced Gradient Descent. In *20th AISTATS*, volume 54 of *PMLR*, pages 1159–1167. PMLR, 2017.

[19] Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Comput. Optim. Appl.*, 2022.

[20] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *37th ICML*, volume 119 of *PMLR*, pages 6083–6093. PMLR, 2020.

[21] Y. Nandwani, A. Pathak, Mausam, and P. Singla. A primal dual formulation for deep learning with constraints. In *NIPS*, volume 32. Curran Associates, Inc., 2019.

[22] J. Nocedal and S. Wright. *Numerical optimization.* Springer, New York, 2006.

[23] H. Rafique, M. Liu, Q. Lin, and T. Yang. Weakly-convex concave minmax optimization: provable algorithms and applications in machine learning. *Optim. Method Softw.*, pages 1–35, 2021.

[24] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis.* Springer Science & Business Media, 2009.

[25] R. R.T. The multiplier method of hestenes and powell applied to convex programming. *J. Optim. Theory Appl.*, 12:555–562, 1973.

[26] A. Ruszczyński. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM J. Control Optim.*, 59(3):2301–2320, 2021.

[27] Q. Tran-Dinh, N. Pham, and L. Nguyen. Stochastic Gauss-Newton algorithms for nonconvex compositional optimization. In *37th ICML*, volume 119 of *PMLR*, pages 9572–9582. PMLR, 2020.

[28] R. Tutunov, M. Li, A. I. Cowen-Rivers, J. Wang, and H. Bou-Ammar. Compositional adam: An adaptive compositional solver. *arXiv:2002.03755v2 [cs.LG] 24 Apr*, 2020.

[29] M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2):419–449, 2017.

[30] M. Wang and J. Liu. A stochastic compositional gradient method using markov samples. In *WSC*, pages 702–713, 2016.

[31] M. Wang, J. Liu, and E. X. Fang. Accelerating stochastic composition optimization. *J. Machine Learn. Res.*, 18(105):1–23, 2017.

[32] X. Wang, S. Ma, and Y. X. Yuan. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Math. Comput.*, 86(306):1793–1820, 2017.

[33] X. Wang and Y. Yuan. An augmented lagrangian trust region method for equality constrained optimization. *Optim. Method Softw.*, 30(3):559–582, 2015.

[34] X. Wang and H. Zhang. An augmented lagrangian affine scaling method for nonlinear programming. *Optim. Methods Softw.*, 30(5):934–964, 2015.

[35] Y. Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM J. Optim.*, 30(2):1664–1692, 2020.

[36] Y. Xu. First-order methods for constrained convex programming based on linearized augmented lagrangian function. *INFORMS J. Optim.*, 3(1):89–117, 2021.

[37] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *J. Sci. Comput.*, 72(2):700–734.

[38] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.*, 6(3):1758–1789, 2013.

[39] J. Yang, A. Orvieto, A. Lucchi, and N. He. Faster single-loop algorithms for minimax optimization without strong concavity. In *25th AISTATS*, volume 151 of *PMLR*, pages 5485–5517. PMLR, 28–30 Mar 2022.

[40] S. Yang, M. Wang, and E. X. Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM J. Optim.*, 29(1):616–659, 2019.

[41] Y. Yu and L. Huang. Fast stochastic variance reduced admm for stochastic composition optimization. In *26th IJCAI*, pages 3364–3370, 2017.

[42] J. Zhang and L. Xiao. A composite randomized incremental gradient method. In *36th ICML*, volume 97 of *PMLR*, pages 7454–7462. PMLR, 2019.

[43] J. Zhang and L. Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM J. Optim.*, 31(2):1131–1157, 2021.

[44] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Math. Program.*, 2021. DOI: https://doi.org/10.1007/s10107-021-01709-z.

[45] Z. Zhang and G. Lan. Optimal algorithms for convex nested stochastic composite optimization. *arXiv::2011.10076v5 [math.OC] 21 Jun*, 2021.