A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization*

Qiankun Shi† Xiao Wang‡ Hao Wang§

August 12, 2022

Abstract

Nonconvex constrained stochastic optimization has emerged in many important application areas. With general functional constraints it minimizes the sum of an expectation function and a convex nonsmooth regularizer. Main challenges arise due to the stochasticity in the random integrand and the possibly nonconvex functional constraints. To cope with these issues we propose a momentum-based linearized augmented Lagrangian method (MLALM) in this paper. A recursive momentum is incorporated to calculate the stochastic gradient and only one sample is taken at each iteration. Meanwhile, to ease the difficulties in keeping the feasibility to general constraints, based on stochastic gradients we build a stochastic approximation to the linearized augmented Lagrangian function to update primal variables, which are further used to update dual variables in a moving average way. Under a nonsingularity condition on constraints and with a nearly feasible initial point, we establish the $\mathcal{O}(\epsilon^{-4})$ oracle complexity of MLALM to find an ϵ-stationary point of the original problem. Numerical experiments on two types of test problems reveal promising performances of the proposed algorithm.

Keywords: Nonconvex optimization, functional constraint, augmented Lagrangian function, stochastic gradient, momentum, oracle complexity

Mathematics Subject Classification 2010: 90C30, 90C06, 65K05, 92C15

1 Introduction

In this paper, we consider the nonconvex constrained stochastic optimization problem

$$\begin{array}{ll} \min\limits_{x\in X} & \{f(x) \equiv \mathbb{E}_\xi[\mathbf{F}(x;\xi)]\} + \chi(x) \\ \text{s.t.} & c_i(x) = 0, \quad i \in \mathcal{E}, \\ & c_i(x) \le 0, \quad i \in \mathcal{I}, \end{array} \tag{1}$$

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

†shiqk@shanghaitech.edu.cn, Peng Cheng Laboratory, Shenzhen, 518066, China; School of Information Science and Technology ShanghaiTech University, Shanghai, 201210, China.

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

§wanghao1@shanghaitech.edu.cn, School of Information Science and Technology ShanghaiTech University, Shanghai, 201210, China.

where $X \subseteq \mathbb{R}^n$ is a closed convex set, ξ is a random variable in the probability space Ξ and independent of x. Here $\mathcal{E}$ and $\mathcal{I}$ are two finite sets of indices, $\mathbf{F} : \mathbb{R}^n \times \Xi \to \mathbb{R}$ and $c_i : \mathbb{R}^n \to \mathbb{R}, i \in \mathcal{E} \cup \mathcal{I}$, are continuously differentiable and possibly nonconvex, and $\chi : \mathbb{R}^n \to \mathbb{R}$ is a proper lower-semicontinuous and convex function. Without loss of generality we presume that both $\mathcal{E}$ and $\mathcal{I}$ are nonempty, the feasible set $\mathcal{X} := \{x \in X \mid c_i(x) = 0, i \in \mathcal{E}; c_i(x) \leq 0, i \in \mathcal{I}\}$ is nonempty and the objective function value of (1) over X is lower bounded. A key feature of problem (1) is that it is normally expensive to compute the expectation or the distribution of ξ cannot be expressed explicitly. Thus the exact function or gradient information of f can be hard to obtain. Those problems widely appear in many application fields. For example, in deep learning constraints are imposed on output of the deep neural networks [25], such as physics-constrained deep learning model [47], constraint-aware deep neural network compression [9], manifold regularized deep learning [30, 36]. Some recent study also highlights the advantages of incorporating various constraints when training deep neural networks [24, 28]. Other applications include, but not limited to, portfolio allocation [2, 35], two/multi-stage modeling [2, 35] and constrained maximum likelihood estimation [8, 15].

The past decade has witnessed great developments in nonconvex stochastic optimization. Since Ghadimi and Lan[16] proposed randomized SGD methods for unconstrained nonconvex optimization, a surge of works have emerged in this line of research. Due to stochastic variances of approximate gradients, SGD methods normally suffer slow convergence [5]. Types of variance reduction techniques have been proposed, trying to reduce the stochastic variance to accelerate SGD methods. Related works include SAG [32], SAGA [13], SVRG [19], SARAH [26] and SPIDER [14]. Moreover, proximal variants aiming for stochastic composite optimization have also been studied [17, 42, 34, 27, 40]. Among those methods, SAG- and SAGA-type methods have high space requirements to store historical gradients at each sample point, while SVRG-, SARAH- and SPIDER-type methods require to compute a (nearly) accurate gradient at a checkpoint from time to time which normally relies on large batch sizes. Recently, a stochastic recursive momentum method [12] attracts attention, in which only one sample is required to estimate the gradient at each iteration, and later a proximal variant is studied in [45] for nonconvex stochastic composite problems. Under the mean-squared smoothness condition, the above proximal algorithm can produce a stochastic ϵ-stationary point with the oracle complexity bounded by $\mathcal{O}(\epsilon^{-3})$, where the oracle complexity refers to the total number of stochastic gradient evaluations.

When it comes to nonconvex optimization with general functional constraints, challenges may arise since the feasibility to constraints can be hard to maintain. Nonconvex constrained optimization in deterministic settings has been studied for decades [41]. Penalty methods and sequential quadratic programming (SQP) methods are two types of most effective approaches for general constrained optimization. Penalty methods normally transform the original constrained problem to a sequence of unconstrained ones by penalizing the constraints into the objective in a term measuring the constraint violation. Among penalty methods, augmented Lagrangian (AL) methods attract much interest due to the fact that the AL function has more advantages in characterizing the optimality conditions for constrained problems and in designing effective algorithms. Nevertheless, classic penalty methods are normally double-loop algorithms in which a penalty function needs to be (approximately) minimized in the inner-loop. Single-loop penalty methods with much simpler subproblems, such as Sℓ_1QP [41], linearized AL methods [38, 39], are thus studied for constrained optimization. On the other hand, SQP methods try to compute search directions by solving a sequence of quadratic programming subproblems. Along with the developments of complexity theories, in the past ten years numerical methods for nonconvex constrained optimization with complexity analysis have been widely studied [6, 20, 21, 22, 31, 33, 44].

For general functional constrained optimization in stochastic settings, such as (1), main concerns lie in that the exact gradient information of the expectation function is expensive sometimes even prohibitive to obtain and meanwhile the feasibility to general constraints may be hard to realize. Proximal point methods [3, 4, 23] transfer problem (1) into a sequence of convex subproblems with proximal terms. These methods usually have multi-loop structure and need to call a subsolver in each inner-loop. For

instance, the inexact constrained proximal point method with ConEx (ICPPC) in [4] transforms the original problem into a sequence of convex subproblems obtained after adding proximal terms and solves each subproblem with the solver ConEx which is designed for convex functional constrained optimization. [37] studies penalty methods based on first- and zeroth-order stochastic approximations for equality constrained optimization, with each subproblem constructed based on ℓ_2 penalty function. Recently stochastic SQP methods have been studied in [1, 10, 11] for equality constrained stochastic optimization, with complexity analysis provided in [10]. Based on the linearized AL function, [43] studies a single-loop primal-dual stochastic gradient method (PDSG) for solving convex stochastic optimization problems. [18] extends PDSG and proposes a single-loop stochastic primal-dual (SPD) method for nonconvex problems with a large number of functional constraints. However, problems studied in aforementioned works contain only one type of constraints, either equality or inequalities. In this paper, we will consider more general problems which allow both equality and inequality constraints. Meanwhile, we will apply a variance reduced strategy to propose an algorithm with lower oracle complexity.

1.1 Contributions

Our main contributions are summarized as follows.

1. We propose a momentum-based linearized augmented Lagrangian method (MLALM) for nonconvex constrained stochastic optimization (1). To overcome difficulties caused by possibly nonconvex constraints in pursuing a feasible solution, we adopt the idea of the linearized AL function to penalize the constraints into the objective, thus leading to a single-loop algorithm framework, which has a much simpler subproblem to solve at each iteration compared with double-loop algorithms such as proximal point methods [4, 23]. The idea on the single-loop algorithm framework is motivated by recent development in [18]. It studies a stochastic primal-dual (SPD) method based on linearized AL function for nonconvex constrained optimization with many functional constraints. However, due to the large sampling size at each iteration when computing stochastic gradients, the total oracle complexity of SPD to find an ϵ-stationary point is relatively higher even when the initial iterate is nearly feasible. So we integrate a recursive momentum which only requires sampling once at each iteration, in order to control variances of stochastic gradients aiming to derive a lower total oracle complexity. Momentum-based methods in [12, 45] apply a similar idea to compute stochastic gradients, but only aiming for unconstrained stochastic optimization. How it works for nonconvex constrained optimization is not mentioned in [12, 45]. To the best of our knowledge, the MLALM method is the first single-loop algorithm based on a variant of momentum for nonconvex stochastic optimization with general functional constraints.

2. We conduct the complexity analysis of the proposed MLALM algorithm. Under a mean-squared smoothness condition and a nonsingularity condition, through analyzing the KKT measure of the output in terms of stationarity, primal feasibility and complementary slackness, we show that the total oracle complexity of MLALM to find an ϵ-stationary point is bounded by $\mathcal{O}(\epsilon^{-4})$ if the initial point is nearly feasible, and $\mathcal{O}(\epsilon^{-5})$ otherwise. These complexity orders are lower than those of SPD [18] in same problem settings. Compared with our algorithm, the stochastic SQP (SSQP) method in [10] relies on an adaptive strategy to update merit parameters and it presumes knowledge of Lipschitz constants for the objective and constraint gradients, which eliminates the direct application of SSQP for nonsmooth problems. Meanwhile, the algorithm framework and theoretical analysis presented in [10] are devoted for equality constrained optimization while they cannot work for problems with more general constraints. In addition, ICPPC [4] requires the strong feasibility assumption which relies on the existence of a strictly feasible point thus restricting the application of the proposed algorithm to problems with general equality constraints. Instead, MLALM allows that the initial point can

be either nearly feasible or not. For both cases we establish the oracle complexity of MLALM accordingly in this paper. This actually broadens the scope of problems that our algorithm can tackle. In addition, except for an approximate primal solution, MLALM can return an approximate dual solution as a byproduct.

3. We report numerical performances of the proposed algorithm MLALM on quadratically constrained nonconvex programs (QCNPs) and multi-class Neyman-Pearson classification problems (mNPCs). We first investigate the impact that the recursive momentum plays on QCNPs. Then we compare MLALM with ICPPC [4] on QCNPs and with the inexact proximal-point penalty (IPPP) method [22] on mNPCs. Numerical results demonstrate that the introduction of momentum indeed brings benefit and the objective function values returned by MLALM decrease faster than other two algorithms while the primal feasibility keeps in similar levels with them.

In Table 1, we list the detailed comparison between MLALM, ICPPC [4], SSQP[10] and SPD[18], regarding the problem type, the KKT measure of the approximate solution, key assumptions and the resulting oracle complexity in terms of stochastic gradient evaluations. For fair comparison, we here only state the related information for ICPPC in their semistochastic settings, namely when constraints are deterministic. The initial near-feasibility means that the initial point is nearly feasible. The nonsingularity condition for MLALM refers to Assumption 5 in this paper while for SPD a similar condition (see Assumption 3.1 in [18]) is imposed when only inequality constraints exist. The strong LICQ condition for SSQP represents that the Jacobian of constraint functions has singular values that are lower bounded away from zero over a set $\mathcal{X}$, which is assumed to contain all iterates for all realizations of the random variable. Note that the nonsingularity condition in [18] and this paper can be inferred by the strong LICQ condition. And in an idealized setting when the merit parameter threshold $\tau_{\min}$ is known, the oracle complexity of SSQP can be bounded by $\mathcal{O}(\epsilon^{-4})$. Otherwise, an adaptive strategy to update the merit parameter is proposed and the resulting complexity bound is $\tilde{O}(\epsilon^{-4})$, under additional assumptions on D_k (see Assumptions 3 and 4 in [10]), which is the search direction obtained based on stochastic gradients at kth iteration with $k \geq 0$. However, those additional assumptions in [10] are relatively stingent and not easy to verify.

Alg.	**Problem**	**KKT measure**	**Assumptions**	**Comp.**
SPD [18]	$\min_{x\in X}\ f(x)+\chi(x)$ s.t. $\ c_i(x)\leq 0, i\in\mathcal{I}$	$\mathbb{E}[\mathbf{d}(\nabla f(x)+\partial\chi(x)+\sum_{i\in\mathcal{I}}\lambda_i\nabla c_i(x), -\mathcal{N}_X(x))]\leq\epsilon$, $\frac{1}{\lvert\mathcal{I}\rvert}\mathbb{E}[\sum_{i\in\mathcal{I}}[c_i(x)]_+]\leq\epsilon$, $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)\rvert]\leq\epsilon$	nonsingularity, initial near-feasibility	$\mathcal{O}(\epsilon^{-5})$
SSQP [10]	$\min_{x\in\mathbb{R}^n}\ f(x)$ s.t. $c_i(x)=0, i\in\mathcal{E}$	$\mathbb{E}[\lVert\nabla f(x)+\sum_{i\in\mathcal{E}}\lambda_i^{true}\nabla c_i(x)\rVert]\leq\epsilon$, $\mathbb{E}[\sqrt{\sum_{i\in\mathcal{E}}\lvert c_i(x)\rvert}]\leq\epsilon$	strong LICQ, $\tau_{\min}$ unknown	$\tilde{\mathcal{O}}(\epsilon^{-4})$
			strong LICQ, $\tau_{\min}$ known	$\mathcal{O}(\epsilon^{-4})$
ICPPC [4]	$\min_{x\in X}\ f(x)+\chi(x)$ s.t. $c_i(x)+\chi_i(x)\leq 0, i\in\mathcal{I}$	$\mathbb{E}[\lVert x-\hat{x}\rVert^2]\leq\epsilon^2$ with $\hat{x}$ feasible, $\mathbb{E}[\mathbf{d}^2(\nabla f(\hat{x})+\partial\chi(\hat{x})+\sum_{i\in\mathcal{I}}\lambda_i(\nabla c_i(\hat{x})+\partial\chi_i(\hat{x})), -\mathcal{N}_X(\hat{x}))]\leq\epsilon^2$, and $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(\hat{x})+\chi_i(\hat{x})\rvert]\leq\epsilon^2$	strong feasibility	$\mathcal{O}(\epsilon^{-4})$
MLALM (this paper)	$\min_{x\in X}\ f(x)+\chi(x)$ s.t. $c_i(x)=0, i\in\mathcal{E}$ $c_i(x)\leq 0, i\in\mathcal{I}$	$\mathbb{E}[\mathbf{d}^2(\nabla f(x)+\partial\chi(x)+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla c_i(x), -\mathcal{N}_X(x))]\leq\epsilon^2$, $\mathbb{E}[\lVert c_{\mathcal{E}}(x)\rVert^2+\lVert[c_{\mathcal{I}}(x)]_+\rVert^2]\leq\epsilon^2$, $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)\rvert]\leq\epsilon$	mean-squared smoothness, nonsingularity, initial near-feasibility	$\mathcal{O}(\epsilon^{-4})$

Table 1: Comparison between algorithms for nonconvex constrained optimization, where $f(x) = \mathbb{E}_\xi[\mathbf{F}(x;\xi)]$, χ and $\chi_i, i\in\mathcal{I}$ are convex but possibly nonsmooth, $\lambda_i, i\in\mathcal{I}$ are nonnegative, λ^{true} is a vector of Lagrange multipliers corresponding to x and $\tau_{\min}$ is the merit parameter threshold in SSQP.

1.2 Notation and preliminaries

Without any specification, we use $\|\cdot\|$ to denote the Euclidean norm of a vector. For brevity, we introduce $[k] := \{1, \ldots, k\}$ and $\xi^{[k]} := \{\xi^1, \ldots, \xi^k\}$ for any positive integer k. For any $u \in \mathbb{R}$, we define its positive and negative parts as $[u]_+ := \max\{0, u\}$ and $[u]_- := \max\{0, -u\}$, respectively. Moreover, for any $u \in \mathbb{R}^n$, $[u]_+$ and $[u]_-$ are referred to as componentwise application of the operator $[\cdot]_+$ and $[\cdot]_-$. The gradient of f at x is denoted by $\nabla f(x)$. By a slight abuse of notation, we define $c_{\mathcal{E}} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{E}|}$ with components being $c_i(\cdot)$, $i \in \mathcal{E}$ and $\nabla c_{\mathcal{E}} : \mathbb{R}^n \to \mathbb{R}^{n \times |\mathcal{E}|}$ with columns being $\nabla c_i(\cdot), i \in \mathcal{E}$. Notations $c_{\mathcal{I}}$ and $\nabla c_{\mathcal{I}}$ are defined in the same way. Given $X, Y \subseteq \mathbb{R}^n$, the distance between them is referred to $\mathbf{d}(X, Y) = \inf_{x \in X, y \in Y} \|x - y\|$. Furthermore, $\mathbb{E}_\xi[\cdot]$ represents the expectation with respect to ξ and $\mathbb{E}_\xi[\cdot \mid \zeta]$ represents the expectation with respect to ξ conditioned on ζ. The inner product of $x, y \in \mathbb{R}^n$ is denoted by $\langle x, y \rangle$.

In general, finding a global or even a local minimizer for nonconvex constrained optimization can be NP-hard. Efforts are thus devoted to finding more trackable solutions. As is known, under certain constraint qualifications a local minimizer satisfies Karush-Kuhn-Tucker (KKT) conditions [41]. Those satisfying KKT conditions are called KKT points. In this paper, without specifying any constraint qualification, we assume the existence of KKT points. To characterize KKT points of (1), we give the following concepts.

Definition 1 *The normal cone to a closed convex set X at a point $\bar{x} \in X$ is defined as*

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

Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous convex function and $x \in dom(h)$. The set of subgradient of h at x is defined as

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

Definition 2 *A point $x^* \in X$ is called a KKT point of (1), if there is a Lagrange multiplier vector $\lambda^* \in \mathbb{R}^{|\mathcal{E} \cup \mathcal{I}|}$ with $\lambda_i^* \geq 0$, $i \in \mathcal{I}$, such that the following conditions are satisfied at (x^*, λ^*):*

$$\mathbf{d}(\nabla f(x^*) + \partial \chi(x^*) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \nabla c_i(x^*), -\mathcal{N}_X(x^*)) = 0; \quad c_{\mathcal{E}}(x) = \mathbf{0}, \ c_{\mathcal{I}}(x) \leq \mathbf{0}; \quad \lambda_i c_i(x) = 0, i \in \mathcal{I}.$$

We next lay out assumptions used throughout the remainder of this paper.

Assumption 1 *The set X is closed and convex. Functions f and $c_i, i \in \mathcal{E} \cup \mathcal{I}$ are continuously differentiable over X with L-Lipschitz continuous gradients. Function χ is proper, lower semicontinuous and convex over X. The objective function value of (1) over X is lower bounded by C^*.*

Assumption 2 *There exist $F, G > 0$ such that for any $x \in X$,*

$$|c_i(x)| \leq F, \ \forall i \in \mathcal{E}\,; \ c_i(x) \leq F, \ \forall i \in \mathcal{I}\,; \tag{2}$$

$$\|\nabla f(x)\| \leq G, \ \|\partial \chi(x)\| \leq G, \text{ and } \|\nabla c_i(x)\| \leq G, \ \forall i \in \mathcal{E} \cup \mathcal{I}. \tag{3}$$

Assumption 3 *$\mathbf{F}(\cdot; \xi)$ is differentiable almost surely for $\xi \in \Xi$ and satisfies*

$$\mathbb{E}_\xi[\|\nabla \mathbf{F}(u; \xi) - \nabla \mathbf{F}(v; \xi)\|^2] \leq L^2 \|u - v\|^2 \quad \forall u, v \in X.$$

Assumption 4 *There exists $\sigma > 0$ such that for any $x \in X$,*

$$\mathbb{E}_\xi[\nabla \mathbf{F}(x; \xi)] = \nabla f(x), \quad \mathbb{E}_\xi[\|\nabla \mathbf{F}(x; \xi) - \nabla f(x)\|^2] \leq \sigma^2,$$

Remark 1 *It is noteworthy that the boundedness in Assumption 2 holds naturally under Assumption 1 when X is compact, which is assumed in [4, 18]. And Assumption 3 refers to the mean-squared smoothness condition [45].*

1.3 Outline

The rest of this paper is organized as follows. In Section 2 we present details of a momentum-based linearized AL method for nonconvex constrained stochastic optimization (1). In Section 3 we investigate theoretical properties of the proposed algorithm and establish its oracle complexities to find an ϵ-stationary point of (1). In Section 4 we report some numerical experimental results and finally we give some conclusional remarks.

2 Momentum-based linearized augmented Lagrangian method for (1)

As is known, the augmented Lagrangian (AL) function plays a crucial role in characterizing optimality conditions for constrained optimization and is widely used in designing effective algorithms. The AL function [29] associated with problem (1) is in the form

$$\mathcal{L}_\beta(x,\lambda) = \phi_\beta(x,\lambda) + \chi(x), \tag{4}$$

where β is a penalty parameter, $\phi_\beta(x,\lambda) = f(x) + \Psi_\beta(x,\lambda)$ and

$$\Psi_\beta(x,\lambda) = \sum_{i\in\mathcal{E}}[\lambda_i c_i(x) + \frac{\beta}{2}c_i^2(x)] + \sum_{i\in\mathcal{I}}\psi_\beta(c_i(x),\lambda_i) \quad \text{with} \quad \psi_\beta(u,v) = \begin{cases} vu + \dfrac{\beta}{2}u^2 & \text{if } \beta u + v \geq 0, \\ -\dfrac{v^2}{2\beta} & \text{otherwise.} \end{cases}$$

It should be noted that $\Psi_\beta(x,\lambda)$ is concave in λ and

$$\nabla_x \Psi_\beta(x,\lambda) = \sum_{i\in\mathcal{E}}(\lambda_i + \beta c_i(x))\nabla c_i(x) + \sum_{i\in\mathcal{I}}[\lambda_i + \beta c_i(x)]_+ \nabla c_i(x). \tag{5}$$

Different from the classic AL methods which try to minimize the AL function in the inner-loop, linearized AL methods [38, 39, 44] construct much simpler subproblems only minimizing an approximation to the AL function at each iteration:

$$\min_{x\in X} \quad \langle \nabla_x \phi_\beta(x^t,\lambda^t), x\rangle + \frac{1}{2\eta_t}\|x - x^t\|^2 + \chi(x),$$

where $\eta_t > 0$. However, due to the problem setting of (1), it is normally expensive sometimes even prohibitive to compute the exact gradient ∇f at an inquiry point $x \in X$. Under this circumstance, we can only get access to a stochastic gradient $\nabla \mathbf{F}(x;\xi)$ by randomly calling a sample ξ. So a stochastic gradient $\nabla \mathbf{\Phi}_\beta(x,\lambda;\xi)$ can be obtained, where

$$\mathbf{\Phi}_\beta(x,\lambda;\xi) := \mathbf{F}(x;\xi) + \Psi_\beta(x,\lambda).$$

The SPD method [18] also adopts the linearized AL function to construct subproblems, but when computing mini-batch stochastic gradients it needs large batch sizes in order to derive desired iteration complexity. Inspired by this, so as to reduce the total oracle complexity, a possible way may be that we control the batch size to compute stochastic gradient in a very low level while not sacrificing the iteration complexity at the meantime.

Before proceeding, let us consider to minimize a continuously differentiable function $h(x) = \mathbb{E}[\mathbf{H}(x\,;\xi)]$ over $\mathbb{R}^n$ with $\xi \in \Xi$. Recall that the Nesterov's accelerated gradient approach reads

$$x^{t+1} = x^t - \eta_t s^t; \quad s^{t+1} = a_t s^t + b_t \nabla h(x^{t+1} - a_t s^t) \text{ with } s^1 = \nabla h(x^1).$$

Nevertheless, since the exact gradient of h cannot be accessed, we have to turn to its stochastic version by randomly taking one sample ξ^{t+1}:

$$\begin{aligned} s^{t+1} &= a_t s^t + b_t \nabla \mathbf{H}(x^{t+1} - a_t s^t; \xi^{t+1}) \\ &\approx a_t s^t + b_t[(1 + \frac{a_t}{\eta_t})\nabla \mathbf{H}(x^{t+1}; \xi^{t+1}) - \frac{a_t}{\eta_t}\nabla \mathbf{H}(x^t; \xi^{t+1})], \end{aligned}$$

where the above approximation uses the linear Lagrange interpolating polynomial. Then, letting $a_t = 1-\alpha_t$ and $b_t = \eta_t = \alpha_t$, we obtain the stochastic gradient estimation in the recursive momentum method [12] and its variant [45]:

$$s^{t+1} = \nabla \mathbf{H}(x^{t+1}; \xi^{t+1}) + (1 - \alpha_t)(s^t - \nabla \mathbf{H}(x^t; \xi^{t+1})), \quad t \geq 1.$$

As shown in [12, 45], the incorporation of the momentum term together with the mean-squared smoothness assumption can help to reduce the oracle complexity of the SGD and the proximal SGD methods. Motivated by this, we extend the idea to the general constrained optimization problem (1) to compute a stochastic approximation to $\nabla_x \phi_\beta$. With $d^1 = \nabla_x \mathbf{\Phi}_\beta(x^1, \lambda^1; \xi^1)$, we define d^t as

$$d^t = \nabla_x \mathbf{\Phi}_\beta(x^t, \lambda^t; \xi^t) + (1 - \alpha_{t-1})(d^{t-1} - \nabla_x \mathbf{\Phi}_\beta(x^{t-1}, \lambda^{t-1}; \xi^t)), \quad t \geq 2. \tag{6}$$

The first term in (6) is a stochastic gradient estimated at x^t, while the second term is the difference between the last approximation and a stochastic gradient estimated at x^{t-1}, aiming to correct the stochastic approximation. With this technique, we can derive reduced oracle complexities as shown in the next section. When $\alpha_{t-1} = 1$, d^t reduces to the standard SGD approximation. In this paper, we choose $\alpha_t \in (0, 1)$, $t \geq 1$. Based on the stochastic approximation d^t in (6), we propose the following scheme to update the primal variable:

$$x^{t+1} = \arg\min_{x \in X} \{\langle d^t, x\rangle + \chi(x) + \frac{1}{2\eta_t}\|x - x^t\|^2\}. \tag{7}$$

Recall that in classic AL methods, the dual variable λ is normally updated through $\lambda^t = \bar{\lambda}^t$, where

$$\bar{\lambda}_i^t = \lambda_i^t + \begin{cases} \beta c_i(x^{t+1}), & i \in \mathcal{E}, \\ \beta \max(-\frac{\lambda_i^t}{\beta}, c_i(x^{t+1})), & i \in \mathcal{I}. \end{cases} \tag{8}$$

However, since in each subproblem of our method it is not directly minimizing the original AL function but its approximation, it seems unnecessary to follow the traditional way (8) to update λ^t. Instead, we adopt an average of λ^t and $\bar{\lambda}^t$:

$$\lambda_i^{t+1} = (1 - \frac{\rho_t}{\beta})\lambda_i^t + \frac{\rho_t}{\beta}\bar{\lambda}_i^t \tag{9}$$

to expect dual variables more trackable.

Algorithm 1 Momentum-based Linearized Augmented Lagrangian Method (MLALM)

Input: Initial point x^1, initial dual point $\lambda^1 = \mathbf{0}$, parameters $\beta \geq \beta_0 > 0,\ \eta_t > 0,\ \rho_t \in (0, \beta),\ \alpha_t \in (0, 1)$.
Output: x^{R+1} where $R \in [T]$ is uniformly chosen at random.

1: **for** $t = 1, \ldots, T$ **do**
2: Sample ξ^t from Ξ and update d^t using (6).
3: Calculate x^{t+1} through (7).
4: Calculate λ^{t+1} through (9).
5: **end for**

Our goal in this paper is to pursue an ϵ-stationary point with the expected KKT measure below certain accuracy.

Definition 3 *Given $\epsilon > 0$, a point $x \in X$ is called ϵ-stationary point of (1), if there is $\lambda \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ with $\lambda_i \geq 0,\ i \in \mathcal{I}$, such that the following conditions are satisfied:*

$$\mathbb{E}[\mathbf{d}^2(\nabla f(x) + \partial\chi(x) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i \nabla c_i(x), -\mathcal{N}_X(x))] \leq \epsilon^2, \tag{10}$$

$$\mathbb{E}[\|c_{\mathcal{E}}(x)\|^2 + \|c_{\mathcal{I}}(x)]_+\|^2] \leq \epsilon^2, \tag{11}$$

$$\mathbb{E}[\sum_{i\in\mathcal{I}} \lambda_i |c_i(x)|] \leq \epsilon, \tag{12}$$

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

3 Theoretical analysis

In this section, we will provide theoretical analysis of the method MLALM. We aim for establishing its oracle complexity to find an ϵ-stationary point of (1). To this end, we need to analyze the KKT measure in terms of stationarity, feasibility and complementary slackness at the output of the algorithm. Without loss of generality, we assume in this section that

$$\rho_t \in (0, \frac{\rho}{T}] \subseteq (0, \beta], \quad t \in [T], \tag{13}$$

where $\rho > 0$ is independent of T and β.

The following two lemmas establish the feasibility and boundedness of $\{\lambda^t, t \in [T+1]\}$ generated by Algorithm 1.

Lemma 1 *For any $t \geq 1$, it holds that $\lambda_i^t \geq 0$, $i \in \mathcal{I}$.*

Proof. It is straightforward to obtain the conclusion by induction from $\lambda^1 = \mathbf{0}$, $\rho_t \in (0, \beta]$ and (9). □

Lemma 2 *Under Assumptions 1-2, it holds that for any $t \geq 1$,*

$$|\lambda_i^t| \leq F\sum_{k=1}^{t-1} \rho_k, \quad \forall i \in \mathcal{E}; \quad \lambda_i^t \leq F\sum_{k=1}^{t-1} \rho_k, \quad \forall i \in \mathcal{I}, \tag{14}$$

and for any $t \in [T]$,

$$|\lambda_i^{t+1} - \lambda_i^t| \leq \rho_t F_c, \quad \forall i \in \mathcal{E}\cup\mathcal{I}, \tag{15}$$

where $F_c := \max(\frac{\rho F}{\beta_0}, F)$ and $\sum_{k=1}^0 \rho_k := 0$.

Proof. On the first hand, by applying $\lambda^1 = \mathbf{0}$ and (9), we have that for any $t \geq 2$,

$$|\lambda_i^t| \leq |\lambda_i^1| + \sum_{k=1}^{t-1} |\lambda_i^{k+1} - \lambda_i^k| \leq \sum_{k=1}^{t-1} \rho_k |c_i(x^{k+1})| \leq F\sum_{k=1}^{t-1} \rho_k, \quad \forall i \in \mathcal{E},$$

$$\lambda_i^t = \lambda_i^1 + \sum_{k=1}^{t-1} (\lambda_i^{k+1} - \lambda_i^k) \leq \sum_{\mathcal{K}} \rho_k c_i(x^{k+1}) \leq F\sum_{k=1}^{t-1} \rho_k, \quad \forall i \in \mathcal{I},$$

where $\mathcal{K} = \{k \in [t-1] \mid \lambda_i^{k+1} \geq \lambda_i^k\}$. We thus obtain (14). On the other hand, for any $i \in \mathcal{E}$, it is easy to obtain

$$|\lambda_i^{t+1} - \lambda_i^t| = \rho_t |c_i(x^{t+1})| \leq \rho_t F_c.$$

Then, for any $i \in \mathcal{I}$, it follows from (9), Lemma 1, Assumption 2 and (14) that

$$|\lambda_i^{t+1} - \lambda_i^t| = \rho_t |\max(-\frac{\lambda_i^t}{\beta}, c_i(x^{t+1}))| \leq \begin{cases} \rho_t F, & \text{if } c_i(x^{t+1}) \geq 0, \\ \frac{\rho_t \lambda_i^t}{\beta} \leq \frac{\rho_t \rho F}{\beta_0}, & \text{otherwise,} \end{cases}$$

which yields (15). □

Lemma 3 *Under Assumptions 1-2, it holds that for any $t \in [T]$,*

$$|\psi_\beta(c_i(x^{t+1}), \lambda_i^{t+1}) - \psi_\beta(c_i(x^{t+1}), \lambda_i^t)| \leq \rho_t F_c^2, \quad i \in \mathcal{I} \tag{16}$$

and

$$\|\nabla_x \Psi_\beta(x, \lambda^{t+1}) - \nabla_x \Psi_\beta(x, \lambda^t)\| \leq m \rho_t F_c G. \tag{17}$$

Proof. For any $t \in [T]$, by the definition of $\psi_\beta(u, v)$, we know that for any $i \in \mathcal{I}$,

$$\psi_\beta(c_i(x^{t+1}), \lambda_i) = \begin{cases} \lambda_i c_i(x^{t+1}) + \frac{\beta}{2} c_i^2(x^{t+1}), & \text{if } \beta c_i(x^{t+1}) + \lambda_i \geq 0, \\ -\frac{(\lambda_i)^2}{2\beta}, & \text{if } \beta c_i(x^{t+1}) + \lambda_i < 0. \end{cases}$$

First, we consider the case of $\beta c_i(x^{t+1}) + \lambda_i^t < 0$, then by Lemma 1 and (9), we have $\lambda_i^{t+1} \leq \lambda_i^t$. Thus $\beta c_i(x^{t+1}) + \lambda_i^{t+1} < 0$. Hence, we have

$$|\psi_\beta(c_i(x^{t+1}), \lambda_i^{t+1}) - \psi_\beta(c_i(x^{t+1}), \lambda_i^t)| = \frac{(\lambda_i^t)^2 - (\lambda_i^{t+1})^2}{2\beta} = \frac{\lambda_i^t + \lambda_i^{t+1}}{2\beta} |\lambda_i^{t+1} - \lambda_i^t| \leq \frac{\rho F}{\beta} |\lambda_i^{t+1} - \lambda_i^t| \leq \rho_t F_c^2.$$

Second, when $\beta c_i(x^{t+1}) + \lambda_i^t \geq 0$, $-\frac{\lambda_i^t}{\beta} \leq c_i(x^{t+1}) \leq F$, then the result holds obviously if $\beta c_i(x^{t+1}) + \lambda_i^{t+1} \geq 0$. If $\beta c_i(x^{t+1}) + \lambda_i^{t+1} < 0$, since $\psi_\beta(u, v)$ is monotonically decreasing in $v \geq 0$ when $u < 0$, we have

$$\begin{aligned} |\psi_\beta(c_i(x^{t+1}), \lambda_i^{t+1}) - \psi_\beta(c_i(x^{t+1}), \lambda_i^t)| &= -\frac{(\lambda_i^{t+1})^2}{2\beta} - \lambda_i^t c_i(x^{t+1}) - \frac{\beta}{2} c_i^2(x^{t+1}) \\ &\leq \lambda_i^{t+1} c_i(x^{t+1}) + \frac{\beta}{2} c_i^2(x^{t+1}) - \lambda_i^t c_i(x^{t+1}) - \frac{\beta}{2} c_i^2(x^{t+1}) \\ &= -c_i(x^{t+1}) |\lambda_i^{t+1} - \lambda_i^t| \leq \frac{\lambda_i^t}{\beta} |\lambda_i^{t+1} - \lambda_i^t| \leq \rho_t F_c^2 \end{aligned}$$

which yields (16). In addition, it indicates from (5) and (15) that for any $t \in [T]$,

$$\begin{aligned} \|\nabla_x \Psi_\beta(x, \lambda^{t+1}) - \nabla_x \Psi_\beta(x, \lambda^t)\| &\leq \|\sum_{i \in \mathcal{E}} (\lambda_i^{t+1} - \lambda_i^t) \nabla c_i(x)\| + \|\sum_{i \in \mathcal{I}} ([\beta c_i(x) + \lambda_i^{t+1}]_+ - [\beta c_i(x) + \lambda_i^t]_+) \nabla c_i(x)\| \\ &\leq G \sum_{i \in \mathcal{E}} |\lambda_i^{t+1} - \lambda_i^t| + G \sum_{i \in \mathcal{I}} |[\beta c_i(x) + \lambda_i^{t+1}]_+ - [\beta c_i(x) + \lambda_i^t]_+| \\ &\leq G \sum_{i \in \mathcal{E} \cup \mathcal{I}} |\lambda_i^{t+1} - \lambda_i^t| \leq m \rho_t F_c G \end{aligned}$$

which is exactly (17). □

The lemma below characterizes the smoothness of $\phi_\beta(x,\lambda)$ with respect to x for fixed λ.

Lemma 4 *Under Assumptions 1-2, it holds that for any $u, v \in X$, $t \in [T+1]$,*

$$\|\nabla_x \phi_\beta(u,\lambda^t) - \nabla_x \phi_\beta(v,\lambda^t)\| \le L_\beta \|u - v\|. \tag{18}$$

Furthermore, if Assumption 3 holds as well, then

$$\mathbb{E}_\xi[\|\nabla_x \mathbf{\Phi}_\beta(u,\lambda^t;\xi) - \nabla_x \mathbf{\Phi}_\beta(v,\lambda^t;\xi)\|^2] \le L_\beta^2 \|u - v\|^2, \tag{19}$$

where $L_\beta = L + m(\beta G^2 + \beta FL + \rho FL)$.

Proof. It follows from Assumptions 1-2, (5) and (14) that for any $u, v \in X$,

$$\begin{aligned}
&\|\nabla_x \Psi_\beta(u,\lambda^t) - \nabla_x \Psi_\beta(v,\lambda^t)\| \\
\le& \sum_{i\in\mathcal{E}} \|(\beta c_i(u) + \lambda_i^t)\nabla c_i(u) - (\beta c_i(v) + \lambda_i^t)\nabla c_i(v)\| \\
&+ \sum_{i\in\mathcal{I}} \|[\beta c_i(u) + \lambda_i^t]_+ \nabla c_i(u) - [\beta c_i(v) + \lambda_i^t]_+ \nabla c_i(v)\| \\
=& \sum_{i\in\mathcal{E}} \|\big[(\beta c_i(u) + \lambda_i^t) - (\beta c_i(v) + \lambda_i^t)\big]\nabla c_i(u) + (\beta c_i(v) + \lambda_i^t)(\nabla c_i(u) - \nabla c_i(v))\| \\
&+ \sum_{i\in\mathcal{I}} \|\big[[\beta c_i(u) + \lambda_i^t]_+ - [\beta c_i(v) + \lambda_i^t]_+\big]\nabla c_i(u) + [\beta c_i(v) + \lambda_i^t]_+(\nabla c_i(u) - \nabla c_i(v))\| \\
\le& \sum_{i\in\mathcal{E}} \big[\beta|c_i(u) - c_i(v)|\|\nabla c_i(u)\| + (\beta c_i(v) + \lambda_i^t)L\|u - v\|\big] \\
&+ \sum_{i\in\mathcal{I}} \big[\beta|c_i(u) - c_i(v)|\|\nabla c_i(u)\| + [\beta c_i(v) + \lambda_i^t]_+ L\|u - v\|\big] \\
\le& \sum_{i\in\mathcal{E}\cup\mathcal{I}} [\beta G^2\|u - v\| + L(\beta F + |\lambda_i^t|)\|u - v\|] \\
\le& m(\beta G^2 + FL(\beta + \rho))\|u - v\|.
\end{aligned}$$

Then, Assumptions 1-2 indicate

$$\|\nabla_x \phi_\beta(u,\lambda^t) - \nabla_x \phi_\beta(v,\lambda^t)\| \le \|\nabla f(u) - \nabla f(v)\| + \|\nabla_x \Psi_\beta(u,\lambda^t) - \nabla_x \Psi_\beta(v,\lambda^t)\| \le L_\beta\|u - v\|.$$

Analogously, it holds from Assumption 3 that

$$\begin{aligned}
&\mathbb{E}_\xi[\|\nabla_x \mathbf{\Phi}_\beta(u,\lambda^t;\xi) - \nabla_x \mathbf{\Phi}_\beta(v,\lambda^t;\xi)\|^2] \\
\le& \mathbb{E}_\xi[\|\nabla \mathbf{F}(u;\xi) - \nabla \mathbf{F}(v;\xi)\|^2] + 2\mathbb{E}_\xi[\|\nabla \mathbf{F}(u;\xi) - \nabla \mathbf{F}(v;\xi)\|]\|\nabla_x \Psi_\beta(u,\lambda^t) - \nabla_x \Psi_\beta(v,\lambda^t)\| \\
&+ \|\nabla_x \Psi_\beta(u,\lambda^t) - \nabla_x \Psi_\beta(v,\lambda^t)\|^2 \\
\le& L^2\|u - v\|^2 + 2mL(\beta G^2 + FL(\beta + \rho))\|u - v\|^2 + (m(\beta G^2 + FL(\beta + \rho)))^2\|u - v\|^2
\end{aligned}$$

which yields (19). □

Recall that d^t is an approximation to $\nabla_x \phi_\beta(x^t, \lambda^t)$. We define ε^t as the error between them:

$$\varepsilon^t := d^t - \nabla_x \phi_\beta(x^t, \lambda^t).$$

Lemma 5 *Under Assumptions 1-2, it holds that*

$$\mathbf{d}^2(\nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) + \partial\chi(x^{t+1}), -\mathcal{N}_X(x^{t+1})) \leq 4(m\rho_t F_c G)^2 + 4\|\varepsilon^t\|^2 + 4(L_\beta^2 + \frac{1}{\eta_t^2})\|x^{t+1} - x^t\|^2. \quad (20)$$

Proof. Optimality conditions for (7) imply that for any $t \geq 1$,

$$\mathbf{d}(d^t + \partial\chi(x^{t+1}) + \frac{1}{\eta_t}(x^{t+1} - x^t), -\mathcal{N}_X(x^{t+1})) = 0. \quad (21)$$

Then it indicates from Jensen's inequality, (17) and (18) that

$$\begin{aligned}
&\mathbf{d}^2(\nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) + \partial\chi(x^{t+1}), -\mathcal{N}_X(x^{t+1})) \\
=& \mathbf{d}^2(d^t + \partial\chi(x^{t+1}) + \frac{1}{\eta_t}(x^{t+1} - x^t) + \nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) - d^t - \frac{1}{\eta_t}(x^{t+1} - x^t), -\mathcal{N}_X(x^{t+1})) \\
\leq& \|\nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) - d^t - \frac{1}{\eta_t}(x^{t+1} - x^t)\|^2 \\
=& \|\nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) - \nabla_x \phi_\beta(x^{t+1}, \lambda^t) + \nabla_x \phi_\beta(x^{t+1}, \lambda^t) - \nabla_x \phi_\beta(x^t, \lambda^t) + \nabla_x \phi_\beta(x^t, \lambda^t) - d^t - \frac{1}{\eta_t}(x^{t+1} - x^t)\|^2 \\
\leq& 4(m\rho_t F_c G)^2 + 4L_\beta^2\|x^{t+1} - x^t\|^2 + 4\|\varepsilon^t\|^2 + \frac{4}{\eta_t^2}\|x^{t+1} - x^t\|^2,
\end{aligned}$$

which completes the proof. □

Lemma 6 *Under Assumptions 1-2, it holds that*

$$(\frac{1}{2\eta_t} - \frac{L_\beta}{2})\|x^{t+1} - x^t\|^2 \leq \mathcal{L}_\beta(x^t, \lambda^t) - \mathcal{L}_\beta(x^{t+1}, \lambda^{t+1}) + \frac{\eta_t}{2}\|\varepsilon^t\|^2 + m\rho_t F_c^2. \quad (22)$$

Proof. Assumption 2 together with (15), (16) implies that

$$\begin{aligned}
\mathcal{L}_\beta(x^{t+1}, \lambda^t) =& \mathcal{L}_\beta(x^{t+1}, \lambda^{t+1}) + \sum_{i\in\mathcal{E}}[(\lambda_i^t c_i(x^{t+1}) + \frac{\beta}{2}c_i^2(x^{t+1})) - (\lambda_i^{t+1} c_i(x^{t+1}) + \frac{\beta}{2}c_i^2(x^{t+1}))] \\
&+ \sum_{i\in\mathcal{I}}[\psi_\beta(c_i(x^{t+1}), \lambda_i^t) - \psi_\beta(c_i(x^{t+1}), \lambda_i^{t+1})] \\
\geq& \mathcal{L}_\beta(x^{t+1}, \lambda^{t+1}) - \sum_{i\in\mathcal{E}}|c_i(x^{t+1})||\lambda_i^t - \lambda_i^{t+1}| - \sum_{i\in\mathcal{I}}\rho_t F_c^2 \\
\geq& \mathcal{L}_\beta(x^{t+1}, \lambda^{t+1}) - m\rho_t F_c^2. \quad (23)
\end{aligned}$$

Note that by optimality conditions for (7), there exists a vector $u \in \partial\chi(x^{t+1})$ such that

$$\langle d^t + u + \frac{1}{\eta_t}(x^{t+1} - x^t), x - x^{t+1}\rangle \geq 0, \quad \forall x \in X.$$

Then by the convexity of χ and the setting $x = x^t$, we have

$$\chi(x^{t+1}) - \chi(x^t) \leq \langle u, x^{t+1} - x^t \rangle \leq -\langle d^t + \frac{1}{\eta_t}(x^{t+1} - x^t), x^{t+1} - x^t \rangle. \tag{24}$$

Thus it together with (23) indicates that

$$\begin{aligned}
\mathcal{L}_\beta(x^{t+1}, \lambda^{t+1}) - \mathcal{L}_\beta(x^t, \lambda^t) &= \mathcal{L}_\beta(x^{t+1}, \lambda^{t+1}) - \mathcal{L}_\beta(x^{t+1}, \lambda^t) + \mathcal{L}_\beta(x^{t+1}, \lambda^t) - \mathcal{L}_\beta(x^t, \lambda^t) \\
&\leq \mathcal{L}_\beta(x^{t+1}, \lambda^t) - \mathcal{L}_\beta(x^t, \lambda^t) + m\rho_t F_c^2 \\
&= \phi_\beta(x^{t+1}, \lambda^t) - \phi_\beta(x^t, \lambda^t) + \chi(x^{t+1}) - \chi(x^t) + m\rho_t F_c^2 \\
&\leq \langle \nabla_x \phi_\beta(x^t, \lambda^t), x^{t+1} - x^t \rangle + \frac{L_\beta}{2}\|x^{t+1} - x^t\|^2 + \chi(x^{t+1}) - \chi(x^t) + m\rho_t F_c^2 \\
&\leq \langle d^t - \varepsilon^t, x^{t+1} - x^t \rangle - \langle d^t + \frac{1}{\eta_t}(x^{t+1} - x^t), x^{t+1} - x^t \rangle + \frac{L_\beta}{2}\|x^{t+1} - x^t\|^2 + m\rho_t F_c^2 \\
&= -\langle \varepsilon^t, x^{t+1} - x^t \rangle + (\frac{L_\beta}{2} - \frac{1}{\eta_t})\|x^{t+1} - x^t\|^2 + m\rho_t F_c^2 \\
&\leq \frac{\eta_t}{2}\|\varepsilon^t\|^2 + \frac{1}{2\eta_t}\|x^{t+1} - x^t\|^2 + (\frac{L_\beta}{2} - \frac{1}{\eta_t})\|x^{t+1} - x^t\|^2 + m\rho_t F_c^2,
\end{aligned}$$

where the second inequality follows from (18), the third inequality comes from (24), and the last inequality is from Young's inequality. Then rearranging the terms, we derive (22). □

The lemma below provides a recursive bound on ε^t.

Lemma 7 *Under Assumptions 1-4, it holds that*

$$\mathbb{E}_{\xi^{[t+1]}}[\|\varepsilon^{t+1}\|^2] \leq (1-\alpha_t)^2 \mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] + 2\alpha_t^2\sigma^2 + 4(1-\alpha_t)^2(m\rho_t F_c G)^2 + 4(1-\alpha_t)^2 L_\beta^2 \mathbb{E}_{\xi^{[t]}}[\|x^{t+1} - x^t\|^2].$$

Proof. Recall that

$$\begin{aligned}
\varepsilon^{t+1} &= d^{t+1} - \nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) \\
&= \nabla_x \mathbf{\Phi}_\beta(x^{t+1}, \lambda^{t+1}; \xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) + (1-\alpha_t)(d^t - \nabla_x \mathbf{\Phi}_\beta(x^t, \lambda^t; \xi^{t+1})) \\
&= \nabla_x \mathbf{\Phi}_\beta(x^{t+1}, \lambda^{t+1}; \xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) + (1-\alpha_t)\varepsilon^t + (1-\alpha_t)(\nabla_x \phi_\beta(x^t, \lambda^t) - \nabla_x \mathbf{\Phi}_\beta(x^t, \lambda^t; \xi^{t+1})).
\end{aligned} \tag{25}$$

Since $\varepsilon^t, x^t, \lambda^t, x^{t+1}$ and λ^{t+1} are independent of ξ^{t+1}, taking expectation with respect to ξ^{t+1} conditioned on $\xi^{[t]}$ yields

$$\begin{aligned}
\mathbb{E}_{\xi^{t+1}}[\langle \nabla_x \mathbf{\Phi}_\beta(x^{t+1}, \lambda^{t+1}; \xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}), \varepsilon^t \rangle \mid \xi^{[t]}] = 0, \\
\mathbb{E}_{\xi^{t+1}}[\langle \nabla_x \mathbf{\Phi}_\beta(x^t, \lambda^t; \xi^{t+1}) - \nabla_x \phi_\beta(x^t, \lambda^t), \varepsilon^t \rangle \mid \xi^{[t]}] = 0.
\end{aligned}$$

Hence, squaring both sides of (25) and taking expectation with respect to ξ^{t+1} conditioned on $\xi^{[t]}$, we have

$$\begin{aligned}
\mathbb{E}_{\xi^{t+1}}[\|\varepsilon^{t+1}\|^2 \mid \xi^{[t]}] \leq (1-\alpha_t)^2\|\varepsilon^t\|^2 + \mathbb{E}_{\xi^{t+1}}[\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1}, \lambda^{t+1}; \xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1}, \lambda^{t+1}) \\
+ (1-\alpha_t)(\nabla_x \phi_\beta(x^t, \lambda^t) - \nabla_x \mathbf{\Phi}_\beta(x^t, \lambda^t; \xi^{t+1}))\|^2 \mid \xi^{[t]}].
\end{aligned}$$

Let us focus on the second term of the right hand side. Note that

$$\mathbb{E}_{\xi^{t+1}}[\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) + (1-\alpha_t)(\nabla_x \phi_\beta(x^t,\lambda^t) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1}))\|^2 \mid \xi^{[t]}]$$

$$=\mathbb{E}_{\xi^{t+1}}[\|\alpha_t(\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1})) + (1-\alpha_t)(\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1})$$

$$-\nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) + \nabla_x \phi_\beta(x^t,\lambda^t) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1}))\|^2 \mid \xi^{[t]}]$$

$$\leq\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1})\|^2 + 2(1-\alpha_t)^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1})$$

$$-\nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) + \nabla_x \phi_\beta(x^t,\lambda^t) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1})\|^2 \mid \xi^{[t]}]$$

$$=\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1})\|^2$$

$$+ 2(1-\alpha_t)^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1})\|^2$$

$$- 4(1-\alpha_t)^2\langle\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1}), \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) - \nabla_x \phi_\beta(x^t,\lambda^t)\rangle$$

$$+ 2(1-\alpha_t)^2\|\nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) - \nabla_x \phi_\beta(x^t,\lambda^t)\|^2 \mid \xi^{[t]}]$$

$$=\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1})\|^2$$

$$+ 2(1-\alpha_t)^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1})\|^2$$

$$- 2(1-\alpha_t)^2\|\nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) - \nabla_x \phi_\beta(x^t,\lambda^t)\|^2 \mid \xi^{[t]}]$$

$$\leq\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1})\|^2$$

$$+2(1-\alpha_t)^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1})\|^2 \mid \xi^{[t]}]$$

$$\leq 2\alpha_t^2\sigma^2 + \mathbb{E}_{\xi^{t+1}}[4(1-\alpha_t)^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^t;\xi^{t+1})\|^2$$

$$+ 4(1-\alpha_t)^2\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^t;\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1})\|^2 \mid \xi^{[t]}],$$

where the last equality follows from

$$\mathbb{E}_{\xi^{t+1}}[\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1}) \mid \xi^{[t]}] = \nabla_x \phi_\beta(x^{t+1},\lambda^{t+1}) - \nabla_x \phi_\beta(x^t,\lambda^t).$$

Hence, we obtain the following relations:

$$\mathbb{E}_{\xi^{t+1}}[\|\varepsilon^{t+1}\|^2 \mid \xi^{[t]}]$$

$$\leq(1-\alpha_t)^2\|\varepsilon^t\|^2 + 2\alpha_t^2\sigma^2 + 4(1-\alpha_t)^2\mathbb{E}_{\xi^{t+1}}[\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^{t+1};\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^t;\xi^{t+1})\|^2 \mid \xi^{[t]}]$$

$$+ 4(1-\alpha_t)^2\mathbb{E}_{\xi^{t+1}}[\|\nabla_x \mathbf{\Phi}_\beta(x^{t+1},\lambda^t;\xi^{t+1}) - \nabla_x \mathbf{\Phi}_\beta(x^t,\lambda^t;\xi^{t+1})\|^2 \mid \xi^{[t]}]$$

$$\leq (1-\alpha_t)^2\|\varepsilon^t\|^2 + 2\alpha_t^2\sigma^2 + 4(1-\alpha_t)^2\|\nabla_x\Psi_\beta(x^{t+1},\lambda^{t+1}) - \nabla_x\Psi_\beta(x^{t+1},\lambda^t)\|^2 + 4(1-\alpha_t)^2L_\beta^2\|x^{t+1}-x^t\|^2$$

$$\leq (1-\alpha_t)^2\|\varepsilon^t\|^2 + 2\alpha_t^2\sigma^2 + 4(1-\alpha_t)^2(m\rho_t F_c G)^2 + 4(1-\alpha_t)^2L_\beta^2\|x^{t+1}-x^t\|^2,$$

where the second inequality follows from (19) and the definition of $\mathbf{\Phi}_\beta(x,\lambda;\xi)$, and the last inequality comes from (17). Then taking expectation with respect to $\xi^{[t]}$ on the both sides of the above inequality yields the result. □

3.1 Stationarity

We now analyze the stationarity measure, i.e. the left side of (10), at $x = x^{R+1}$ with λ defined componentwise by

$$\lambda_i = \begin{cases} \beta c_i(x^{R+1}) + \lambda_i^{R+1}, & i \in \mathcal{E}, \\ [\beta c_i(x^{R+1}) + \lambda_i^{R+1}]_+, & i \in \mathcal{I}. \end{cases} \tag{26}$$

In the following, we call a point $x \in X$ is δ-feasible to (1) if

$$\|c_{\mathcal{E}}(x)\|^2 + \|c_{\mathcal{I}}(x)\|^2 \leq \delta^2.$$

Theorem 1 *Under Assumptions 1-4, set $\beta = T^{1/4}$ and*

$$\rho_t \equiv \frac{\rho}{T},\ \eta_t \equiv \frac{\eta}{3L_\beta T^{1/4}},\ \alpha_t \equiv \frac{4\alpha\eta^2}{3T^{1/4}(3T^{1/4}-\eta)}, \quad \forall t \in [T]$$

with $\rho \in (0, T^{5/4}]$, $\eta \in (0, T^{1/4}]$, $\alpha \in [1, \frac{3T^{1/4}(3T^{1/4}-\eta)}{4\eta^2})$ are constants independent of T. If x^1 is δ-feasible with $\delta = \mathcal{O}(T^{-1/2})$, it holds that with λ defined through (26),

$$\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial\chi(x^{R+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla c_i(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(T^{-1/2}). \tag{27}$$

Proof. Summing up (20) over $t = 1,\ldots,T$ and then dividing it by T, we obtain

$$\frac{1}{T}\sum_{t=1}^T \mathbf{d}^2(\nabla_x\phi_\beta(x^{t+1},\lambda^{t+1}) + \partial\chi(x^{t+1}), -\mathcal{N}_X(x^{t+1})) \leq \frac{4(m\rho F_c G)^2}{T^2} + \frac{4}{T}\sum_{t=1}^T\|\varepsilon^t\|^2 + \frac{4}{T}\sum_{t=1}^T(L_\beta^2 + \frac{1}{\eta_t^2})\|x^{t+1}-x^t\|^2. \tag{28}$$

Let us consider the term related with $\|x^{t+1}-x^t\|^2$ in (28) first. It is easy to attain from Lemma 6 and $\eta_t = \frac{\eta}{3L_\beta T^{1/4}}$ that

$$\frac{L_\beta}{2}(\frac{3T^{1/4}}{\eta} - 1)\|x^{t+1}-x^t\|^2 = (\frac{1}{2\eta_t} - \frac{L_\beta}{2})\|x^{t+1}-x^t\|^2 \leq \mathcal{L}_\beta(x^t,\lambda^t) - \mathcal{L}_\beta(x^{t+1},\lambda^{t+1}) + \frac{\eta\|\varepsilon^t\|^2}{6L_\beta T^{1/4}} + m\rho_t F_c^2,$$

thus

$$\|x^{t+1}-x^t\|^2 \leq \frac{2\eta(\mathcal{L}_\beta(x^t,\lambda^t) - \mathcal{L}_\beta(x^{t+1},\lambda^{t+1}))}{L_\beta(3T^{1/4}-\eta)} + \frac{\eta^2\|\varepsilon^t\|^2}{3L_\beta^2 T^{1/4}(3T^{1/4}-\eta)} + \frac{2m\eta\rho_t F_c^2}{L_\beta(3T^{1/4}-\eta)} \tag{29}$$

which further yields

$$(L_\beta^2 + \frac{1}{\eta_t^2})\|x^{t+1}-x^t\|^2 = L_\beta^2(\frac{9T^{1/2}}{\eta^2} + 1)\|x^{t+1}-x^t\|^2 \leq \frac{2L_\beta(\frac{9T^{1/2}}{\eta} + \eta)}{3T^{1/4}-\eta}(\mathcal{L}_\beta(x^t,\lambda^t) - \mathcal{L}_\beta(x^{t+1},\lambda^{t+1}))$$

$$+\frac{9T^{1/2}+\eta^2}{3T^{1/4}(3T^{1/4}-\eta)}\|\varepsilon^t\|^2+\frac{2m\rho_t L_\beta F_c^2(\frac{9T^{1/2}}{\eta}+\eta)}{3T^{1/4}-\eta}.$$

Since $T\geq 1$, we have $\frac{9T^{1/2}+\eta^2}{T^{1/4}(3T^{1/4}-\eta)}\leq 5$. Then taking expectation with respect to $\xi^{[t]}$ on both sides of the above inequality and summing up over $t=1,\ldots,T$, we obtain from $\sum_{t=1}^T\rho_t\leq\rho$ that

$$\frac{4}{T}\sum_{t=1}^T(L_\beta^2+\frac{1}{{\eta_t}^2})\mathbb{E}_{\xi^{[t]}}[\|x^{t+1}-x^t\|^2]\leq\frac{40L_\beta}{\eta T^{3/4}}(\mathcal{L}_\beta(x^1,\lambda^1)-\mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})]+m\rho F_c^2)+\frac{20}{3T}\sum_{t=1}^T\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]. \tag{30}$$

We now focus on the second term in the right hand side of (28). By Lemma 7 and $0<\alpha_t<1$, we have

$$\begin{aligned}
\mathbb{E}_{\xi^{[t+1]}}[\|\varepsilon^{t+1}\|^2]\leq&(1-\alpha_t)^2\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]+2\alpha_t^2\sigma^2+4(1-\alpha_t)^2(m\rho_t F_c G)^2+4(1-\alpha_t)^2L_\beta^2\mathbb{E}_{\xi^{[t]}}[\|x^{t+1}-x^t\|^2]\\
\leq&(1-\alpha_t)^2(1+\frac{4\eta^2}{3T^{1/4}(3T^{1/4}-\eta)})\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]+2\alpha_t^2\sigma^2+\frac{4m^2\rho^2F_c^2G^2}{T^2}\\
&+\frac{8\eta L_\beta}{3T^{1/4}-\eta}\mathbb{E}_{\xi^{[t]}}[\mathcal{L}_\beta(x^t,\lambda^t)-\mathcal{L}_\beta(x^{t+1},\lambda^{t+1})]+\frac{8m\eta\rho L_\beta F_c^2}{T(3T^{1/4}-\eta)}\\
<&(1-\alpha_t)\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]+2\alpha_t^2\sigma^2+\frac{4m^2\rho^2F_c^2G^2}{T^2}\\
&+\frac{8\eta L_\beta}{3T^{1/4}-\eta}\mathbb{E}_{\xi^{[t]}}[\mathcal{L}_\beta(x^t,\lambda^t)-\mathcal{L}_\beta(x^{t+1},\lambda^{t+1})]+\frac{8m\eta\rho L_\beta F_c^2}{T(3T^{1/4}-\eta)},
\end{aligned}$$

where the second inequality follows from (29) and $(1-\alpha_t)^2<1$, and the last inequality is due to the definition of α_t. Summing the above inequality over $t=1,\ldots,T$ and using $\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]\leq\sigma^2$, we obtain

$$\sum_{t=1}^T\alpha_t\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]\leq(1+2\alpha_t^2T)\sigma^2+\frac{4m^2\rho^2F_c^2G^2}{T}+\frac{8\eta L_\beta}{3T^{1/4}-\eta}(\mathcal{L}_\beta(x^1,\lambda^1)-\mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})]+m\rho F_c^2).$$

Then dividing the whole inequality by $\sum_{t=1}^T\alpha_t$ and due to the constant setting of α_t for $t\in[T]$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]\leq&(\frac{3T^{1/4}(3T^{1/4}-\eta)}{4\alpha\eta^2T}+\frac{8\alpha\eta^2}{3T^{1/4}(3T^{1/4}-\eta)})\sigma^2+\frac{3m^2\rho^2F_c^2G^2(3T^{1/4}-\eta)}{\alpha\eta^2T^{7/4}}\\
&+\frac{6L_\beta}{\alpha\eta T^{3/4}}(\mathcal{L}_\beta(x^1,\lambda^1)-\mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})]+m\rho F_c^2).
\end{aligned}\tag{31}$$

Plugging (30) and (31) into (28) with the definition of λ in (26), we obtain

$$\begin{aligned}
&\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1})+\partial\chi(x^{R+1})+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla c_i(x^{R+1}),-\mathcal{N}_X(x^{R+1}))]\\
=&\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla_x\phi_\beta(x^{t+1},\lambda^{t+1})+\partial\chi(x^{t+1}),-\mathcal{N}_X(x^{t+1}))]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{4m^2\rho^2F_c^2G^2}{T^2} + \frac{40L_\beta}{\eta T^{3/4}}(\mathcal{L}_\beta(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})] + m\rho F_c^2) + \frac{32}{3T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] \\
&\leq \frac{4m^2\rho^2F_c^2G^2}{T^2} + \frac{104L_\beta}{\eta T^{3/4}}(\mathcal{L}_\beta(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})] + m\rho F_c^2) \\
&\quad + 32((\frac{T^{1/4}(3T^{1/4}-\eta)}{4\alpha\eta^2T} + \frac{8\alpha\eta^2}{9T^{1/4}(3T^{1/4}-\eta)})\sigma^2 + \frac{m^2\rho^2F_c^2G^2(3T^{1/4}-\eta)}{\alpha\eta^2T^{7/4}}).
\end{aligned} \tag{32}
$$

Now we concentrate on the upper bound of $\mathcal{L}_\beta(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})]$. Since $\lambda^1 = \mathbf{0}$, we have

$$
\begin{aligned}
&\mathcal{L}_\beta(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})] \\
&= f(x^1) + \chi(x^1) + \frac{\beta}{2}\|c_\mathcal{E}(x^1)\|^2 + \frac{\beta}{2}\|[c_\mathcal{I}(x^1)]_+\|^2 \\
&\quad - \mathbb{E}_{\xi^{[T]}}[f(x^{t+1}) + \chi(x^{t+1}) + \sum_{i\in\mathcal{E}}[\lambda_i^{T+1}c_i(x^{t+1}) + \frac{\beta}{2}c_i^2(x^{t+1})] + \sum_{i\in\mathcal{I}}\psi_\beta(c_i(x^{t+1}),\lambda_i^{T+1})] \\
&\leq f(x^1) + \chi(x^1) - C^* + \frac{\beta}{2}\|c_\mathcal{E}(x^1)\|^2 + \frac{\beta}{2}\|[c_\mathcal{I}(x^1)]_+\|^2 \\
&\quad - \mathbb{E}_{\xi^{[T]}}[\sum_{i\in\mathcal{E}}[\lambda_i^{T+1}c_i(x^{t+1}) + \frac{\beta}{2}c_i^2(x^{t+1})] + \sum_{i\in\mathcal{I}}\psi_\beta(c_i(x^{t+1}),\lambda_i^{T+1})] \\
&\leq f(x^1) + \chi(x^1) - C^* + \frac{\beta}{2}\|c_\mathcal{E}(x^1)\|^2 + \frac{\beta}{2}\|[c_\mathcal{I}(x^1)]_+\|^2 + \mathbb{E}_{\xi^{[T]}}[\sum_{i\in\mathcal{E}}|\lambda_i^{T+1}c_i(x^{t+1})| + \sum_{i\in\mathcal{I}}\frac{(\lambda_i^{T+1})^2}{2\beta}] \\
&\leq f(x^1) + \chi(x^1) - C^* + \frac{\beta}{2}\|c_\mathcal{E}(x^1)\|^2 + \frac{\beta}{2}\|[c_\mathcal{I}(x^1)]_+\|^2 + \sum_{i\in\mathcal{E}}\rho F^2 + \sum_{i\in\mathcal{I}}\frac{\rho^2F^2}{2\beta} \\
&\leq f(x^1) + \chi(x^1) - C^* + \frac{\beta}{2}(\|c_\mathcal{E}(x^1)\|^2 + \|[c_\mathcal{I}(x^1)]_+\|^2) + m\rho F_c^2,
\end{aligned} \tag{33}
$$

where C^* is the lower bound of the objective function value over X, the second inequality comes from $\psi_\beta(u,v) \geq -\frac{v^2}{2\beta}$, and the third inequality holds due to (14). It is worthy to note that the order of $\mathcal{L}_\beta(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})]$ is affected by the constraint violation of the initial point x^1. Since x^1 is δ-feasible with $\delta = T^{-1/8}$, we have

$$
\frac{\beta}{2}\|c_\mathcal{E}(x^1)\|^2 + \frac{\beta}{2}\|[c_\mathcal{I}(x^1)]_+\|^2 \leq \frac{\beta}{2}(\sum_{i\in\mathcal{E}}|c_i(x^1)| + \sum_{i\in\mathcal{I}}[c_i(x^1)]_+)^2 \leq \frac{\beta\delta^2}{2} = \frac{1}{2}
$$

which indicates

$$
\mathcal{L}_\beta(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_\beta(x^{T+1},\lambda^{T+1})] \leq f(x^1) + \chi(x^1) - C^* + \frac{1}{2} + m\rho F_c^2 = \mathcal{O}(1).
$$

Then substituting (33) into (32) yields

$$\begin{aligned}
&\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial\chi(x^{R+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i \nabla c_i(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] \\
\leq & \frac{4m^2\rho^2 F_c^2 G^2}{T^2} + \frac{104L_\beta}{\eta T^{3/4}}(f(x^1) + \chi(x^1) - C^* + \frac{1}{2} + 2m\rho F_c^2) \\
& + 32((\frac{T^{1/4}(3T^{1/4}-\eta)}{4\alpha\eta^2 T} + \frac{8\alpha\eta^2}{9T^{1/4}(3T^{1/4}-\eta)})\sigma^2 + \frac{m^2\rho^2 F_c^2 G^2(3T^{1/4}-\eta)}{\alpha\eta^2 T^{7/4}}) \\
= & \mathcal{O}(\frac{1}{T^2}) + \mathcal{O}(\frac{L_\beta}{T^{3/4}}) + \mathcal{O}(\frac{1}{T^{1/2}}) + \mathcal{O}(\frac{1}{T^{3/2}}),
\end{aligned}$$

which implies the conclusion from $L_\beta = L + m(\beta G^2 + \beta FL + \rho FL) = \mathcal{O}(\beta) = \mathcal{O}(T^{1/4})$. □

For a general initial point without the near-feasibility requirement, we can obtain the following corollary.

Corollary 1 *Under Assumptions 1-4, set $\beta = T^{1/5}$ and*

$$\rho_t \equiv \frac{\rho}{T},\ \eta_t \equiv \frac{\eta}{3L_\beta T^{1/5}},\ \alpha_t \equiv \frac{4\alpha\eta^2}{3T^{1/5}(3T^{1/5}-\eta)},\quad \forall k \in [T]$$

with $\rho \in (0, T^{6/5}]$, $\eta \in (0, T^{1/5}]$, $\alpha \in [1, \frac{3T^{1/5}(3T^{1/5}-\eta)}{4\eta^2})$ are constants independent of T. Then it holds that with λ defined through (26)*,*

$$\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial\chi(x^{R+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i \nabla c_i(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(T^{-1/2.5}). \tag{34}$$

3.2 Primal feasibility

As in general it is impossible to find a feasible solution for a nonconvex constrained optimization problem, an additional regularity assumption is necessary to guarantee the near-feasibility of a solution.

Assumption 5 *(**N**on**S**ingularity **C**ondition, NSC) There exists a parameter $\nu > 0$ such that*

$$\nu\sqrt{\|c_\mathcal{E}(x^t)\|^2 + \|[c_\mathcal{I}(x^t)]_+\|^2} \leq \mathbf{d}(\nabla c_\mathcal{E}(x^t)c_\mathcal{E}(x^t) + \nabla c_\mathcal{I}(x^t)[c_\mathcal{I}(x^t)]_+, -\mathcal{N}_X(x^t)),\quad \forall t \in [T+1]. \tag{35}$$

NSC is also assumed in Assumption 4 [22]. This condition and its variants have been used in analyzing the complexity of AL methods in recent years [18, 20, 22, 31]. As discussed in [22], NSC is closely related to the well-known Kurdyka-Lojasiewicz(KL) condition for the feasibility problem, i.e. minimizing the constraint violation of (1). And under this assumption we can find a δ-feasible initial point by applying existing algorithms [46]. From another perspective, NSC is similar to the condition used in [11], which can be seen as a stronger version of the linear independence constraint qualification (LICQ), i.e. the Jacobian of constraint function has singular values that are uniformly lower bounded by a positive real number over a set. The next theorem shows the near-feasibility of x^{R+1} under NSC.

Theorem 2 *Under conditions of Theorem 1 and Assumption 5, it holds that*

$$\mathbb{E}_{R;\xi^{[T]}}[\|c_\mathcal{E}(x^{R+1})\|^2 + \|[c_\mathcal{I}(x^{R+1})]_+\|^2] = \mathcal{O}(T^{-1/2}).$$

Proof. For simplicity we define

$$\tilde{\lambda}_i^t = \begin{cases} \beta c_i(x^t) + \lambda_i^t, & i \in \mathcal{E}, \\ [\beta c_i(x^t) + \lambda_i^t]_+, & i \in \mathcal{I}. \end{cases}$$

It is apparent that there exists a $v^{t+1} \in \partial\chi(x^{t+1})$ such that

$$\begin{aligned}&\mathbf{d}(\nabla f(x^{t+1}) + v^{t+1} + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1}))\\ =&\,\mathbf{d}(\nabla f(x^{t+1}) + \partial\chi(x^{t+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1})).\end{aligned}$$

Then Assumption 5 indicates that for any $t \in [T]$,

$$\begin{aligned}&\|c_{\mathcal{E}}(x^{t+1})\|^2 + \|[c_{\mathcal{I}}(x^{t+1})]_+\|^2\\ \le&\frac{1}{\nu^2}\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{t+1})c_{\mathcal{E}}(x^{t+1}) + \nabla c_{\mathcal{I}}(x^{t+1})[c_{\mathcal{I}}(x^{t+1})]_+, -\mathcal{N}_X(x^{t+1}))\\ \le&\frac{1}{\nu^2\beta^2}\mathbf{d}^2(\sum_{i\in\mathcal{E}}\beta c_i(x^{t+1})\nabla c_i(x^{t+1}) + \sum_{i\in\mathcal{I}}[\beta c_i(x^{t+1})]_+\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1}))\\ \le&\frac{4}{\nu^2\beta^2}[\mathbf{d}^2(\nabla f(x^{t+1}) + v^{t+1} + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1})) + \|\nabla f(x^{t+1})\|^2\\ &\qquad + \|v^{t+1}\|^2 + \sum_{i\in\mathcal{E}\cup\mathcal{I}} |\lambda_i^{t+1}|\|\nabla c_i(x^{t+1})\|^2]\\ \le&\frac{4}{\nu^2\beta^2}[\mathbf{d}^2(\nabla f(x^{t+1}) + \partial\chi(x^{t+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1})) + (2+m^2\rho^2F^2)G^2]. \qquad (36)\end{aligned}$$

Taking expectation with respect to R and $\xi^{[T]}$ on both sides of (36) and by the setting $\beta = T^{1/4}$ and λ in (26), we have

$$\begin{aligned}&\mathbb{E}_{R;\xi^{[T]}}[\|c_{\mathcal{E}}(x^{R+1})\|^2 + \|[c_{\mathcal{I}}(x^{R+1})]_+\|^2] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\,\|c_{\mathcal{E}}(x^{t+1})\|^2 + \|[c_{\mathcal{I}}(x^{t+1})]_+\|^2]\\ \le&\frac{4}{\nu^2\beta^2}\{(2+m^2\rho^2F^2)G^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla f(x^{t+1}) + \partial\chi(x^{t+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1}))]\}\\ =&\frac{4}{\nu^2\beta^2}\{(2+m^2\rho^2F^2)G + \mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial\chi(x^{R+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i\nabla c_i(x^{R+1}), -\mathcal{N}_X(x^{R+1}))]\}\\ =&\mathcal{O}(\beta^{-2}) = \mathcal{O}(T^{-1/2})\end{aligned}$$

which completes the proof. □

Corollary 2 *Under conditions of Corollary 1 and Assumption 5, it holds that*

$$\mathbb{E}_{R;\xi^{[T]}}[\|c_{\mathcal{E}}(x^{R+1})\|^2 + \|[c_{\mathcal{I}}(x^{R+1})]_+\|^2] = \mathcal{O}(T^{-1/2.5}).$$

Remark 2 *Without the NSC assumption, we cannot guarantee the near-feasibility of x^{R+1} in expectation, but following the proof in Theorem 2 it is easy to have*

$$\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{R+1})c_{\mathcal{E}}(x^{R+1}) + \nabla c_{\mathcal{I}}(x^{R+1})[c_{\mathcal{I}}(x^{R+1})]_+, -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(T^{-1/2}),$$

Hence x^{R+1} can be regarded as an approximate stationary solution of

$$\min_{x\in X} \quad \frac{1}{2}\|c_{\mathcal{E}}(x)\|^2 + \frac{1}{2}\|[c_{\mathcal{I}}(x)]_+\|^2.$$

3.3 Complementary slackness

In the following theorem we characterize the complementary slackness of x^{R+1} and λ defined in (26).

Theorem 3 *Under conditions of Theorem 1 and Assumption 5, it holds that*

$$\mathbb{E}_{R;\xi^{[T]}}[\sum_{i\in\mathcal{I}}\lambda_i|c_i(x^{R+1})|] = \mathcal{O}(T^{-1/4}),$$

where $\lambda_i = [\beta c_i(x^{R+1}) + \lambda_i^{R+1}]_+, i \in \mathcal{I}$.

Proof. First, in light of $|c_i(x^{t+1})| = [c_i(x^{t+1})]_+ + [c_i(x^{t+1})]_-$ for $i \in \mathcal{I}$, we will analyze these two parts respectively in the following. For simplicity, we define $\tilde{\lambda}_i^{t+1} = [\beta c_i(x^{t+1}) + \lambda_i^{t+1}]_+,\ \forall i \in \mathcal{I}$. Note that if $c_i(x^{t+1}) \le -\frac{\lambda_i^{t+1}}{\beta}$, $\tilde{\lambda}_i^{t+1} = 0$. Then we define index sets $\mathcal{I}_1 := \{i \mid c_i(x^{t+1}) \ge 0\}$ and $\mathcal{I}_2 := \{i \mid -\frac{\lambda_i^{t+1}}{\beta} \le c_i(x^{t+1}) < 0\}$. Thus we obtain

$$\begin{aligned}
\sum_{i\in\mathcal{I}}\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})| &= \sum_{i\in\mathcal{I}_1}[\beta c_i^2(x^{t+1}) + \lambda_i^{t+1}c_i(x^{t+1})] + \sum_{i\in\mathcal{I}_2}[-c_i(x^{t+1})(\beta c_i(x^{t+1}) + \lambda_i^{t+1})] \\
&\le \sum_{i\in\mathcal{I}_1}[\beta c_i^2(x^{t+1}) + \lambda_i^{t+1}c_i(x^{t+1})] + \sum_{i\in\mathcal{I}_2}\frac{|\lambda_i^{t+1}|^2}{\beta} \qquad (37) \\
&\le \beta\sum_{i\in\mathcal{I}_1}c_i^2(x^{t+1}) + \rho F\sum_{i\in\mathcal{I}_1}c_i(x^{t+1}) + \frac{m\rho^2F^2}{\beta},
\end{aligned}$$

where the first inequality comes from $-u(bu+a) \le \frac{a^2}{b}, \forall u \in \mathbb{R}$ with $b > 0$, the second inequality is obtained by (14), $|\lambda_i^{t+1}| \le \rho F,\ \forall i \in \mathcal{I}$. Then taking expectation with respect to R and $\xi^{[T]}$ on both sides of (37), with λ defined in (26) we obtain

$$\begin{aligned}
\mathbb{E}_{R;\xi^{[T]}}[\sum_{i\in\mathcal{I}}\lambda_i|c_i(x^{R+1})|] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[T]}}[\sum_{i\in\mathcal{I}}\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})|] \\
&\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[T]}}[\beta\sum_{i\in\mathcal{I}_1}c_i^2(x^{t+1}) + \rho F\sum_{i\in\mathcal{I}_1}c_i(x^{t+1})] + \frac{m\rho^2F^2}{\beta} \\
&= \beta\mathbb{E}_{R;\xi^{[T]}}[\|[c_{\mathcal{I}}(x^{R+1})]_+\|^2] + \rho F|\mathcal{I}|\mathbb{E}_{R;\xi^{[T]}}[\|[c_{\mathcal{I}}(x^{R+1})]_+\|] + \frac{m\rho^2F^2}{\beta} \\
&= \mathcal{O}(\beta^{-1}) = \mathcal{O}(T^{-1/4}),
\end{aligned}$$

where the third equality comes from Theorem 2 and $(\mathbb{E}[u])^2 \le \mathbb{E}[u^2]$ for a random variable $u \in \mathbb{R}$. □

Corollary 3 *Under conditions of Corollary 1 and Assumption 5, it holds that*

$$\mathbb{E}_{R;\xi^{[T]}}[\sum_{i\in\mathcal{I}}\lambda_i|c_i(x^{R+1})|] = \mathcal{O}(T^{-1/5}),$$

where $\lambda_i = [\beta c_i(x^{R+1}) + \lambda_i^{R+1}]_+, i \in \mathcal{I}$.

3.4 Oracle complexity

From previous Theorems 1-3, when the initial point is δ-feasible with $\delta = T^{-1/8}$, it is easy to obtain that to find an ϵ-stationary point of (1), the maximum iteration number of MLALM should be $T = \mathcal{O}(\epsilon^{-4})$. Recall that computation of the stochastic gradient at each iteration requires only one sample. Then the oracle complexity of Algorithm 1 is also bounded by $\mathcal{O}(\epsilon^{-4})$. Moreover, if without the near-feasibility requirement on the initial point, following Corollaries 1-3 we can obtain that the oracle complexity of MLALM to find an ϵ-stationary point is bounded by $\mathcal{O}(\epsilon^{-5})$. We next summarize the above results in the following theorem and corollary omitting the straightforward proofs.

Theorem 4 *Under conditions of Theorem 1 and Assumption 5, the oracle complexity of Algorithm 1 to find an ϵ-stationary point of* (1) *is bounded by $\mathcal{O}(\epsilon^{-4})$.*

Corollary 4 *Under conditions of Corollary 1 and Assumption 5, the oracle complexity of Algorithm 1 to find an ϵ-stationary point of* (1) *is bounded by $\mathcal{O}(\epsilon^{-5})$.*

4 Numerical Simulations

In this section, we would like to conduct some numerical experiments to demonstrate the practical performance of the proposed method MLALM, i.e. Algorithm 1. All the experiments were implemented in Matlab 2021b running on a 64-bit Linux machine with a 4.90 Ghz Intel Core i7-12700K CPU and 32GB of memory.

4.1 Quadratically Constrained Nonconvex Program

In this subsection, we test the proposed method on solving quadratically constrained nonconvex programs [18]:

$$\begin{aligned} \min_{x\in X} \quad & f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + \frac{1}{2}\|H_i x - c_i\|^2) \\ s.t. \quad & f_j(x) = \frac{1}{2}x^T Q_j x + a_j^T x \le b_j, \quad j = 1, \ldots, M, \end{aligned} \tag{38}$$

where $X = [-10, 10]^n$. For each $i \in [N]$, we generate $H_i \in \mathbb{R}^{p\times n}$ randomly with elements independently following the standard Gaussian distribution. For each $j \in [M]$, we generate diagonal matrices $Q_j \in \mathbb{R}^{n\times n}$ with elements uniformly and randomly chosen from the interval $[0.5, 1]$, i.e. following U$[0.5, 1]$ and vectors $a_j \sim$ U$[0.1, 1.1]^n$. Then we generate a random point $x_* \sim$ U$(0, 1)^n$ and set $c_i = H_i x_*$, $i \in [N]$, $b_j = \frac{1}{2}x_*^T Q_j x_* + a_j^T x_*$, $j \in [M]$. Note that x_* is feasible to (38) and $f(x_*) = 0$, so x_* is the optimal solution of (38).

We first detect the impact of α_t on the numerical performance in (6). We choose α_t from $\{0.1, 0.2, 0.4, 0.6, 0.8\}$. And for all of them, we set $n = 100$, $p = 5$, $N = M = 1000$, the initial point $x^1 = \mathbf{0}$, maximum iteration number $T = 2000$, penalty parameter $\beta = T^{1/4}$, $\rho_t = 6.6$ and stepsize $\eta_t \in \{0.05/T^{1/4}, 0.06/T^{1/4}\}$. Each subfigure in Figures 1 and 2 shows the average performance over 10 runs of the algorithm for each α_t. The left one in each figure reports the trend of the objective function value while the right one is for the constraint violation $\sum_{j=1}^{M}[f_j(x) - b_j]_+$. We can see that for each η_t, α_t with the best performance is often not the largest or smallest, but a certain intermediate value, both for the objective function value and for the constraint violation. Furthermore, comparing the best α_t in Figures 1 and 2, we see that as η_t increases, the value of the best α_t increases accordingly. To a certain degree, these observations verify the positive correlation between α_t and η_t as shown in the parameter settings of Theorem 1. On the other hand, when η_t increases, it is more often that larger α_t will show relatively better performances, seen from

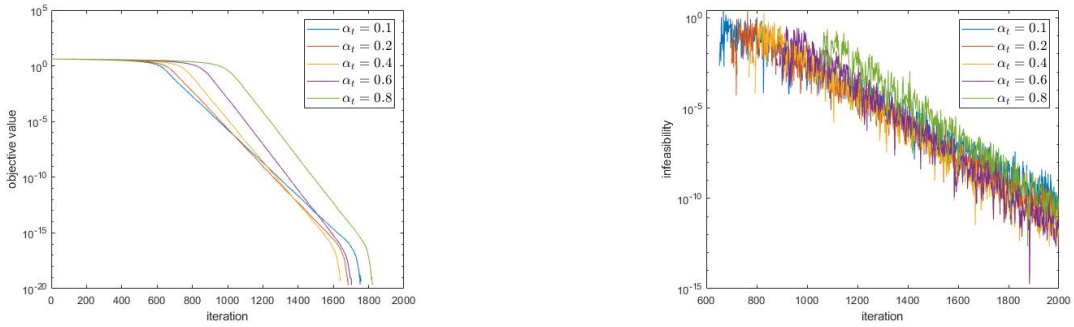


Figure 1: The impact of α_t on MLALM with $\eta_t = 0.05/T^{1/4}$

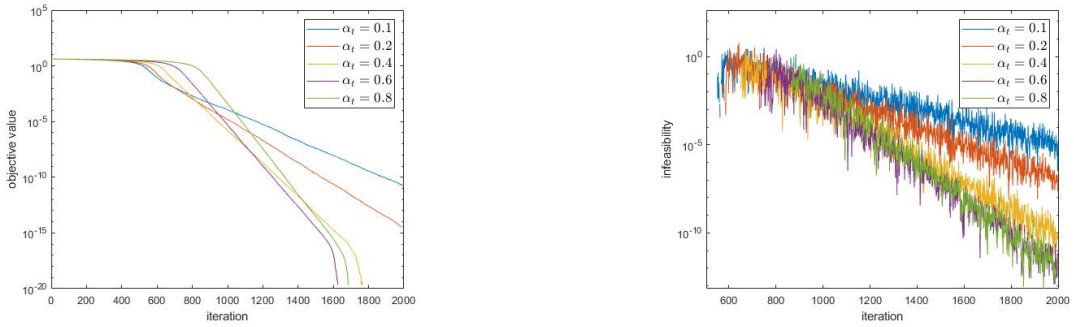


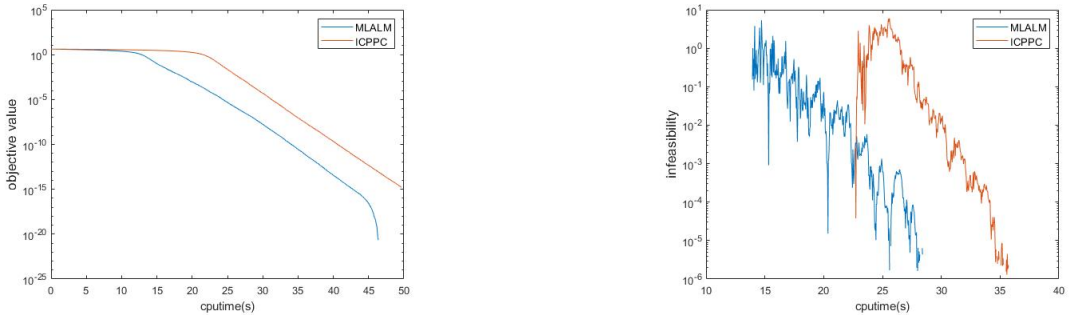
Figure 2: The impact of α_t on MLALM with $\eta_t = 0.06/T^{1/4}$

$\alpha_t = 0.1, 0.2$ in Figure 2. We can get some insight from theoretical perspective as shown in Theorem 1 that α_t has a positive lower bound, so it is not allowed for too small values in theoretical analysis.

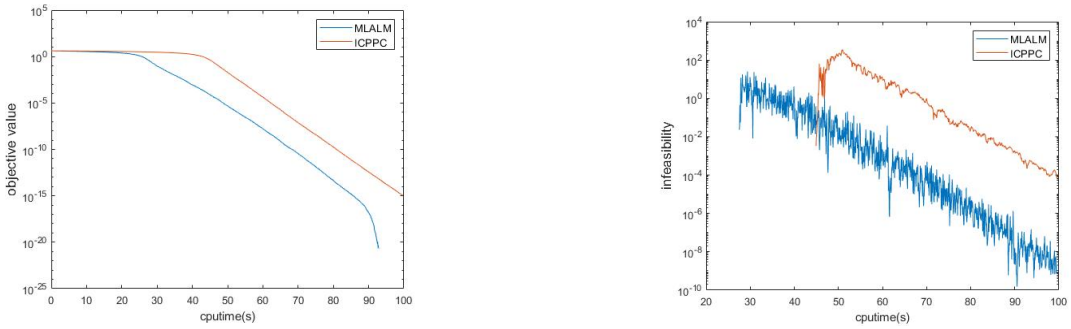
We next compare MLALM with ICPPC [4] for solving (38). We set $n \in \{100, 200\}$, $p = 5$, $N = 10000$, $M \in \{5000, 10000\}$. For both algorithms, we choose initial point $x^1 = \mathbf{0}$, maximum iteration number $T = 10000$. For ICPPC, we set $t_0 = 2$, $\mathcal{M} = 0.1M$ and $(\mu_0, n) = (50, 100), (100, 200)$. All other parameters are set as required by [4]. It is worthy to mention that we set the maximum inner-iteration number of ICPPC as 1 according to its good performance in numerical tests. For MLALM, we set $\eta_t = 0.05/t^{1/4}$ for $n = 100$, $\eta_t = 0.04/t^{1/4}$ for $n = 200$, $\alpha_t = 0.5$, $\beta = T^{1/4}$, $\rho_t = 10$. Figure 3 shows the performances of the two algorithms regarding objective function values and constraint violations on QCNP problems under different scenarios. All the results are reported with average over 10 runs of each algorithm. It can be observed that within the same CPU time MLALM can find the turning point to decrease objective function values faster than ICPPC, while achieving lower constraint violations along with the algorithm process.

4.2 Multi-class Neyman-Pearson classification problems

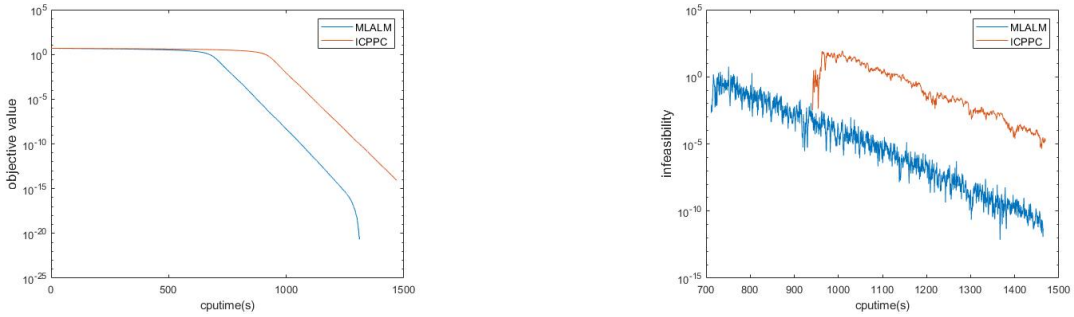
In this subsection, we consider multi-class Neyman-Pearson classification (mNPC) problems [22]. The mNPC problem focuses on learning K models x_k, $k \in [K]$ in order to predict the class of a potential data point ξ as $\arg\max_{k\in[K]} x_k^T \xi$. Specifically, the optimization problem is to minimize the loss on one class



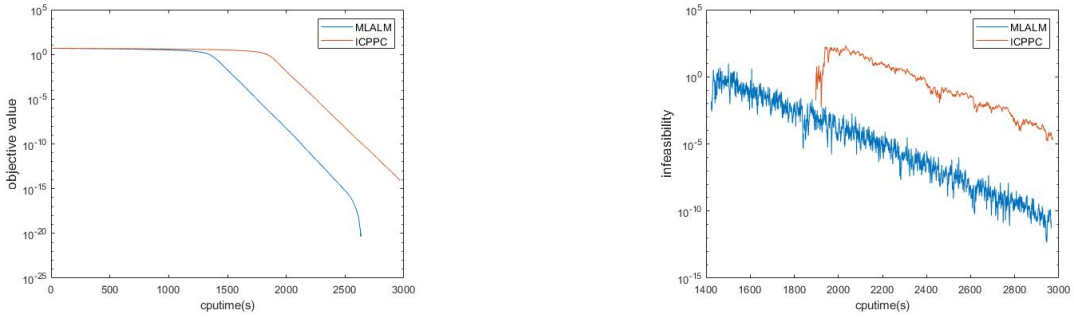
$n = 100, M = 5000$



$n = 100, M = 10000$



$n = 200, M = 5000$



$n = 200, M = 10000$

Figure 3: Comparison between MLALM and ICPPC for solving QCNP problems

while controlling its value on others:

$$\begin{aligned} \min_{\|x_k\|\le\lambda,\, k\in[K]} \quad & \frac{1}{|\mathcal{D}_1|}\sum_{l>1}\sum_{\xi\in\mathcal{D}_1} h(x_1^T\xi - x_l^T\xi) \\ s.t. \quad & \frac{1}{|\mathcal{D}_k|}\sum_{l\neq k}\sum_{\xi\in\mathcal{D}_k} h(x_k^T\xi - x_l^T\xi) \le \gamma_k, \quad k = 2,\ldots,K, \end{aligned} \tag{39}$$

where $h(z) = 1/(1+e^z)$ is the loss function and $\mathcal{D}_k$ represents the training data of the k-th class. We use two datasets from LibSVM [7]: *covtype* ($K = 7$) and *mnist* ($K = 10$). Besides, we set $\gamma_k = 0.5(K-1)$ and $\lambda = 0.3$.

We compare MLALM with IPPP [22]. Codes for IPPP were from authors of [22]. For MLALM, we set the maximum iteration number $T = 10000$ and $\alpha_t = 0.1$ for both datasets. For *covtype*, we set $\eta_t = 0.01/t^{1/4}$, $\beta = 5T^{1/4}$, $\rho_t = 10^{-3}$; For *mnist*, we set $\eta_t = 0.005/t^{1/4}$, $\beta = T^{1/4}$, $\rho_t = 10^{-5}$. Figures 4 and 5 show performances of these two algorithms for solving mNPC problems on each dataset, respectively. We can observe that the objective function value by MLALM decreases faster compared with IPPP, while the constraint violations are basically at the same level.

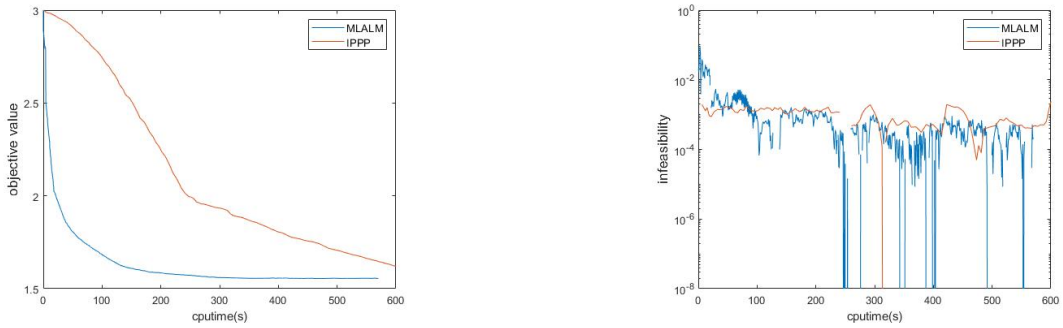


Figure 4: Comparison between MLALM and IPPP on dataset *covtype*

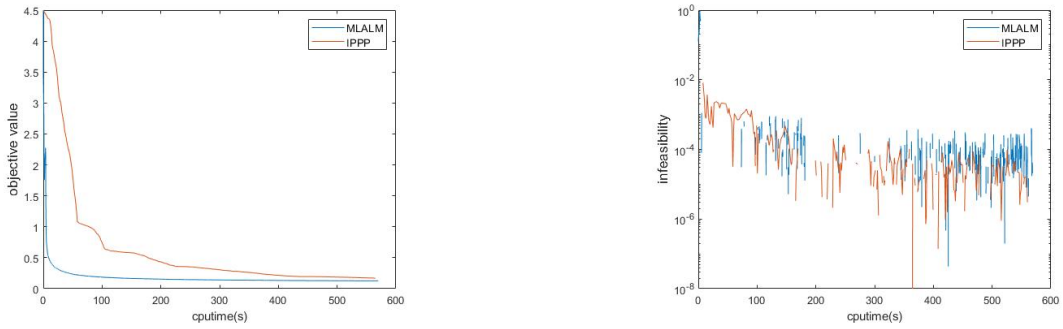


Figure 5: Comparison between MLALM and IPPP on dataset *mnist*

5 Conclusion

We study in this paper a single-loop linearized augmented Lagrangian method (MLALM) based on a variant of momentum for nonconvex optimization with an expectation function in the objective and with

deterministic functional constraints. To tackle the possibly nonconvex constraints, by leveraging the linearized AL function, we construct its stochastic approximation to update the primal variable, whose latest information is further used to update the dual variable. Furthermore, motivated by recent developments of momentum-type methods for unconstrained optimization, we incorporate a momentum step to compute the stochastic gradient and sample only once at each iteration. Under an initial near-feasibility condition and a nonsigularity condition, we establish that the oracle complexity to find an ϵ-stationary point is bounded by $\mathcal{O}(\epsilon^{-4})$, which matches the best complexity of current stochastic approximation methods for nonconvex constrained stochastic optimization. But different from those methods, MLALM merely solves much simpler subproblems and can handle problems with both equality and inequality constraints. Moreover, numerical experiments on two types of test problems show promising performances of our algorithm.

Acknowledgements

We would like to thank Dr. Qihang Lin and Dr. Yangyang Xu for kindly sharing their codes on IPPP [22].

References

[1] 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.

[2] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Math. Program.*, 171(1):217–282, 2018.

[3] D. Boob, Q. Deng, and G. Lan. Level constrained first order methods for function constrained optimization. *arXiv:2205.08011*, 2022.

[4] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Math. Program.*, pages 1436–4646, 2022.

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

[6] C. Cartis, N. I. Gould, and P. L. Toint. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Math. Program.*, 144(1):93–106, 2014.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[8] N. Chatterjee, Y.-H. Chen, P. Maas, and R. J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Am. Stat. Assoc.*, 111(513):107–117, 2016.

[9] C. Chen, F. Tung, N. Vedula, and G. Mori. Constraint-aware deep neural network compression. In *ECCV*, pages 400–415, 2018.

[10] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *arXiv:2112.14799*, 2021.

[11] F. E. Curtis, D. P. Robinson, and B. Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv:2107.03512*, 2021.

[12] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019.

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

[14] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NeurIPS*, volume 31. Curran Associates, Inc., 2018.

[15] C. J. Geyer. Constrained maximum likelihood exemplified by isotonic convex logistic regression. *J. Am. Stat. Assoc.*, 86(415):717–724, 1991.

[16] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

[17] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1):267–305, 2016.

[18] L. Jin and X. Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Comput. Optim. Appl.*, 2022.

[19] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NeurIPS*, volume 26. Curran Associates, Inc., 2013.

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

[21] Z. Li and Y. Xu. Augmented lagrangian based first-order methods for convex-constrained programs with weakly-convex objective. *INFROMS J. Optim.*, 3(4):373–397, 2021.

[22] 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(1):175–224, 2022.

[23] R. Ma, Q. Lin, and T. Yang. Proximally constrained methods for weakly convex optimization with weakly convex constraints. *arXiv:1908.01871*, 2019.

[24] P. Márquez-Neila, M. Salzmann, and P. Fua. Imposing hard constraints on deep networks: Promises and limitations. *arXiv:1706.02025*, 2017.

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

[26] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, volume 70, pages 2613–2621. PMLR, 2017.

[27] N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res.*, 21(110):1–48, 2020.

[28] S. N. Ravi, T. Dinh, V. S. Lokhande, and V. Singh. Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. In *AAAI*, volume 33, pages 4772–4779, 2019.

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

[30] S. K. Roy, Z. Mhammedi, and M. Harandi. Geometry aware constrained optimization techniques for deep learning. In *CVPR*, pages 4460–4469, 2018.

[31] M. F. Sahin, A. eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019.

[32] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 83(1):112–162, 2017.

[33] G. Scutari, F. Facchinei, and L. Lampariello. Parallel and distributed methods for constrained nonconvex optimization—part i: Theory. *IEEE Trans. Signal Process.*, 65(8):1929–1944, 2017.

[34] F. Shang, L. Jiao, K. Zhou, J. Cheng, Y. Ren, and Y. Jin. Asvrg: Accelerated proximal svrg. In *ACML*, volume 95, pages 815–830. PMLR, 2018.

[35] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[36] V. S. Tomar and R. C. Rose. Manifold regularized deep neural networks. In *INTERSPEECH*, 2014.

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

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

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

[40] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019.

[41] S. Wright and J. Nocedal. *Numerical optimization*. Springer, 2006.

[42] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, 24(4):2057–2075, 2014.

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

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

[45] Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *arXiv:2006.00425*, 2020.

[46] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *J. Sci. Comput.*, 72(2):700–734, 2017.

[47] Y. Zhu, N. Zabaras, P. S. Koutsourelakis, and P. Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.*, 394:56–81, 2019.