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

Lingzi Jin † Xiao Wang ‡

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 the inner layer function values we compute stochastic gradients of the nested function using a 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 ordinary differential equation (ODE). We analyze the KKT measure at the output by the STEP method with constant parameters and establish its iteration and sample complexities to find an ϵ-stationary point, ensuring that expected stationarity, feasibility as well as complementary slackness are below accuracy ϵ. To leverage the benefit of the (near) initial feasibility in the STEP method, we propose a two-stage framework incorporating a feasibility-seeking phase, aiming to locate a nearly feasible initial point. Moreover, to enhance the adaptivity of the STEP algorithm, we propose an adaptive variant by adaptively adjusting its parameters, along with a complexity analysis. Numerical results on a risk-averse portfolio optimization problem and an orthogonal nonnegative matrix decomposition reveal the effectiveness of the proposed algorithms.

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

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

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 $\Lambda : \mathbb{R}^n \to \mathbb{R}$ 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 $\phi \in \Xi_l$. Without loss of generality, we assume that the feasible set $\mathcal{X} := \{x \in X \mid g_i(x) \le 0, i = 1,\dots,m\}$ is nonempty. We also assume that Λ is simple in the sense that for any $\alpha > 0$ and $\bar{x} \in \mathbb{R}^n$, the proximal operator $\operatorname{argmin}_{x\in X}\left\{\Lambda(x) + \|x-\bar{x}\|_2^2/(2\alpha)\right\}$ can be efficiently computed. This premise is widely used in the literature related to composite optimization, such as [4, 45]. It holds true in many scenarios and interested readers are referred to [32] for more details. Problem (1.1) covers a wide range of applications, such as the single-layer stochastic optimization (corresponding to identity mapping f and $\bar{n} = 1$), risk-averse optimization [7, 43, 45], reinforcement

*Part of this research work was supported by the National Natural Science Foundation of China (No. 12271278) and the Major Key Project of PCL (No. PCL2022A05)

†ling-zi.jin@connect.polyu.hk, Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China; Peng Cheng Laboratory, Shenzhen, 518066, China.

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

learning [45] and meta-learning [7], where the explicit constraints can be added to characterize some prior knowledge or field knowledge as [27]. We present in Appendix A some specific application examples that can be formalized into (1.1).

Challenges often arise when solving problem (1.1). Due to the nested 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 of nonconvex constraints makes it impractical to keep feasibility through simple projection. Study on stochastic composition optimization has attracted much interest. In [43] 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 the 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 the strongly convex case. Later, [45] 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. [53] further generalizes SCGD type methods for two-layer optimization to algorithms for multi-layer problems with complexity theories analyzed. Meanwhile, in [42] an adaptive solver is designed for unconstrained (possibly nonconvex) smooth problems by integrating Adam [19] 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, [15] proposes a single time-scale algorithm for two-layer nested stochastic smooth nonconvex optimization with convex set constraints, which is further generalized into multi-level cases by [1]. Later, Ruszczyński [37] adds linear corrections to path-averaged inner function estimates on the basis of [15] and proposes a stochastic subgradient algorithm for nonconvex nonsmooth multi-level composition optimization with convex set constraints, but no convergence in terms of the gradient norm is analyzed until being provided in [1] later. Based on the estimations of the inner function in [37] and quasi-gradient updates of x in basic SCGD [43], [7] further discusses the validity of the linear correction in [37] 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 [10] and [45], while the latter one shows the convergence rate of $\mathcal{O}\left(K^{-1}\right)$ (resp. $\mathcal{O}\left(K^{-1/2}\right)$) in the strongly convex (resp. general convex) case, which matches the optimal rate for single-layer stochastic optimization. [58] 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 [11,17,23,55,57]. Prox-linear type algorithms are studied in [41, 59] for nonsmooth stochastic compositional optimization with a 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 [44]. In a different way, [60] 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. A similar idea is used in [10] 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 minimize $\Gamma(x) + \Lambda(x)$ over $\mathbb{R}^n$ as mentioned in [6]. Note that all the literature above assumes that the feasible region is easy to project onto. In [55] 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 a simple projection, without considering the stochastic nested structure in the objective of (1.1), there has been a surge of works in past decades [28]. And motivated by recent progress in convex constrained stochastic optimization [4, 20, 49], some research effort has been devoted to more general constrained stochastic optimization, such as penalty methods for nonconvex equality constrained stochastic optimization [46], 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 [25, 34, 51], SQP type methods for nonconvex smooth constrained optimization with deterministic equality constraints [3, 9] and so on.

A stochastic primal-dual method (SPD) for nonconvex optimization with many nonconvex constraints has been studied in a recent paper [18]. At each iteration, SPD [18] 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 a nested structure due to the absence of unbiased stochastic gradients.

1.1 Contributions

To tackle the challenges posed by the nested structure and potential nonconvexity of the objective and constraints in the nonconvex constrained composition optimization problem (1.1), we propose a STochastic nEsted Primal-dual (STEP) algorithm. The goal is to find an ϵ-stationary point, where the expected norm of the KKT (Karush-Kuhn-Tucker) measure in terms of stationarity, primal feasibility, and complementary slackness, measured in Euclidean norm, is below the desired accuracy level ϵ. By 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 gain insights into asymptotic dynamics of the stochastic update schemes, we analyze a deterministic continuous-time system. Under mild conditions, we establish that the iteration and sample complexities of the proposed algorithm with constant parameters to find an ϵ-stationary point are bounded by $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(\epsilon^{-6})$ (Theorem 3.7). When the initial guess of the primal variable is (nearly) feasible, we can reduce the previous orders to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-5})$, respectively. Building upon this result, we propose a two-stage algorithm called STEP+, which can find an ϵ-stationary point within $\mathcal{O}(\epsilon^{-3})$ iterations and $\mathcal{O}(\epsilon^{-5})$ samples (Corollary 3.8). Furthermore, if (1.1) is reduced to the case with merely convex set constraints, i.e., $m = 0$, the previous orders can be further improved to $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$ respectively (Remark 3.1). Notably, the sample complexity is consistent with the best existing result depicted in Table 1. We also propose an adaptive STEP algorithm, adaSTEP, which uses adaptively updated parameters, and address its complexities. Finally we report the promising numerical performances of the proposed algorithms on solving a risk-averse portfolio optimization problem and an orthogonal nonnegative matrix decomposition. To the best of our knowledge, this is the first work on algorithms for stochastic composition optimization with nonconvex constraints.

In Table 1 we provide an overview on our algorithms in this paper and related ones in the literature for nonconvex stochastic composition optimization. It includes information on total sample complexities and problem types as well as associated assumptions for the listed problems. Here, we would like to highlight some key differences between our problem settings and those listed in the table. The problem (1.1) we consider here contains 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 [15] and STABLE [5] and the full space $\mathbb{R}^n$ in other literature, the feasibility of iterates with respect to the general set $\mathcal{X}$ can be more challenging to maintain, which raises additional difficulties when solving (1.1). Furthermore, unlike existing work that assumes $\Lambda \equiv 0$ or Λ is an indicator function, we allow for a more general nonsmooth regularizer, which broadens the scope of problems that our algorithm can handle. 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.

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 $J_g = \nabla g = (\nabla g_1, \ldots, \nabla g_m)^T$. Given $x, y \in \mathbb{R}^n$, define the inner product $\langle x, y \rangle := x^T y$ and Hadamard product $x \odot y := (x_1 y_1, \ldots, x_n y_n)^T$. Without specification we use $\|\cdot\|$ to denote the Euclidean norm and $\|\cdot\|_D$ to denote $\sqrt{\langle \cdot, D \cdot \rangle}$ for a given positive semidefinite matrix $D \in \mathbb{R}^{n \times n}$. Given a point $x \in \mathbb{R}^n$ and a nonempty set $Y \subseteq \mathbb{R}^n$, $\mathbf{d}(x, Y) := \inf_{y \in Y} \|x - y\|$ refers to the distance between x and Y. The distance between two nonempty sets $X, Y \subseteq \mathbb{R}^n$ is denoted by $\mathbf{d}(X, Y) := \inf_{x \in X, y \in Y} \|x - y\|$. For brevity, we denote the square of distance as $\mathbf{d}^2(\cdot, \cdot) := [\mathbf{d}(\cdot, \cdot)]^2$. Given a

Algorithm	Complexity	$\mathcal{X}$	Λ	Smoothness	Ass. ($\mathcal{SFO}$)
a-SCGD [43]	$\mathcal{O}\left(\epsilon^{-7}\right)$	$\mathbb{R}^n$	$\times$	$\oplus$	fourth moment
ASC-PG [45]	$\mathcal{O}\left(\epsilon^{-4.5}\right)$	$\mathbb{R}^n$	$\times$	$\oplus$	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 [58]	$\mathcal{O}\left(\epsilon^{-3}\right)$	$\mathbb{R}^n$	✓	$\oplus$	standard
STABLE [5]	$\mathcal{O}\left(\epsilon^{-4}\right)$	convex	$\times$	$\oplus$	fourth moment
NASA [15]	$\mathcal{O}\left(\epsilon^{-4}\right)$	convex	$\times$	$\bigcirc$	standard
STEP (Remark 3.1)	$\mathcal{O}\left(\epsilon^{-4}\right)$	convex	✓	$\bigcirc$	standard
STEP (Theorem 3.7)	$\mathcal{O}\left(\epsilon^{-6}\right)$ $(x^0 \notin \mathcal{X})$ $\mathcal{O}\left(\epsilon^{-5}\right)$ $(x^0 \in \mathcal{X})$	nonconvex	✓	$\bigcirc$	standard
STEP+ (Corollary 3.8)	$\mathcal{O}\left(\epsilon^{-5}\right)$	nonconvex	✓	$\bigcirc$	standard

Table 1: Complexities of different algorithms for minimizing the nonconvex objective function $\Gamma(x)+\Lambda(x)$ 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 moments 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 $\phi \in \Xi_l, \xi \in \Xi_u$.

nonempty set $X \subseteq \mathbb{R}^n$, $\mathbf{1}_X$ 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 index for vectors, and the subscript $_k$ for scalars.

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 dual variables through a proper parameter setting scheme, then analyze the theoretical bounds on the KKT measure in terms of stationarity, feasibility as well as complementary slackness, and then deduce the iteration and the sample complexities accordingly. We also propose a two-stage STEP algorithm with complexity analysis. In Section 4 we present an adaptive STEP algorithm by adopting parameters that are updated adaptively, and present the complexity analysis accordingly. In Section 5, numerical experiments on a risk-adverse portfolio optimization problem and an orthogonal nonnegative matrix decomposition are reported. 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 challenging to obtain a global or even a local minimizer for nonconvex constrained optimization. As a result, the focus often shifts towards seeking more tractable solutions. It has been shown that under certain constraint qualifications [28], a local minimizer satisfies first-order necessary conditions, known as Karush-Kuhn-Tucker (KKT) conditions. Points satisfying KKT conditions are referred to as KKT points. In this paper we do not explicitly specify the constraint qualifications. Instead, we assume the existence of a KKT point of (1.1). Before giving its definition, we need following concepts. The normal cone [16] of a convex set X at a point $x \in X$ is denoted by

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

For the proper lower-semicontinuous convex function $\Lambda : \mathbb{R}^n \to \mathbb{R}$, its subdifferential [36] at x is defined as

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

And each element of $\partial\Lambda(x)$ is called a subgradient of Λ 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*

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

As is well known, the augmented Lagrangian (AL) function plays an important role in characterizing the optimality conditions for constrained optimization and helping design effective algorithms. For any $x \in X$, the AL function [35] associated with (1.1) can be defined as [1]

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

where $\beta > 0$,

$$\mathcal{D}_\beta(x, z) := \Gamma(x) + \Psi_\beta(x, z), \quad \Psi_\beta(x, z) := \sum_{i=1}^m \psi_\beta(g_i(x), z_i) \text{ with } \psi_\beta(u, v) = \begin{cases} uv + \frac{\beta}{2}u^2, & \text{if } \beta u + v \geq 0, \\ -\frac{v^2}{2\beta}, & \text{otherwise.} \end{cases} \tag{2.3}$$

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 [43], we use an auxiliary variable y to track the inner layer function value and update it in a moving average way:

$$y^{k+1} = (1 - \eta_k) y^k + \frac{\eta_k}{P_{k,1}} \sum_{\phi \in \mathcal{P}_{k,1}} H\left(x^k; \phi\right) \tag{2.4}$$

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

$$\overline{\nabla} \Gamma^k := \left[\frac{1}{P_{k,2}} \sum_{\phi \in \mathcal{P}_{k,2}} \nabla H\left(x^k; \phi\right)\right]^T \left[\frac{1}{J_k} \sum_{\xi \in \mathcal{J}_k} \nabla F\left(y^{k+1}; \xi\right)\right]. \tag{2.5}$$

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

$$\overline{\nabla} G^k := \nabla_x \Psi_\beta\left(x^k, z^k\right) = \sum_{i=1}^m \left[\beta g_i\left(x^k\right) + z_i^k\right]_+ \nabla g_i\left(x^k\right). \tag{2.6}$$

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 $h\left(x^k\right)$, $(\overline{\nabla} \Gamma^k + \overline{\nabla} G^k)$ will be a good approximation to $\nabla_x \mathcal{D}_\beta(x^k, z^k)$. Then it is straightforward to have

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

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

$$x^{k+1} = \underset{x \in X}{\operatorname{argmin}} \left\{ \left\langle \overline{\nabla} \Gamma^k + \overline{\nabla} G^k, x \right\rangle + \frac{1}{2\alpha_k} \left\|x - x^k\right\|^2 + \Lambda(x) \right\}. \tag{2.7}$$

It is noteworthy that different from classical double-loop AL methods, only an approximation to the original AL function is minimized in (2.7). This idea is motivated by linearized AL methods, such as [47,48,50], which can make

[1] According to formulas (7)-(9) in [35], the associated AL function of (1.1) is defined as

$$L_r(x, z) := \Gamma(x) + \Lambda(x) + \mathbf{1}_X(x) + \frac{1}{4r} \sum_{i=1}^m \left([z_i + 2r g_i(x)]_+^2 - z_i^2\right)$$

with $r > 0$. Since for any $i \in [m]$,

$$\frac{1}{4r}\left([z_i + 2r g_i(x)]_+^2 - z_i^2\right) = \begin{cases} z_i g_i(x) + r[g_i(x)]^2, & \text{if } 2r g_i(x) + z_i \geq 0, \\ -\frac{z_i^2}{4r}, & \text{otherwise,} \end{cases}$$

we can simplify the AL function as (2.2) by defining $\beta = 2r$ and restricting $\operatorname{dom} \mathcal{L}_\beta(\cdot, z)$ to be X.

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 (2.7) 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 to update the dual variable:

$$\begin{aligned} z^{k+1} &= \frac{\rho_k}{\beta}\bar{z}^{k+1} + \left(1 - \frac{\rho_k}{\beta}\right) z^k \\ &= \frac{\rho_k}{\beta}\left(z^k + \beta \max\left(-\frac{z^k}{\beta}, g\left(x^{k+1}\right)\right)\right) + \left(1 - \frac{\rho_k}{\beta}\right) z^k \\ &= z^k + \rho_k \cdot \max\left(-\frac{z^k}{\beta}, g\left(x^{k+1}\right)\right), \end{aligned} \tag{2.8}$$

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 (2.7) and (2.8), let us consider a corresponding continuous-time system in deterministic setting. We take a special case of (1.1) as an example, where $X = \mathbb{R}^n$ and $\Lambda \equiv 0$. 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:

$$\dot{x}(t) = -\alpha\left(\nabla h(x(t))^T \nabla f(y(t)) + \nabla_x \Psi_\beta(x(t), z)\right), \tag{2.9}$$

$$\dot{z}(t) = \rho \nabla_z \Psi_\beta(x, z(t)), \tag{2.10}$$

where $\alpha > 0, \rho > 0$ and $y(t)$ is an approximation to $h(x(t))$. By [13], (2.9) and (2.10) can be regarded as the continuous analogues of (2.7) and (2.8), respectively. From (2.9) and the Young's inequality, it holds that

$$\begin{aligned} \frac{d}{dt}\mathcal{L}_\beta(x(t), z) &= \left[\nabla_x \mathcal{L}_\beta(x(t), z)\right]^T \dot{x}(t) \\ &= -\alpha\left[\nabla_x \mathcal{L}_\beta(x(t), z)\right]^T \left[\nabla_x \mathcal{L}_\beta(x(t), z) + \nabla h(x(t))^T \nabla f(y(t)) - \nabla h(x(t))^T \nabla f(h(x(t)))\right] \\ &\le -\alpha \left\|\nabla_x \mathcal{L}_\beta(x(t), z)\right\|^2 + \alpha \left\|\nabla_x \mathcal{L}_\beta(x(t), z)\right\| \left\|\nabla h(x(t))^T \nabla f(y(t)) - \nabla h(x(t))^T \nabla f(h(x(t)))\right\| \\ &\le -\frac{\alpha}{2}\left\|\nabla_x \mathcal{L}_\beta(x(t), z)\right\|^2 + \frac{\alpha}{2}\left\|\nabla h(x(t))^T \nabla f(y(t)) - \nabla h(x(t))^T \nabla f(h(x(t)))\right\|^2. \end{aligned} \tag{2.11}$$

Then $\frac{d}{dt}\mathcal{L}_\beta(x(t), z)$ can be nonpositive when the second term of the right hand side of (2.11) 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 (2.10) that $\mathcal{L}_\beta(x, z(t))$ keeps monotonicity with respect to t, due to the fact that

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

We are now ready to present the stochastic nested primal-dual method for (1.1) in Algorithm 2.1. To simplify the convergence analysis, we will adopt constant parameters in Algorithm 2.1. A variant of algorithm with adaptively updated parameters will be introduced in Section 4.

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 (2.4).
4: Compute $\overline{\nabla}\Gamma^k$ and $\overline{\nabla}G^k$ through (2.5) and (2.6), respectively.
5: Compute x^{k+1} through (2.7).
6: Compute z^{k+1} through (2.8).
7: **end for**

It is straightforward to obtain the nonnegativity of z^k from $z^0 = \mathbf{0}$, $\rho_k \subseteq (0, \beta]$ and (2.8). So we state the lemma with the proof omitted.

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

Since the iteration process of STEP is random, we aim for an ϵ-stationary point of (1.1) with the expected KKT measure below a given tolerance $\epsilon > 0$.

DEFINITION **2.2.** *Let $x \in X$ be generated by a random process. Then x is called an ϵ-stationary point of* (1.1) *for a given $\epsilon > 0$, if there exists $z \in \mathbb{R}^m_+$ such that*

$$\mathbb{E}\left[\mathbf{d}\left(\nabla\Gamma(x) + \partial\Lambda(x) + \sum_{i=1}^{m} z_i \nabla g_i(x) + \mathcal{N}_X(x), \mathbf{0}\right)\right] \leq \epsilon, \tag{2.12}$$

$$\mathbb{E}\left[\left\|[g(x)]_+\right\|\right] \leq \epsilon, \tag{2.13}$$

$$\mathbb{E}\left[\|z \odot g(x)\|\right] \leq \epsilon, \tag{2.14}$$

where the expectation is taken with respect to all the random variables related to the generation of x.

3 Theoretical analysis

Our focus in this section is to study theoretical properties of the STEP algorithm. The goal is to 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. The objective function value of* (1.1) *over X is lower bounded by C^*. 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 $y, \bar{y} \in \mathbb{R}^{\bar{n}}$,*

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

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

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

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

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

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

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

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 [15], while the work on constrained stochastic optimization [3] also assumes (3.3). In addition, the boundedness on $[g_i]_+$ and $\|\partial\Lambda\|$ is also required in [3] and ASC-PG [45] respectively. In [49] 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 $y \in \mathbb{R}^{\bar{n}}$,*

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

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

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

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

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

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 $k \in [K+1]$,*

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

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

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

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

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

Together with $\sum_{t=0}^{k-1}\rho_t \leq \rho k/K \leq 2\rho$ 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 $k = 0, 1, \ldots, K$,*

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

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

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

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

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

which yields (3.7).

Secondly, it follows from (2.6), (3.3) and (3.4) that for any $x \in \mathbb{R}^n$,

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

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

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

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

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 (2.5) that

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

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 $k=0,\ldots,K$,*

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

where σ_{Γ_k} is defined in Lemma 3.4.

Proof. For any $k\geq 0$, it follows from optimality conditions for (2.7) that

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

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

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

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

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

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 (2.7) that there exists a vector $v \in \mathcal{N}_X\left(x^{k+1}\right)$ satisfying

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

which by the convexity of Λ and (2.1) indicates

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

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

$$\mathcal{L}_\beta\left(x^{k+1}, z^k\right)$$

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

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

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

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 (2.3) that for any $u \in [0, G]$ with $G > 0$ and $(v_1, v_2) \in \mathbb{R}_+^2$,

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

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

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

where equalities in the second and third cases use the monotonically decreasing property of $\psi_\beta(u, v)$ in $v \geq 0$ 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 $k \in [K]$,*

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

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

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

$$+\left(1+\gamma_k^{-1}\right)(1-\eta_k)^2 L_{h,0}^2\left\|x^k-x^{k-1}\right\|^2.$$

Notably, (2.4) 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

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

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

Proof. It follows from Lemma 3.5 that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right)+\partial\Lambda\left(x^{R+1}\right)+\sum_{i=1}^m\left[\beta g_i\left(x^{R+1}\right)+z_i^{R+1}\right]_+\nabla g_i\left(x^{R+1}\right)+\mathcal{N}_X\left(x^{R+1}\right),\mathbf{0}\right)\right] \\
=&\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{k+1}\right)+\partial\Lambda\left(x^{k+1}\right)+\sum_{i=1}^m\left[\beta g_i\left(x^{k+1}\right)+z_i^{k+1}\right]_+\nabla g_i\left(x^{k+1}\right)+\mathcal{N}_X\left(x^{k+1}\right),\mathbf{0}\right)\right] \\
\leq&\frac{3}{K}\sum_{k=1}^K\left(L_\Gamma+L_\beta+\frac{1}{\alpha_k}\right)^2\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right]+\frac{6L_f^2}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right] \\
&+\frac{3L_{g,0}^2m}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|^2\right]+\frac{6}{K}\sum_{k=1}^K\sigma_{\Gamma_k}^2 \\
\leq&\frac{3}{K}\sum_{k=1}^K\mathbb{E}\left[\left(L_\Gamma+L_\beta+\frac{1}{\bar{\alpha}}\right)^2\left\|x^{k+1}-x^k\right\|^2+2L_f^2\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right] \\
&+\frac{3L_{g,0}^2m}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|^2\right]+6\sigma_\Gamma^2,
\end{aligned} \tag{3.22}$$

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

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

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 $k\in[K]$:

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

Summing up the above inequality over $k=1,\ldots,K$ leads to

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

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

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

and

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

Applying (3.27)-(3.28) to relax the coefficients of $\left\|x^{k+1}-x^k\right\|^2$ and $\left\|y^{k+1}-h\left(x^k\right)\right\|^2$ on the left side of (3.26), we obtain

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

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

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 Assumptions 3.1, 3.2 and Lemma 3.2 that

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

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

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

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

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

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

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

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

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), and 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 observed from (3.34), we can see that

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

This result provides a theoretical explanation that y^{k+1} can be close to $h(x^k)$ in expectation when K is sufficiently large. It aligns with our expectation discussed for (2.11). Besides, (3.32) indicates that the value $\mathcal{L}_\beta(x^0,z^0)$ plays a crucial effect in determining orders obtained in (3.34) and (3.21). Notably, when the initial point x^0 is feasible for (1.1), $\mathcal{L}_\beta(x^0,z^0)$ becomes independent of parameter β. Therefore, improved results can be derived as shown in the corollary below. The proof is straightforward, following a similar analysis to that of Theorem 3.1, and thus we omit it here.

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

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

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

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

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 convex set constrained optimization problem*

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

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 letting

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

we can obtain

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

Then the iteration and sample complexities to find an ϵ-stationary point of (3.35), i.e. satisfying (2.12), are bounded by $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$, respectively, where the sample complexity is same as NASA [15].

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 problem, to guarantee near-feasibility we give another assumption.

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

$$\nu \left\|[g(x)]_+\right\| \leq \mathbf{d}\left(\left(J_g(x)\right)^T [g(x)]_+ + \mathcal{N}_X(x), \mathbf{0}\right) \tag{3.36}$$

holds for all x^k with $k \in [K+1]$.

In the 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 $\{\nabla g_i(x) \mid i \in [m], g_i(x) > 0\}$. For general inequality constraints, it has been verified in [24] and [18] that the multi-class Neyman-Pearson classification (mNPC) problems and constrained form of weighted maximin dispersion problem satisfy the NSC condition, respectively. Regarding general equality constraints, the NSC condition has been shown to hold for generalized eigenvalue problems and clustering [38], and also hold for linear equality constraints with box/ball constraint [22]. In the case $X = \mathbb{R}^n$, the NSC condition can be implied by the strong LICQ used in [9], which requires the singular values of $J_g(x)$ be bounded below and away from zero.

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

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

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

$$\begin{aligned}\left\|\left[g\left(x^k\right)\right]_+\right\| &\leq \frac{1}{\nu\beta}\mathbf{d}\left(\beta\left(J_g\left(x^k\right)\right)^T \left[g\left(x^k\right)\right]_+ + \mathcal{N}_X\left(x^k\right), \mathbf{0}\right) \\ &\leq \frac{1}{\nu\beta}\left\|\sum_{i=1}^{m}\left[\beta g_i\left(x^k\right)\right]_+ \nabla g_i\left(x^k\right) + v^k\right\|\end{aligned}$$

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

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

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

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

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

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

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

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

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

which further yields

$$\mathbb{E}\left[\sum_{i=1}^m\bar{z}_i\left[g_i\left(x^{R+1}\right)\right]_-\right]=\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\sum_{i=1}^m\left[\beta g_i\left(x^{k+1}\right)+z_i^{k+1}\right]_+\left[g_i\left(x^{k+1}\right)\right]_-\right]\le\frac{mG^2\rho^2}{\beta}=\mathcal{O}\left(K^{-\frac{1}{4}}\right).\tag{3.37}$$

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

$$\mathbb{E}\left[\sum_{i=1}^m z_i^{R+1}\left[g_i\left(x^{R+1}\right)\right]_+\right]\le\mathbb{E}\left[2G\rho\sum_{i=1}^m\left[g_i\left(x^{R+1}\right)\right]_+\right]\le\mathbb{E}\left[2G\rho\sqrt{m}\left\|\left[g\left(x^{R+1}\right)\right]_+\right\|\right]=\mathcal{O}\left(K^{-\frac{1}{4}}\right),\tag{3.38}$$

where the equality follows from Theorem 3.3 and $\left(\mathbb{E}\left[u\right]\right)^2\le\mathbb{E}[u^2]$ for any random variable $u\in\mathbb{R}$. Combining (3.38) with $\beta=K^{1/4}$ and Theorem 3.3, we have

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

which together with (3.37) 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 $\|\cdot\|\le\|\cdot\|_1$, 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 (2.4) and (2.5).

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

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 $\left(\mathbb{E}\left[u\right]\right)^2\le\mathbb{E}\left[u^2\right]$ for any random variable u. Regarding the sample complexity, as the total number of samples is $\sum_{k=0}^K\left(J_k+P_{k,1}+P_{k,2}\right)$, 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, Algorithm 2.1 and its theoretical convergence can be extended to the case of equality-constrained optimization with some modifications, as demonstrated in [39]. The numerical performance of the generalized algorithm in the equality-constrained case is also validated in Section 5.2, providing empirical evidence of its effectiveness.*

It is noteworthy that feasible points for certain types of nonconvex constraints can be obtained at a relatively low cost. For example, let us consider the orthogonal nonnegative decomposition [12,30,33], where the constraints are $U \in \mathbb{R}^{m\times r}$ and $U^TU = I_r$, with integers $r < m$. In this case, satisfying the constraints can be achieved by setting $U = (e_1, \ldots, e_m)$, where $e_i \in \mathbb{R}^m$ has only one nonzero entry equal to 1 at the i-th coordinate. For more general constraints, although a feasible point can be hard to find, it is manageable under certain conditions to locate an approximately feasible initial point, with which previous analysis in Corollaries 3.2, 3.4 and 3.6 still hold. We here provide a two-stage algorithm combining an initial feasibility-seeking phase with the optimization phase as indicated by Algorithm 2.1. We give the outline of the two-stage algorithm (Algorithm 3.1).

Algorithm 3.1 STEP+

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

1: Call the gradient descent method [31] on $\min_x[\|[g(x)]_+\|^2/2 + \mathbf{1}_X(x)]$ to find a point $x^0 \in X$ such that $\mathbf{d}([J_g(x^0)]^T[g(x^0)]_+ + \mathcal{N}_X(x^0), \mathbf{0}) \le K^{-1/6}$.

2: Call the Algorithm 2.1 with initial point x^0.

Corollary 3.8. (Iteration and sample complexities) *Assume that* (3.36) *holds at x^0. Under conditions of Corollary 3.2, given $\epsilon > 0$, the STEP+ method can find an ϵ-stationary point of* (1.1) *after $\mathcal{O}(\epsilon^{-3})$ iterations with associated sample complexity bounded by $\mathcal{O}(\epsilon^{-5})$.*

Proof. It follows from Assumptions 3.1 and 3.2 that the gradient of $\|[g(x)]_+\|^2/2$ is Lipschitz continuous with modulus $(GL_{g,1}+L^2_{g,0})$. Together with the convexity of X and Proposition 2.5 of [31], it implies that $[\|[g(x)]_+\|^2/2+ \mathbf{1}_X(x)]$ is weakly-convex. Besides, its value is lower bounded by 0. By setting $K = \lceil \epsilon^{-3} \rceil$, we obtain from Table 2 of [31] that a (full) gradient descent method can find a point $x^0 \in X$ such that $\mathbf{d}([J_g(x^0)]^T[g(x^0)]_+ + \mathcal{N}_X(x^0), \mathbf{0}) \le K^{-1/6}$ within $\tilde{\mathcal{O}}((K^{-1/6})^{-2}) = \tilde{\mathcal{O}}(\epsilon^{-1})$ iterations and evaluations of $(g, \nabla g)$. Applying (3.36), we further have

$$\|[g(x^0)]_+\|^2 \le \left[\frac{1}{\nu}\mathbf{d}([J_g(x^0)]^T[g(x^0)]_+ + \mathcal{N}_X(x^0), \mathbf{0})\right]^2 = \mathcal{O}(K^{-1/3}).$$

It together with (3.32) and $\beta = K^{1/3}$ yields that the iteration and sample complexities are $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-5})$, respectively, for the second stage similar to Corollaries 3.2, 3.4 and 3.6. By summing up the complexities in two stages, we obtain the result. □

4 Adaptive STEP

In the section, we will propose an adaptive variant of algorithm, adaSTEP, in Algorithm 4.1.

Algorithm 4.1 adaSTEP

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

Output: x^{R+1} and $\bar{z}^{R+1} := [\beta_R g(x^{R+1}) + z^{R+1}]_+$ 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 (2.4).

4: Compute $\overline{\nabla}\Gamma^k$ through (2.5), and $\overline{\nabla} G^k := \nabla_x \Psi_{\beta_k}(x^k, z^k)$, respectively.

5: Compute x^{k+1} through

$$x^{k+1} = \underset{x \in X}{\operatorname{argmin}} \left\{ \left\langle \overline{\nabla}\Gamma^k + \overline{\nabla} G^k, x \right\rangle + \frac{1}{2} \left\| x - x^k \right\|_{D_k}^2 + \Lambda(x) \right\}, \tag{4.1}$$

where $D_k := \operatorname{diag}(s^k) + \alpha_k^{-1} \mathrm{I}$ with

$$s^k := \mu \left(\sum_{t=0}^{k} \frac{(\overline{\nabla}\Gamma^t + \overline{\nabla} G^t)^2}{\left(\max(1, \|\overline{\nabla}\Gamma^t + \overline{\nabla} G^t\|) \right)^2} \right)^{1/4}. \tag{4.2}$$

6: Compute $z^{k+1} := z^k + \rho_k \max(-\frac{z^k}{\beta_k}, g(x^{k+1}))$.

7: **end for**

In Algorithm 4.1 we adopt a set of adaptive parameters motivated by [49]. In (4.2), u^2 and $u^{1/4}$ denote componentwise square and fourth-root for any non-negative vector u. For any k, we have the following bound for the i-th component of s^k:

$$0 \leq s_i^k \leq \mu \left(\sum_{t=0}^{k} \frac{((\overline{\nabla}\Gamma^t + \overline{\nabla} G^t)_i)^2}{\left(\max(1, \|\overline{\nabla}\Gamma^t + \overline{\nabla} G^t\|) \right)^2} \right)^{1/4} \leq \mu (k+1)^{1/4},$$

which implies that $\|s^k\| \leq \mu \sqrt{n} (k+1)^{1/4}$. Note that D_k is adaptive to the stochastic gradient and dependent on $\mu > 0$. Thus Algorithm 4.1 can also adjust its adaptiveness by properly choosing the value of μ.

In Algorithm 4.1, the nonnegativity of z^k still holds due to $z^0 = \mathbf{0}, \rho_k \subseteq (0, \beta_k)$. We assume that $\{\beta_k\}$ is an increasing sequence, and

$$\rho_k \in \left(0, \frac{\rho}{K}\right] \subseteq (0, \beta_0], \quad k = 0, \ldots, K,$$

where $\rho, \beta_0 > 0$ are independent of K. Then Lemma 3.2 holds obviously and Lemma B.1 is true. As the main proof sketch is similar to the analysis in the previous section, we present the main results below and provide the detailed proofs in Appendix C.

THEOREM **4.1.** *Under Assumptions 3.1-3.3, set*

$$\begin{aligned} &\beta_k = \beta_0 (k+1)^{1/4},\ \rho_k \equiv \frac{\rho}{K},\ P_{k,1} \equiv P_1 := \lceil K^{1/4} \rceil,\ P_{k,2} \equiv P_2 := \lceil K^{1/2} \rceil,\ J_k \equiv J := \lceil K^{1/2} \rceil, \\ &\alpha_k := [2L_\Gamma + 2L_{\beta_k} + (12 + L_{f,1}^2) L_{h,0}^2]^{-1},\ \eta_k := 2\alpha_k \max\left(L_f^2, 8L_{h,0}^2\right),\ \gamma_k := 8L_{h,0}^2 \alpha_k \end{aligned} \tag{4.3}$$

for $k \geq 0$, where

$$\bar{\alpha}_1 := [2\max(L_f^2, 8L_{h,0}^2)]^{-1},\ \beta_0 \geq \frac{\bar{\alpha}_1^{-1} - 2L_\Gamma - (12 + L_{f,1}^2) L_{h,0}^2 - 2mG\rho L_{g,1}}{mL_{g,0}^2 + mGL_{g,1}},\ \beta_0 > 0.$$

Then it holds that

$$\mathbb{E}\left[\mathbf{d}^2 \left(\nabla\Gamma\left(x^{R+1}\right) + \partial\Lambda\left(x^{R+1}\right) + \sum_{i=1}^{m} \bar{z}_i^{R+1} \nabla g_i\left(x^{R+1}\right) + \mathcal{N}_X\left(x^{R+1}\right), \mathbf{0} \right) \right] = \mathcal{O}\left(K^{-\frac{1}{2}}\right), \tag{4.4}$$

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

THEOREM **4.2.** *Under the conditions of Theorem 4.1 and Assumption 3.4, it holds that*

$$\mathbb{E}\left[\left\|\left[g\left(x^{R+1}\right)\right]_+\right\|^2\right] = \mathcal{O}\left(K^{-\frac{1}{2}}\right) \tag{4.5}$$

and

$$\mathbb{E}\left[\left\|\tilde{z}^{R+1} \odot g\left(x^{R+1}\right)\right\|\right] = \mathcal{O}\left(K^{-\frac{1}{4}}\right). \tag{4.6}$$

We now summarize Theorems 4.1 and 4.2 into the theorem below characterizing both iteration and sample complexities of Algorithm 4.1 to find an ϵ-stationary point of (1.1).

THEOREM **4.3.** *Under the conditions of Theorem 4.1 and Assumption 3.4 and given $\epsilon > 0$, Algorithm 4.1 can find an ϵ-stationary point of* (1.1) *after $\mathcal{O}(\epsilon^{-4})$ iterations with associated sample complexity bounded by $\mathcal{O}(\epsilon^{-6})$.*

Remark 4.1. *As can be observed from* (6.17) *and* (6.18)*, the complexity of Algorithm 4.1 is still affected by β_{K+1}, which is in the order of $\mathcal{O}(K^{1/4})$, and it cannot be improved by requiring the feasibility of x^0. Although in* (4.3)*, we set ρ_k and batch-sizes as constants dependent on K to simplify the analysis, similar results still hold for varying ρ_k and batch-sizes, e.g. $\rho_k := \rho/(k+1)^2$, $P_{k,1} := \lceil (k+1)^{1/4} \rceil$, $P_{k,2} = J_k := \lceil (k+1)^{1/2} \rceil$.*

5 Numerical experiments

5.1 Risk-averse portfolio optimization

In this section, we would like to carry out some numerical experiments for solving the following risk-averse portfolio optimization problem [7, 55, 58]:

$$\min_{x \in \Delta^n} \Gamma(x) := -\mathbb{E}_\phi\left[r_\phi(x)\right] + \lambda \operatorname{Var}\left[r_\phi(x)\right] \quad \text{s.t. } Ax \le b, \tag{5.1}$$

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 (5.1) into the form (1.1) by defining $h(x) = \mathbb{E}_\phi\left[H(x;\phi)\right]$ with $H(x;\phi) = [r_\phi(x), r^2_\phi(x)]^T$ and $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(y) = -y_1 + \lambda y_2 - \lambda y_1^2$. As indicated by [40], the (non)linear constraints have been commonly imposed in portfolio optimization, such as bounds on individual asset positions ($Ax \le b$). The reason to consider linear constraints here is that the algorithm com-SVR-ADMM to be compared, which currently is the only algorithm in the literature for functional constrained stochastic compositional optimization problems, is designed to solve (1.1) with linear constraints. 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 [55]) 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, 58]. 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

$$\mathbb{E}_\phi\left[r_\phi(x)\right] := \frac{1}{P}\sum_{\phi=1}^P R_\phi^T x, \quad \mathbb{E}_\phi\left[r^2_\phi(x)\right] := \frac{1}{P}\sum_{\phi=1}^P \left(R_\phi^T x\right)^2. \tag{5.2}$$

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 (5.1). For both com-SVR-ADMM and STEP, based on the result evaluated by $fmincon$, and the theoretical parameter settings in [55] and in this paper, we have fine-tuned parameters for each of the two algorithms when solving (5.1). Below are the specific settings for both algorithms.

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

- 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.

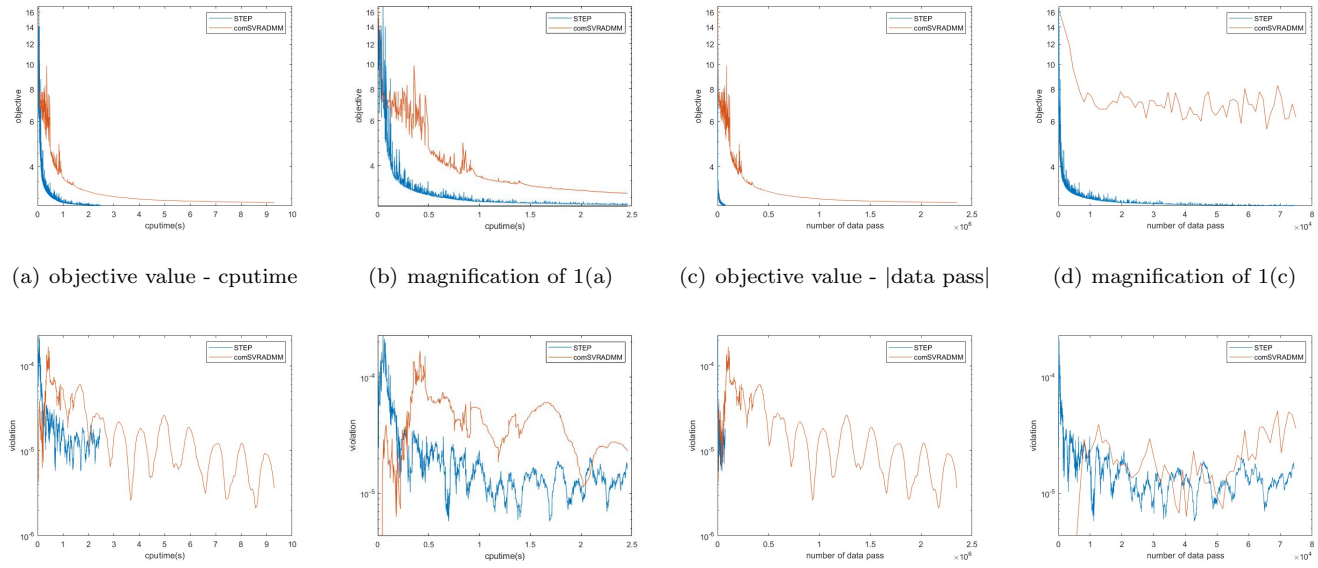
- STEP:

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

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

The maximum number of iterations for both is set to 2000.

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 and the number of data-pass $\sum_{k=0}^{K}(J_k + P_{k,1} + P_{k,2})$ in 10 repeated tests on four industry portfolios. And for each industry portfolio we only show the performances under a single set of randomly generated constraints (A, b) based on the observation that both algorithms perform similarly under other constraints randomly generated by same method. Noting that the risk-averse portfolio optimization problem in our test is a convex optimization problem, we choose the objective value and constraint violation to measure the numerical performances [49, 50]. From Figures 1-4, we can see that objective function values by STEP decrease faster. Although the constraint violations by both algorithms are gradually approaching similar values, com-SVR-ADMM oscillates more dramatically than STEP.



(a) objective value - cputime (b) magnification of 1(a) (c) objective value - |data pass| (d) magnification of 1(c)

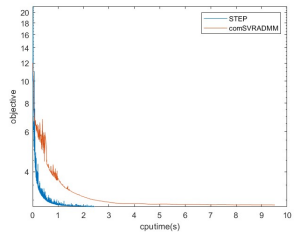
(e) violation - cputime (f) magnification of 1(e) (g) violation - |data pass| (h) magnification of 1(g)

Figure 1: Comparison on 49-Industrial Portfolio Dataset for solving (5.1)

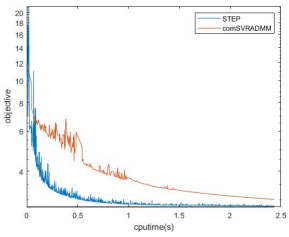
5.2 Orthogonal nonnegative matrix decomposition

In this subsection we consider the following orthogonal nonnegative matrix decomposition problem [12, 30, 33]:

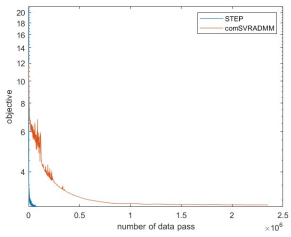
$$\min_{(U,V)\in\mathbb{R}^{m\times r}_+\times\mathbb{R}^{r\times n}_+} \left\|\bar{X} - UV\right\|_F^2, \text{ s.t. } U^TU = I_r, \tag{5.3}$$



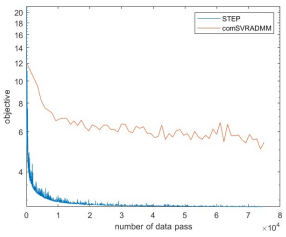
(a) objective value - cputime



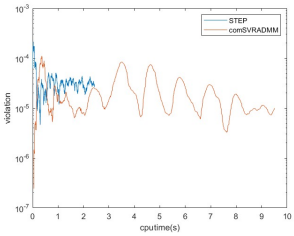
(b) magnification of 2(a)



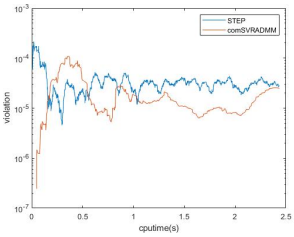
(c) objective value - |data pass|



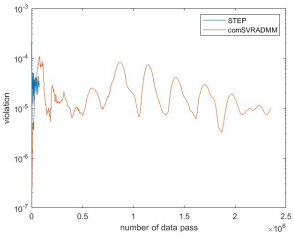
(d) magnification of 2(c)



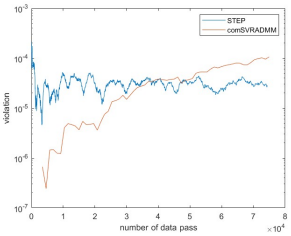
(e) violation - cputime



(f) magnification of 2(e)

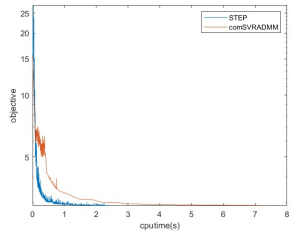


(g) violation - |data pass|

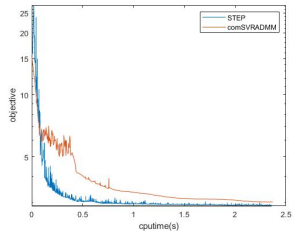


(h) magnification of 2(g)

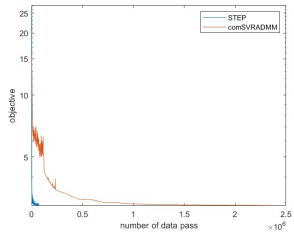
Figure 2: Comparison on 48-Industrial Portfolio Dataset for solving (5.1)



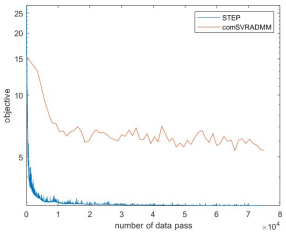
(a) objective value - cputime



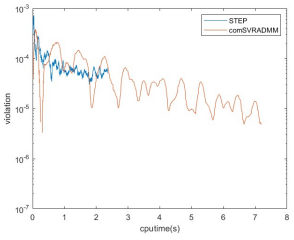
(b) magnification of 3(a)



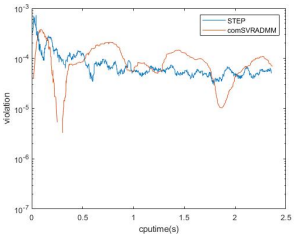
(c) objective value - |data pass|



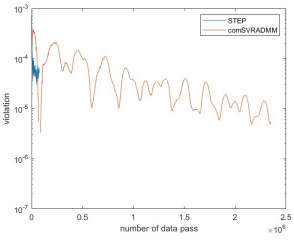
(d) magnification of 3(c)



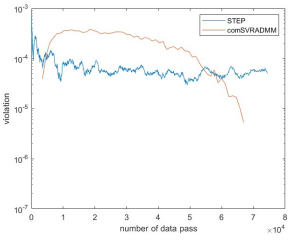
(e) violation - cputime



(f) magnification of 3(e)

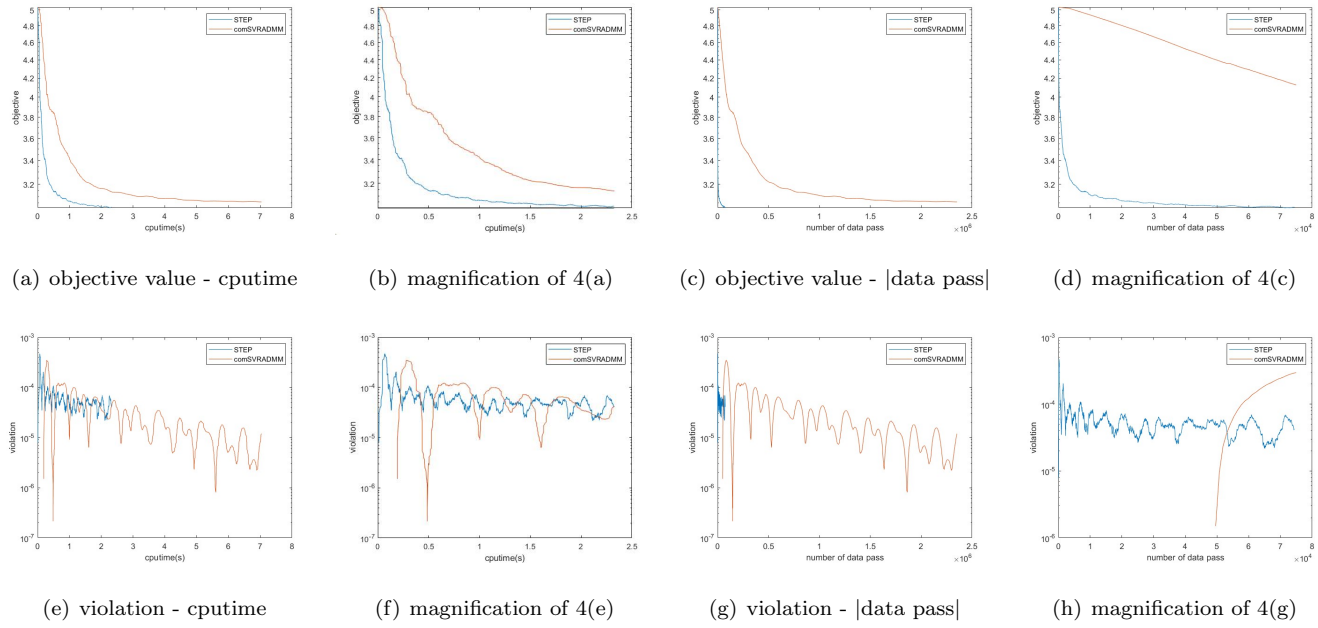


(g) violation - |data pass|



(h) magnification of 3(g)

Figure 3: Comparison on 38-Industrial Portfolio Dataset for solving (5.1)



(a) objective value - cputime (b) magnification of 4(a) (c) objective value - |data pass| (d) magnification of 4(c)

(e) violation - cputime (f) magnification of 4(e) (g) violation - |data pass| (h) magnification of 4(g)

Figure 4: Comparison on 30-Industrial Portfolio Dataset for solving (5.1)

where $\bar{X} \in \mathbb{R}_+^{m\times n}$ is an unknown matrix estimated by $\bar{X} = \mathbb{E}[X]$ based on some unbiased samples $\{X\}$ [15], $I_r \in \mathbb{R}^{r\times r}$ with $r < \min\{m, n\}$ is the identity matrix. Note that nonnegativity and orthogonality can guarantee the sparsity of matrix U ([29, Lemma 1]). Then there is no need to add a regularizer such as $\sum_{i,j} |U_{i,j}|$ in (5.3). Although a feasible point of (5.3) can be easily obtained, we focus on performances of algorithms initialized with an infeasible point in this section. All the implementations are conducted in MATLAB version R2022a on a desktop computer of 64GB RAM and 12th Gen Intel(R) Core(TM) i7-12700 2.10 GHz.

Orthogonal nonnegative matrix decomposition (5.3) is a nonlinear constrained optimization problem with the stochastic composition objective function $\Gamma(U, V) := f(h(U, V))$, where $f(Y) := \|Y\|_F^2$ and $h(U, V) := \mathbb{E}[H(U, V; X)]$ with $H(U, V; X) := UV - X$. Noting that $\nabla_U \Gamma(U, V) = 2(UV - \mathbb{E}[X])V^T$ and $\nabla_V \Gamma(U, V) = 2U^T(UV - \mathbb{E}[X])$ are linear in X, the unbiased estimation of $\nabla\Gamma$ can be obtained by $2YV^T$ and $2U^TY$ with $Y = \sum_{X\in\mathcal{P}_1}(UV - X)/P_1 \approx \mathbb{E}[X]$. But in the numerical tests on (5.3), we can still observe the convergence of STEP with the biased estimation

$$\overline{\nabla}\Gamma_U^k = 2Y(V^k)^T, \overline{\nabla}\Gamma_V^k = 2(U^k)^TY, \text{ with } Y^{k+1} = (1-\eta_k)Y^k + \frac{\eta_k}{P_{k,1}} \sum_{X\in\mathcal{P}_{k,1}} (U^kV^k - X), \eta_k \in (0, 1),$$

which complies with the convergence analysis in Section 3. We also notice that in later iterations it is more stable compared with the case when the estimation is unbiased (i.e. $\eta_k \equiv 1$). Hence, we keep the biased estimation for STEP in formal comparison.

Except STEP, the existing methods [12, 29, 30, 33, 54] for (5.3) only consider the case with known $\bar{X}$. Hence, we first apply a sample averaging approximation (SAA) on $\mathbb{E}[X]$, i.e. replacing $\mathbb{E}[X]$ by $\tilde{X}_N = \sum_{i\in[N]} X_i/N$ with sample size $N = 100$. The OPNMF[3] method studied in [54] aims for solving the orthonormal projective nonnegative matrix factorization, which is equivalent to (5.3) [12, 29]. We thus compare STEP and adaSTEP with OPNMF on the above SAA problem. We choose two datasets to compare: Iris (provided by MATLAB) and tr23[4], with synthetic noise randomly generated according to the rules in [33]. The dimensions of the problem are set as $(m, n, r) = (4, 150, 3)$ for dataset Iris and $(m, n, r) = (204, 5832, 6)$ for dataset tr23, respectively. Moreover, for both datasets the noise for each entry is drawn from a normal distribution with zero mean and fixed standard deviation $\varepsilon := 0.01$, since the performance of OPNMF will be significantly influenced by deviation larger than 0.01 according to figure 4 in [33]. As a multiplicative method, OPNMF is free from cumbersome parameter adjustment.

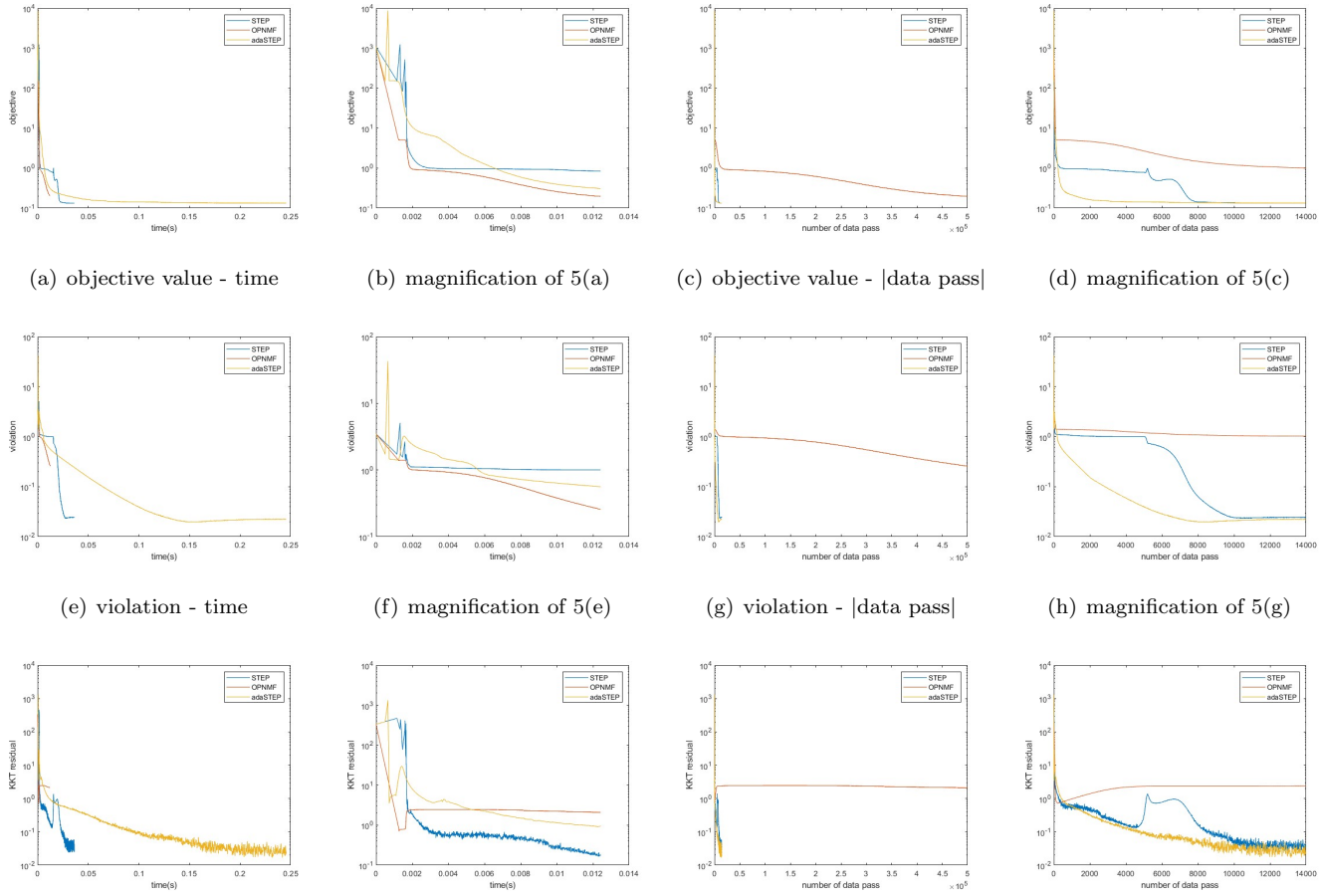
[3]Code available at https://folk.ntnu.no/yangzh/pnmf/index.html

[4]tr23 is one of the well-known preprocessed document databases in evaluating the performance of CLUTO's clustering algorithms. It can be downloaded from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download/

It only needs sufficient iterations and a tiny increment on the denominator of its update formulas to keep numerical stability. For STEP and adaSTEP, parameter settings on two datasets are set as below.

- Iris. $K = 5000, \beta_k = 2(k+1)^{0.25}, \eta_k = (k+1)^{-0.25}, \rho_k = (k+1)^{-0.25}K^{-1}, P_{k,1} = \lceil (k+1)^{0.1} \rceil$,
  - STEP: $\alpha_k = 8.658 \times 10^{-3}$;
  - adaSTEP: $\alpha_k = \frac{3.463\times 10^{-2}}{(k+1)^{0.25}}, \mu = 1$
- tr23. $K = 10000, \beta_k = 20(k+1)^{0.25}, \eta_k = (k+1)^{-0.25}, , \rho_k = (k+1)^{-0.25}K^{-1}, P_{k,1} = \lceil (k+1)^{0.1} \rceil$,
  - STEP: $\alpha_k = 9.388 \times 10^{-4}$,
  - adaSTEP: $\alpha_k = \frac{1.684\times 10^{-3}}{(k+1)^{0.25}}, \mu = 1$

for $k = 0, \ldots, K-1$. The maximum iteration numbers of OPNMF on two datasets are also set as 5000 and 10000, respectively. And all algorithms are initialized at the same point which is randomly generated by $U^0 = \text{rand(m,r)}$ and $V^0 = (U^0)^T \tilde{X}_N$.



(a) objective value - time (b) magnification of 5(a) (c) objective value - |data pass| (d) magnification of 5(c)

(e) violation - time (f) magnification of 5(e) (g) violation - |data pass| (h) magnification of 5(g)

(i) KKT residual - time (j) magnification of 5(i) (k) KKT residual - |data pass| (l) magnification of 5(k)

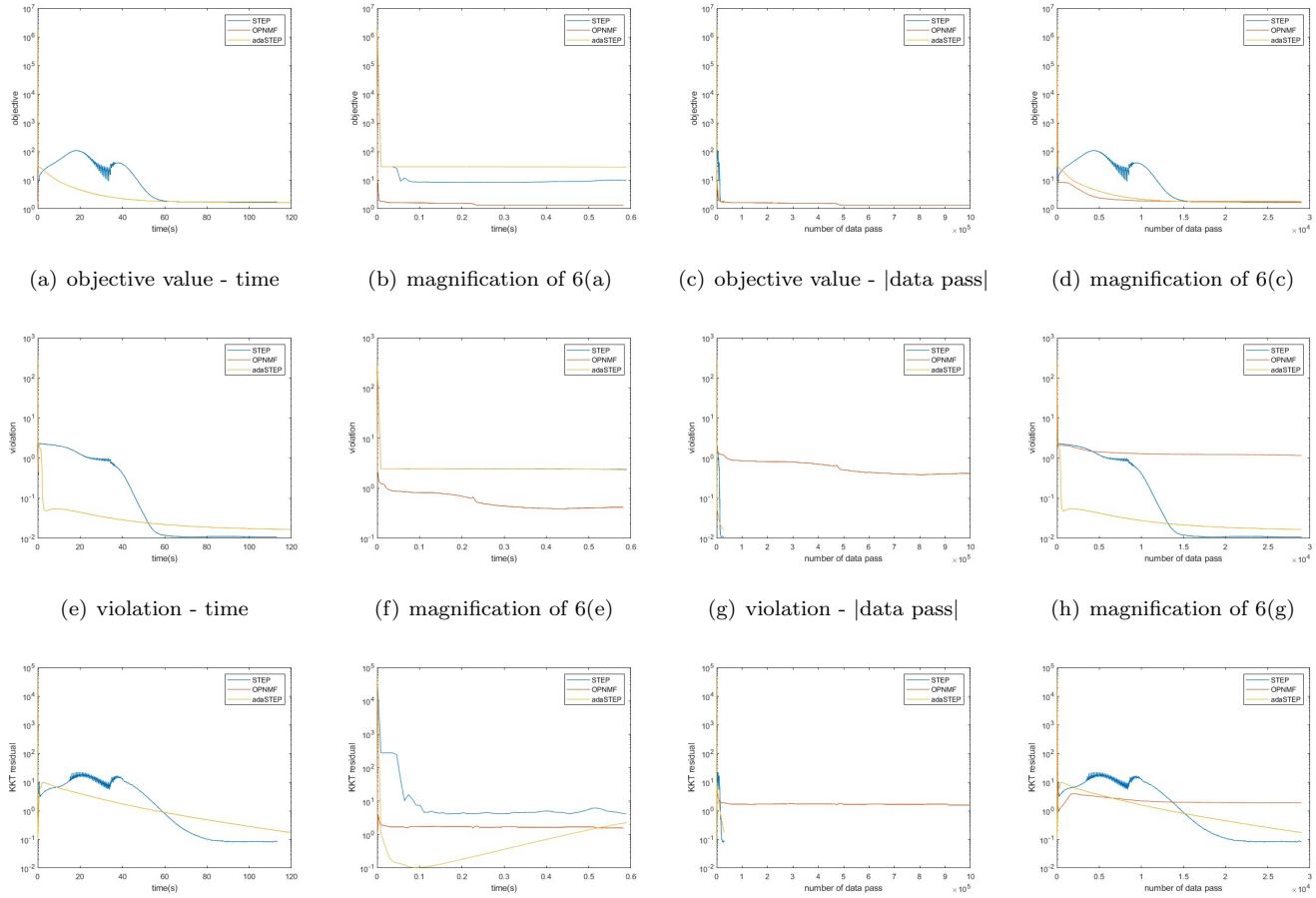
Figure 5: Comparison on Iris Dataset for solving (5.3)

The numerical test results are reported in Figures 5 and 6. The objective function value refers to $\|U^k V^k - \tilde{X}_N\|_F^2$, the constraint violation refers to $\|(U^k)^T U^k - I_r\|_F$, and the KKT residual refers to

$$\min_{Z \in \mathbb{R}^{r\times r}} \left(\mathbf{d}_F^2 \left(\nabla_U \Gamma(U^k, V^k) + 2U^k Z + \mathcal{N}_{\mathbb{R}_+^{m\times r}}(U^k), \mathbf{0} \right) + \mathbf{d}_F^2 \left(\nabla_V \Gamma(U^k, V^k) + \mathcal{N}_{\mathbb{R}_+^{r\times n}}(V^k), \mathbf{0} \right) \right)^{\frac{1}{2}},$$

where $\mathbf{d}_F^2(\mathcal{S}, \mathbf{0}) := \min_{U\in\mathcal{S}} \|U\|_F^2$ for any matrix set $\mathcal{S}$, and $V^k := (U^k)^T \tilde{X}_N$ for OPNMF. As can be observed from Figures 5 and 6, STEP performs better w.r.t. the KKT residual although due to the simple matrix operations the

multiplicative method OPNMF is faster. Moreover, the STEP method achieves better approximate solution within the same number of data passes. Moreover, compared with the STEP method, its adaptive variant converges to the similar level with stabler performance in whole process and faster decrease in early iterations, while it costs more time on adjusting stepsizes adaptively. Actually, it is the adaptiveness in D_k that allows Algorithm 4.1 converges with larger α_k: we find that the performance of Algorithm 4.1 with $\mu = 0$ is inferior to the adaSTEP in Figures 5 and 6 under the same α_k.



(a) objective value - time (b) magnification of 6(a) (c) objective value - |data pass| (d) magnification of 6(c)

(e) violation - time (f) magnification of 6(e) (g) violation - |data pass| (h) magnification of 6(g)

(i) KKT residual - time (j) magnification of 6(i) (k) KKT residual - |data pass| (l) magnification of 6(k)

Figure 6: Comparison on tr23 Dataset for solving (5.3)

6 Conclusion

In this paper we propose a stochastic nested primal-dual (STEP) method for nonconvex constrained composition optimization, whose objective involves a nested 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. Additionally, we propose and analyze a two-stage STEP algorithm with the first stage used to find a nearly feasible point. We also propose an adaptive STEP algorithm, allowing parameters updated adaptively, and establish

its iteration and oracle complexity accordingly. Finally, we conduct some numerical experiments on risk-averse portfolio optimization and orthogonal nonnegative matrix decomposition. The results reveal promising numerical performances of proposed algorithms.

Acknowledgements

We would like to thank Prof. Xiantao Xiao for his valuable comments on an earlier version of this paper. We also would like to thank two anonymous reviewers for their insightful comments and suggestions that have greatly improved this paper.

References

[1] 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.

[2] I. E. Bardakci, A. Jalilzadeh, C. Lagoa, and U. V. Shanbhag. Probability maximization via minkowski functionals: convex representations and tractable resolution. *Math. Program.*, 199(1):595–637, May 2023.

[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.*, 197:215–279, 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] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.*, 67(20):5239–5269, 2019.

[9] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, Jun 2023. DOI: https://doi.org/10.1007/s10107-023-01981-1.

[10] 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.

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

[12] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 126–135. Association for Computing Machinery, 2006.

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

[14] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842. PMLR, 2015.

[15] 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.

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

[17] 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.

[18] L. Jin and X. Wang. A stochastic primal-dual algorithm for a class of nonconvex constrained optimization. *Comput. Optim. Appl.*, 83:143–180, 2022.

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

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

[21] Q. Li, Z. Zhu, G. Tang, and M. B. Wakin. The geometry of equality-constrained global consensus problems. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7928–7932, 2019.

[22] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In *24th AISTATS*, volume 130 of *PMLR*, pages 2170–2178. PMLR, 13–15 Apr 2021.

[23] 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.

[24] Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Comput. Optim. Appl.*, 82:175–224, 2022.

[25] 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.

[26] H. Markowitz. Portfolio selection. *J. Finance*, 7(1):77–91, 1952.

[27] 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.

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

[29] J. Pan and M. K. Ng. Orthogonal nonnegative matrix factorization by sparsity and nuclear norm optimization. *SIAM J. Matrix Anal. Appl.*, 39(2):856–875, 2018.

[30] J. Pan, M. K. Ng, Y. Liu, X. Zhang, and H. Yan. Orthogonal nonnegative tucker decomposition. *SIAM J. Sci. Comput.*, 43(1):B55–B81, 2021.

[31] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui. Catalyst acceleration for gradient-based non-convex optimization. *arXiv:1703.10993*, 2018.

[32] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2014.

[33] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur. Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. *Neurocomputing*, 141:15–25, 2014.

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

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

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

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

[38] M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. *NeurIPS*, 2019.

[39] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *optimization-online.org*, 2022.

[40] R. Stubbs and D. Vandenbussche. Constraint attribution. *J. Portfolio Manage.*, 36:48–59, 2010.

[41] 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.

[42] 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.

[43] 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.

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

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

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

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

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

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

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

[51] 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.

[52] S. Yang, X. Li, and G. Lan. Data-driven minimax optimization with expectation constraints. *Operations Research*, 2024.

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

[54] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Trans. Neural Netw.*, 21(5):734–749, 2010.

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

[56] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. *Dive into Deep Learning.* Cambridge University Press, 2023.

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

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

[59] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Math. Program.*, 195:649–691, 2022.

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

Appendix

A. Application examples

We next present several application examples that can be formulated into (1.1).

Chance constrained optimization. Shanbhag et al. [2] consider a type of chance constrained optimization problems

$$\max_{x\in X} \Gamma(x) := \mathbb{P}\{\zeta \in \mathcal{K} \mid c(x;\zeta) \geq \mathbf{0}\}, \tag{6.1}$$

where $X \subseteq \mathbb{R}^n$ and $\mathcal{K} \subseteq \mathbb{R}^d$ are convex and closed, $\zeta : \Xi \to \mathbb{R}^d$ is a d-dimensional random vector following a prescribed probability distribution, and $c : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$. Problems in the form of (6.1) are common in robust portfolio selection problems and set covering problems. It is proved in [2] that under certain conditions (6.1) is equivalent to minimizing $f(\mathbb{E}_{\tilde{p}}[H(x;\xi)])$ over X with f being smooth and $H(\cdot;\xi)$ being nonconvex and possibly nonsmooth. Furthermore, under certain conditions they extend the results for $c(x;\zeta) = 1 - |\zeta^\top x|^2$ to $c(x;\zeta) = 1 - |\zeta^\top g(x)|^2$ by adding auxiliary variables obtaining

$$\max_{x\in X} \mathbb{P}\{\zeta \in \mathcal{K} \mid 1 - |\zeta^\top g(x)|^2 \geq \mathbf{0}\} \Leftrightarrow \min f(\mathbb{E}_{\tilde{p}}[H(y;\xi)]), \text{ s.t. } y = g(x), x \in X.$$

For general chance constrained optimization problem

$$\min_{x\in X} \Gamma(x), \text{ s.t. } \mathbb{P}\{\zeta \in \mathcal{K} \mid c(x;\zeta) \geq \mathbf{0}\} \geq p, \tag{6.2}$$

the results in [2] can also be applied to analyze the equivalence between (6.2) and its expectation-valued counterpart

$$\min_{x\in X} \Gamma(x), \text{ s.t. } f(\mathbb{E}_{\tilde{p}}[H(y;\xi)]) \leq \theta(p),$$

where f is smooth and decreasing on $\mathbb{R}_{++}$.

Risk-averse portfolio optimization. A class of risk-averse portfolio optimization problems, given by

$$\min_{x\in\mathbb{R}^n} \Gamma(x) := -\mathbb{E}_\phi[r(x;\phi)] + \lambda \operatorname{Var}[r(x;\phi)], \tag{6.3}$$

has been commonly used to measure the performance of algorithms for (un)constrained stochastic composition optimization [7,55,58], where $\Gamma(x) = f(h(x))$ with $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(y) = -y_1 + \lambda y_2 - \lambda y_1^2$ and $h : \mathbb{R}^n \to \mathbb{R}^2$ defined by $h(x) = \mathbb{E}_\phi[H(x;\phi)]$ and $H(x;\phi) = [r_\phi(x), r_\phi^2(x)]^T$. In [7,55,58], various theoretical analyses have been conducted for different cases of the problem. More specifically, [7] considers the unconstrained case, [55] focuses on the linear constrained case, while [58] addresses the unconstrained case with a convex nonsmooth regularizer, which includes the case of convex set constraints. However, in all of these studies, the unconstrained problem (6.3) is tested numerically. In the formulation of problem (6.3), the variable x represents a portfolio, where the j-th entry corresponds to the amount of investment or the percentage of total investment allocated to asset j. The term $r(x;\phi)$ represents the random return under portfolio x, and $\operatorname{Var}[r(x;\phi)]$ is the variance of $r(x;\phi)$. The parameter $\lambda \in \mathbb{R}_{++}$ denotes the mean-variance trade-off. According to modern portfolio theory [26], it is natural to require $x \in \Delta^n := \{x \in \mathbb{R}^n_+ \mid \sum_{j=1}^n x_j = 1\}$. Furthermore, as indicated by [40], the (in)direct use of constraints has been commonly imposed by asset owners, regulators, risk managers, trading desks, and the portfolio managers in quantitative portfolio management as follows:

$$\min_{x\in\Delta^n} \Gamma(x) := -\mathbb{E}_\phi[r(x;\phi)] + \lambda \operatorname{Var}[r(x;\phi)], \text{ s.t. } g(x) \leq \mathbf{0}.$$

Here x is the decision variable with each component x_j representing the percentage of the total investment allocated to asset j, $j = 1,\ldots,n$. And $g : \mathbb{R}^n \to \mathbb{R}^m$ includes various constraints, such as bounds on individual asset positions ($Ax \leq b$), controls on volatility ($\operatorname{Var}[r(x;\phi)] \leq \theta$) and controls on the extent of re-balancing ($c^\top|x - x_0| \leq b$ with $|\cdot|$ applied componentwisely). The functional constraints may also be induced by group sparsity requirements. Consider

$$\min_{x\in\Delta^n} \Gamma(x) := -\mathbb{E}_\phi[r(x;\phi)] + \lambda \operatorname{Var}[r(x;\phi)] + \mu \sum_{k=1}^N \|x_k\|,$$

where $\mu > 0$, $x = (x_1, \ldots, x_N)$ with $x_k \in \mathbb{R}^{n_k}$, $\sum_{k=1}^N n_k = n$. By adding a auxiliary variable $y \in \mathbb{R}^N$, we obtain an equivalent problem in the form of (1.1):

$$\min_{x\in\Delta^n, y\in\mathbb{R}^N} \Gamma(x) := -\mathbb{E}_\phi[r(x;\phi)] + \lambda \operatorname{Var}[r(x;\phi)] + \mu\|y\|_1 \text{ s.t. } \|x_k\|^2 = y_k^2, k \in [N].$$

Low-rank matrix estimation. To address the low-rank decomposition of an unknown matrix $\bar{X}$ based on unbiased samples X (i.e., $\mathbb{E}[X] = \bar{X}$), Ghadimi et al. [15] propose a formulation for low-rank matrix estimation problems as follows:

$$\min_{(U,V)\in S} f\left(\mathbb{E}[X] - UV^T\right), \tag{6.4}$$

where $S \subseteq \mathbb{R}^{n\times k} \times \mathbb{R}^{n\times k}$ with $n \gg k$, and $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a nonlinear loss function. This class of problems arises in matrix completion and robust principal component analysis with random sampling [8]. With slight modifications on objective function, [8] and [21] consider (6.4) with different S. Specifically, in [8], S is set as the Cartesian product of $\mathbb{R}^{n\times k}$ and a collection of some sparse matrix such as $S_\alpha := \{U \in \mathbb{R}^{m_1\times m_2} \mid \|S_{i,\cdot}\| \le \alpha m_2, \|S_{\cdot,j}\| \le \alpha m_1, \forall i, j\}$ with some positive integers m_1, m_2 and $\alpha \in (0, 1]$. In [21], it proposes the distributed low-rank decomposition by adding equality constraints $U_i = U$ for $i \in [N]$ and separating $(\mathbb{E}(X), V^T)$ in column. Additionally, in certain applications such as independent component analysis, clustering and deep learning, including recommendation systems, it may be necessary to enforce $U = [u_1, \ldots, u_k]$ with $u_i^\top u_j = 0$ for all $i \neq j$ and $\|u_i\| = 1$ for all i as well as $U \ge 0$ and $V \ge 0$. These requirements lead to a model in the form of (1.1), as indicated by Ge et al. [14] and Pan et al. [30].

Constrained minimax optimization. In [52], Yang et al. consider a minimax optimization problem with expectation constraints

$$\begin{aligned}\min_{x\in X}\max_{y\in Y}\ & \mathbb{E}_w[c(x,y;w)],\\ \text{s.t. }& \mathbb{E}_{\xi_i}[h_i(x;\xi_i)] \le 0,\ i = 1,2,\ldots,m_1,\\ & \mathbb{E}_{\zeta_j}[g_j(y;\zeta_j)] \le 0,\ j = 1,2,\ldots,m_2.\end{aligned}$$

If $N := |Y \cap \{y \mid \mathbb{E}_{\zeta_j}[g_j(y;\zeta_j)] \le 0,\ j = 1,2,\ldots,m_2\}| < \infty$, by defining

$$H(x;w) := \begin{pmatrix} c(x,y_1;w)\\ \vdots \\ c(x,y_N;w)\end{pmatrix} \text{ with } \{y_1,\ldots,y_N\} = Y \cap \{y \mid \mathbb{E}_{\zeta_j}[g_j(y;\zeta_j)] \le 0,\ j = 1,2,\ldots,m_2\},$$
$$h(x) := \mathbb{E}_w[H(x;w)],\ f(u) := \max\{u_1,\ldots,u_N\},$$

we can transform the original minimax problem into the form of (1.1).

Recurrent neural network (RNN) with perplexity loss function and regularity requirements. In the field of natural language processing (NLP), perplexity is commonly used to measure the performance of an NLP model [56]. Perplexity is defined as the exponential of the averaged cross-entropy loss. Given a set of data $\{x_t^n, y_t^n, t \in [T], n \in [N]\}$, the training process aims for finding the best model parameter θ^* that minimizes $\exp[\frac{1}{NT}\sum_{n=1}^N\sum_{t=1}^T\langle -y_t^n, \log(r(x_t^n,\theta))\rangle]$. Here $r(x_t^n,\theta)$ refers to the output of the model with parameter θ given the input x_t^n, and y_t^n is the given label under one-hot encoding. If the model is based on the classical Elman RNN and certain regularity requirements are imposed on the model parameters, an optimization problem in the form of (1.1) can be obtained:

$$\begin{aligned}\min_{\theta:=(W,U,V,b,c)}\ & \exp\left(\frac{1}{NT}\sum_{n=1}^N\sum_{t=1}^T\langle -y_t^n, \log(r_t^n)\rangle\right) + \Lambda(\theta),\\ \text{s.t.}\ & r_t^n = Vh_t^n + c,\ h_t^n = \sigma(Wh_{t-1}^n + Ux_t^n + b),\ t \in [T], n \in [N],\end{aligned}$$

where $h_0^n := \mathbf{0}$ for all $n \in [N]$.

B. Auxiliary lemmas

The auxiliary lemmas presented in this appendix are used in Section 4. Let $\{x^k\}$ and $\{z^k\}$ be generated by Algorithm 4.1.

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

$$\left|z_i^{k+1} - z_i^k\right| \leq \frac{\rho G}{K}\left(\frac{2\rho}{\beta_k} + 1\right), \quad \forall i \in [m],$$
$$\|\nabla_x \Psi_{\beta_k}\left(x, z^k\right) - \nabla_x \Psi_{\beta_k}\left(x^k, z^k\right)\| \leq L_{\beta_k}\|x - x^k\|, \quad \forall x \in \mathbb{R}^n,$$

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

It is straightforward to obtain Lemma B.2 following a similar analysis to Lemma 3.5.

LEMMA **B.2.** *Under Assumptions 3.1-3.3, it holds that for any $k = 0, \ldots, K$,*

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

where σ_{Γ_k} is defined in Lemma 3.4.

LEMMA **B.3.** *Under Assumptions 3.1-3.3, it holds that for any $k = 0, \ldots, K$,*

$$\begin{aligned}
&\left(\frac{1}{2\alpha_k} - \frac{L_\Gamma + L_{\beta_k}}{2}\right)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
\leq & \mathbb{E}\left[\mathcal{L}_{\beta_k}\left(x^k, z^k\right) - \mathcal{L}_{\beta_{k+1}}\left(x^{k+1}, z^{k+1}\right)\right] + \mathbb{E}\left[\sum_{i=1}^m \max\left(G, \frac{z_i^k}{\beta}, \frac{z_i^{k+1}}{\beta}\right)\left|z_i^k - z_i^{k+1}\right|\right] \\
& + \alpha_k L_f^2 \mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2\right] + \alpha_k \sigma_{\Gamma_k}^2 + mG^2 \max\left(\frac{1}{2}, \frac{\rho^2}{\beta_0^2}, 2\rho^2\right)\left[\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}} + \beta_{k+1} - \beta_k\right]. \qquad (6.5)
\end{aligned}$$

Proof. Similar to Lemma 3.6, we can obtain

$$\begin{aligned}
&\left(\frac{1}{2\alpha_k} - \frac{L_\Gamma + L_{\beta_k}}{2}\right)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
\leq & \mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2_{\mathrm{diag}(s^k)}\right] + \left(\frac{1}{2\alpha_k} - \frac{L_\Gamma + L_{\beta_k}}{2}\right)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
\leq & \mathbb{E}\left[\mathcal{L}_{\beta_k}\left(x^k, z^k\right) - \mathcal{L}_{\beta_k}\left(x^{k+1}, z^{k+1}\right)\right] + \mathbb{E}\left[\sum_{i=1}^m \max\left(G, \frac{z_i^k}{\beta}, \frac{z_i^{k+1}}{\beta}\right)\left|z_i^k - z_i^{k+1}\right|\right] \\
& + \alpha_k L_f^2 \mathbb{E}\left[\left\|y^{k+1} - h\left(x^k\right)\right\|^2\right] + \alpha_k \sigma_{\Gamma_k}^2. \qquad (6.6)
\end{aligned}$$

We now focus on the term $\mathcal{L}_{\beta_{k+1}}\left(x^{k+1}, z^{k+1}\right) - \mathcal{L}_{\beta_k}\left(x^{k+1}, z^{k+1}\right)$. Note

$$\mathcal{L}_{\beta_{k+1}}\left(x^{k+1}, z^{k+1}\right) - \mathcal{L}_{\beta_k}\left(x^{k+1}, z^{k+1}\right) = \sum_{i=1}^m \left[\psi_{\beta_{k+1}}\left(g_i\left(x^{k+1}\right), z_i^{k+1}\right) - \psi_{\beta_k}\left(g_i\left(x^{k+1}\right), z_i^{k+1}\right)\right]. \qquad (6.7)$$

It can be derived from $\beta_{k+1} > \beta_k > 0$ and the definition of $\psi_\beta(u, v)$ in (2.3) that for any $u \in (-\infty, G]$ with $G > 0$ and $v \in [0, 2G\rho]$:

(i) if $\beta_k u + v \geq 0$, then

$$\psi_{\beta_{k+1}}(u, v) - \psi_{\beta_k}(u, v) \leq \left(uv + \frac{\beta_{k+1}}{2}u^2\right) - \left(uv + \frac{\beta_k}{2}u^2\right) = \frac{u^2}{2}(\beta_{k+1} - \beta_k) \leq G^2 \max\left(\frac{1}{2}, \frac{2\rho^2}{\beta_0^2}\right)(\beta_{k+1} - \beta_k),$$

where the last inequality comes from $u \in [-v/\beta_k, G] \subseteq [-2G\rho/\beta_0, G]$;

(ii) if $\beta_k u + v < 0$, then $u < 0$ and $\beta_{k+1}u + v < \beta_k u + v < 0$, which implies that

$$\psi_{\beta_{k+1}}(u, v) - \psi_{\beta_k}(u, v) = \frac{v^2}{2}\left(\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}}\right) \leq 2G^2\rho^2\left(\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}}\right).$$

Hence for any $i \in [m]$ we have

$$\psi_{\beta_{k+1}}\left(g_i\left(x^{k+1}\right), z_i^{k+1}\right) - \psi_{\beta_k}\left(g_i\left(x^{k+1}\right), z_i^{k+1}\right) \leq G^2 \max\left(\frac{1}{2}, \frac{\rho^2}{\beta_0^2}, 2\rho^2\right)\left[\frac{1}{\beta_k} - \frac{1}{\beta_{k+1}} + \beta_{k+1} - \beta_k\right],$$

which, together with (6.6) and (6.7), yields the result. □

C. Proofs of theorems

C1. Proof of Theorem 4.1

Proof. It follows from Lemma B.2 that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right)+\partial\Lambda\left(x^{R+1}\right)+\sum_{i=1}^{m}\left[\beta_R g_i\left(x^{R+1}\right)+z_i^{R+1}\right]_+\nabla g_i\left(x^{R+1}\right)+\mathcal{N}_X\left(x^{R+1}\right),\mathbf{0}\right)\right]\\
\leq&\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left(L_\Gamma+L_{\beta_k}+\frac{1}{\alpha_k}+\|s^k\|\right)^2\left\|x^{k+1}-x^k\right\|^2+2L_f^2\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
&+\frac{3L_{g,0}^2m}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|^2\right]+\frac{6}{K}\sum_{k=1}^{K}\sigma_{\Gamma_k}^2.
\end{aligned}\tag{6.8}$$

Moreover, through a similar analysis to (3.24) and (3.25) we derive

$$\frac{3L_{g,0}^2m}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|^2\right]+\frac{6}{K}\sum_{k=1}^{K}\sigma_{\Gamma_k}^2=\mathcal{O}\left(K^{-1/2}\right).$$

Hence, to estimate the upper bound of (6.8) we only need to analyze the first term in the right hand side.

Multiplying (3.19) by $(1+\alpha_kL_f^2)$ and plugging it into (6.5), we obtain from Lemma 3.2 that for any $k\in[K]$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{2\alpha_k}-\frac{L_\Gamma+L_{\beta_k}}{2}\right)\left\|x^{k+1}-x^k\right\|^2+\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\mathbb{E}\left[\mathcal{L}_{\beta_k}\left(x^k,z^k\right)-\mathcal{L}_{\beta_{k+1}}\left(x^{k+1},z^{k+1}\right)\right]+G\max\left(1,\frac{2\rho}{\beta_k}\right)\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]+\alpha_k\sigma_{\Gamma_k}^2\\
&+\left(1+\alpha_kL_f^2\right)\left(1+\gamma_k\right)\left(1-\eta_k\right)^2\mathbb{E}\left[\left\|y^k-h\left(x^{k-1}\right)\right\|^2\right]+\left(1+\alpha_kL_f^2\right)\left(1+\gamma_k^{-1}\right)\left(1-\eta_k\right)^2L_{h,0}^2\mathbb{E}\left[\left\|x^k-x^{k-1}\right\|^2\right]\\
&+\left(1+\alpha_kL_f^2\right)\left(1+\gamma_k\right)\eta_k^2\frac{\sigma_{h,0}^2}{P_1}+mG^2\max\left(\frac{1}{2},\frac{\rho^2}{\beta_0^2},2\rho^2\right)\left[\frac{1}{\beta_k}-\frac{1}{\beta_{k+1}}+\beta_{k+1}-\beta_k\right].
\end{aligned}$$

Summing up the above inequality over $k=1,\ldots,K$ implies

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{2\alpha_K}-\frac{L_\Gamma+L_{\beta_K}}{2}\right)\left\|x^{K+1}-x^K\right\|^2+\left\|y^{K+1}-h\left(x^K\right)\right\|^2\right]\\
&+\sum_{k=1}^{K-1}\mathbb{E}\left[\left(\frac{1}{2\alpha_k}-\frac{L_\Gamma+L_{\beta_k}}{2}-\left(1+\alpha_{k+1}L_f^2\right)\left(1+\gamma_{k+1}^{-1}\right)\left(1-\eta_{k+1}\right)^2L_{h,0}^2\right)\left\|x^{k+1}-x^k\right\|^2\right]\\
&+\sum_{k=1}^{K-1}\mathbb{E}\left[\left(1-\left(1+\alpha_{k+1}L_f^2\right)\left(1+\gamma_{k+1}\right)\left(1-\eta_{k+1}\right)^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\mathbb{E}\left[\mathcal{L}_{\beta_1}\left(x^1,z^1\right)-\mathcal{L}_{\beta_{K+1}}\left(x^{K+1},z^{K+1}\right)\right]+G\max\left(1,\frac{2\rho}{\beta_0}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]+\sigma_\Gamma^2\sum_{k=1}^{K}\alpha_k\\
&+\left(1+\alpha_1L_f^2\right)\left(1+\gamma_1\right)\left(1-\eta_1\right)^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]+\left(1+\alpha_1L_f^2\right)\left(1+\gamma_1^{-1}\right)\left(1-\eta_1\right)^2L_{h,0}^2\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]\\
&+\sum_{k=1}^{K}\left(1+\alpha_kL_f^2\right)\left(1+\gamma_k\right)\eta_k^2\frac{\sigma_{h,0}^2}{P_1}+mG^2\max\left(\frac{1}{2},\frac{\rho^2}{\beta_0^2},2\rho^2\right)\left[\frac{1}{\beta_1}-\frac{1}{\beta_{K+1}}+\beta_{K+1}-\beta_1\right].
\end{aligned}\tag{6.9}$$

Under parameter settings in (4.3) and the definition of L_{β_k}, we can deduce $\alpha_k\leq\bar{\alpha}_1$ for any $k\geq0$, which together with $L_f=L_{h,0}L_{f,1}$ defined in Lemma 3.1 implies that $0<\eta_k\leq1$ for any $k\geq0$, and for any $k\geq-1$,

$$\begin{aligned}
\left(1+\alpha_{k+1}L_f^2\right)\left(1+\gamma_{k+1}\right)\left(1-\eta_{k+1}\right)^2-1\leq&\left(\left(1+\frac{\eta_{k+1}}{2}\right)\left(1-\eta_{k+1}\right)\right)^2-1\leq1-\frac{\eta_{k+1}}{2}-\frac{\eta_{k+1}^2}{2}-1\\
&\leq-\frac{\eta_{k+1}}{2}=-\alpha_{k+1}\max\left(L_f^2,8L_{h,0}^2\right)
\end{aligned}\tag{6.10}$$

and

$$0 \leq \left(1+\alpha_{k+1}L_f^2\right)\left(1+\gamma_{k+1}^{-1}\right)\left(1-\eta_{k+1}\right)^2 L_{h,0}^2 \leq \left(1+\alpha_{k+1}L_f^2\right)\left(1+\gamma_{k+1}^{-1}\right) L_{h,0}^2$$
$$\leq \left(1+\frac{L_{f,1}^2}{8}+L_f^2\bar{\alpha}_1\right)L_{h,0}^2+\frac{1}{8\alpha_{k+1}}$$
$$\leq \left(\frac{3}{2}+\frac{L_{f,1}^2}{8}\right)L_{h,0}^2+\frac{1}{8\alpha_{k+1}} \leq \frac{1}{4\alpha_{k+1}}-\frac{L_\Gamma+L_{\beta_{k+1}}}{4}. \tag{6.11}$$

At the same time, it follows from $\alpha_k^{-1}-L_\Gamma-L_{\beta_k} = L_\Gamma+L_{\beta_k}+(12+L_{f,1}^2)L_{h,0}^2 = (L_\Gamma+2mG\rho L_{g,1}+(12+L_{f,1}^2)L_{h,0}^2)+(mL_{g,0}^2+mGL_{g,1})\beta_0(k+1)^{1/4}$ that for any $k\geq 1$,

$$\frac{\alpha_{k+1}^{-1}-L_\Gamma-L_{\beta_{k+1}}}{\alpha_k^{-1}-L_\Gamma-L_{\beta_k}} \leq \left(\frac{k+2}{k+1}\right)^{1/4} \leq \frac{6}{5}, \tag{6.12}$$

which together with (6.11) implies that

$$\left(1+\alpha_{k+1}L_f^2\right)\left(1+\gamma_{k+1}^{-1}\right)\left(1-\eta_{k+1}\right)^2 L_{h,0}^2-\left(\frac{1}{2\alpha_k}-\frac{L_\Gamma+L_{\beta_k}}{2}\right) \leq -\frac{1}{6}\left(\frac{1}{\alpha_{k+1}}-(L_\Gamma+L_{\beta_{k+1}})\right) \tag{6.13}$$

Applying (6.10) and (6.12)-(6.13) to relax the coefficients of $\left\|x^{k+1}-x^k\right\|^2$ and $\left\|y^{k+1}-h\left(x^k\right)\right\|^2$ on the left side of (6.9), we obtain

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\left[\left(\frac{1}{6\alpha_{k+1}}-\frac{L_\Gamma+L_{\beta_{k+1}}}{6}\right)\left\|x^{k+1}-x^k\right\|^2+\alpha_{k+1}\max\left(L_f^2,8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\mathbb{E}\left[\mathcal{L}_{\beta_1}\left(x^1,z^1\right)-\mathcal{L}_{\beta_{K+1}}\left(x^{K+1},z^{K+1}\right)\right]+\left(1+\alpha_1L_f^2\right)\left(1+\gamma_1^{-1}\right)\left(1-\eta_1\right)^2L_{h,0}^2\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]\\
&+\left(1+\alpha_1L_f^2\right)\left(1+\gamma_1\right)\left(1-\eta_1\right)^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]+G\left(1+\frac{2\rho}{\beta_0}\right)\sum_{k=1}^K\mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_1\right]\\
&+\sum_{k=1}^K\left(1+\alpha_kL_f^2\right)\left(1+\gamma_k\right)\eta_k^2\frac{\sigma_{h,0}^2}{P_1}+\sigma_\Gamma^2\sum_{k=1}^K\alpha_k+mG^2\max\left(\frac{1}{2},\frac{\rho^2}{\beta_0^2},2\rho^2\right)\left[\frac{1}{\beta_1}-\frac{1}{\beta_{K+1}}+\beta_{K+1}-\beta_1\right].
\end{aligned} \tag{6.14}$$

Similar to (3.30), we can attain

$$\mathcal{L}_{\beta_{K+1}}\left(x^{K+1},z^{K+1}\right) \geq C^*-\frac{2mG^2\rho^2}{\beta_0}. \tag{6.15}$$

And for $(1+\alpha_1L_f^2)(1+\gamma_1^{-1})(1-\eta_1)^2L_{h,0}^2\mathbb{E}[\|x^1-x^0\|^2]+\mathbb{E}[\mathcal{L}_\beta(x^1,z^1)]$, it follows from Lemmas 3.2 and B.3 that

$$\begin{aligned}
\left(\frac{1}{2\alpha_0}-\frac{L_\Gamma+L_{\beta_0}}{2}\right)\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right] \leq&\mathbb{E}\left[\mathcal{L}_{\beta_0}\left(x^0,z^0\right)-\mathcal{L}_{\beta_1}\left(x^1,z^1\right)\right]+\alpha_0\sigma_\Gamma^2+\alpha_0L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]\\
&+G\left(1+\frac{2\rho}{\beta_0}\right)\mathbb{E}\left[\left\|z^1-z^0\right\|_1\right]+mG^2\max\left(\frac{1}{2},\frac{\rho^2}{\beta_0^2},2\rho^2\right)\left[\frac{1}{\beta_0}-\frac{1}{\beta_1}+\beta_1-\beta_0\right].
\end{aligned}$$

Then it together with (6.11) and (6.12) yields

$$\begin{aligned}
&\left(1+\alpha_1L_f^2\right)\left(1+\gamma_1^{-1}\right)\left(1-\eta_1\right)^2L_{h,0}^2\mathbb{E}\left[\left\|x^1-x^0\right\|^2\right]+\mathbb{E}\left[\mathcal{L}_{\beta_1}\left(x^1,z^1\right)\right]\\
\leq&\mathcal{L}_{\beta_0}\left(x^0,z^0\right)+\alpha_0\sigma_\Gamma^2+G\left(1+\frac{2\rho}{\beta_0}\right)\mathbb{E}\left[\left\|z^1-z^0\right\|_1\right]+\alpha_0L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]\\
&+mG^2\max\left(\frac{1}{2},\frac{\rho^2}{\beta_0^2},2\rho^2\right)\left[\frac{1}{\beta_0}-\frac{1}{\beta_1}+\beta_1-\beta_0\right]\\
\leq&\Gamma\left(x^0\right)+\Lambda\left(x^0\right)+\frac{\beta_0}{2}G^2m+\alpha_0\sigma_\Gamma^2+2G^2m\rho\left(1+\frac{2\rho}{\beta_0}\right)+\alpha_0L_f^2\mathbb{E}\left[\left\|y^1-h\left(x^0\right)\right\|^2\right]
\end{aligned}$$

$$+ mG^2 \max\left(\frac{1}{2}, \frac{\rho^2}{\beta_0^2}, 2\rho^2\right)\left[\frac{1}{\beta_0} - \frac{1}{\beta_1} + \beta_1 - \beta_0\right], \tag{6.16}$$

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

$$\mathcal{L}_{\beta_0}\left(x^0, z^0\right) = \Gamma\left(x^0\right) + \Lambda\left(x^0\right) + \frac{\beta_0}{2}\sum_{i=1}^{m}\left(\left[g_i\left(x^0\right)\right]_+\right)^2 \leq \Gamma\left(x^0\right) + \Lambda\left(x^0\right) + \frac{\beta_0}{2}G^2 m. \tag{6.17}$$

Hence, plugging (6.15) and (6.16) into (6.14) we obtain

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{1}{6\alpha_{k+1}} - \frac{L_\Gamma + L_{\beta_{k+1}}}{6}\right)\left\|x^{k+1} - x^k\right\|^2 + \alpha_{k+1}\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) + \Lambda\left(x^0\right) - C^* + \frac{2G^2\rho^2 m}{\beta_0} + \frac{\beta_0}{2}G^2 m + \sigma_\Gamma^2\sum_{k=0}^{K}\alpha_k + 2G^2 m\rho\left(1 + \frac{2\rho}{\beta_0}\right) + \sum_{k=1}^{K}\left(1 + \alpha_k L_f^2\right)\left(1 + \gamma_k\right)\eta_k^2\frac{\sigma_{h,0}^2}{P_1} \\
&+ \left[\left(1 + \alpha_1 L_f^2\right)\left(1 + \gamma_1\right)\left(1 - \eta_1\right)^2 + \alpha_0 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_0}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|_1\right] \\
&+ mG^2\max\left(\frac{1}{2}, \frac{\rho^2}{\beta_0^2}, 2\rho^2\right)\left[\frac{1}{\beta_0} - \frac{1}{\beta_{K+1}} + \beta_{K+1} - \beta_0\right] \\
\leq& \Gamma\left(x^0\right) + \Lambda\left(x^0\right) - C^* + \frac{2G^2\rho^2 m}{\beta_0} + \frac{\beta_0}{2}G^2 m + \sigma_\Gamma^2\sum_{k=0}^{K}\alpha_k + 2G^2 m\rho\left(1 + \frac{2\rho}{\beta_0}\right) + \sum_{k=1}^{K}\left(1 + \alpha_k L_f^2\right)\left(1 + \gamma_k\right)\eta_k^2\frac{\sigma_{h,0}^2}{P_1} \\
&+ 2\mathbb{E}\left[\left\|y^0 - h\left(x^0\right)\right\|^2\right] + 2\eta_0^2\frac{\sigma_{h,0}^2}{P_1} + G\left(1 + \frac{2\rho}{\beta_0}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|_1\right] \\
&+ mG^2\max\left(\frac{1}{2}, \frac{\rho^2}{\beta_0^2}, 2\rho^2\right)\left[\frac{1}{\beta_0} - \frac{1}{\beta_{K+1}} + \beta_{K+1} - \beta_0\right] \\
\leq& \Gamma\left(x^0\right) + \Lambda\left(x^0\right) - C^* + \frac{2G^2\rho^2 m}{\beta_0} + \frac{\beta_0}{2}G^2 m + \sigma_\Gamma^2\sum_{k=0}^{K}\alpha_k + 2G^2 m\rho\left(1 + \frac{2\rho}{\beta_0}\right) \\
&+ 2\left\|y^0 - h\left(x^0\right)\right\|^2 + 2\eta_0^2\frac{\sigma_{h,0}^2}{P_1} + G^2 m\rho\left(1 + \frac{2\rho}{\beta_0}\right)^2 + \sum_{k=1}^{K}3\eta_k^2\frac{\sigma_{h,0}^2}{P_1} \\
&+ mG^2\max\left(\frac{1}{2}, \frac{\rho^2}{\beta_0^2}, 2\rho^2\right)\left[\frac{1}{\beta_0} - \frac{1}{\beta_{K+1}} + \beta_{K+1} - \beta_0\right] \\
=& \mathcal{O}\left(1 + K^{3/4}\left(\frac{1}{P_2} + \frac{1}{J}\right) + \frac{K^{1/2}}{P_1} + K^{1/4}\right),
\end{aligned} \tag{6.18}$$

where the second inequality holds due to $0 < \eta_k \leq 1$, $(1 + \alpha_1 L_f^2)(1 + \gamma_1)(1 - \eta_1)^2 + \alpha_0 L_f^2 \leq 2$ from (6.10) and $\alpha_0 \leq \bar{\alpha}_1$, and

$$\begin{aligned}
\mathbb{E}\left[\left\|y^1 - h\left(x^0\right)\right\|^2\right] =& \mathbb{E}\left[\left\|(1 - \eta_0)\left(y^0 - h\left(x^0\right)\right) + \eta_0\left(\frac{1}{P_{0,1}}\sum_{\phi\in\mathcal{P}_{0,1}}H\left(x^0;\phi\right) - h\left(x^0\right)\right)\right\|^2\right] \\
&\leq (1 - \eta_0)^2\left\|y^0 - h\left(x^0\right)\right\|^2 + \eta_0^2\frac{\sigma_{h,0}^2}{P_1},
\end{aligned}$$

the third inequality follows from

$$\sum_{k=1}^{K}\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|_1\right] \leq \sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\frac{\rho G}{K}\left(\frac{2\rho}{\beta_0} + 1\right)\right] \leq Gm\rho\left(\frac{2\rho}{\beta_0} + 1\right)$$

by Lemma B.1, and $(1 + \alpha_k L_f^2)(1 + \gamma_k) \leq (3/2)^2 < 3$ from $\gamma_k = 8L_{h,0}^2\alpha_k, \alpha_k \leq \bar{\alpha}_1$, and the last equality follows from (3.23) and (4.3).

Therefore, (6.18) implies that

$$
\begin{aligned}
&\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left(L_\Gamma+L_{\beta_k}+\frac{1}{\alpha_k}+\|s^k\|\right)^2\left\|x^{k+1}-x^k\right\|^2+2L_f^2\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
\leq&\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\Bigg[\max\left(\frac{\left(L_\Gamma+L_{\beta_k}+\frac{1}{\alpha_k}+\|s^k\|\right)^2}{\frac{1}{6\alpha_{k+1}}-\frac{L_\Gamma+L_{\beta_{k+1}}}{6}},\frac{2}{\alpha_{k+1}}\right)\cdot\\
&\left(\left(\frac{1}{6\alpha_{k+1}}-\frac{L_\Gamma+L_{\beta_{k+1}}}{6}\right)\left\|x^{k+1}-x^k\right\|^2+\alpha_{k+1}\max\left(L_f^2,8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right)\Bigg]\\
\leq&\frac{9}{\alpha_{K+1}K}\left(3+\frac{\mu\sqrt{n}}{m\beta_0(L_{g,0}^2+GL_{g,1})}\right)^2\cdot\\
&\sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{1}{6\alpha_{k+1}}-\frac{L_\Gamma+L_{\beta_{k+1}}}{6}\right)\left\|x^{k+1}-x^k\right\|^2+\alpha_{k+1}\max\left(L_f^2,8L_{h,0}^2\right)\left\|y^{k+1}-h\left(x^k\right)\right\|^2\right]\\
=&\mathcal{O}\left(K^{-3/4}+\left(\frac{1}{P_2}+\frac{1}{J}\right)+\frac{1}{K^{1/4}P_1}+K^{-1/2}\right)=\mathcal{O}\left(K^{-\frac{1}{2}}\right),
\end{aligned}
\tag{6.19}
$$

where the second inequality follows from

$$
\|s^k\|\leq\mu\sqrt{n}(k+1)^{1/4}\leq\frac{\mu\sqrt{n}}{m\beta_0(L_{g,0}^2+GL_{g,1})}L_{\beta_k}\leq\frac{\mu\sqrt{n}}{m\beta_0(L_{g,0}^2+GL_{g,1})}(L_\Gamma+L_{\beta_{k+1}}),
$$

$L_\Gamma+L_{\beta_k}+\alpha_k^{-1}<L_\Gamma+L_{\beta_{k+1}}+\alpha_{k+1}^{-1}$, $L_\Gamma+L_{\beta_{k+1}}\leq(2\alpha_{k+1})^{-1}$ and $\alpha_{k+1}\geq\alpha_{K+1}$, and the last equality holds due to $P_2=J\geq\sqrt{K},P_1\geq K^{1/4}$. Then plugging (6.19) into (6.8) yields the result. □

C2. Proof of Theorem 4.2

Proof. Similar to Theorem 3.3, for any $k\geq1$ we have

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

Hence, it indicates from (4.4) and $\beta_k=\beta_0(k+1)^{1/4}$ that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left[g\left(x^{R+1}\right)\right]_+\right\|^2\right]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\left[g\left(x^{k+1}\right)\right]_+\right\|^2\right]\\
\leq&\sum_{k=1}^{K}\frac{4}{\nu^2\beta_k^2K}\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{k+1}\right)+\sum_{i=1}^{m}\left[\beta_kg_i\left(x^{k+1}\right)+z_i^{k+1}\right]_+\nabla g_i\left(x^{k+1}\right)+\partial\Lambda\left(x^{k+1}\right)+\mathcal{N}_X\left(x^{k+1}\right),\mathbf{0}\right)\right]\\
&+\sum_{k=1}^{K}\frac{4}{\nu^2\beta_k^2K}\left[4m^2L_{g,0}^2G^2\rho^2+L_{f,0}^2L_{h,0}^2+G_\Lambda^2\right]\\
\leq&\frac{4}{\nu^2\beta_0^2}\mathbb{E}\left[\mathbf{d}^2\left(\nabla\Gamma\left(x^{R+1}\right)+\sum_{i=1}^{m}\left[\beta_kg_i\left(x^{R+1}\right)+z_i^{R+1}\right]_+\nabla g_i\left(x^{R+1}\right)+\partial\Lambda\left(x^{R+1}\right)+\mathcal{N}_X\left(x^{R+1}\right),\mathbf{0}\right)\right]\\
&+\frac{4}{\nu^2\beta_0^2\sqrt{K}}\left[4m^2L_{g,0}^2G^2\rho^2+L_{f,0}^2L_{h,0}^2+G_\Lambda^2\right]\\
=&\mathcal{O}\left(K^{-\frac{1}{2}}+K^{-\frac{1}{2}}\right)=\mathcal{O}\left(K^{-\frac{1}{2}}\right),
\end{aligned}
$$

where the second inequality uses $\sum_{k=1}^{K}\frac{1}{\sqrt{k+1}}\leq2(\sqrt{K+1}-1)\leq2\sqrt{K}$. We thus obtain (4.5).

Furthermore, it is straightforward to derive (4.6) following a similar proof to Theorem 3.5. □