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

Qiankun Shi
School of Computer Science and Engineering Sun Yat-Sen University, Guangzhou, 510006, China;
Peng Cheng Laboratory, Shenzhen, 518066, China,
shiqk@mail2.sysu.edu.cn

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

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

July 17, 2024

Abstract

Nonconvex constrained stochastic optimization has emerged in many important application areas. Subject to general functional constraints it minimizes the sum of an expectation function and a nonsmooth regularizer. Main challenges arise due to the stochasticity in the random integrand and the possibly nonconvex functional constraints. To address these issues we propose a momentum-based linearized augmented Lagrangian method (MLALM). MLALM adopts a single-loop framework and incorporates a recursive momentum scheme to compute the stochastic gradient, which enables the construction of a stochastic approximation to the augmented Lagrangian function. We provide an analysis of global convergence of MLALM. Under mild conditions and with unbounded penalty parameters, we show that the sequences of average stationarity measure and constraint violations are convergent in expectation. Under a constraint qualification assumption the sequences of average constraint violation and complementary slackness measure converge to zero in expectation. We also explore properties of those related metrics when penalty parameters are bounded. Furthermore, we investigate oracle complexities of MLALM in terms of total number of stochastic gradient evaluations to find an ϵ-stationary point and an ϵ-KKT point when assuming the constraint qualification. 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, global convergence, oracle complexity

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

1 Introduction

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

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

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. For any fixed $\xi \in \Xi$, $\mathbf{F}(\cdot;\xi) : \mathbb{R}^n \to \mathbb{R}$ and $c_i(\cdot) : \mathbb{R}^n \to \mathbb{R}, i \in \mathcal{E}\cup\mathcal{I}$, are continuously

differentiable and possibly nonconvex, and $h : \mathbb{R}^n \to \mathbb{R}$ is proper, lower-semicontinuous and convex. Without loss of generality we presume that both $\mathcal{E}$ and $\mathcal{I}$ are nonempty, the feasible set $\{x \in X : c_i(x) = 0, i \in \mathcal{E}; c_i(x) \leq 0, i \in \mathcal{I}\}$ is nonempty, and the objective function value of (1) over X is lower bounded by C^*. For problem (1), it can be expensive to compute the expectation or the distribution of ξ may not be expressed explicitly. Thus the exact function or gradient information of f can be hard to obtain. This type of problems widely appear in various application fields. For example, in deep learning, constraints are imposed on output of the deep neural networks [32] to enforce specific behaviors or properties, such as physics-constrained deep learning model [57], constraint-aware deep neural network compression [10], manifold regularized deep learning [39, 45]. Some recent study has also highlighted the advantages of incorporating various constraints when training deep neural networks [28, 37]. Other applications include, but not limited to, portfolio allocation [3, 44], two/multi-stage modeling [3, 44] and constrained maximum likelihood estimation [9, 17].

The past decade has witnessed great developments in nonconvex stochastic optimization. Since Ghadimi and Lan [18] proposed randomized SGD methods for unconstrained nonconvex optimization, a surge of works have emerged in this area of research. However, due to stochastic variances of approximate gradients, SGD methods often suffer from slow convergence [6]. To address this issue, several types of variance reduction techniques have been proposed. Related works include SAG [41], SAGA [15], SVRG [23], SARAH [33] and SPIDER [16]. Moreover, proximal variants aiming for stochastic composite optimization have also been studied [19, 52, 43, 34, 50]. 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 [14] attracts attention, in which only one sample is required to estimate the gradient at each iteration. Later a proximal variant was studied in [56] for nonconvex stochastic composite problems. Under the mean-squared smoothness condition, the aforementioned 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.

Nonconvex optimization with general functional constraints can be challenging since the feasibility to these constraints can be hard to maintain. Nonconvex constrained optimization in deterministic settings has been studied for decades [51]. Penalty methods and sequential quadratic programming (SQP) methods are two of most effective approaches for general constrained optimization. Penalty methods normally transform the original constrained problem into 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 [51], linearized AL methods [48, 49], have thus been 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, numerical methods for nonconvex constrained optimization with complexity analysis have been widely studied in the past ten years, including [7, 24, 25, 26, 40, 42, 55].

For general functional constrained optimization in stochastic settings, such as (1), the main concerns lie in that computing the exact gradient information of the expectation function can be expensive, sometimes even prohibitive, and that maintaining the feasibility to general constraints can be challenging. Proximal point methods [5, 4, 27] 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. A level constrained proximal gradient (LCPG) method is developed in [5] for deterministic constrained optimization, by constructing a sequence of relatively easier subproblems with an increasing constraint level. The authors also extend LCPG method to stochastic (LCSPG) and variance-reduced (LCSVRG) variants when the objective takes either expectation or finite-sum form. [47] 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 [2, 11, 12] for equality constrained stochastic optimization, with complexity analysis provided in [11]. Based on the linearized AL function, [54] studies a single-loop primal-dual stochastic gradient method (PDSG) for solving convex stochastic optimization problems. [21] extends PDSG and proposes a single-loop stochastic primal-dual (SPD) method for non-

Algorithm	Problem	Stationarity measure	Assumptions	Comp.
SPD [21]	$\min_{x\in X}\ f(x)+h(x)$ s.t. $c_i(x)\le 0, i\in\mathcal{I}$	$\mathbb{E}[\mathbf{d}(\nabla f(x)+\partial h(x)+\sum_{i\in\mathcal{I}}\lambda_i\nabla c_i(x), -\mathcal{N}_X(x))]\le\epsilon$, $\frac{1}{\lvert\mathcal{I}\rvert}\mathbb{E}[\sum_{i\in\mathcal{I}}[c_i(x)]_+]\le\epsilon$, $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)\rvert]\le\epsilon$	nonsingularity	$\mathcal{O}(\epsilon^{-6})$ $\mathcal{O}(\epsilon^{-5})$ (if initial feasible)
SSQP [11]	$\min_{x\in\mathbb{R}^n}\ f(x)$ s.t. $c_i(x)=0, i\in\mathcal{E}$	$\mathbb{E}[\lVert\nabla f(x)+\sum_{i\in\mathcal{E}}\lambda_i^{true}\nabla c_i(x)\rVert]\le\epsilon$, $\mathbb{E}[\sqrt{\sum_{i\in\mathcal{E}}\lvert c_i(x)\rvert}]\le\epsilon$	strong LICQ	$\tilde{\mathcal{O}}(\epsilon^{-4})$ $\mathcal{O}(\epsilon^{-4})$ (if $\tau_{\min}$ known)
ICPPC [4]	$\min_{x\in X}\ f(x)+h(x)$ s.t. $c_i(x)+h_i(x)\le 0, i\in\mathcal{I}$	$\mathbb{E}[\lVert x-\hat{x}\rVert^2]\le\epsilon^2$ with $\hat{x}$ feasible, $\mathbb{E}[\mathbf{d}^2(\nabla f(\hat{x})+\partial h(\hat{x})+\sum_{i\in\mathcal{I}}\lambda_i(\nabla c_i(\hat{x})+\partial h_i(\hat{x})), -\mathcal{N}_X(\hat{x}))]\le\epsilon^2$, and $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(\hat{x})+h_i(\hat{x})\rvert]\le\epsilon^2$	strong feasibility	$\mathcal{O}(\epsilon^{-4})$
LCSPG [5]	$\min_{x\in X}\ f(x)+h(x)$ s.t. $c_i(x)+h_i(x)\le 0, i\in\mathcal{I}$	$\mathbb{E}[\mathbf{d}^2(\nabla f(x)+\partial h(x)+\sum_{i\in\mathcal{I}}\lambda_i(\nabla c_i(x)+\partial h_i(x)), 0)]\le\epsilon^2$, and $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)+h_i(x)\rvert]\le\epsilon^2$, where x is feasible and $\lambda\ge 0$	uniform MFCQ	$\mathcal{O}(\epsilon^{-4})$
MLALM (this paper)	$\min_{x\in X}\ f(x)+h(x)$ s.t. $c_i(x)=0, i\in\mathcal{E}$ $c_i(x)\le 0, i\in\mathcal{I}$	$\mathbb{E}[\mathbf{d}^2(\nabla f(x)+\partial h(x)+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla c_i(x), -\mathcal{N}_X(x))]\le\epsilon^2$, $\mathbb{E}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x)c_{\mathcal{E}}(x)+\nabla c_{\mathcal{I}}(x)[c_{\mathcal{I}}(x)]_+, -\mathcal{N}_X(x))]\le\epsilon^2$	mean-squared smoothness	$\mathcal{O}(\epsilon^{-4})$ $\mathcal{O}(\epsilon^{-3})$ (if initial $\sqrt{\epsilon}$-feasible)
		$\mathbb{E}[\mathbf{d}^2(\nabla f(x)+\partial h(x)+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla c_i(x), -\mathcal{N}_X(x))]\le\epsilon^2$, $\mathbb{E}[\lVert c_{\mathcal{E}}(x)\rVert^2+\lVert[c_{\mathcal{I}}(x)]_+\rVert^2]\le\epsilon^2$, $\mathbb{E}[\sum_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)\rvert]\le\epsilon$	mean-squared smoothness constraint qualification	$\mathcal{O}(\epsilon^{-4})$ $\mathcal{O}(\epsilon^{-3})$ (if initial $\sqrt{\epsilon}$-feasible)

Table 1: Comparison between algorithms for nonconvex constrained optimization, where $f(x)=\mathbb{E}_\xi[\mathbf{F}(x;\xi)]$, h and $h_i, i\in\mathcal{I}$ are convex but possibly nonsmooth, $\lambda_i, i\in\mathcal{I}$ are nonnegative, λ^{true} is a vector of Lagrange multipliers corresponding to x, $\tau_{\min}$ is the merit parameter threshold in SSQP, and "initial $\sqrt{\epsilon}$-feasible" means the initial point satisfies $\|c_{\mathcal{E}}(x^1)\|^2+\|[c_{\mathcal{I}}(x^1)]_+\|^2\le\epsilon$, while "initial feasible" means the initial point is feasible to (1). The "nonsingularity" condition refers to Assumption 3.1 in [21]. The "uniform MFCQ" condition for LCSPG requires all the feasible points of the considered problem satisfy MFCQ. The "strong LICQ" condition for SSQP represents that the Jacobian of constraint functions have singular values that are lower bounded away from zero over a set containing all iterates for all realizations of the random variable. The "mean-squared smoothness" and "constraint qualification" condition for MLALM refer to Assumptions 3 and 5 in this paper, respectively.

convex problems with a large number of functional constraints. [22] proposes a stochastic nested primal-dual method for a class of nonconvex constrained composition optimization whose objective is a composition of two expected-value functions.

1.1 Contributions

Our main contributions in this paper are summarized as follows.

1. We propose a Momentum-based Linearized Augmented Lagrangian Method (MLALM) for solving nonconvex constrained stochastic optimization problem (1). The presence of potentially nonconvex constraints poses challenges in finding feasible solutions. To cope with this issue, we adopt the idea of the linearized augmented Lagrangian (AL) function. This approach allows us to propose a single-loop algorithm framework and simplifies the subproblem at each iteration significantly, in contrast to double-loop algorithms like proximal point methods [4, 27]. The integration of the momentum technique within a single-loop algorithm framework is motivated by the prior work [21]. The SPD method proposed in [21] is based on the linearized AL function and aims for nonconvex constrained optimization with a large number of functional constraints. However, it requires large sampling sizes to compute stochastic gradients, resulting in relatively higher total oracle complexity to find an approximate solution, even when the initial iterate is feasible. To mitigate this issue, we employ a recursive momentum technique that only necessitates a small sampling size at each iteration, effectively controlling the variances of stochastic gradients.

2. We investigate the global convergence properties of MLALM. Our analysis reveals that, as the penalty parameter tends to infinity, the sequence of average stationarity measure in expectation converges to zero, and

the average constraint violation sequence also exhibits convergence (refer to Theorem 1). Under a constraint qualification assumption (Assumption 5), we establish that the sequences of average constraint violation and average complementary slackness measure converge to zero (refer to Theorem 2). Additionally, we analyze the properties of MLALM when penalty parameters are bounded (refer to Theorem 3). In contrast to recent research on stochastic SQP methods for inequality constrained optimization, such as [29, 30, 31], which necessitate probabilistic conditions on the accuracy of gradient estimates, the related assumptions enforced in this paper can be considerably more relaxed (refer to Assumptions 3 and 4). Another closely related study, [13], investigates the convergence properties of an adaptive stochastic SQP algorithm for problems with deterministic equality and inequality constraints. However, the analysis in [13] relies on the occurrence of an event where the merit parameter sequence eventually becomes sufficiently small yet remains bounded away from zero.

3. We conduct an oracle complexity analysis for MLALM. Under certain conditions, through analyzing the measure of the output in terms of stationarity, constraint violations and complementary slackness, we show that the oracle complexities of MLALM to find an ϵ-stationary point (Definition 1) and to find an ϵ-KKT point (Definition 2) under the constraint qualification are both in order $\mathcal{O}(\epsilon^{-4})$. If the initial point is nearly feasible, previous orders can be reduced to $\mathcal{O}(\epsilon^{-3})$. The SPD method proposed in [21] achieves an ϵ-KKT point with an oracle complexity of $\mathcal{O}(\epsilon^{-6})$ (resp. $\mathcal{O}(\epsilon^{-5})$) without (resp. with) requiring an initial-feasibility condition. Both SPD and MLALM adopt the idea of using a linearized augmented Lagrangian function. However, the incorporation of momentum in MLALM enables us to achieve improved oracle complexities under the same problem settings and under the mean-squared smoothness assumption. The stochastic SQP method proposed in [11], which is designed for equality constrained optimization, relies on an adaptive strategy to update merit parameters and assumes prior knowledge of Lipschitz constants for the objective and constraint gradients. This assumption poses a challenge for the direct application of stochastic SQP to nonsmooth problems. Similarly, the ICPPC algorithm [4] and LCSPG [5], designed for inequality constrained optimization, requires a strong feasibility assumption, depending on the availability of a strictly feasible point. The algorithm framework and theoretical analysis presented in [4, 5] and [11] are specifically tailored for problems with only inequality or equality constraints. In contrast, MLALM aims for more general problems. A more detailed comparison between MLALM, stochastic SQP (SSQP), ICPPC and LCSPG is provided in Table 1.

4. We present the numerical performance analysis of the proposed algorithm MLALM on two problem classes: quadratically constrained nonconvex programs (QCNPs) and multi-class Neyman-Pearson classification problems (mNPCs). We first investigate the impact of the recursive momentum on QCNPs, to better understand how the introduction of momentum affects the algorithm's performance. We then compare MLALM with ICPPC [4] and LCSVRG (a variance-reduced variant of LCSPG) for QCNPs and with SPD [21] and ICPPC for mNPCs. Numerical results reveal that the use of momentum brings benefits and delivers competitive performance.

1.2 Notation and preliminaries

We use $\|\cdot\|$ to denote the Euclidean norm of a vector without any specification. For brevity, we introduce $[k] := \{1, \ldots, k\}$ 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 $[\cdot]_-$, respectively. The gradient of f at x is denoted by $\nabla f(x)$. With 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, given random variables ξ and ζ, $\mathbb{E}_\xi[\cdot]$ represents the expectation with respect to ξ and $\mathbb{E}_\xi[\cdot \mid \zeta]$ represents the expectation conditioned on ζ. The inner product of $x, y \in \mathbb{R}^n$ is denoted by $\langle x, y \rangle$. The normal cone to a closed convex set X at a point $\bar{x} \in X$ is defined as

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

And its dual cone is denoted by $\mathcal{N}_X^*(\bar{x})$. Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, lower-semicontinuous and convex. The set of subgradient of h at $x \in \mathrm{dom}(h)$ is defined as

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

In general, finding a global or even a local minimizer for nonconvex constrained optimization can be NP-hard. Efforts are thus devoted to seeking more trackable solutions. Under certain constraint qualification, a local minimizer of (1) satisfies necessary conditions, e.g. KKT conditions. A point satisfying these conditions is called a KKT point. We assume in this paper that, there exist a KKT point $x^* \in X$ and a vector $\lambda^* \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ with $\lambda_i^* \geq 0,\ i \in \mathcal{I}$, such that the KKT conditions are satisfied:

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

In practical computations, however, it is inevitable that the iteration may be trapped at an infeasible stationary point of the problem:

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

From the optimality condition for (2), the following stationary holds:

$$\mathbf{d}(\nabla c_{\mathcal{E}}(x)c_{\mathcal{E}}(x) + \nabla c_{\mathcal{I}}(x)[c_{\mathcal{I}}(x)]_+, -\mathcal{N}_X(x)) = 0.$$

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

Assumption 1 *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 h is proper, lower semicontinuous and convex over X. Moreover, the objective function value of (1) over X is lower bounded by C^*.*

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

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

Assumption 3 *$\mathbf{F}(\cdot;\xi)$ is continuously differentiable for each $\xi \in \Xi$ 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, 5, 21]. Assumption 3 refers to the mean-squared smoothness condition, also known for L-average smoothness, is widely used in stochastic variance reduction-based methods [46, 56, 1]. It is slightly stronger than Lipschitz continuity of the expected value function f by Jensen's inequality. In [30, Assumption 3], a similar yet stronger assumption to Assumption 3 is made, where it requires $\mathbf{F}$ be thrice differentiable and that $\nabla^2\mathbf{F}(x,\xi)$ be uniformly bounded over a set containing all iterates and for all ξ. Besides, all constraint functions are assumed thrice continuously differentiable in [30].*

1.3 Outline

The rest of this paper is organized as follows. In Section 2 we introduce a momentum-based linearized augmented method for solving nonconvex constrained stochastic optimization problem (1). In Section 3 we present auxiliary lemmas that are required in subsequent sections. In Section 4 we investigate global convergence properties of the proposed algorithm. In Section 5 we establish oracle complexities to find an ϵ-stationary point and an ϵ-KKT point, respectively. In Section 6 we report some numerical experimental results and finally we give some conclusional remarks.

2 Momentum-based linearized augmented Lagrangian method

As is well-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 associated with problem (1) is in the form as described in [38]:

$$\mathcal{L}_\beta(x,\lambda) := \phi_\beta(x,\lambda) + h(x),$$

where $\beta > 0$ 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 is easy to check that

$$\nabla_x \Psi_\beta(x,\lambda) = \sum_{i\in\mathcal{E}}(\lambda_i + \beta c_i(x))\nabla c_i(x) + \sum_{i\in\mathcal{I}}[\lambda_i + \beta c_i(x)]_+ \nabla c_i(x). \tag{3}$$

Different from classical AL methods which try to minimize the AL function in the inner-loop, the linearized AL methods [48, 49, 55] construct a much simpler subproblem that minimizes an approximation to the AL function around current iterate x:

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

where $\eta > 0$. 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. As a result, we obtain a stochastic gradient $\nabla \mathbf{\Phi}_\beta(x,\lambda;\xi)$, where

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

The SPD method [21] also adopts the linearized AL function to construct subproblems, but it requires large batch sizes when computing mini-batch stochastic gradients in order to derive desired iteration complexity. A natural way to reduce the total oracle complexity is to try adopting smaller batch size per iteration.

Before proceeding, let us consider the problem of minimizing a continuously differentiable function $f(x) = \mathbb{E}[\mathbf{F}(x\,;\xi)]$ over $\mathbb{R}^n$ with $\xi \in \Xi$. Recall Nesterov's accelerated gradient approach, which reads

$$x^{t+1} = x^t - \eta_t s^t; \quad s^{t+1} = a_t s^t + b_t \nabla f(x^{t+1} - a_t s^t), \quad t \geq 1,$$

with $s^1 = \nabla f(x^1)$. Nevertheless, since the exact gradient of f cannot be accessed, we have to turn to its stochastic approximation by randomly picking one sample ξ^{t+1}:

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

where the above expression 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 [14] and its variant [56]:

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

As shown in [14, 56], under the mean-squared smoothness assumption momentum-based approaches can help to reduce the oracle complexity of SGD and proximal SGD methods. Motivated by this, we extend the idea to the general constrained optimization problem (1). Let $\{\beta_t\}$ be a sequence of penalty parameters for $t \geq 1$. We define

$$d^t = \begin{cases} \frac{1}{|\mathcal{J}_t|}\sum_{j\in\mathcal{J}_t} \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^t), & t = 1, \\ \frac{1}{|\mathcal{J}_t|}\sum_{j\in\mathcal{J}_t} \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^t) + (1-\alpha_{t-1})(d^{t-1} - \frac{1}{|\mathcal{J}_t|}\sum_{j\in\mathcal{J}_t} \nabla_x \mathbf{\Phi}_{\beta_{t-1}}(x^{t-1},\lambda^{t-1};\xi_j^t)), & t \geq 2, \end{cases} \tag{4}$$

where $\{\xi_j^t, j \in \mathcal{J}_t\}$ is a batch of samples selected randomly uniformly and independently from Ξ. Obviously, d^t is an approximation to $\nabla\phi_{\beta_t}(x^t, \lambda^t)$. The first term in the second line of (4) is a stochastic gradient estimated at x^t, while the second term is the difference between d^{t-1} and a stochastic gradient estimated at x^{t-1}. This correction aims to improve the accuracy of the stochastic approximation. We define the error

$$\varepsilon^t := d^t - \nabla_x \phi_{\beta_t}(x^t, \lambda^t), \quad t \geq 1.$$

When $\alpha_{t-1} = 1$, the expression for d^t reduces to the gradient approximation in the vanilla mini-batch SGD method. In this paper, we choose $\alpha_t \in (0, 1)$ for $t \geq 1$. To a certain extent, α_{t-1} can replace the batch $\mathcal{J}_t$ to reduce variance, thereby reducing oracle complexity with a smaller batch. Based on the stochastic approximation d^t, as defined in (4), we propose the following scheme to update the primal variable:

$$x^{t+1} = \arg\min_{x \in X}\{\langle d^t, x\rangle + h(x) + \frac{1}{2\eta_t}\|x - x^t\|^2\}, \tag{5}$$

where $\eta_t > 0$, $t \geq 1$. To expect λ more trackable we propose to update λ^t through:

$$\lambda_i^{t+1} = \lambda_i^t + \begin{cases} \rho_t c_i(x^{t+1}), & i \in \mathcal{E}, \\ \rho_t \max\{-\frac{\lambda_i^t}{\beta_t}, c_i(x^{t+1})\}, & i \in \mathcal{I}, \end{cases} \tag{6}$$

where $\rho_t \in (0, \beta_t)$.

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

Require: $x^1 \in \mathbb{R}^n$, $\lambda^1 \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ with $\lambda_i^1 \geq 0, i \in \mathcal{I}$, a non-decreasing positive sequence $\{\beta_t\}$, and parameters $\{\eta_t > 0\}$, $\{\rho_t \in (0, \beta_t)\}$, $\{\alpha_t \in (0, 1)\}$.

1: **for** $t = 1, 2, \ldots$ **do**
2: Calculate d^t through (4).
3: Calculate x^{t+1} through (5).
4: Calculate λ^{t+1} through (6).
5: **end for**

3 Auxiliary lemmas

In this section, we introduce auxiliary lemmas that will be useful in subsequent sections for global convergence and oracle complexity analysis. Let $\{x^t\}$ and $\{\lambda^t\}$ be 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 $\lambda_i^1 \geq 0$, $\forall i \in \mathcal{I}$, $\rho_t \in (0, \beta_t)$ and (6). □

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

$$|\lambda_i^t| \leq |\lambda_i^1| + C\sum_{k=1}^{t-1}\rho_k, \quad \forall i \in \mathcal{E}; \quad \lambda_i^t \leq \lambda_i^1 + C\sum_{k=1}^{t-1}\rho_k, \quad \forall i \in \mathcal{I}, \tag{7}$$

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

$$|\lambda_i^{t+1} - \lambda_i^t| \leq \rho_t\tilde{C}, \quad \forall i \in \mathcal{E}\cup\mathcal{I}, \tag{8}$$

where $\tilde{C} := \max(\frac{|\lambda_i^1| + C\sum_{k=1}^{\infty}\rho_k}{\beta_1}, C)$ and $\sum_{k=1}^{0}\rho_k := 0$.

Proof. Firstly, by applying $\lambda^1 = \mathbf{0}$ and (6), we have that for any $t \geq 2$,

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

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

where $\mathcal{K} = \{k \in [t-1] \mid \lambda_i^{k+1} \ge \lambda_i^k\}$. We thus obtain (7). Subsequently, for any $i \in \mathcal{E}$, it is easy to obtain

$$|\lambda_i^{t+1} - \lambda_i^t| \le \rho_t |c_i(x^{t+1})| \le \rho_t \tilde{C}.$$

Then, for any $i \in \mathcal{I}$, it follows from (6), Lemma 1, Assumption 2 and (7) that

$$|\lambda_i^{t+1} - \lambda_i^t| = \rho_t |\max(-\frac{\lambda_i^t}{\beta_t}, c_i(x^{t+1}))| \le \begin{cases} \rho_t C, & \text{if } c_i(x^{t+1}) \ge 0, \\ \frac{\rho_t \lambda_i^t}{\beta_t} \le \rho_t \tilde{C}, & \text{otherwise}, \end{cases}$$

which yields (8). □

Lemma 3 *Under Assumptions 1-2, it holds that for any $t \ge 1$ and $\beta \ge \beta_1$,*

$$|\psi_\beta(c_i(x^{t+1}), \lambda_i^{t+1}) - \psi_\beta(c_i(x^{t+1}), \lambda_i^t)| \le \rho_t \tilde{C}^2, \quad i \in \mathcal{I}, \tag{9}$$

and

$$\|\nabla_x \Psi_\beta(x^{t+1}, \lambda^{t+1}) - \nabla_x \Psi_\beta(x^{t+1}, \lambda^t)\| \le m\rho_t \tilde{C} G, \tag{10}$$

where $m := |\mathcal{E} \cup \mathcal{I}|$.

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

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

Firstly, we consider the case when $\beta c_i(x^{t+1}) + \lambda_i^t < 0$. By Lemma 1 and (6), we have $\lambda_i^{t+1} \le \lambda_i^t$. Thus $\beta c_i(x^{t+1}) + \lambda_i^{t+1} < 0$. Hence, the following relations hold true:

$$\begin{aligned} |\psi_\beta(c_i(x^{t+1}), \lambda_i^{t+1}) - \psi_\beta(c_i(x^{t+1}), \lambda_i^t)| = \frac{(\lambda_i^t)^2 - (\lambda_i^{t+1})^2}{2\beta} &= \frac{\lambda_i^t + \lambda_i^{t+1}}{2\beta} |\lambda_i^{t+1} - \lambda_i^t| \\ &\le \frac{\lambda_i^1 + C\sum_{k=1}^t \rho_k}{\beta_1} |\lambda_i^{t+1} - \lambda_i^t| \\ &\le \rho_t \tilde{C}^2. \end{aligned}$$

Secondly, when $\beta c_i(x^{t+1}) + \lambda_i^t \ge 0$, we have $-\frac{\lambda_i^t}{\beta} \le c_i(x^{t+1}) \le C$, then (9) is derived obviously if $\beta c_i(x^{t+1}) + \lambda_i^{t+1} \ge 0$. If $\beta c_i(x^{t+1}) + \lambda_i^{t+1} < 0$, since $\psi_\beta(u, v)$ is monotonically decreasing in $v \ge 0$ when $u < 0$, it follows that

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

which yields (9). In addition, (3), together with (8), indicates that for any $t \ge 1$,

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

$$\leq G\sum_{i\in\mathcal{E}}|\lambda_i^{t+1}-\lambda_i^t|+G\sum_{i\in\mathcal{I}}|[\beta c_i(x^{t+1})+\lambda_i^{t+1}]_+-[\beta c_i(x^{t+1})+\lambda_i^t]_+|$$
$$\leq G\sum_{i\in\mathcal{E}\cup\mathcal{I}}|\lambda_i^{t+1}-\lambda_i^t|\leq m\rho_t\tilde{C}G$$

which is exactly (10). □

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 $u,v\in X$, $t\geq 1$ and $\beta\geq\beta_1$,*

$$\|\nabla_x\phi_\beta(u,\lambda^t)-\nabla_x\phi_\beta(v,\lambda^t)\|\leq L_\beta\|u-v\|. \tag{11}$$

Furthermore, if Assumption 3 holds as well, then

$$\mathbb{E}_\xi[\|\nabla_x\mathbf{\Phi}_\beta(u,\lambda^t;\xi)-\nabla_x\mathbf{\Phi}_\beta(v,\lambda^t;\xi)\|^2]\leq L_\beta^2\|u-v\|^2, \tag{12}$$

where $L_\beta:=\beta\tilde{L}$ with $\tilde{L}:=\frac{L+mCL\sum_{k=1}^\infty\rho_k}{\beta_1}+mG^2+mCL$.

Proof. It follows from Assumptions 1-2, (3) and (7) that for any $u,v\in X$,

$$\begin{aligned}
&\|\nabla_x\Psi_\beta(u,\lambda^t)-\nabla_x\Psi_\beta(v,\lambda^t)\|\\
\leq&\sum_{i\in\mathcal{E}}\|(\beta c_i(u)+\lambda_i^t)\nabla c_i(u)-(\beta c_i(v)+\lambda_i^t)\nabla c_i(v)\|\\
&+\sum_{i\in\mathcal{I}}\|[\beta c_i(u)+\lambda_i^t]_+\nabla c_i(u)-[\beta c_i(v)+\lambda_i^t]_+\nabla c_i(v)\|\\
=&\sum_{i\in\mathcal{E}}\|\big[(\beta c_i(u)+\lambda_i^t)-(\beta c_i(v)+\lambda_i^t)\big]\nabla c_i(u)+(\beta c_i(v)+\lambda_i^t)(\nabla c_i(u)-\nabla c_i(v))\|\\
&+\sum_{i\in\mathcal{I}}\|\big[[\beta c_i(u)+\lambda_i^t]_+-[\beta c_i(v)+\lambda_i^t]_+\big]\nabla c_i(u)+[\beta c_i(v)+\lambda_i^t]_+(\nabla c_i(u)-\nabla c_i(v))\|\\
\leq&\sum_{i\in\mathcal{E}}\big[\beta|c_i(u)-c_i(v)|\|\nabla c_i(u)\|+(\beta c_i(v)+\lambda_i^t)L\|u-v\|\big]\\
&+\sum_{i\in\mathcal{I}}\big[\beta|c_i(u)-c_i(v)|\|\nabla c_i(u)\|+[\beta c_i(v)+\lambda_i^t]_+L\|u-v\|\big]\\
\leq&\sum_{i\in\mathcal{E}\cup\mathcal{I}}[\beta G^2\|u-v\|+L(\beta C+|\lambda_i^t|)\|u-v\|]\\
\leq&m(\beta G^2+CL(\beta+\sum_{k=1}^{t-1}\rho_k))\|u-v\|.
\end{aligned}$$

Then, Assumptions 1 and 2 indicate

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

where the last inequality is due to $\beta\geq\beta_1$. Analogously, it holds from Assumption 3 that

$$\begin{aligned}
&\mathbb{E}_\xi[\|\nabla_x\mathbf{\Phi}_\beta(u,\lambda^t;\xi)-\nabla_x\mathbf{\Phi}_\beta(v,\lambda^t;\xi)\|^2]\\
\leq&\mathbb{E}_\xi[\|\nabla\mathbf{F}(u;\xi)-\nabla\mathbf{F}(v;\xi)\|^2]+2\mathbb{E}_\xi[\|\nabla\mathbf{F}(u;\xi)-\nabla\mathbf{F}(v;\xi)\|]\|\nabla_x\Psi_\beta(u,\lambda^t)-\nabla_x\Psi_\beta(v,\lambda^t)\|\\
&+\|\nabla_x\Psi_\beta(u,\lambda^t)-\nabla_x\Psi_\beta(v,\lambda^t)\|^2\\
\leq&L^2\|u-v\|^2+2mL(\beta G^2+CL(\beta+\sum_{k=1}^{t-1}\rho_k))\|u-v\|^2+(m(\beta G^2+CL(\beta+\sum_{k=1}^{t-1}\rho_k)))^2\|u-v\|^2
\end{aligned}$$

which yields (12). □

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

$$\mathbf{d}^2(\nabla_x \phi_{\beta_t}(x^{t+1}, \lambda^{t+1}) + \partial h(x^{t+1}), -\mathcal{N}_X(x^{t+1})) \leq 4(m\rho_t \tilde{C} G)^2 + 4\|\varepsilon^t\|^2 + 4(L_{\beta_t}^2 + \frac{1}{\eta_t^2})\|x^{t+1} - x^t\|^2, \tag{13}$$

where $L_{\beta_t} := \beta_t \tilde{L}$.

Proof. Optimality conditions for (5) imply that for any $t \geq 1$,

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

Then from (10) and (11) we obtain

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

which completes the proof. □

Lemma 6 *Under Assumptions 1-2, the following relation holds true:*

$$\Psi_{\beta_{t+1}}(x, \lambda) \leq \Psi_{\beta_t}(x, \lambda) + \frac{\beta_{t+1} - \beta_t}{2} \sum_{i \in \mathcal{E} \cup \mathcal{I}} c_i^2(x). \tag{14}$$

Furthermore, we have

$$(\frac{1}{2\eta_t} - \frac{L_{\beta_t}}{2})\|x^{t+1} - x^t\|^2 \leq \mathcal{L}_{\beta_t}(x^t, \lambda^t) - \mathcal{L}_{\beta_{t+1}}(x^{t+1}, \lambda^{t+1}) + \frac{\beta_{t+1} - \beta_t}{2} mC^2 + \frac{\eta_t}{2}\|\varepsilon^t\|^2 + m\rho_t \tilde{C}^2. \tag{15}$$

Proof. The key to prove (14) is to verify

$$\psi_{\beta_{t+1}}(c_i(x), \lambda_i) \leq \psi_{\beta_t}(c_i(x), \lambda_i) + \frac{\beta_{t+1} - \beta_t}{2} c_i^2(x), \ \forall i \in \mathcal{I}.$$

According to the definition of ψ_β and $\beta_{t+1} \geq \beta_t$, one has

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

For the second case, $\beta_{t+1} c_i(x) + \lambda_i < 0$, we have

$$-\frac{\lambda_i^2}{2\beta_{t+1}} - \lambda_i c_i(x) - \frac{\beta_t}{2} c_i^2(x) \leq \lambda_i c_i(x) + \frac{\beta_{t+1}}{2} c_i^2(x) - \lambda_i c_i(x) - \frac{\beta_t}{2} c_i^2(x) = \frac{\beta_{t+1} - \beta_t}{2} c_i^2(x).$$

Further, if $\beta_t c_i(x) + \lambda_i < 0$, then $\lambda_i^2 < \beta_t \beta_{t+1} c_i^2(x)$. It holds that

$$\frac{\lambda_i^2}{2\beta_t} - \frac{\lambda_i^2}{2\beta_{t+1}} \leq \frac{\beta_{t+1} - \beta_t}{2} c_i^2(x).$$

Thus, (14) can be derived. Moreover, Assumption 2 together with (8) and (9) implies that

$$\begin{aligned}\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t) =& \mathcal{L}_{\beta_t}(x^{t+1},\lambda^{t+1}) + \sum_{i\in\mathcal{E}}[(\lambda_i^t c_i(x^{t+1}) + \frac{\beta_t}{2}c_i^2(x^{t+1})) - (\lambda_i^{t+1}c_i(x^{t+1}) + \frac{\beta_t}{2}c_i^2(x^{t+1}))] \\ &+ \sum_{i\in\mathcal{I}}[\psi_{\beta_t}(c_i(x^{t+1}),\lambda_i^t) - \psi_{\beta_t}(c_i(x^{t+1}),\lambda_i^{t+1})] \\ \geq& \mathcal{L}_{\beta_t}(x^{t+1},\lambda^{t+1}) - \sum_{i\in\mathcal{E}}|c_i(x^{t+1})||\lambda_i^t - \lambda_i^{t+1}| - \sum_{i\in\mathcal{I}}\rho_t\tilde{C}^2 \\ \geq& \mathcal{L}_{\beta_t}(x^{t+1},\lambda^{t+1}) - m\rho_t\tilde{C}^2. \end{aligned} \tag{16}$$

Note that by optimality conditions for (5), there exists a vector $u \in \partial h(x^{t+1})$ 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 h and setting $x = x^t$, we obtain

$$h(x^{t+1}) - h(x^t) \leq \langle u, x^{t+1} - x^t\rangle \leq -\langle d^t + \frac{1}{\eta_t}(x^{t+1} - x^t), x^{t+1} - x^t\rangle. \tag{17}$$

Thus it indicates

$$\begin{aligned}&\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t) - \mathcal{L}_{\beta_t}(x^t,\lambda^t) \\ =& \phi_{\beta_t}(x^{t+1},\lambda^t) - \phi_{\beta_t}(x^t,\lambda^t) + h(x^{t+1}) - h(x^t) \\ \leq& \langle\nabla_x\phi_{\beta_t}(x^t,\lambda^t), x^{t+1} - x^t\rangle + \frac{L_{\beta_t}}{2}\|x^{t+1} - x^t\|^2 + h(x^{t+1}) - h(x^t) \\ \leq& \langle d^t - \varepsilon^t, x^{t+1} - x^t\rangle - \langle d^t + \frac{1}{\eta_t}(x^{t+1} - x^t), x^{t+1} - x^t\rangle + \frac{L_{\beta_t}}{2}\|x^{t+1} - x^t\|^2 \\ =& -\langle\varepsilon^t, x^{t+1} - x^t\rangle + (\frac{L_{\beta_t}}{2} - \frac{1}{\eta_t})\|x^{t+1} - x^t\|^2 \\ \leq& \frac{\eta_t}{2}\|\varepsilon^t\|^2 + \frac{1}{2\eta_t}\|x^{t+1} - x^t\|^2 + (\frac{L_{\beta_t}}{2} - \frac{1}{\eta_t})\|x^{t+1} - x^t\|^2 \\ \leq& \frac{\eta_t}{2}\|\varepsilon^t\|^2 + (\frac{L_{\beta_t}}{2} - \frac{1}{2\eta_t})\|x^{t+1} - x^t\|^2, \end{aligned} \tag{18}$$

where the first inequality follows from (11), the second inequality comes from (17), and the third inequality is due to Young's inequality. Thus it together with (16) yields that

$$\begin{aligned}\mathcal{L}_{\beta_t}(x^{t+1},\lambda^{t+1}) - \mathcal{L}_{\beta_t}(x^t,\lambda^t) =& \mathcal{L}_{\beta_t}(x^{t+1},\lambda^{t+1}) - \mathcal{L}_{\beta_t}(x^{t+1},\lambda^t) + \mathcal{L}_{\beta_t}(x^{t+1},\lambda^t) - \mathcal{L}_{\beta_t}(x^t,\lambda^t) \\ \leq& \frac{\eta_t}{2}\|\varepsilon^t\|^2 + (\frac{L_{\beta_t}}{2} - \frac{1}{2\eta_t})\|x^{t+1} - x^t\|^2 + m\rho_t\tilde{C}^2.\end{aligned}$$

Then rearranging the terms and together with (14) derives (15). □

From now on and for simplicity, we introduce $\tilde{\lambda}^t \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$, $t \geq 1$, defined componentwise by

$$\tilde{\lambda}_i^t = \begin{cases}\beta_{t-1}c_i(x^t) + \lambda_i^t, & i \in \mathcal{E}, \\ [\beta_{t-1}c_i(x^t) + \lambda_i^t]_+, & i \in \mathcal{I}.\end{cases} \tag{19}$$

Lemma 7 *Under Assumptions 1-2, it holds that for any $T \geq 1$,*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{t+1})c_{\mathcal{E}}(x^{t+1}) + \nabla c_{\mathcal{I}}(x^{t+1})[c_{\mathcal{I}}(x^{t+1})]_+, -\mathcal{N}_X(x^{t+1})) \\
\leq &\frac{4}{\beta_1^2}\frac{1}{T}\sum_{t=1}^{T}\mathbf{d}^2(\nabla f(x^{t+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}}\tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}) + \partial h(x^{t+1}), -\mathcal{N}_X(x^{t+1})) \\
&+ \frac{4(2+m^2(\|\lambda_1\| + C\sum_{k=1}^{T}\rho_k)^2)G^2}{T}\sum_{t=1}^{T}\frac{1}{\beta_t^2}.
\end{aligned}$$

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

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

Then for any $t \geq 1$, one has

$$\begin{aligned}
&\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{t+1})c_{\mathcal{E}}(x^{t+1}) + \nabla c_{\mathcal{I}}(x^{t+1})[c_{\mathcal{I}}(x^{t+1})]_+, -\mathcal{N}_X(x^{t+1})) \\
\leq &\frac{1}{\beta_t^2}\mathbf{d}^2(\sum_{i\in\mathcal{E}}\beta_t c_i(x^{t+1})\nabla c_i(x^{t+1}) + \sum_{i\in\mathcal{I}}[\beta_t c_i(x^{t+1})]_+\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1})) \\
\leq &\frac{4}{\beta_t^2}[\mathbf{d}^2(\nabla f(x^{t+1}) + v^{t+1} + \sum_{i\in\mathcal{E}\cup\mathcal{I}}\tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}), -\mathcal{N}_X(x^{t+1})) + \|\nabla f(x^{t+1})\|^2 \\
&\quad + \|v^{t+1}\|^2 + (\sum_{i\in\mathcal{E}\cup\mathcal{I}}|\lambda_i^{t+1}|\|\nabla c_i(x^{t+1})\|)^2] \\
\leq &\frac{4}{\beta_t^2}[\mathbf{d}^2(\nabla_x\phi_{\beta_t}(x^{t+1},\lambda^{t+1}) + \partial h(x^{t+1}), -\mathcal{N}_X(x^{t+1})) + (2+m^2(\|\lambda^1\| + C\sum_{k=1}^{t}\rho_k)^2)G^2]. \qquad (20)
\end{aligned}$$

Summing up the above inequality over $t = 1, \ldots, T$ and dividing it by T yields the conclusion from $\beta_t \geq \beta_1$ for any $t \geq 1$. □

The lemma below provides a recursive bound on the error ε^t. For notation simplicity, we denote in the following that:

$$\xi^t = \{\xi_j^t, j \in \mathcal{J}_t\}, \quad \xi^{[t]} = \{\xi^1, \ldots, \xi^t\}, \quad t \geq 1.$$

Lemma 8 *Under Assumptions 1-4, it holds that for any $t \geq 1$,*

$$\mathbb{E}_{\xi^{[t+1]}}[\|\varepsilon^{t+1}\|^2] \leq (1-\alpha_t)^2\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] + \frac{1}{|\mathcal{J}_{t+1}|}(2\alpha_t^2\sigma^2 + 2(1-\alpha_t)^2L^2\mathbb{E}_{\xi^{[t]}}[\|x^{t+1}-x^t\|^2]). \qquad (21)$$

Proof. Recall that

$$\begin{aligned}
\varepsilon^{t+1} =& d^{t+1} - \nabla_x\phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) \\
=& \frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}}\nabla_x\mathbf{\Phi}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1};\xi_j^{t+1}) - \nabla_x\phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) \\
&+ (1-\alpha_t)(d^t - \frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}}\nabla_x\mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^{t+1}))
\end{aligned}$$

$$= \frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}} \nabla_x \mathbf{\Phi}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1};\xi_j^{t+1}) - \nabla_x \phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1})$$

$$+ (1-\alpha_t)\varepsilon^t + (1-\alpha_t)(\nabla_x \phi_{\beta_t}(x^t,\lambda^t) - \frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}} \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^{t+1})). \tag{22}$$

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} yields

$$\mathbb{E}_{\xi^{t+1}}[\langle \frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}} \nabla_x \mathbf{\Phi}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1};\xi_j^{t+1}) - \nabla_x \phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}), \varepsilon^t\rangle] = 0,$$

$$\mathbb{E}_{\xi^{t+1}}[\langle \frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}} \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^{t+1}) - \nabla_x \phi_{\beta_t}(x^t,\lambda^t), \varepsilon^t\rangle] = 0.$$

Hence, squaring both sides of (22) and then taking expectation with respect to ξ^{t+1}, we have

$$\begin{aligned}
&\mathbb{E}_{\xi^{t+1}}[\|\varepsilon^{t+1}\|^2] \\
\leq&(1-\alpha_t)^2\|\varepsilon^t\|^2 + \mathbb{E}_{\xi^{t+1}}[\|\frac{1}{|\mathcal{J}_{t+1}|}\sum_{j\in\mathcal{J}_{t+1}} (\nabla_x \mathbf{\Phi}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1};\xi_j^{t+1}) - \nabla_x \phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) \\
&+ (1-\alpha_t)(\nabla_x \phi_{\beta_t}(x^t,\lambda^t) - \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^{t+1})))\|^2] \\
=&(1-\alpha_t)^2\|\varepsilon^t\|^2 + \mathbb{E}_{\xi^{t+1}}[\frac{1}{|\mathcal{J}_{t+1}|^2}\sum_{j\in\mathcal{J}_{t+1}} \|\nabla_x \mathbf{\Phi}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1};\xi_j^{t+1}) - \nabla_x \phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) \\
&+ (1-\alpha_t)(\nabla_x \phi_{\beta_t}(x^t,\lambda^t) - \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^{t+1}))\|^2].
\end{aligned} \tag{23}$$

Let us focus on the second term in R.H.S. of (23). Note that for any $j \in \mathcal{J}_{t+1}$,

$$\begin{aligned}
&\mathbb{E}_{\xi^{t+1}}[\|\nabla_x \mathbf{\Phi}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1};\xi_j^{t+1}) - \nabla_x \phi_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) + (1-\alpha_t)(\nabla_x \phi_{\beta_t}(x^t,\lambda^t) - \nabla_x \mathbf{\Phi}_{\beta_t}(x^t,\lambda^t;\xi_j^{t+1}))\|^2] \\
=&\mathbb{E}_{\xi^{t+1}}[\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1}) + (1-\alpha_t)(\nabla f(x^t) - \nabla F(x^t;\xi_j^{t+1}))\|^2] \\
=&\mathbb{E}_{\xi^{t+1}}[\|\alpha_t(\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1})) + (1-\alpha_t)(\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1}) + \nabla f(x^t) - \nabla F(x^t;\xi_j^{t+1}))\|^2] \\
\leq&\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1})\|^2 + 2(1-\alpha_t)^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1}) + \nabla f(x^t) - \nabla F(x^t;\xi_j^{t+1})\|^2] \\
=&\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1})\|^2 + 2(1-\alpha_t)^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla F(x^t;\xi_j^{t+1})\|^2 \\
&-4(1-\alpha_t)^2\langle \nabla F(x^{t+1};\xi_j^{t+1}) - \nabla F(x^t;\xi_j^{t+1}), \nabla f(x^{t+1}) - \nabla f(x^t)\rangle + 2(1-\alpha_t)^2\|\nabla f(x^{t+1}) - \nabla f(x^t)\|^2] \\
=&\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1})\|^2 + 2(1-\alpha_t)^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla F(x^t;\xi_j^{t+1})\|^2 \\
&-2(1-\alpha_t)^2\|\nabla f(x^{t+1}) - \nabla f(x^t)\|^2] \\
\leq&\mathbb{E}_{\xi^{t+1}}[2\alpha_t^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla f(x^{t+1})\|^2 + 2(1-\alpha_t)^2\|\nabla F(x^{t+1};\xi_j^{t+1}) - \nabla F(x^t;\xi_j^{t+1})\|^2] \\
\leq&2\alpha_t^2\sigma^2 + 2(1-\alpha_t)^2L^2\|x^{t+1} - x^t\|^2.
\end{aligned}$$

Hence, we obtain the following relation:

$$\mathbb{E}_{\xi^{t+1}}[\|\varepsilon^{t+1}\|^2] \leq (1-\alpha_t)^2\|\varepsilon^t\|^2 + \frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|} + \frac{2(1-\alpha_t)^2}{|\mathcal{J}_{t+1}|}L^2\|x^{t+1} - x^t\|^2.$$

Taking expectation over $\xi^{[t+1]}$ yields the desired result. □

Interestingly, parameter α_t and batch $\mathcal{J}_{t+1}$ are somewhat intertwined in our approach. More specifically, when $\alpha_t = 1$, the gradient estimate d^t turns to the vanilla mini-batch SGD approximation, where the batch size is normally chosen large enough to reduce stochastic variances. However, in this paper, we focus on the case where $0 < \alpha_t < 1$. In this scenario, the current error ε^t is controlled by previous error ε^{t-1} and corrections accumulated in past iterations. We can prove that, under appropriate parameter settings like (33), the term $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]$ tends to zero even without the use of batches $\mathcal{J}_t$ (refer to (31) and (37)). However, it can help ensure the boundedness of $\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]$, which is crucial for the global convergence analysis (refer to (39)).

4 Global convergence analysis

In this section, we conduct a global convergence analysis for MLALM. To this end, we assume that the parameters used in Algorithm 1 satisfy the conditions below:

$$\eta_t L_{\beta_t} \leq \frac{1}{2}, \quad 8\eta_1\eta_t L^2 \leq \alpha_t < 1, \quad \eta_{t+1} \leq \eta_t. \tag{24}$$

Lemma 9 *Suppose that Assumptions 1-4 and (24) hold. Then for any $T \geq 1$ the following is true:*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\alpha_t\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] \leq \frac{2\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2] + 4\sum_{t=1}^{T}\alpha_t^2\sigma^2}{T}\\ + \frac{16\eta_1 L^2}{T}(\mathcal{L}_{\beta_1}(x^1,\lambda^1) - C^* + C\|\lambda^1\|_1 + \frac{mC^2}{2}\sum_{t=1}^{T}\frac{\beta_{t+1}-\beta_t}{|\mathcal{J}_{t+1}|} + 2m\tilde{C}^2\sum_{t=1}^{T}\rho_t),\end{aligned} \tag{25}$$

where C^* is the lower bound of the objective function of (1) over X.

Proof. From Lemma 6 it follows that

$$\|x^{t+1}-x^t\|^2 \leq \frac{2\eta_t}{1-\eta_t L_{\beta_t}}(\mathcal{L}_{\beta_t}(x^t,\lambda^t) - \mathcal{L}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) + \frac{\beta_{t+1}-\beta_t}{2}mC^2 + \frac{\eta_t}{2}\|\varepsilon^t\|^2 + m\rho_t\tilde{C}^2). \tag{26}$$

We then substitute the above relation into (21) indicating

$$\begin{aligned}&\mathbb{E}_{\xi^{[t+1]}}[\|\varepsilon^{t+1}\|^2]\\ &\leq (1-\alpha_t)^2\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] + \frac{1}{|\mathcal{J}_{t+1}|}(2\alpha_t^2\sigma^2 + 2(1-\alpha_t)^2L^2\mathbb{E}_{\xi^{[t]}}[\|x^{t+1}-x^t\|^2])\\ &\leq (1-\alpha_t)\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] + \frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|}\\ &\quad + \frac{4\eta_1L^2}{|\mathcal{J}_{t+1}|(1-\eta_tL_{\beta_t})}\mathbb{E}_{\xi^{[t]}}[\mathcal{L}_{\beta_t}(x^t,\lambda^t) - \mathcal{L}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1}) + \frac{\beta_{t+1}-\beta_t}{2}mC^2 + \frac{\eta_t}{2}\|\varepsilon^t\|^2 + m\rho_t\tilde{C}^2],\end{aligned}$$

where the last inequality is due to the condition $0 < \alpha_t < 1$. By summing the above inequality over $t = 1, \ldots, T$, we obtain from $\frac{1}{1-\eta_tL_{\beta_t}} \leq 2$ (thanks to $\eta_tL_{\beta_t} \leq \frac{1}{2}$) and $\alpha_t \geq 8\eta_1\eta_tL^2$ that

$$\begin{aligned}\sum_{t=1}^{T}\frac{\alpha_t}{2}\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] \leq \sum_{t=1}^{T}(\alpha_t - 4\eta_1\eta_tL^2)\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] \leq \mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2] + 2\sum_{t=1}^{T}\alpha_t^2\sigma^2\\ + 8\eta_1L^2(\mathcal{L}_{\beta_1}(x^1,\lambda^1) - \mathbb{E}_{\xi^{[T]}}[\mathcal{L}_{\beta_{T+1}}(x^{T+1},\lambda^{T+1})] + \frac{mC^2}{2}\sum_{t=1}^{T}\frac{\beta_{t+1}-\beta_t}{|\mathcal{J}_{t+1}|} + m\tilde{C}^2\sum_{t=1}^{T}\frac{\rho_t}{|\mathcal{J}_{t+1}|}).\end{aligned} \tag{27}$$

Moreover, we can upper bound $\mathcal{L}_{\beta_1}(x^1,\lambda^1)-\mathbb{E}_{\xi^{[T]}}[\mathcal{L}_{\beta_{T+1}}(x^{T+1},\lambda^{T+1})]$ by

$$
\begin{aligned}
&\mathcal{L}_{\beta_1}(x^1,\lambda^1)-\mathbb{E}_{\xi^{[T]}}[\mathcal{L}_{\beta_{T+1}}(x^{T+1},\lambda^{T+1})]\\
=&\mathcal{L}_{\beta_1}(x^1,\lambda^1)-\mathbb{E}_{\xi^{[T]}}[f(x^{T+1})+h(x^{T+1})+\sum_{i\in\mathcal{E}}[\lambda_i^{T+1}c_i(x^{T+1})+\frac{\beta_t}{2}c_i^2(x^{T+1})]+\sum_{i\in\mathcal{I}}\psi_{\beta_t}(c_i(x^{T+1}),\lambda_i^{T+1})]\\
\le&\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*-\mathbb{E}_{\xi^{[T]}}[\sum_{i\in\mathcal{E}}[\lambda_i^{T+1}c_i(x^{T+1})+\frac{\beta_t}{2}c_i^2(x^{T+1})]+\sum_{i\in\mathcal{I}}\psi_{\beta_t}(c_i(x^{T+1}),\lambda_i^{T+1})]\\
\le&\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+\mathbb{E}_{\xi^{[T]}}[\sum_{i\in\mathcal{E}}|\lambda_i^{T+1}c_i(x^{T+1})|+\sum_{i\in\mathcal{I}}\frac{(\lambda_i^{T+1})^2}{2\beta_t}]\\
\le&\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C(\sum_{i\in\mathcal{E}}(|\lambda_i^1|+C\sum_{t=1}^T\rho_t)+\frac{1}{2\beta_1}\sum_{i\in\mathcal{I}}(\lambda_i^1+C\sum_{t=1}^T\rho_t)^2\\
\le&\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+m\tilde{C}^2\sum_{t=1}^T\rho_t,
\end{aligned}
\tag{28}
$$

where the second inequality comes from $\psi_{\beta_t}(u,v)\ge-\frac{v^2}{2\beta_t}$, and the third inequality holds due to (7). Plugging the above inequality into (27) and dividing the whole inequality by T yield the desired result. □

Lemma 10 *Under the conditions of Lemma 9, suppose that $\{\alpha_t\}$ and $\{\eta_t\}$ are non-increasing sequences. Then it holds that for any $T\ge1$,*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla f(x^{t+1})+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\tilde{\lambda}_i^{t+1}c_i(x^{t+1})+\partial h(x^{t+1}),-\mathcal{N}_X(x^{t+1}))]\\
\le&\frac{4m^2\tilde{C}^2G^2}{T}\sum_{t=1}^T\rho_t^2+\frac{28(\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+2\sum_{t=1}^T\alpha_t^2\sigma^2)}{\alpha_TT}\\
&+\frac{1}{T}(\frac{224\eta_1L^2}{\alpha_T}+\frac{20}{\eta_T})(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+\frac{m\beta_{T+1}C^2}{2}+2m\tilde{C}^2\sum_{t=1}^T\rho_t).
\end{aligned}
\tag{29}
$$

Proof. By using the definition of ϕ_β and summing up (13) for $t=1,\dots,T$, we can subsequently divide the resulting expression by T to obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\mathbf{d}^2(\nabla f(x^{t+1})+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\tilde{\lambda}_i^{t+1}c_i(x^{t+1})+\partial h(x^{t+1}),-\mathcal{N}_X(x^{t+1}))\\
\le&\frac{4m^2\tilde{C}^2G^2}{T}\sum_{t=1}^T\rho_t^2+\frac{4}{T}\sum_{t=1}^T\|\varepsilon^t\|^2+\frac{4}{T}\sum_{t=1}^T(L_{\beta_t}^2+\frac{1}{\eta_t^2})\|x^{t+1}-x^t\|^2.
\end{aligned}
\tag{30}
$$

On the one hand, since $\{\alpha_t\}$ is non-increasing and $|\mathcal{J}_t|\ge1$, by Lemma 9 the second term in R.H.S. of (30) can be upper bounded by using

$$
\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]\le\frac{2\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+4\sum_{t=1}^T\alpha_t^2\sigma^2}{\alpha_TT}+\frac{16\eta_1L^2}{\alpha_TT}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+\frac{m\beta_{T+1}C^2}{2}+2m\tilde{C}^2\sum_{t=1}^T\rho_t).
\tag{31}
$$

On the other hand, it follows from (26) that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(L_{\beta_t}^2+\frac{1}{\eta_t^2})\|x^{t+1}-x^t\|^2\\
\leq&\frac{1}{T}\sum_{t=1}^{T}\frac{2(1+\eta_t^2L_{\beta_t}^2)}{\eta_t(1-\eta_tL_{\beta_t})}(\mathcal{L}_{\beta_t}(x^t,\lambda^t)-\mathcal{L}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1})+\frac{\beta_{t+1}-\beta_t}{2}mC^2+\frac{\eta_t}{2}\|\varepsilon^t\|^2+m\rho_t\tilde{C}^2)\\
\leq&\frac{5}{\eta_TT}\sum_{t=1}^{T}(\mathcal{L}_{\beta_t}(x^t,\lambda^t)-\mathcal{L}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1})+\frac{\beta_{t+1}-\beta_t}{2}mC^2+m\rho_t\tilde{C}^2)+\frac{5}{2T}\sum_{t=1}^{T}\|\varepsilon^t\|^2\\
\leq&\frac{5}{\eta_TT}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+\frac{m\beta_{T+1}C^2}{2}+2m\tilde{C}^2\sum_{t=1}^{T}\rho_t)+\frac{5}{2T}\sum_{t=1}^{T}\|\varepsilon^t\|^2,
\end{aligned}\tag{32}$$

where the second inequality follows from the fact that $\frac{1+u^2}{1-u}<\frac{5}{2}$ when $0<u<\frac{1}{2}$ and $\eta_T\leq\eta_t$, the last comes from (28). Therefore, we obtain the conclusion. □

4.1 Unbounded penalty parameters

In this subsection, we detect the global convergence of MLALM with unbounded penalty parameters. The theorem below shows global convergence properties of the sequences of average stationarity measure and average constraint violation in expectation, respectively.

Theorem 1 *Suppose that Assumptions 1-4 hold, and the parameters used in Algorithm 1 satisfy* $\beta_t=\beta_0t^\iota$ *and*

$$\rho_t=\frac{\rho}{t^\theta},\ \eta_t=\frac{\eta}{t^\iota\max\{L,\tilde{L}\}},\ \alpha_t=\frac{8\alpha\eta^2}{t^\iota},\quad t\geq1,\tag{33}$$

where $\rho,\beta_0>0$, $0<\eta\leq\min\{\frac{1}{2\beta_0},\frac{\sqrt{2}}{4}\}$, $\theta\in(1,\infty)$, $\iota\in(0,\frac{1}{2})$ *and* $\alpha\in[1,\frac{1}{8\eta^2})$ *are given constants. Then the followings are true:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla f(x^{t+1})+\partial h(x^{t+1})+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\tilde{\lambda}_i^{t+1}\nabla c_i(x^{t+1}),-\mathcal{N}_X(x^{t+1}))]=0,\tag{34}$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{t+1})c_{\mathcal{E}}(x^{t+1})+\nabla c_{\mathcal{I}}(x^{t+1})[c_{\mathcal{I}}(x^{t+1})]_+,-\mathcal{N}_X(x^{t+1}))]=0.\tag{35}$$

Furthermore, if $|\mathcal{J}_t|=t^q$ *with* $q>1$, *suppose that there exists* $C_u>0$ *such that* $\mathbb{E}[f(x^t)+h(x^t)]\leq C_u$ *for any* $t\geq1$. *Then*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|c_{\mathcal{E}}(x^t)\|^2+\|[c_{\mathcal{I}}(x^t)]_+\|^2]\ \textit{exists and is finite.}\tag{36}$$

Proof. We can verify that the settings defined in (33) satisfy the required conditions specified in (24). It is noteworthy that the values of $\alpha\in[1,\frac{1}{8\eta^2}]$ are derived from the inequality $\alpha_t\leq1$. Additionally, the range for η, i.e., $0<\eta\leq\min\{\frac{1}{2\beta_0},\frac{\sqrt{2}}{4}\}$ is determined based on the conditions $\eta_t\beta_t\leq\frac{1}{2}$ and $\frac{1}{8\eta^2}>1$. From the upper bounds as shown in (29) and Lemma 7, the key to prove (34) and (35) is to derive

$$\frac{\sum_{t=1}^{T}\rho_t^2}{T}\to0,\quad\frac{\sum_{t=1}^{T}\alpha_t^2}{\alpha_TT}\to0,\quad\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\beta_t^2}\to0\quad\text{and}\quad\frac{1}{T}(\frac{\eta_1}{\alpha_T}+\frac{1}{\eta_T})(\beta_{T+1}+\sum_{t=1}^{T}\rho_t)\to0,\tag{37}$$

as T increases to infinity. These can be achieved under parameter settings of the theorem.

We next prove (36). Recall that it is derived in (18) that

$$\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t)-\mathcal{L}_{\beta_t}(x^t,\lambda^t)\leq \frac{\eta_t}{2}\|\varepsilon^t\|^2+(\frac{L_{\beta_t}}{2}-\frac{1}{2\eta_t})\|x^{t+1}-x^t\|^2$$

which yields

$$\|x^{t+1}-x^t\|^2\leq \frac{2\eta_t}{1-\eta_t L_{\beta_t}}(\mathcal{L}_{\beta_t}(x^t,\lambda^t)-\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t)+\frac{\eta_t}{2}\|\varepsilon^t\|^2).$$

Plugging the above inequality into (21) implies

$$\begin{aligned}
\mathbb{E}_{\xi^{[t+1]}}[\|\varepsilon^{t+1}\|^2] \leq & (1-\alpha_t)^2\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]+\frac{1}{|\mathcal{J}_{t+1}|}(2\alpha_t^2\sigma^2+2(1-\alpha_t)^2L^2t^\iota\mathbb{E}_{\xi^{[t]}}[\|x^{t+1}-x^t\|^2])\\
\leq & (1-\alpha_t)^2(1+\frac{2\eta_t^2L^2t^\iota}{1-\eta_tL_{\beta_t}})\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]+\frac{4(1-\alpha_t)^2\eta_tL^2t^\iota}{1-\eta_tL_{\beta_t}}(\mathbb{E}_{\xi^{[t]}}[\mathcal{L}_{\beta_t}(x^t,\lambda^t)-\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t)])+\frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|}\\
\leq & \mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]+\frac{4(1-\alpha_t)^2\eta_tL^2t^\iota}{1-\eta_tL_{\beta_t}}(\mathbb{E}_{\xi^{[t]}}[\mathcal{L}_{\beta_t}(x^t,\lambda^t)-\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t)])+\frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|},
\end{aligned}$$

where the last inequality uses $(1-\alpha_t)^2(1+\frac{2\eta_t^2L^2t^\iota}{1-\eta_tL_{\beta_t}})\leq 1$ thanks to $\frac{2\eta_t^2L^2t^\iota}{1-\eta_tL_{\beta_t}}\leq 4\eta_1\eta_tL^2\leq\alpha_t$. Moreover, dividing both sides of the above inequality by β_t, rearranging the terms and taking the expectation lead to

$$\begin{aligned}
A_t:=&\mathbb{E}_{\xi^{[t+1]}}[\frac{1}{\beta_t}\mathcal{L}_{\beta_t}(x^{t+1},\lambda^t)+\frac{1-\eta_tL_{\beta_t}}{4(1-\alpha_t)^2\eta_t\beta_tL^2t^\iota}\|\varepsilon^{t+1}\|^2]\\
\leq&\mathbb{E}_{\xi^{[t]}}[\frac{1}{\beta_t}\mathcal{L}_{\beta_t}(x^t,\lambda^t)+\frac{1-\eta_tL_{\beta_t}}{4(1-\alpha_t)^2\eta_t\beta_tL^2t^\iota}(\|\varepsilon^t\|^2+\frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|})]=:B_t.
\end{aligned}$$

Thus, $\{\frac{1}{T}\sum_{p=1}^T\sum_{t=1}^p(A_t-B_t)\}_{T\geq 1}$ is a non-increasing sequence as T increases to infinity. Recalling the definition of $\mathcal{L}_\beta$, it is obvious that

$$\begin{aligned}
&\frac{1}{T}\sum_{p=1}^T\sum_{t=1}^p(A_t-B_t)\\
=&\frac{1}{T}\sum_{p=1}^T\sum_{t=1}^p\frac{\mathbb{E}_{\xi^{[t+1]}}[f(x^{t+1})+h(x^{t+1})-f(x^t)-h(x^t)]}{\beta_t}\\
&+\frac{1}{T}\sum_{p=1}^T\mathbb{E}_{\xi^{[p]}}[\sum_{t=1}^p\frac{\Psi_{\beta_t}(x^{t+1},\lambda^t)-\Psi_{\beta_t}(x^t,\lambda^t)}{\beta_t}-\frac{1}{2}(\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2)]\\
&+\frac{1}{2T}\sum_{p=1}^T\mathbb{E}_{\xi^{[p]}}[\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2]\\
&+\frac{1}{T}\sum_{p=1}^T\sum_{t=1}^p\frac{1-\eta_tL_{\beta_t}}{4(1-\alpha_t)^2\eta_t\beta_tL^2t^\iota}(\mathbb{E}_{\xi^{[t+1]}}[\|\varepsilon^{t+1}\|^2-\|\varepsilon^t\|^2]-\frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|}).
\end{aligned}\tag{38}$$

Firstly, under assumptions of this theorem, we can infer that the following inequality holds for any $p\geq 1$:

$$|\sum_{t=1}^p\mathbb{E}[f(x^{t+1})+h(x^{t+1})-f(x^t)-h(x^t)]|=|\mathbb{E}[f(x^{p+1})+h(x^{p+1})]-f(x^1)-h(x^1)|\leq C_u+C^*.$$

Then, due to the fact that the sequence $\{\beta_t\}_{t\geq 1}$ is monotonically increasing and $\lim_{t\to\infty}\frac{1}{\beta_t}=0$, by utilizing Dirichlet's Test we deduce that

$$\lim_{p\to\infty}\sum_{t=1}^{p}\frac{\mathbb{E}[f(x^{t+1})+h(x^{t+1})-f(x^t)-h(x^t)]}{\beta_t} \quad \text{exists.}$$

Therefore, the arithmetic mean (the first item on R.H.S. of (38)) converges as $T\to\infty$. Secondly, by Lemma A.1 we are able to prove that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{p=1}^{T}\sum_{t=1}^{p}\frac{\mathbb{E}[\Psi_{\beta_t}(x^{t+1},\lambda^t)-\Psi_{\beta_t}(x^t,\lambda^t)]}{\beta_t}-\frac{1}{2}\mathbb{E}[\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2] \quad \text{exists.}$$

Thirdly, for the last item in R.H.S. of (38), it implies from (23) that for any $t\geq 1$,

$$\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]\leq \mathbb{E}_{\xi^{[t-1]}}[\|\varepsilon^{t-1}\|^2]+\frac{4\sigma^2}{|\mathcal{J}_t|}\leq \mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+\sum_{k=2}^{t}\frac{4\sigma^2}{|\mathcal{J}_k|}=\sigma^2+\sum_{k=2}^{t}\frac{4\sigma^2}{k^q}. \tag{39}$$

Then it derives that

$$|\sum_{t=1}^{p}(\mathbb{E}[\|\varepsilon^{t+1}\|^2-\|\varepsilon^t\|^2]-\frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|})|=|\mathbb{E}[\|\varepsilon^{p+1}\|^2]-\|\varepsilon^1\|^2-\sum_{t=1}^{p}\frac{2\alpha_t^2\sigma^2}{|\mathcal{J}_{t+1}|}|$$

is bounded following the settings of α_t and $\mathcal{J}_{t+1}$. At the meanwhile, we can verify that $\{\frac{1-\eta_t L_{\beta_t}}{4(1-\alpha_t)^2\eta_t\beta_t L^2 t^c}\}_{t\geq 1}$ is monotonically decreasing and converging to zero, as t increases to infinity. Therefore, the last term of (38) is also convergent by applying Dirichlet's test. In summary, our analysis has demonstrated that the first, second, and fourth terms in R.H.S. of (38) converge to finite values as T increases to infinity. Consequently, the sequence $\{\frac{1}{T}\sum_{p=1}^{T}\sum_{t=1}^{p}(A_t-B_t)\}_{T\geq 1}$ is uniformly lower bounded by a finite value, and since it is non-increasing, it must converge as T tends to infinity. Therefore, the third term in R.H.S. of (38), which represents the average of constraint violations across all previous T iterates, also converges to a finite value as T approaches infinity. This completes the proof of (36). □

Like in most penalty methods for nonconvex constrained optimization, the iterates generated by MLALM may get trapped around an infeasible stationary point, when no constraint qualification is assumed. Hence, to further analyze feasibility and the complementary slackness of iterates, it is necessary to impose a constraint qualification. Various constraint qualification conditions have been used in the literature for nonconvex constrained optimization. Given that MLALM operates as a stochastic approximation method, it becomes necessary to analyze its theoretical performance in an average sense by considering the average of relevant stationarity measures over all previous iterates. However, it is important to note that the iterates produced by MLALM can be infeasible during the algorithmic process. Therefore, it is crucial to establish a broader region beyond the feasible region where a constraint qualification holds. Drawing motivation from [20] and [35, 36, 53], we introduce the following assumption regarding a constraint qualification.

Assumption 5 *There exist positive constants δ and Z such that for any $t\geq 1$ the linear system*

$$\begin{aligned}\delta\cdot\operatorname{sgn}(c_i(x^t))+\nabla c_i(x^t)^T z=0, \qquad & i\in\mathcal{E}: c_i(x^t)\neq 0;\\ \delta+\nabla c_i(x^t)^T z\leq 0, \qquad & i\in\mathcal{I}: c_i(x^t)>0\end{aligned} \tag{40}$$

has a solution $z^t\in-\mathcal{N}_X^*(x^t)$ with $\|z^t\|\leq Z$.

Remark 2 *The constraint qualification assumed in Assumption 5 can be regarded as an extended variant of MFCQ which was originally proposed for smooth nonconvex constrained optimization [51]. In the case of infeasible methods, especially stochastic approximation methods for nonconvex constrained optimization, a constraint qualification (or nonsingularity condition) is often imposed on infeasible iterates in the literature. The necessity of a nonsingularity condition is evident in works such as [24, 26, 40], where it is utilized to analyze the complexity of penalty methods for nonconvex constrained optimization. Compared with the constraint qualification assumed in [30, Assumption 4] for*

smooth deterministic constrained stochastic optimization, we replace $c_i(x^t)$ with $\delta \cdot sgn(c_i(x^t))$ for infeasible equality constraints and with δ for infeasible inequality constraints in the linear system (40). When $\mathcal{I} = \emptyset$ and $X = \mathbb{R}^n$, the condition required by Assumption 5 is weaker than the strong LICQ in [2], where it assumes that $\nabla c(x)^T$ has full row rank and its singular values are uniformly lower bounded away from zero over a convex and compact set containing all iterates. Under Assumption 5, we are able to prove the average of expected Lagrange multiplier vectors $\tilde{\lambda}^t, t \geq 1$, as defined in (19), is upper bounded in the lemma below.

Lemma 11 *Under the conditions and parameter settings of Theorem 1 with $\iota \in (0, \frac{1}{3}]$, suppose that Assumption 5 holds, then $\{\mathbb{E}_{\xi^{[T]}}[\frac{1}{T}\sum_{t=1}^{T}\|\tilde{\lambda}^t\|]\}_{T\geq 1}$ is uniformly bounded.*

Proof. The detailed proof is presented in Appendix B. □

The following theorem provides a characterization of global convergence, in which the average of expected constraint violation and complementary slackness converge to zero, respectively.

Theorem 2 *Under the same conditions and parameter settings of Theorem 1 with $\iota \in (0, \frac{1}{3}]$, suppose that Assumption 5 holds, then*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|c_{\mathcal{E}}(x^{t+1})\|^{\frac{1}{2}} + \|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{1}{2}}] = 0, \tag{41}$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\sum_{i\in\mathcal{I}}(\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})|)^{\frac{1}{4}}] = 0. \tag{42}$$

Proof. As demonstrated in Lemma 11, there exists a constant $\tilde{\Lambda} > 0$ such that

$$\mathbb{E}_{\xi^{[T]}}[\frac{1}{T}\sum_{t=1}^{T}\|\tilde{\lambda}^{t+1}\|] \leq \tilde{\Lambda} \quad \forall T \geq 1.$$

Recall that as shown in Lemma 2, there exists $\Lambda > 0$ such that $\|\lambda^t\| \leq \Lambda$, $t \geq 1$. Then by the definition of $\tilde{\lambda}^t$, we obtain

$$\mathbb{E}_{\xi^{[T]}}[\frac{1}{T}\sum_{t=1}^{T}\beta_t(\|c_{\mathcal{E}}(x^{t+1})\| + \|[c_{\mathcal{I}}(x^{t+1})]_+\|)] \leq \tilde{\Lambda} + \Lambda \quad \forall T \geq 1. \tag{43}$$

Therefore, the following relation holds true:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[(\|c_{\mathcal{E}}(x^{t+1})\| + \|[c_{\mathcal{I}}(x^{t+1})]_+\|)^{\frac{1}{2}}] \\
=&\mathbb{E}_{\xi^{[T]}}[\frac{1}{T}\sum_{t=1}^{T}\beta_t^{-\frac{1}{2}}\beta_t^{\frac{1}{2}}(\|c_{\mathcal{E}}(x^{t+1})\| + \|[c_{\mathcal{I}}(x^{t+1})]_+\|)^{\frac{1}{2}}] \\
\leq&(\frac{1}{T}\sum_{t=1}^{T}\beta_t^{-1})^{\frac{1}{2}}\mathbb{E}_{\xi^{[T]}}[(\frac{1}{T}\sum_{t=1}^{T}\beta_t(\|c_{\mathcal{E}}(x^{t+1})\| + \|[c_{\mathcal{I}}(x^{t+1})]_+\|))^{\frac{1}{2}}] \leq (\frac{1}{T}\sum_{t=1}^{T}\beta_t^{-1})^{\frac{1}{2}}\sqrt{\tilde{\Lambda}+\Lambda}
\end{aligned} \tag{44}$$

which indicates (41) from $a^{\frac{1}{2}} + b^{\frac{1}{2}} \leq \sqrt{2}(a+b)^{\frac{1}{2}}$ and the setting of β_t. The remainder of the proof is to show (42). Note that if $c_i(x^{t+1}) \leq -\frac{\lambda_i^{t+1}}{\beta_t}$, $\tilde{\lambda}_i^{t+1} = 0$. Then by using notations

$$\mathcal{I}_1 := \{i \in \mathcal{I} : c_i(x^{t+1}) > 0\}, \quad \mathcal{I}_2 := \{i \in \mathcal{I} : -\frac{\lambda_i^t}{\beta_t} \leq c_i(x^{t+1}) \leq 0\}, \tag{45}$$

we derive

$$\begin{aligned}
\sum_{i\in\mathcal{I}}(\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})|)^{\frac{1}{4}} &= \sum_{i\in\mathcal{I}_1}(\beta_t(c_i(x^{t+1}))^2+\lambda_i^{t+1}c_i(x^{t+1}))^{\frac{1}{4}} + \sum_{i\in\mathcal{I}_2}(-\lambda_i^{t+1}c_i(x^{t+1})-\beta_t(c_i(x^{t+1}))^2)^{\frac{1}{4}} \\
&\le \sum_{i\in\mathcal{I}_1}(\beta_t^{\frac{1}{4}}c_i^{\frac{1}{2}}(x^{t+1})+(\lambda_i^{t+1}c_i(x^{t+1}))^{\frac{1}{4}}) + \sum_{i\in\mathcal{I}_2}(-\lambda_i^{t+1}c_i(x^{t+1}))^{\frac{1}{4}} \\
&\le \sum_{i\in\mathcal{I}_1}\beta_t^{\frac{1}{4}}c_i^{\frac{1}{2}}(x^{t+1}) + \sum_{i\in\mathcal{I}_1\cup\mathcal{I}_2}(\lambda_i^{t+1}|c_i(x^{t+1})|)^{\frac{1}{4}} \\
&\le \beta_t^{\frac{1}{4}}\sum_{i\in\mathcal{I}_1}c_i^{\frac{1}{2}}(x^{t+1}) + \Lambda^{\frac{1}{4}}\sum_{i\in\mathcal{I}_1}c_i^{\frac{1}{4}}(x^{t+1}) + \Lambda^{\frac{1}{2}}|\mathcal{I}_2|\beta_t^{-\frac{1}{4}} \\
&\le |\mathcal{I}_1|^{\frac{3}{4}}\beta_t^{\frac{1}{4}}\|c_{\mathcal{I}_1}(x^{t+1})\|^{\frac{1}{2}} + |\mathcal{I}_1|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}\|c_{\mathcal{I}}(x^{t+1})\|^{\frac{1}{4}} + \Lambda^{\frac{1}{2}}|\mathcal{I}_2|\beta_t^{-\frac{1}{4}} \\
&\le |\mathcal{I}|^{\frac{3}{4}}\beta_t^{\frac{1}{4}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{1}{2}} + |\mathcal{I}|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{1}{4}} + \Lambda^{\frac{1}{2}}|\mathcal{I}|\beta_t^{-\frac{1}{4}}, \qquad (46)
\end{aligned}$$

where the last inequality holds by Jensen's inequality. Then taking expectation with respect to $\xi^{[t]}$ on both sides of (46) and sum-averaging over the first T iterations yield

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\sum_{i\in\mathcal{I}}(\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})|)^{\frac{1}{4}}] \\
\le &\frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}_{\xi^{[t]}}[|\mathcal{I}|^{\frac{3}{4}}\beta_t^{\frac{1}{4}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{1}{2}} + |\mathcal{I}|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{1}{4}} + \Lambda^{\frac{1}{2}}|\mathcal{I}|\beta_t^{-\frac{1}{4}}]) \\
\le &|\mathcal{I}|^{\frac{3}{4}}(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\beta_t\|[c_{\mathcal{I}}(x^{t+1})]_+\|])^{\frac{1}{2}}(\frac{1}{T}\sum_{t=1}^{T}\beta_t^{-\frac{1}{2}})^{\frac{1}{2}} + |\mathcal{I}|^{\frac{7}{8}}\Lambda^{\frac{1}{4}}(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{1}{2}}])^{\frac{1}{2}} \\
&+ \frac{|\mathcal{I}|\Lambda^{\frac{1}{2}}}{T}\sum_{t=1}^{T}\beta_t^{-\frac{1}{4}},
\end{aligned}$$

where the second equality comes from Cauchy–Schwarz inequality and $(\mathbb{E}[u])^2 \le \mathbb{E}[u^2]$ for a random variable $u > 0$. Then the desired result is concluded by (43) and (44). □

Remark 3 *Assumption 5 plays a crucial role in the convergence analysis of MLALM. To verify the frequency that Assumption 5 holds at iterates, we provide an example in Appendix C and report the experiment results for different cases. In addition, more discussions on CQ conditions and potential extensions are presented in Appendix D.*

4.2 Bounded penalty parameters

In this subsection, we analyze convergence properties of MLALM when the penalty parameters are bounded. Without loss of generality, we assume that $\beta_t \equiv \beta$ for all $t \ge 1$.

Theorem 3 *Suppose that Assumptions 1-4 hold, and the parameters used in Algorithm 1 satisfy (33) and $\beta_t \equiv \beta$, $t \ge 1$. Then (34) holds and there exists a positive constant δ_1 (independent of β) such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{t+1})c_{\mathcal{E}}(x^{t+1}) + \nabla c_{\mathcal{I}}(x^{t+1})[c_{\mathcal{I}}(x^{t+1})]_+, -\mathcal{N}_X(x^{t+1}))] \le \delta_1\beta^{-2}. \qquad (47)$$

Furthermore, assume that $|\mathcal{J}_t| = t^q$, $t \geq 1$, with $q > 1$, and there exists $C_u > 0$ such that $\mathbb{E}[f(x^t) + h(x^t)] \leq C_u$ for any $t \geq 1$. Then there exists a positive constant δ_2 (independent of β) such that for all sufficiently large T,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|c_{\mathcal{E}}(x^t)\|^2 + \|[c_{\mathcal{I}}(x^t)]_+\|^2] \leq C_{vio}^1 + \delta_2\beta^{-1}, \tag{48}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\sum_{i\in\mathcal{I}}(\tilde{\lambda}_i^{t+1})^{\frac{1}{2}}|c_i(x^{t+1})|] \leq |\mathcal{I}|^{\frac{1}{4}}(\beta^{\frac{1}{2}}(C_{vio}^1 + \delta_2\beta^{-1})^{3/4} + \Lambda^{\frac{1}{2}}(C_{vio}^1 + \delta_2\beta^{-1})^{1/2} + \beta^{-1}\Lambda^{3/2}), \tag{49}$$

where $C_{vio}^1 = \|c_{\mathcal{E}}(x^1)\|^2 + \|[c_{\mathcal{I}}(x^1)]_+\|^2$.

Proof. In analogy to Theorem 1, by analyzing the upper bound in (29) and following the parameter setting, we can obtain (34). Lemma 7 can be applied to derive (47).

To prove (48), it suffices to analyze the bound of (38) when $\beta_t \equiv \beta$ for all $t \geq 1$. Following the analysis to (38) in previous theorem and by Lemma A.2, we can obtain

$$\begin{aligned}0 \geq \frac{1}{T}\sum_{p=1}^{T}\sum_{t=1}^{p}(A_t - B_t) = &\frac{1}{2T}\sum_{p=1}^{T}\mathbb{E}_{\xi^{[p]}}[\|c_{\mathcal{E}}(x^{p+1})\|^2 + \|[c_{\mathcal{I}}(x^{p+1})]_+\|^2] \\ &- \frac{1}{2}(\|c_{\mathcal{E}}(x^1)\|^2 + \|[c_{\mathcal{I}}(x^1)]_+\|^2) + \mathcal{O}(\beta^{-1}).\end{aligned}$$

Due to the non-increasing property of $A_t - B_t$ we derive (48).

To prove (49), it follows from the definitions of $\mathcal{I}_1$ and $\mathcal{I}_2$ in (45) that

$$\begin{aligned}\sum_{i\in\mathcal{I}}(\tilde{\lambda}_i^{t+1})^{\frac{1}{2}}|c_i(x^{t+1})| = &\sum_{i\in\mathcal{I}_1\cup\mathcal{I}_2}([\lambda_i^{t+1} + \beta c_i(x^{t+1})]_+)^{\frac{1}{2}}|c_i(x^{t+1})| \\ \leq &\beta^{\frac{1}{2}}\sum_{i\in\mathcal{I}_1}(c_i(x^{t+1}))^{\frac{3}{2}} + \sum_{i\in\mathcal{I}_1}(\lambda_i^{t+1})^{\frac{1}{2}}c_i(x^{t+1}) + \sum_{i\in\mathcal{I}_2}(\lambda_i^{t+1})^{\frac{1}{2}}|c_i(x^{t+1})| \\ \leq &|\mathcal{I}_1|^{\frac{1}{4}}(\beta^{\frac{1}{2}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{3}{2}} + \Lambda^{\frac{1}{2}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|) + |\mathcal{I}_2|^{\frac{1}{4}}\beta^{-1}\Lambda^{\frac{3}{2}} \\ \leq &|\mathcal{I}|^{\frac{1}{4}}(\beta^{\frac{1}{2}}\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{3}{2}} + \Lambda^{\frac{1}{2}}\|[c_{\mathcal{I}}(x^{t+1})]_+\| + \beta^{-1}\Lambda^{\frac{3}{2}}).\end{aligned}$$

Therefore, taking the expectation of the above formula and the average over $t = 1, \ldots, T$ yields

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\sum_{i\in\mathcal{I}}(\tilde{\lambda}_i^{t+1})^{\frac{1}{2}}|c_i(x^{t+1})|] \\ \leq &|\mathcal{I}|^{\frac{1}{4}}(\frac{\beta^{\frac{1}{2}}}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|^{\frac{3}{2}}] + \frac{\Lambda^{\frac{1}{2}}}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|] + \beta^{-1}\Lambda^{\frac{3}{2}}) \\ \leq &|\mathcal{I}|^{\frac{1}{4}}(\frac{\beta^{\frac{1}{2}}}{T}\sum_{t=1}^{T}(\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|^2])^{\frac{3}{4}} + \frac{\Lambda^{\frac{1}{2}}}{T}\sum_{t=1}^{T}(\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|^2])^{\frac{1}{2}} + \beta^{-1}\Lambda^{\frac{3}{2}}) \\ \leq &|\mathcal{I}|^{\frac{1}{4}}(\beta^{\frac{1}{2}}(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|^2])^{\frac{3}{4}} + \Lambda^{\frac{1}{2}}(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|[c_{\mathcal{I}}(x^{t+1})]_+\|^2])^{\frac{1}{2}} + \beta^{-1}\Lambda^{\frac{3}{2}}) \\ \leq &|\mathcal{I}|^{\frac{1}{4}}(\beta^{\frac{1}{2}}(C_{vio}^1 + \delta_2\beta^{-1})^{\frac{3}{4}} + \Lambda^{\frac{1}{2}}(C_{vio}^1 + \delta_2\beta^{-1})^{\frac{1}{2}} + \beta^{-1}\Lambda^{\frac{3}{2}}),\end{aligned}$$

where the second and third inequalities come from Jensen's inequality. □

Remark 4 *It is noteworthy that Theorem 3 does not assume Assumption 5. However, if we do impose this assumption, we can actually obtain stronger results, which can be seen in Subsection 5.2.*

5 Oracle complexity analysis

In order to investigate oracle complexities of MLALM, we limit the maximum number of iterations to a fixed integer, denoted by T with a little abuse of notation, and set the batch size

$$|\mathcal{J}_t| = 1, \quad t = 1, \ldots, T.$$

We randomly select an iterate x^{R+1} as the output, where R follows a uniform distribution over $\{1, \ldots, T\}$. To characterize the output of the algorithm, we define two types of approximate solutions for problem (1), and we will analyze oracle complexities of MLALM to find those solutions accordingly.

Definition 1 *Given $\epsilon > 0$, we call $x \in X$ an ϵ-stationary point of (1), if there exists $\lambda \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ with $\lambda_i \geq 0,\ i \in \mathcal{I}$, such that*

$$\mathbb{E}[\mathbf{d}^2(\nabla f(x) + \partial h(x) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i \nabla c_i(x), -\mathcal{N}_X(x))] \leq \epsilon^2, \tag{50}$$

$$\mathbb{E}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x) c_{\mathcal{E}}(x) + \nabla c_{\mathcal{I}}(x)[c_{\mathcal{I}}(x)]_+, -\mathcal{N}_X(x))] \leq \epsilon^2. \tag{51}$$

Definition 2 *Given $\epsilon > 0$, we call $x \in X$ an ϵ-KKT point of (1), if there exists $\lambda \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ with $\lambda_i \geq 0,\ i \in \mathcal{I}$, such that (50) holds and*

$$\mathbb{E}[\|c_{\mathcal{E}}(x)\|^2 + \|c_{\mathcal{I}}(x)]_+\|^2] \leq \epsilon^2, \quad \mathbb{E}[\sum_{i\in\mathcal{I}} \lambda_i |c_i(x)|] \leq \epsilon.$$

5.1 Towards an ϵ-stationary point

In this subsection, we will analyze the complexity of MLALM for finding an ϵ-stationary point of (1). Towards this end, we need to first estimate the stationarity measure, i.e., L.H.S. of (50), at x^{R+1}.

Lemma 12 *Under Assumptions 1-4 and (24), set $\rho_t \equiv \frac{\rho}{T}$ and positive parameters $\beta_t \equiv \beta_1,\ \eta_t \equiv \eta_1,\ \alpha_t \equiv \alpha_1,\ t \geq 1$, then it holds that with $\tilde{\lambda}$ defined through (19),*

$$\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial h(x^{R+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{R+1} \nabla c_i(x^{R+1}), -\mathcal{N}_X(x^{R+1}))]$$

$$\leq \frac{4m^2\rho^2\tilde{C}^2G^2}{T^2} + \frac{1}{T}(\frac{224\eta_1 L^2}{\alpha_1} + \frac{20}{\eta_1})(\mathcal{L}_{\beta_1}(x^1, \lambda^1) - C^* + C\|\lambda^1\|_1 + 2m\rho\tilde{C}^2) + \frac{28(\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2] + 2\sum_{t=1}^T \alpha_t^2\sigma^2)}{\alpha_1 T}.$$

Proof. Similar to the proof of Lemma 10, we first give an upper bound on the stationarity measure as shown in (50). By taking expectation on both sides of (13) and average over $t = 1, \ldots, T$, we obtain

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla_x \phi_{\beta_t}(x^{t+1}, \lambda^{t+1}) + \partial h(x^{t+1}), -\mathcal{N}_X(x^{t+1}))]$$

$$\leq \frac{4(m\rho\tilde{C}G)^2}{T^2} + \frac{4}{T}\sum_{t=1}^T \mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] + \frac{4}{T}\sum_{t=1}^T (L_{\beta_t}^2 + \frac{1}{\eta_t^2})\mathbb{E}_{\xi^{[t]}}[\|x^{t+1} - x^t\|^2]. \tag{52}$$

From (25) with $\alpha_t = \alpha_1$ and $\beta_t = \beta_1$, it follows that

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] \leq \frac{2\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2] + 4\sum_{t=1}^T \alpha_t^2\sigma^2}{\alpha_1 T} + \frac{16\eta_1 L^2}{\alpha_1 T}(\mathcal{L}_{\beta_1}(x^1, \lambda^1) - C^* + C\|\lambda^1\|_1 + 2m\tilde{C}^2\sum_{t=1}^T \rho_t). \tag{53}$$

For the last term in R.H.S. of (52), it is easy to attain from (26) that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(L_{\beta_t}^2+\frac{1}{\eta_t^2})\mathbb{E}_{\xi^{[t]}}[\|x^{t+1}-x^t\|^2]\\
\leq&\frac{1}{T}\sum_{t=1}^{T}\frac{2(1+\eta_t^2L_{\beta_t}^2)}{\eta_t(1-\eta_tL_{\beta_t})}\mathbb{E}_{\xi^{[t]}}[(\mathcal{L}_{\beta_t}(x^t,\lambda^t)-\mathcal{L}_{\beta_{t+1}}(x^{t+1},\lambda^{t+1})+\frac{\eta_t}{2}\|\varepsilon^t\|^2+m\rho_t\tilde{C}^2)]\\
\leq&\frac{5}{\eta_1T}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+2m\rho\tilde{C}^2)+\frac{5}{2T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2],
\end{aligned}\tag{54}$$

where the last inequality is due to $\frac{1+u^2}{1-u}\leq\frac{5}{2}$ for $0<u\leq\frac{1}{2}$ and $\eta_t=\eta_1$ for all $t\in[T]$. Then, plugging (54) and (53) into (52) together with the definition of $\tilde{\lambda}$ in (19), we obtain

$$\begin{aligned}
&\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1})+\partial h(x^{R+1})+\sum_{i\in\mathcal{E}\cup\mathcal{I}}\tilde{\lambda}_i^{R+1}\nabla c_i(x^{R+1}),-\mathcal{N}_X(x^{R+1}))]\\
=&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\mathbf{d}^2(\nabla_x\phi_{\beta_t}(x^{t+1},\lambda^{t+1})+\partial h(x^{t+1}),-\mathcal{N}_X(x^{t+1}))]\\
\leq&\frac{4m^2\rho^2\tilde{C}^2G^2}{T^2}+\frac{20}{\eta_1T}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+2m\rho\tilde{C}^2)+\frac{14}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2]\\
\leq&\frac{4m^2\rho^2\tilde{C}^2G^2}{T^2}+\frac{1}{T}(\frac{224\eta_1L^2}{\alpha_1}+\frac{20}{\eta_1})(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+2m\rho\tilde{C}^2)+\frac{28(\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+2\sum_{t=1}^{T}\alpha_t^2\sigma^2)}{\alpha_1T}
\end{aligned}$$

which yields the conclusion. □

As in general it may be intractable to find a feasible solution for a nonconvex constrained optimization problem. Hence, we need to characterize the infeasible stationarity measure.

Lemma 13 *Under the conditions of Lemma 12, it holds that*

$$\begin{aligned}
&\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{R+1})c_{\mathcal{E}}(x^{R+1})+\nabla c_{\mathcal{I}}(x^{R+1})[c_{\mathcal{I}}(x^{R+1})]_+,-\mathcal{N}_X(x^{R+1}))]\\
\leq&\frac{4}{\beta_1^2}((2+m^2(\|\lambda^1\|+\rho C)^2)G^2+\frac{4m^2\rho^2\tilde{C}^2G^2}{T^2}\\
&+\frac{1}{T}(\frac{224\eta_1L^2}{\alpha_1}+\frac{20}{\eta_1})(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+C\|\lambda^1\|_1+2m\rho\tilde{C}^2)+\frac{28(\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+2\sum_{t=1}^{T}\alpha_t^2\sigma^2)}{\alpha_1T}).
\end{aligned}$$

Proof. By applying (20) and taking expectation with respect to R and $\xi^{[T]}$ on both sides, we obtain the conclusion from the setting of β_t. □

Selecting appropriate parameters is crucial as the ultimate oracle complexities depend on these choices. To ensure (24), we set the parameters as follows:

$$\beta_t=\beta_0T^{\iota},\ \eta_t=\frac{\eta}{T^{\iota}\max\{L,\tilde{L}\}},\ \alpha_t=\frac{8\alpha\eta^2}{T^{\tau}},\tag{55}$$

where $\beta_0\geq0,\tau\leq2\iota,0<\eta\leq\min\{\frac{1}{2\beta_0},\frac{\sqrt{2}}{4}\}$, $\alpha\in[1,\frac{1}{8\eta^2}]$ are given constants that are independent of T. Then the upper bound shown in Lemma 12 is of order

$$\begin{aligned}
&\mathcal{O}(T^{-2}+T^{-1}(T^{\tau-\iota}+T^{\iota})(\mathcal{L}_{\beta_1}(x^1,\lambda^1)+1)+T^{\tau-1}\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+T^{-\tau})\\
=&\ \mathcal{O}(T^{-2}+T^{\iota-1}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)+1)+T^{\tau-1}\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+T^{-\tau}).
\end{aligned}\tag{56}$$

Obviously, when $\tau = 2\iota = \frac{1}{2}$, above order can reach the lowest order. We summarize above analysis into the following theorem.

Theorem 4 *Under Assumptions 1-4, suppose that*

$$\rho_t \equiv \frac{\rho}{T},\ \beta_t = \beta_0 T^{1/4},\ \eta_t = \frac{\eta}{T^{1/4}\max\{L, \tilde{L}\}},\ \alpha_t = \frac{8\alpha\eta^2}{T^{1/2}},\quad \forall t \in [T]$$

with $\rho \in (0, T^{5/4}]$, $\beta_0 \geq 0$, $0 < \eta \leq \min\{\frac{1}{2\beta_0}, \frac{\sqrt{2}}{4}\}$, $\alpha \in [1, \frac{1}{8\eta^2}]$ *being constants independent of* T. *Then it holds that with* $\tilde{\lambda}$ *defined through* (19),

$$\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla f(x^{R+1}) + \partial h(x^{R+1}) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \tilde{\lambda}_i^{R+1}\nabla c_i(x^{R+1}), -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(T^{-1/2}), \tag{57}$$

$$\mathbb{E}_{R;\xi^{[T]}}[\mathbf{d}^2(\nabla c_{\mathcal{E}}(x^{R+1})c_{\mathcal{E}}(x^{R+1}) + \nabla c_{\mathcal{I}}(x^{R+1})[c_{\mathcal{I}}(x^{R+1})]_+, -\mathcal{N}_X(x^{R+1}))] = \mathcal{O}(T^{-1/2}). \tag{58}$$

Consequently, the oracle complexity of MLALM to reach an ϵ*-stationary point of* (1) *is of order* $\mathcal{O}(\epsilon^{-4})$.

Proof. It is straightforward to obtain (57) by Lemma 12 together with the analysis to (56). Under the parameter setting (55), the upper bound in Lemma 13 is of order $\mathcal{O}(T^{-2\iota}+T^{-\iota-1}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)+1)+T^{\tau-1-2\iota}\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]+T^{-\tau-2\iota})$. Then (58) can be derived from $\iota = \frac{1}{4}$. Hence, to achieve an ϵ-stationary point of (1), the maximum number of iterations T should be of order $\mathcal{O}(\epsilon^{-4})$. Since the computation of the stochastic gradient only requires sampling once per iteration, as shown in (4), the total number of stochastic oracle calls is in the order of $\mathcal{O}(T)$, which is of order $\mathcal{O}(\epsilon^{-4})$. □

It is important to note that Lemmas 12 and 13 demonstrate that the term $\mathcal{L}_{\beta_1}(x^1,\lambda^1)$ explicitly appears in the upper bounds. Then the term $\frac{\beta_1}{2}(\|c_{\mathcal{E}}(x^1)\|^2 + \|[c_{\mathcal{I}}(x^1)]_+\|^2)$ directly affects these upper bounds. In particular, we need to consider the impact of potentially large values of β_1 on the complexity order derived in (56). However, when the initial point is sufficiently close to the feasible region, such that the aforementioned term is of order $\mathcal{O}(1)$, the influence of large β_1 on the complexity order can be reduced. Besides, $\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2]$ also affects the order in (56). Therefore, if we sample T^ι times at the initial point, one has $\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2] = \mathcal{O}(T^{-\iota})$. In this case, under the parameter setting (55) the upper bound in Lemma 12 is of order

$$\mathcal{O}(T^{-2} + T^{\iota-1} + T^{\tau-1-\iota} + T^{-\tau}).$$

Thus to achieve the lowest order $\mathcal{O}(T^{-\frac{2}{3}})$ we can choose $\tau = 2\iota = \frac{2}{3}$. We can also obtain that the upper bound in Lemma 13 is of order $\mathcal{O}(T^{-\frac{2}{3}})$. We thus derive the following corollary, with the proof omitted.

Corollary 1 *Under Assumptions 1-4, suppose that*

$$\rho_t \equiv \frac{\rho}{T},\ \beta_t = \beta_0 T^{1/3},\ \eta_t = \frac{\eta}{T^{1/3}\max\{L, \tilde{L}\}},\ \alpha_t = \frac{8\alpha\eta^2}{T^{2/3}},\quad \forall t \in [T]$$

with $\rho \in (0, T^{5/4}]$, $\beta_0 \geq 0$, $0 < \eta \leq \min\{\frac{1}{2\beta_0}, \frac{\sqrt{2}}{4}\}$, $\alpha \in [1, \frac{1}{8\eta^2}]$ *being constants independent of* T. *If* $\|c_{\mathcal{E}}(x^1)\|^2 + \|[c_{\mathcal{I}}(x^1)]_+\|^2 = \mathcal{O}(T^{-1/3})$ *and* $\mathbb{E}_{\xi^1}[\|\varepsilon^1\|^2] = \mathcal{O}(T^{-1/3})$, *the oracle complexity to reach an* ϵ*-stationary point of* (1) *is of order* $\mathcal{O}(\epsilon^{-3})$.

5.2 Towards an ϵ-KKT point

In this subsection, we will analyze the oracle complexity of MLALM towards an ϵ-KKT point of (1). With the help of Assumption 5 we can obtain the following lemma characterizing the near-feasibility of x^{R+1} and the complementary slackness in expectation.

Theorem 5 *Suppose that Assumption 5 and the conditions of Theorem 4 hold, then*

$$\mathbb{E}_{R;\xi^{[T]}}[\|c_{\mathcal{E}}(x^{R+1})\|^2 + \|[c_{\mathcal{I}}(x^{R+1})]_+\|^2] = \mathcal{O}(T^{-1/2}), \tag{59}$$

$$\mathbb{E}_{R;\xi^{[T]}}[\sum_{i\in\mathcal{I}}\tilde{\lambda}_i^{R+1}|c_i(x^{R+1})|] = \mathcal{O}(T^{-1/4}). \tag{60}$$

Consequently, the oracle complexity of MLALM to reach an ϵ-KKT point of (1) is of order $\mathcal{O}(\epsilon^{-4})$.

Proof. Under Assumption 5, it is easy to obtain

$$c_i(x^t)^{\mathrm{T}}\nabla c_i(x^t)^{\mathrm{T}}z^t = -\delta|c_i(x^t)|, \quad i\in\mathcal{E}: c_i(x^t)\neq 0,$$

$$[c_i(x^t)]_+^{\mathrm{T}}\nabla c_i(x^t)^{\mathrm{T}}z^t \leq -\delta[c_i(x^t)]_+, \quad i\in\mathcal{I}: c_i(x^t) > 0,$$

which further yields

$$c_{\mathcal{E}}(x^t)^{\mathrm{T}}\nabla c_{\mathcal{E}}(x^t)^{\mathrm{T}}z + [c_{\mathcal{I}}(x^t)]_+^{\mathrm{T}}\nabla c_{\mathcal{I}}(x^t)^{\mathrm{T}}z \leq -\delta(\|c_{\mathcal{E}}(x^t)\|_1 + \|[c_{\mathcal{I}}(x^t)]_+\|_1).$$

Moreover, since $v^{\mathrm{T}}z\leq 0$ for all $v\in\mathcal{N}_X(x^t)$, we have

$$c_{\mathcal{E}}(x^t)^{\mathrm{T}}\nabla c_{\mathcal{E}}(x^t)^{\mathrm{T}}z + [c_{\mathcal{I}}(x^t)]_+^{\mathrm{T}}\nabla c_{\mathcal{I}}(x^t)^{\mathrm{T}}z^t + v^{\mathrm{T}}z \leq -\delta(\|c_{\mathcal{E}}(x^t)\|_1 + \|[c_{\mathcal{I}}(x^t)]_+\|_1).$$

Hence, it holds that for any $v\in\mathcal{N}_X(x^t)$,

$$\begin{aligned}\delta(\|c_{\mathcal{E}}(x^t)\|_1 + \|[c_{\mathcal{I}}(x^t)]_+\|_1) \leq& |c_{\mathcal{E}}(x^t)^{\mathrm{T}}\nabla c_{\mathcal{E}}(x^t)^{\mathrm{T}}z^t + [c_{\mathcal{I}}(x^t)]_+^{\mathrm{T}}\nabla c_{\mathcal{I}}(x^t)^{\mathrm{T}}z^t + v^{\mathrm{T}}z^t|\\ \leq& \|\nabla c_{\mathcal{E}}(x^t)c_{\mathcal{E}}(x^t) + \nabla c_{\mathcal{I}}(x^t)[c_{\mathcal{I}}(x^t)]_+ + v\|\|z^t\|.\end{aligned}$$

Due to $\|z^t\|\leq Z$ and the arbitrariness of $v\in\mathcal{N}_X(x^t)$, we obtain

$$\frac{\delta}{Z}(\|c_{\mathcal{E}}(x^t)\|_1 + \|[c_{\mathcal{I}}(x^t)]_+\|_1) \leq \mathbf{d}(\nabla c_{\mathcal{E}}(x^t)c_{\mathcal{E}}(x^t) + \nabla c_{\mathcal{I}}(x^t)[c_{\mathcal{I}}(x^t)]_+, -\mathcal{N}_X(x^t)).$$

Therefore, (59) can be derived from (58).

In light of $|c_i(x^{t+1})| = [c_i(x^{t+1})]_+ + [c_i(x^{t+1})]_-$ for $i\in\mathcal{I}$ and using the notations

$$\mathcal{I}_1^{t+1} := \{i \mid c_i(x^{t+1})\geq 0\}, \quad \mathcal{I}_2^{t+1} := \{i \mid -\frac{\lambda_i^{t+1}}{\beta_t} \leq c_i(x^{t+1}) < 0\},$$

we can derive the following relations

$$\begin{aligned}\sum_{i\in\mathcal{I}}\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})| =& \sum_{i\in\mathcal{I}_1^{t+1}}[\beta_t c_i^2(x^{t+1}) + \lambda_i^{t+1}c_i(x^{t+1})] + \sum_{i\in\mathcal{I}_2^{t+1}}[-c_i(x^{t+1})(\beta_t c_i(x^{t+1}) + \lambda_i^{t+1})]\\ \leq& \sum_{i\in\mathcal{I}_1^{t+1}}[\beta_t c_i^2(x^{t+1}) + \lambda_i^{t+1}c_i(x^{t+1})] + \sum_{i\in\mathcal{I}_2^{t+1}}\frac{|\lambda_i^{t+1}|^2}{\beta_t}\\ \leq& \beta_t\sum_{i\in\mathcal{I}_1^{t+1}}c_i^2(x^{t+1}) + \rho C\sum_{i\in\mathcal{I}_1^{t+1}}c_i(x^{t+1}) + \frac{m\rho^2C^2}{\beta_t},\end{aligned} \tag{61}$$

where the first inequality comes from $-u(bu+a)\leq\frac{a^2}{b}, \forall u\in\mathbb{R}$ with $b>0$, and the second inequality is due to $|\lambda_i^{t+1}|\leq\lambda_i^1+\rho C,\ \forall i\in\mathcal{I}$ by (7) and $\rho_t=\frac{\rho}{T}$. Then taking expectation with respect to R and $\xi^{[T]}$ on both sides of (61) yields

$$\begin{aligned}\mathbb{E}_{R;\xi^{[T]}}[\sum_{i\in\mathcal{I}}\tilde{\lambda}_i^{R+1}|c_i(x^{R+1})|] =& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[T]}}[\sum_{i\in\mathcal{I}}\tilde{\lambda}_i^{t+1}|c_i(x^{t+1})|]\\ \leq& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi^{[T]}}[\beta_t\sum_{i\in\mathcal{I}_1^{t+1}}c_i^2(x^{t+1}) + \rho C\sum_{i\in\mathcal{I}_1^{t+1}}c_i(x^{t+1})] + \frac{m\rho^2C^2}{\beta_t}\end{aligned}$$

$$= \beta_1 \mathbb{E}_{R;\xi^{[T]}}[\|[c_{\mathcal{I}}(x^{R+1})]_+\|^2] + \rho C|\mathcal{I}|\mathbb{E}_{R;\xi^{[T]}}[\|[c_{\mathcal{I}}(x^{R+1})]_+\|] + \frac{m\rho^2 C^2}{\beta_1}$$

$$= \mathcal{O}(T^{-1/4}),$$

where the third equality comes from (59) and $(\mathbb{E}[u])^2 \leq \mathbb{E}[u^2]$ for a random variable $u \in \mathbb{R}$.

Consequently, the oracle complexity of MLALM to reach an ϵ-KKT point of (1) is $\mathcal{O}(T)$ which is in order $\mathcal{O}(\epsilon^{-4})$. □

In analogy to Corollary 1, when the initial point is sufficiently close to the feasible region of (1), it is straightforward to obtain the following results.

Corollary 2 *Suppose that Assumption 5 and the conditions of Corollary 1 hold, then the oracle complexity of MLALM to reach an ϵ-KKT point of (1) is in order $\mathcal{O}(\epsilon^{-3})$.*

6 Numerical Simulations

In this section, we conduct a series of numerical experiments to evaluate the practical performance of MLALM, as outlined in Algorithm 1. All experiments were performed using Matlab 2021b on a 64-bit Linux machine equipped with a 4.90 GHz Intel Core i7-12700K CPU and 32GB of memory.

6.1 Quadratically Constrained Nonconvex Program

In this subsection, we test the proposed method on solving quadratically constrained nonconvex programs [21]:

$$\begin{aligned} \min_{x\in X} \quad & f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + \frac{1}{2}\|H_i x - c_i\|^2) \\ \text{s.t.} \quad & f_j(x) = \frac{1}{2}x^T Q_j x + a_j^T x \leq b_j, \quad j = 1, \ldots, M, \end{aligned} \tag{62}$$

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 $j \in [M]$, $Q_j \in \mathbb{R}^{n\times n}$ is sum of a random matrix and a diagonal matrix with elements randomly selected from $\mathrm{U}[-1, 1]$, and a_j is randomly generated following $\mathrm{U}[0.1, 1.1]^n$. Then we generate a random point $x_* \sim \mathrm{U}(0, 1)^n$ and set $c_i = H_i x_*$, $i \in [N]$, $b_j = \frac{1}{2}x_*^T Q_j x_* + a_j^T x_*$, $j \in [M]$. It is worth noting that x_* is feasible for (62) and has an objective value of $f(x_*) = 0$. Therefore, x_* is the optimal solution for problem (62). In this subsection, we present the average performance over 10 independent runs of the algorithm in each scenario.

To assess the impact of the parameter α_t on the numerical performance of (4), we implement experiments by solving (62). We select α_t from the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. In all experiments, we set $n = M = 50$, $p = 5$, $N = 1000$, and initialize the algorithm with $x^1 = \mathbf{0}$. The maximum number of iterations is set to $T = 2000$, and the penalty parameter is chosen as $\beta_t = T^{1/4}$. Additionally, we set $\rho_t = T^{1/4}$ and consider two step sizes, $\eta_t \in \{0.15/T^{1/4}, 0.17/T^{1/4}\}$. To obtain reliable results, in Figures 1 and 2 the left subplot presents the trend of average objective function values at all previous iterates, while the right one illustrates the average of constraint violation $\sum_{j=1}^{M}[f_j(x) - b_j]_+$ over past iterates. We observe that, for a given step size η_t, the best-performing value of α_t tends to lie within an intermediate range, rather than being the largest or smallest option. Additionally, by comparing the optimal α_t values depicted in Figures 1 and 2, we notice that as the step size η_t increases, the optimal α_t value also increases. This aligns with the positive correlation between α_t and η_t, as indicated in the parameter settings of Theorem 1. They highlight the importance of selecting an appropriate α_t value within an intermediate range to achieve optimal performance in terms of both the objective function value and the constraint violation.

We next compare the performance of MLALM with that of ICPPC [4] and LCSVRG [5] for solving problem (62). We adopt the following experimental settings. For those algorithms, we set $n \in \{50, 100\}$, $p = 5$, $N = 1000$, and $M \in \{50, 100\}$. The initial point is chosen as $x^1 = \mathbf{0}$, and the maximum number of iterations is set to $T = 2000$. For ICPPC, we set $t_0 = 2$, $\mathcal{M} = 0.1M$. All other parameters for ICPPC are set according to the requirements specified in [4]. It is important to note that for ICPPC we set the maximum umber of inner iterations as 2, based on its favorable performance observed in numerical tests. LCSVRG is the variance-reduced variant of LCPG [5], aiming

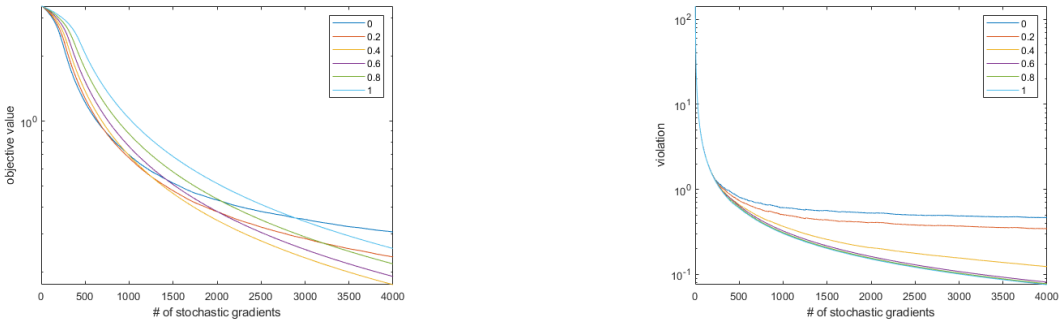


Figure 1: The impact of α_t on MLALM with $\eta_t = 0.15/T^{1/4}$

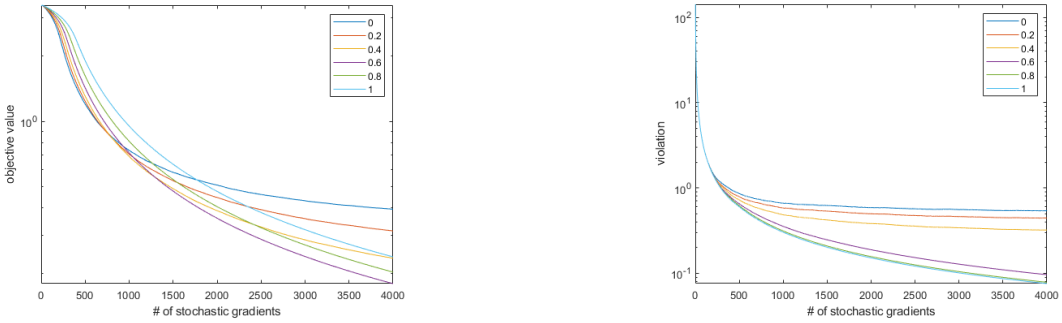


Figure 2: The impact of α_t on MLALM with $\eta_t = 0.17/T^{1/4}$

to solve finite-sum optimization problems. For LCSVRG, we set its batch size $|\mathcal{J}_t| = 30$, maximum outer iteration number $T = 50$, ensuring that the total complexity of computing stochastic gradient equals to that of the other two algorithms, i.e. $(1000 + 50 \times 30 \times 2)$. Regarding MLALM, we set $\eta_t = 0.05/T^{1/4}$, $\alpha_t = 0.5$, $\beta_t = T^{1/4}$, and $\rho_t = 10$. Figure 3 illustrates the performance of both algorithms in terms of objective function values and constraint violations on QCNP problems under different scenarios. All reported results are the average values obtained from 10 independent runs of each algorithm. The observations from the figures indicate that, within the same number of stochastic gradient evaluations, MLALM reduces the constraint violations much faster throughout the algorithm's progress, although the speed to reduce the objective function value by both algorithms is comparable.

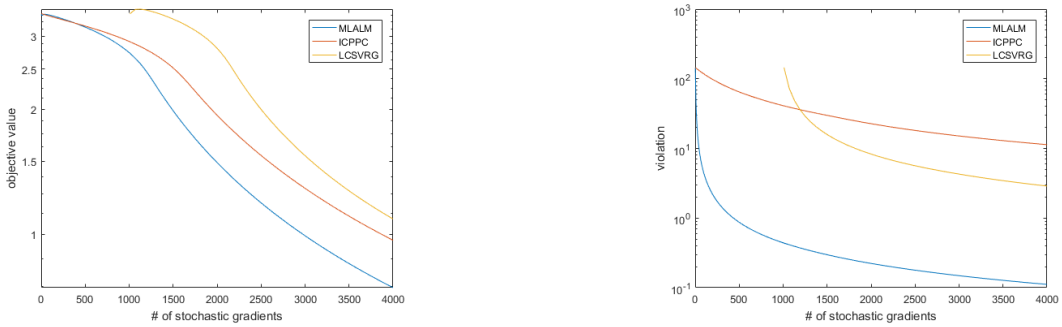
6.2 Multi-class Neyman-Pearson classification problems

In this subsection, we consider multi-class Neyman-Pearson classification (mNPC) problems [26]. The mNPC problem focuses on the task of learning K models x_k, where $k \in [K]$, in order to predict the class of a potential data point ξ by selecting the model that maximizes the inner product $x_k^T \xi$. More precisely, the optimization problem aims to minimize the loss associated with one specific class while controlling the loss values for the remaining classes. The problem formulation can be expressed as follows:

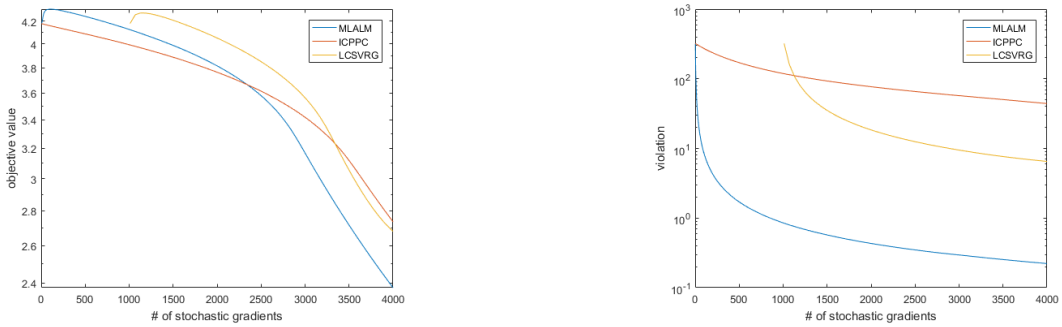
$$\begin{aligned} \min_{\|x_k\| \le \lambda,\, k \in [K]} \quad & \frac{1}{|\mathcal{D}_1|} \sum_{l>1} \sum_{\xi \in \mathcal{D}_1} h(x_1^T \xi - x_l^T \xi) \\ \text{s.t.} \quad & \frac{1}{|\mathcal{D}_k|} \sum_{l \neq k} \sum_{\xi \in \mathcal{D}_k} h(x_k^T \xi - x_l^T \xi) \le \gamma_k, \quad k = 2, \ldots, K, \end{aligned} \tag{63}$$

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 [8]: *covtype* $(K = 7)$ and *mnist* $(K = 10)$. Besides, we set $\gamma_k = 0.5(K - 1)$ and $\lambda = 0.3$.

We now compare MLALM with SPD [21] and ICPPC [4]. For these three algorithms, we set the maximum number of stochastic gradient computations is 4000. For MLALM, we set parameter $\alpha_t = 0.7$ for both datasets.



$n = 50, M = 50$



$n = 100, M = 100$

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

Other parameters are set same for both MLALM and SPD. Specifically, for *covtype* we set $\eta_t = 0.01/t^{1/4}$, $\beta_t = 5T^{1/4}$, $\rho_t = 0.67$, while for *mnist* we set $\eta_t = 0.005/t^{1/4}$, $\beta_t = T^{1/4}$, $\rho_t = 0.0067$. It is worth noting that the number of stochastic gradients calculated by SPD at tth iteration is $t^{1/4}$. Figures 4 and 5 present the performances of MLALM and SPD on the respective datasets for solving mNPC problems. The parameters of the ICPPC algorithm for the two datasets are selected as follows: For *covtype*, we set $\theta_t = 0.67$, $\tau_t = 2.5$ and $\eta_t = 2.6 \times 10^{-4}$ in subproblem solver ConEx; for *mnist*, we set $\theta_t = 0.67$, $\tau_t = 2.5$ and $\eta_t = 0.003$. The inner-iteration number is set to 2. All reported results are the average values obtained from 5 independent runs of each algorithm. From the figures we observe that for the *covtype* dataset, MLALM demonstrates a faster reduction in the objective function value compared to the other algorithms, while ICPPC maintains a lower level of constraint violations. Regarding the *mnist* dataset, MLALM exhibits superior performance in both the objective function value and the constraint violations. Furthermore, it is worth mentioning that MLALM outperforms SPD significantly, indicating that the incorporation of momentum brings notable benefits to the numerical performance.

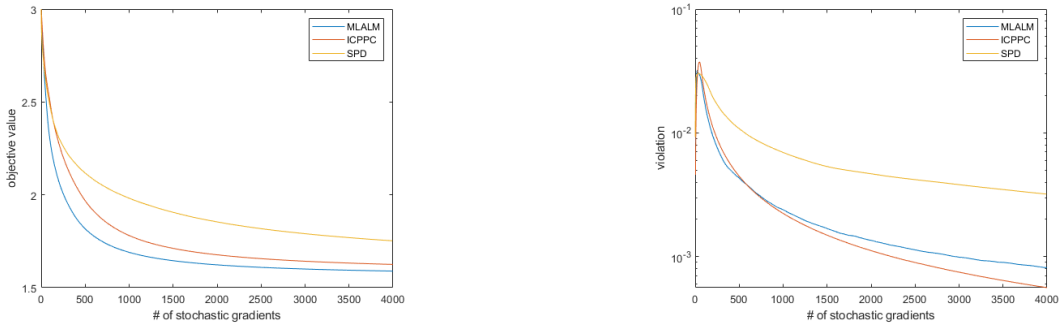


Figure 4: Comparison between MLALM, ICPPC and SPD on dataset *covtype*

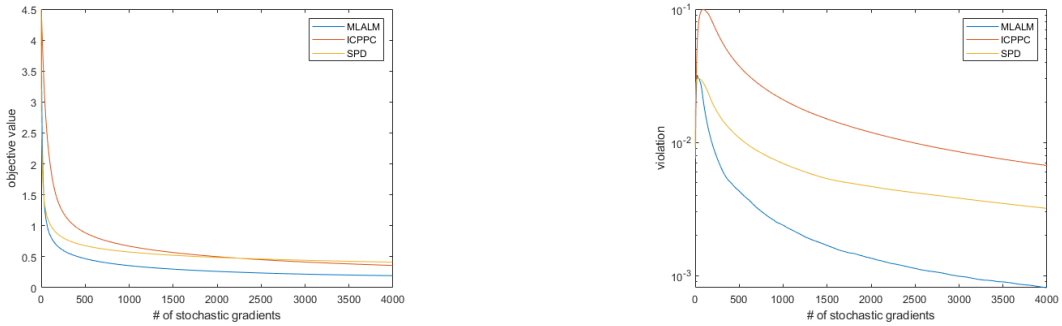


Figure 5: Comparison between MLALM, ICPPC and SPD on dataset *mnist*

7 Conclusion

In this paper, we present MLALM, a single-loop linearized augmented Lagrangian method for nonconvex optimization problems with an expectation function in the objective and deterministic functional constraints. To address potentially nonconvex constraints, we leverage the linearized augmented Lagrangian function to construct a stochastic approximation, enabling updates to the primal variable. Inspired by recent developments in momentum-type methods for unconstrained optimization in the literature, we introduce a momentum step to compute the stochastic gradient at each iteration. We conduct a comprehensive analysis of the global convergence properties of the proposed algorithm. Our analysis reveals that, under appropriate parameter settings and with unbounded increasing penalty parameters, the sequence of average stationarity measure in expectation converges to zero. Additionally, the sequence of average constraint violations exhibits convergence in expectation. Moreover, under a constraint qualification assumption, both the sequences of average expected constraint violations and complementary slackness measures converge to zero. We also investigate the properties of MLALM when penalty parameters are bounded. Furthermore, we analyze the oracle complexity of the algorithm in achieving an ϵ-stationary point and an ϵ-KKT point. Specifically, the oracle complexity to reach an ϵ-stationary point is of the order $\mathcal{O}(\epsilon^{-4})$. Under the constraint qualification assumption, the proposed algorithm can reach an ϵ-KKT point with oracle complexity bounded by $\mathcal{O}(\epsilon^{-4})$. When the initial point is nearly feasible within a certain accuracy, the complexity orders can be improved to $\mathcal{O}(\epsilon^{-3})$ accordingly. Finally, we implement numerical experiments on two types of test problems to evaluate the performance of our algorithm. The experimental results demonstrate promising performance and validate the effectiveness of our proposed approach.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 12271278), the Major Key Project of PCL (No. PCL2022A05) and the Natural Science Foundation of Shanghai (No. 21ZR1442800). The authors would like to thank Dr. Qi Deng for kindly sharing codes of LCPG.

References

[1] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Math. Program.*, 199:165–214, 2023.

[2] 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.

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

[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] D. Boob, Q. Deng, and G. Lan. Level constrained first order methods for function constrained optimization. *Math. Program*, 2024.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Math. Program*, page 431–483, 2023.

[12] 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.

[13] F. E. Curtis, D. P. Robinson, and B. Zhou. Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. *arXiv:2302.14790 [math.OC]*, 2023.

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

[15] 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.

[16] 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.

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

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

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

[20] Z. Jia and B. Grimmer. First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points. *arXiv:2212.00927 [math.OC]*, 2022.

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

[22] L. Jin and X. Wang. Stochastic nested primal-dual method for nonconvex constrained composition optimization. *Math. Comput.*, 2024.

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

[24] 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.

[25] 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.

[26] 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.

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

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

[29] S. Na, M. Anitescu, and M. Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Math. Program.*, pages 1–71, 2022.

[30] S. Na, M. Anitescu, and M. Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Math. Program*, 2023.

[31] S. Na and M. W. Mahoney. Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv preprint arXiv:2205.13687*, 2022.

[32] 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.

[33] 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.

[34] 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.

[35] G. D. Pillo and L. Grippo. On the exactness of a class of nondifferentiable penalty functions. *Journal of Optimization Theory and Applications*, 57:399–410, 1988.

[36] G. D. Pillo and L. Grippo. Exact penalty functions in constrained optimization. *SIAM Journal on Control and Optimization*, 27(6):1333–1360, 1989.

[37] 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.

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

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

[40] 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.

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

[42] 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.

[43] 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.

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

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

[46] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Math. Program.*, 191:1005–1071, 2022.

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

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

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

[50] 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.

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

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

[53] M. Xu, J. J. Ye, and L. Zhang. Smoothing augmented lagrangian method for nonsmooth constrained optimization problems. *Journal of Global Optimization*, 62(4):675–694, 2015.

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

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

[56] Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *J. Optim Theory Appl.*, 196(1):266–297, 2023.

[57] 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.

Appendix

A. Auxiliary lemmas

Lemma A.1 *Under the same conditions of Theorem 1, it holds that*

$$\lim_{p\to\infty}\sum_{t=1}^{p}\frac{\Psi_{\beta_t}(x^{t+1},\lambda^t)-\Psi_{\beta_t}(x^t,\lambda^t)}{\beta_t}-\frac{1}{2}(\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2)\ \textit{exists.}$$

Proof. By the definition of Ψ_{β_t}, we have

$$\begin{aligned}
&\sum_{t=1}^{p}\frac{\Psi_{\beta_t}(x^{t+1},\lambda^t)-\Psi_{\beta_t}(x^t,\lambda^t)}{\beta_t}-\frac{1}{2}(\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2)\\
=&\sum_{t=1}^{p}\sum_{i\in\mathcal{E}}(\frac{\lambda_i^t}{\beta_t}c_i(x^{t+1})+\frac{1}{2}c_i^2(x^{t+1})-\frac{\lambda_i^t}{\beta_t}c_i(x^t)-\frac{1}{2}c_i^2(x^t))-\frac{1}{2}\|c_{\mathcal{E}}(x^{p+1})\|^2\\
&+\sum_{t=1}^{p}\sum_{i\in\mathcal{I}}\frac{\psi_{\beta_t}(c_i(x^{t+1}),\lambda_i^t)-\psi_{\beta_t}(c_i(x^t),\lambda_i^t)}{\beta_t}-\frac{1}{2}\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2. \qquad (64)
\end{aligned}$$

We next consider the terms in (64) related with equality constraints and inequality constraints in the following part (a) and (b) separately.

(a) Note that

$$\sum_{t=1}^{p}\sum_{i\in\mathcal{E}}(\frac{\lambda_i^t}{\beta_t}c_i(x^{t+1})+\frac{1}{2}c_i^2(x^{t+1})-\frac{\lambda_i^t}{\beta_t}c_i(x^t)-\frac{1}{2}c_i^2(x^t))-\frac{1}{2}\|c_{\mathcal{E}}(x^{p+1})\|^2$$

$$=\frac{1}{2}(\|c_{\mathcal{E}}(x^{p+1})\|^2-\|c_{\mathcal{E}}(x^1)\|^2)+\sum_{i\in\mathcal{E}}\sum_{t=1}^{p}\frac{\lambda_i^t}{\beta_t}(c_i(x^{t+1})-c_i(x^t))-\frac{1}{2}\|c_{\mathcal{E}}(x^{p+1})\|^2$$

$$=-\frac{1}{2}\|c_{\mathcal{E}}(x^1)\|^2+\sum_{i\in\mathcal{E}}\sum_{t=1}^{p}\frac{\lambda_i^t}{\beta_t}(c_i(x^{t+1})-c_i(x^t)). \tag{65}$$

For the second term in above equation:

$$\sum_{i\in\mathcal{E}}\sum_{t=1}^{p}\frac{\lambda_i^t}{\beta_t}(c_i(x^{t+1})-c_i(x^t))=\sum_{i\in\mathcal{E}}\frac{\lambda_i^p}{\beta_p}c_i(x^{p+1})+\sum_{i\in\mathcal{E}}\sum_{t=1}^{p-1}(\frac{\lambda_i^t}{\beta_t}-\frac{\lambda_i^{t+1}}{\beta_{t+1}})c_i(x^{t+1}), \tag{66}$$

it is noteworthy that for any $i\in\mathcal{E}$, $\{\frac{\lambda_i^p}{\beta_p}c_i(x^{p+1})\}_{p\geq 1}$ converges to zero by the boundedness of λ^t and $c_i(x^t)$ for all $i\in\mathcal{E}$, $t\geq 1$, and

$$\begin{aligned}\sum_{t=1}^{p-1}(\frac{\lambda_i^t}{\beta_t}-\frac{\lambda_i^{t+1}}{\beta_{t+1}})c_i(x^{t+1})=&\sum_{t=1}^{p-1}(\frac{\lambda_i^t}{\beta_t}-\frac{\lambda_i^{t+1}}{\beta_t}+\frac{\lambda_i^{t+1}}{\beta_t}-\frac{\lambda_i^{t+1}}{\beta_{t+1}})c_i(x^{t+1})\\ =&\sum_{t=1}^{p-1}\frac{-\rho_t c_i^2(x^{t+1})}{\beta_t}+\sum_{t=1}^{p-1}(\frac{1}{\beta_t}-\frac{1}{\beta_{t+1}})\lambda_i^{t+1}c_i(x^{t+1})\\ =&\sum_{t=1}^{p-1}\frac{-\rho_t c_i^2(x^{t+1})}{\beta_t}+\sum_{t=1}^{p-1}\frac{(t+1)^{\frac{\iota}{2}}}{\beta_0 t^\iota}\frac{1}{(t+1)^{\frac{\iota}{2}}}(1-(\frac{t}{t+1})^\iota)\lambda_i^{t+1}c_i(x^{t+1}).\end{aligned} \tag{67}$$

Here, $|\sum_{t=1}^{p-1}-\rho_t c_i^2(x^{t+1})|$ is bounded following the setting of ρ_t. Since $\iota<1$, one has $1-(\frac{t}{t+1})^\iota<\frac{1}{t+1}$, then $|\sum_{t=1}^{p-1}\frac{1}{(t+1)^{\frac{\iota}{2}}}(1-(\frac{t}{t+1})^\iota)\lambda_i^{t+1}c_i(x^{t+1})|$ is bounded due to the boundedness of λ^t, $t\geq 1$. Then owing to the fact that both sequences $\{\frac{1}{\beta_t}\}$ and $\{\frac{(t+1)^{\frac{\iota}{2}}}{\beta_0 t^\iota}\}_{t\geq 1}$ are monotonically decreasing and converging to zero, it follows from Dirichlet's Test that $\sum_{t=1}^{p-1}(\frac{\lambda_i^t}{\beta_t}-\frac{\lambda_i^{t+1}}{\beta_{t+1}})c_i(x^{t+1})$ converges as $p\to\infty$, so does (66). Therefore, it is straightforward to obtain the convergence of L.H.S. of (65) as $p\to\infty$.

(b) We now consider the remaining part in (64) related with inequality constraints. Our goal is to prove its convergence as $p\to\infty$. Towards this end, for any $i\in\mathcal{I}$ we consider the following index sets:

$$\begin{aligned}\mathcal{T}_1&=\{t\in[p-1]:\beta_t c_i(x^{t+1})+\lambda_i^t\geq 0,\ \beta_{t+1}c_i(x^{t+1})+\lambda_i^{t+1}\geq 0\};\\ \mathcal{T}_2&=\{t\in[p-1]:\beta_t c_i(x^{t+1})+\lambda_i^t\geq 0,\ \beta_{t+1}c_i(x^{t+1})+\lambda_i^{t+1}<0\};\\ \mathcal{T}_3&=\{t\in[p-1]:\beta_t c_i(x^{t+1})+\lambda_i^t<0,\ \beta_{t+1}c_i(x^{t+1})+\lambda_i^{t+1}<0\};\\ \mathcal{T}_4&=\{t\in[p-1]:\beta_t c_i(x^{t+1})+\lambda_i^t<0,\ \beta_{t+1}c_i(x^{t+1})+\lambda_i^{t+1}\geq 0\}.\end{aligned} \tag{68}$$

Here we omit their dependence upon i. Obviously, $\cup_{j\in[4]}\mathcal{T}_j=[p-1]$. One can verify that $\mathcal{T}_4=\emptyset$. This is because whenever $\beta_t c_i(x^{t+1})+\lambda_i^t<0$, $c_i(x^{t+1})<0$ and $\lambda^{t+1}=\lambda^t-\frac{\rho_t\lambda^t}{\beta_t}<\lambda_t$ by (6), thus there must have

$\beta_{t+1}c_i(x^{t+1}) + \lambda^{t+1} < \beta_t c_i(x^{t+1}) + \lambda^t < 0$. Then the following relations hold from the definition of ψ_β:

$$\begin{aligned}
&\sum_{t=1}^{p} \frac{\psi_{\beta_t}(c_i(x^{t+1}), \lambda_i^t) - \psi_{\beta_t}(c_i(x^t), \lambda_i^t)}{\beta_t} - \frac{1}{2}([c_i(x^{p+1})]_+)^2 \\
=&\sum_{t=1}^{p} \frac{\psi_{\beta_t}(c_i(x^{t+1}), \lambda_i^t)}{\beta_t} - \sum_{t=0}^{p-1} \frac{\psi_{\beta_{t+1}}(c_i(x^{t+1}), \lambda_i^{t+1})}{\beta_{t+1}} - \frac{1}{2}([c_i(x^{p+1})]_+)^2 \\
=&\frac{\psi_{\beta_p}(c_i(x^{p+1}), \lambda_i^p)}{\beta_p} - \frac{\psi_{\beta_1}(c_i(x^1), \lambda_i^1)}{\beta_1} + \sum_{t=1}^{p-1}(\frac{\psi_{\beta_t}(c_i(x^{t+1}), \lambda_i^t)}{\beta_t} - \frac{\psi_{\beta_{t+1}}(c_i(x^{t+1}), \lambda_i^{t+1})}{\beta_{t+1}}) - \frac{1}{2}([c_i(x^{p+1})]_+)^2 \\
=&\frac{\lambda_i^p \max\{c_i(x^{p+1}), -\frac{\lambda_i^p}{\beta_p}\}}{\beta_p} + \frac{(\max\{c_i(x^{p+1}), -\frac{\lambda_i^p}{\beta_p}\})^2}{2} - \frac{1}{2}([c_i(x^{p+1})]_+)^2 - \frac{\psi_{\beta_1}(c_i(x^1), \lambda_i^1)}{\beta_1} \\
&+ \sum_{t\in\mathcal{T}_1}(\frac{\lambda_i^t}{\beta_t} - \frac{\lambda_i^{t+1}}{\beta_{t+1}})c_i(x^{t+1}) + \sum_{t\in\mathcal{T}_2}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2} + \frac{\lambda_i^t}{\beta_t}c_i(x^{t+1}) + \frac{1}{2}\|c_i(x^{t+1})\|^2) + \sum_{t\in\mathcal{T}_3}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2} - \frac{(\lambda_i^t)^2}{2\beta_t^2}).
\end{aligned} \tag{69}$$

Without loss of generality, we assume that as p approaches infinity, $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ are infinite sets. It is easy to see that the sum of the first three terms in R.H.S. of (69) converges to zero as p approaches infinity. Therefore, to demonstrate the convergence of L.H.S. of (69), it suffices to prove that the terms pertaining to sets $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ converge as $p \to \infty$.

(b1) Our first claim is that $\sum_{t\in\mathcal{T}_1}(\frac{\lambda_i^t}{\beta_t} - \frac{\lambda_i^{t+1}}{\beta_{t+1}})c_i(x^{t+1})$ converges as p approaches infinity. This claim can be shown by following the same analysis as in the previous proof for (67).

(b2) For $t \in \mathcal{T}_2$, $\frac{\lambda_i^{t+1}}{\beta_{t+1}} < -c_i(x^{t+1}) \leq \frac{\lambda_i^t}{\beta_t}$. It indicates

$$\begin{aligned}
\frac{(\lambda_i^{t+1})^2}{\beta_{t+1}^2} - \frac{(\lambda_i^t)^2}{\beta_t^2} \leq \frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2} + \frac{\lambda_i^t}{\beta_t}c_i(x^{t+1}) + \frac{1}{2}\|c_i(x^{t+1})\|^2 \leq& \frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2} - \frac{\lambda_i^t\lambda_i^{t+1}}{\beta_t\beta_{t+1}} + \frac{(\lambda_i^t)^2}{2\beta_t^2} \\
=&\frac{1}{2}(\frac{\lambda_i^{t+1}}{\beta_{t+1}} - \frac{\lambda_i^t}{\beta_t})^2.
\end{aligned} \tag{70}$$

Then we multiply both sides of the above inequality by $\frac{\beta_{t+1}}{t^{\frac{\iota}{2}}}$ and subsequently sum up the resulting inequality over $\mathcal{T}_2$, obtaining

$$\sum_{t\in\mathcal{T}_2} \frac{(\lambda_i^{t+1})^2}{\beta_{t+1}t^{\frac{\iota}{2}}} - \frac{\beta_{t+1}(\lambda_i^t)^2}{\beta_t^2 t^{\frac{\iota}{2}}} \leq \sum_{t\in\mathcal{T}_2} \frac{\beta_{t+1}}{t^{\frac{\iota}{2}}}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2} + \frac{\lambda_i^t}{\beta_t}c_i(x^{t+1}) + \frac{1}{2}\|c_i(x^{t+1})\|^2) \leq \sum_{t\in\mathcal{T}_2} \frac{\beta_{t+1}}{2t^{\frac{\iota}{2}}}(\frac{\lambda_i^{t+1}}{\beta_{t+1}} - \frac{\lambda_i^t}{\beta_t})^2. \tag{71}$$

By Lemma 2 and the setting of ρ_t, there exists $\Lambda > 0$ such that $\|\lambda^t\| \leq \Lambda$ for any $t \geq 1$. And for any $t \in \mathcal{T}_2$, $\lambda_i^t \geq \lambda_i^{t+1} = \lambda_i^t + \rho_t c_i(x^{t+1}) \geq (1 - \frac{\rho_t}{\beta_t})\lambda_i^t$, thus $\frac{\lambda_i^{t+1}}{\beta_{t+1}} < \frac{\lambda_i^t}{\beta_t}$. Then due to $(1 + \frac{1}{t})^{2\iota} < 1 + \frac{1}{t}$ by $\iota < \frac{1}{2}$, it can be implied that

$$\begin{aligned}
0 > \sum_{t\in\mathcal{T}_2} \frac{(\lambda_i^{t+1})^2}{\beta_{t+1}t^{\frac{\iota}{2}}} - \frac{\beta_{t+1}(\lambda_i^t)^2}{\beta_t^2 t^{\frac{\iota}{2}}} \geq& \sum_{t\in\mathcal{T}_2}(\frac{(1 - \frac{\rho_t}{\beta_t})^2}{\beta_{t+1}t^{\frac{\iota}{2}}} - \frac{\beta_{t+1}}{\beta_t^2 t^{\frac{\iota}{2}}})\Lambda^2 \\
=&\frac{1}{\beta_0}\sum_{t\in\mathcal{T}_2} \frac{1}{t^{\frac{\iota}{2}}(t+1)^\iota}((1 - \frac{\rho_t}{\beta_t})^2 - \frac{(t+1)^{2\iota}}{t^{2\iota}})\Lambda^2 \\
\geq&\frac{1}{\beta_0}\sum_{t\in\mathcal{T}_2} \frac{1}{t^{\frac{\iota}{2}}(t+1)^\iota}(1 - \frac{2\rho_t}{\beta_t} - 1 - \frac{1}{t})\Lambda^2
\end{aligned}$$

$$\geq \frac{1}{\beta_0}\sum_{t\in\mathcal{T}_2}\frac{1}{t^{\frac{\iota}{2}}(t+1)^{\iota}}(-\frac{2\rho}{\beta_0 t^{\theta+\iota}}-\frac{1}{t})\Lambda^2$$

$$= \frac{1}{\beta_0}\sum_{t\in\mathcal{T}_2}\frac{1}{t^{1+\frac{\iota}{2}}(t+1)^{\iota}}(-\frac{2\rho}{\beta_0 t^{\theta+\iota-1}}-1)\Lambda^2, \tag{72}$$

which is lower bounded by a finite value, since $\theta+\iota>1$. On the other hand, it follows from $\beta_{t+1}\geq\beta_t\geq\beta_1$, $\|\lambda^t\|\leq\Lambda$ and $\lambda_i^{t+1}\geq(1-\frac{\rho_t}{\beta_t})\lambda_i^t$ that

$$\begin{aligned}
\sum_{t\in\mathcal{T}_2}\frac{\beta_{t+1}}{2t^{\frac{\iota}{2}}}(\frac{\lambda_i^{t+1}}{\beta_{t+1}}-\frac{\lambda_i^t}{\beta_t})^2 &\leq \sum_{t\in\mathcal{T}_2}\frac{\beta_{t+1}}{2t^{\frac{\iota}{2}}}(\frac{(1-\frac{\rho_t}{\beta_t})}{\beta_{t+1}}\lambda_i^t-\frac{\lambda_i^t}{\beta_t})^2\\
&\leq \sum_{t\in\mathcal{T}_2}\frac{1}{2\beta_{t+1}t^{\frac{\iota}{2}}}((1-\frac{\rho_t}{\beta_t})-\frac{\beta_{t+1}}{\beta_t})^2\Lambda^2\\
&= \sum_{t\in\mathcal{T}_2}\frac{1}{2\beta_{t+1}t^{\frac{\iota}{2}}}((1-\frac{\rho}{\beta_0 t^{\theta+\iota}})-(1+\frac{1}{t})^{\iota})^2\Lambda^2\\
&\leq \sum_{t\in\mathcal{T}_2}\frac{1}{2\beta_{t+1}t^{\frac{\iota}{2}}}(\frac{\rho}{\beta_0 t^{\theta+\iota}}+\frac{1}{t})^2\Lambda^2,
\end{aligned}$$

which is upper bounded by a finite value since $\theta>1$. Therefore, by (71), $\sum_{t\in\mathcal{T}_2}\frac{\beta_{t+1}}{t^{\frac{\iota}{2}}}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2}+\frac{\lambda_i^t}{\beta_t}c_i(x^{t+1})+\frac{1}{2}\|c_i(x^{t+1})\|^2)$ is bounded, then from $\frac{t^{\frac{\iota}{2}}}{\beta_{t+1}}\to 0$ as $t\to\infty$ we derive that $\sum_{t\in\mathcal{T}_2}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2}+\frac{\lambda_i^t}{\beta_t}c_i(x^{t+1})+\frac{1}{2}\|c_i(x^{t+1})\|^2)$ converges as $p\to\infty$.

(b3) For $t\in\mathcal{T}_3$, by (6) we have $\lambda_i^{t+1}=(1-\frac{\rho_t}{\beta_t})\lambda_i^t$. Then it indicates from the monotonically increasing property of $\{\beta_t\}$ that $\frac{\lambda_i^{t+1}}{\beta_{t+1}}\leq\frac{\lambda_i^t}{\beta_t}$, and from (72) that $\sum_{t\in\mathcal{T}_3}\frac{\beta_{t+1}}{t^{\frac{\iota}{2}}}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2}-\frac{(\lambda_i^t)^2}{2\beta_t^2})$ is bounded. Furthermore, since $\{\frac{t^{\frac{\iota}{2}}}{\beta_{t+1}}\}_{t\geq 1}$ is a monotonically decreasing sequence converging to zero, then $\sum_{t\in\mathcal{T}_3}(\frac{(\lambda_i^{t+1})^2}{2\beta_{t+1}^2}-\frac{(\lambda_i^t)^2}{2\beta_t^2})$ converges as $p\to\infty$. □

Lemma A.2 *Under the same conditions of Theorem 3, it holds that*

$$|\sum_{t=1}^{p}\frac{\Psi_\beta(x^{t+1},\lambda^t)-\Psi_\beta(x^t,\lambda^t)}{\beta}-\frac{1}{2}(\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2)|=C_{vio}^1+\mathcal{O}(\beta^{-1}),$$

where $C_{vio}^1=\frac{1}{2}\|c_{\mathcal{E}}(x^1)\|^2+\frac{1}{2}\|[c_{\mathcal{I}}(x^1)]_+\|^2$.

Proof. Similar to Lemma A.1, to prove the conclusion it suffices to provide estimate bounds on terms in (64) with $\beta_t\equiv\beta$ for all $t\geq 1$, that is

$$\begin{aligned}
&\sum_{t=1}^{p}\frac{\Psi_\beta(x^{t+1},\lambda^t)-\Psi_\beta(x^t,\lambda^t)}{\beta}-\frac{1}{2}(\|c_{\mathcal{E}}(x^{p+1})\|^2+\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2)\\
=&\sum_{t=1}^{p}\sum_{i\in\mathcal{E}}(\frac{\lambda_i^t}{\beta}c_i(x^{t+1})+\frac{1}{2}c_i^2(x^{t+1})-\frac{\lambda_i^t}{\beta}c_i(x^t)-\frac{1}{2}c_i^2(x^t))-\frac{1}{2}\|c_{\mathcal{E}}(x^{p+1})\|^2\\
&+\sum_{t=1}^{p}\sum_{i\in\mathcal{I}}\frac{\psi_\beta(c_i(x^{t+1}),\lambda_i^t)-\psi_\beta(c_i(x^t),\lambda_i^t)}{\beta}-\frac{1}{2}\|[c_{\mathcal{I}}(x^{p+1})]_+\|^2.
\end{aligned}$$

Since the analysis is essentially the same as Lemma A.1 except replacing β_t with constant β, we simply state the main ideas here.

(a) Following (65)-(66), we obtain

$$\sum_{t=1}^{p}\sum_{i\in\mathcal{E}}(\frac{\lambda_i^t}{\beta}c_i(x^{t+1})+\frac{1}{2}c_i^2(x^{t+1})-\frac{\lambda_i^t}{\beta}c_i(x^t)-\frac{1}{2}c_i^2(x^t))-\frac{1}{2}\|c_{\mathcal{E}}(x^{p+1})\|^2=-\frac{1}{2}\|c_{\mathcal{E}}(x^1)\|^2+\mathcal{O}(\beta^{-1}).$$

(b) With a slight abuse of notation, corresponding to inequality constraints we still use $\mathcal{T}_i$, $i=1,\ldots,4$ as defined in (68), except $\beta_t\equiv\beta$ in this scenario. It is easy to check from the definition of $\mathcal{T}_1$ together with (8) and the setting of ρ_t that for any $i\in\mathcal{I}$,

$$\begin{aligned}|\sum_{t\in\mathcal{T}_1}(\frac{\lambda_i^t}{\beta}-\frac{\lambda_i^{t+1}}{\beta})c_i(x^{t+1})|\leq&\beta^{-1}\sum_{t\in\mathcal{T}_1}|\lambda_i^t-\lambda_i^{t+1}||c_i(x^{t+1})|\\ \leq&\beta^{-1}\sum_{t\in\mathcal{T}_1}|\lambda_i^t-\lambda_i^{t+1}|\cdot\max\{\frac{\lambda_i^t}{\beta},\frac{\lambda_i^{t+1}}{\beta},C\}=\mathcal{O}(\beta^{-1}).\end{aligned}$$

From (70), Lemma 2 and the setting of ρ_t, it follows that for any $i\in\mathcal{I}$,

$$|\sum_{t\in\mathcal{T}_2}(\frac{(\lambda_i^{t+1})^2}{2\beta^2}+\frac{\lambda_i^t}{\beta}c_i(x^{t+1})+\frac{1}{2}\|c_i(x^{t+1})\|^2)|=\mathcal{O}(\beta^{-2}),$$

$$|\sum_{t\in\mathcal{T}_3}(\frac{(\lambda_i^{t+1})^2}{2\beta^2}-\frac{(\lambda_i^t)^2}{2\beta^2})|=\mathcal{O}(\beta^{-2}).$$

Hence, we obtain from (69) that

$$\sum_{t=1}^{p}\sum_{i\in\mathcal{I}}\frac{\psi_\beta(c_i(x^{t+1}),\lambda_i^t)-\psi_{\beta_t}(c_i(x^t),\lambda_i^t)}{\beta_t}-\frac{1}{2}([c_i(x^{p+1})]_+)^2=-\frac{1}{2}\|[c_{\mathcal{I}}(x^1)]_+\|^2+\mathcal{O}(\beta^{-1}).$$

Consequently, the conclusion of this lemma can be derived. □

B. Proof of Lemma 11

Let us prove the conclusion by contradiction. For notation simplicity, we define $a_k:=\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|\widehat{\lambda}^t\|]$, $k\geq 1$ with

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

We assume that $\limsup_{k\to\infty}a_k=+\infty$, that is, for any $M>0$, there exists an infinite subsequence $\{a_k\}_{k\in\mathcal{K}}$ such that $a_k\geq M$, for all $k\in\mathcal{K}$. In view of the first-order optimality condition of the subproblem (5), there exists $u^{t+1}\in\partial h(x^{t+1})$ such that

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

Subsequently, dividing both sides of the above inequality by a_k and recalling $d^t=\nabla f(x^t)+\nabla c(x^t)\widehat{\lambda}^t+\varepsilon^t$ we have for any $x\in X$, $t>0$, $k\in\mathcal{K}$,

$$\frac{1}{a_k}\langle\nabla f(x^t)+\nabla c(x^t)\widehat{\lambda}^t+\varepsilon^t+u^{t+1}+\frac{1}{\eta_t}(x^{t+1}-x^t),x-x^{t+1}\rangle\geq 0.$$

Equivalently, there exists $v^{t+1}\in\mathcal{N}_X(x^{t+1})$ such that

$$\frac{\nabla f(x^t)}{a_k}+\frac{\nabla c(x^t)\widehat{\lambda}^t}{a_k}+\frac{\varepsilon^t}{a_k}+\frac{u^{k+1}}{a_k}+\frac{x^{t+1}-x^t}{\eta_t a_k}+v^{t+1}=0,\quad\forall t>0;\forall k\in\mathcal{K}.\tag{73}$$

Following Assumption 5, for any $t \geq 1$, there exists $z^{t+1} \in -\mathcal{N}_X^*(x^{t+1})$ with $\|z^{t+1}\| \leq Z$ such that

$$\begin{aligned} \nabla c_i(x^{t+1})^{\mathrm{T}} z^{t+1} &= -\delta \cdot \mathrm{sgn}(c_i(x^{t+1})), \quad && i \in \mathcal{E} : c_i(x^{t+1}) \neq 0; \\ \nabla c_i(x^{t+1})^{\mathrm{T}} z^{t+1} &\leq -\delta, && i \in \mathcal{I} : c_i(x^{t+1}) > 0. \end{aligned} \tag{74}$$

Then taking inner product with z^{t+1} on both sides, summing (73) over $[k]$, dividing the result by k, and concurrently taking the expectation yield

$$\begin{aligned} 0 =& \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}(\frac{\nabla f(x^t)}{a_k} + \frac{\nabla c(x^t)\widehat{\lambda}^t}{a_k} + \frac{\varepsilon^t}{a_k} + \frac{u^{t+1}}{a_k} + \frac{x^{t+1}-x^t}{\eta_t a_k} + v^{t+1})] \\ =& \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}(\frac{\nabla f(x^t)}{a_k} + \frac{\varepsilon^t}{a_k} + \frac{u^{t+1}}{a_k} + \frac{(\nabla c(x^t) - \nabla c(x^{t+1}))\widehat{\lambda}^t}{a_k} + \frac{x^{t+1}-x^t}{\eta_t a_k})] \\ &+ \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}(\frac{\nabla c(x^{t+1})\widehat{\lambda}^t}{a_k} + v^{t+1})]. \end{aligned} \tag{75}$$

It is easy to prove that for all $k \in \mathcal{K}$, one has

$$|\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\frac{(z^{t+1})^{\mathrm{T}}\nabla f(x^t)}{a_k}]| \leq \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|z^{t+1}\|\|\nabla f(x^t)\|] \leq \frac{ZG}{M},$$

and

$$|\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\frac{(z^{t+1})^{\mathrm{T}}u^{t+1}}{a_k}]| \leq \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|z^{t+1}\|\|u^{t+1}\|] \leq \frac{ZG}{M}.$$

Besides, we can also verify that

$$\begin{aligned} |\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\frac{(z^{t+1})^{\mathrm{T}}\varepsilon^t}{a_k}]| &\leq \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|z^{t+1}\|\|\varepsilon^t\|] \\ &\leq \frac{Z}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\sqrt{\frac{1}{k}\sum_{t=1}^{k}\|\varepsilon^t\|^2}] \\ &\leq \frac{Z}{a_k} \cdot \sqrt{\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|\varepsilon^t\|^2]}. \end{aligned}$$

With the setting of α_t, β_t and ρ_t in (33), we have

$$\sum_{t=1}^{k}\alpha_t^2 = \sum_{t=1}^{k}\frac{64\alpha^2\eta^4}{t^{2\iota}} \leq \frac{64\alpha^2\eta^4}{1-2\iota}(k^{1-2\iota}-1) + 64\alpha^2\eta^4, \quad \sum_{t=1}^{k}\rho_t \leq \sum_{t=1}^{k}\rho_0 t^{-\theta}.$$

Since $\iota < \frac{1}{2}$, $\sum_{t=1}^{k}(\beta_{t+1}-\beta_t)^2$, $\sum_{t=1}^{k}\rho_t^2$, $\sum_{t=1}^{k}\rho_t$ are upper bounded by a positive constant, and $\sum_{t=1}^{k}\alpha_t^2$ is upper bounded by $\mathcal{O}(k^{1-2\iota})$. Therefore, from (31), there exists a positive constant C_{eps} such that

$$\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|\varepsilon^t\|^2] = \frac{1}{k}\sum_{t=1}^{k}\mathbb{E}_{\xi^{[t]}}[\|\varepsilon^t\|^2] \leq \frac{C_{eps}^2}{3}k^{\iota-1}(k^{1-2\iota}+k^{\iota}+1) \leq C_{eps}^2. \tag{76}$$

Hence, we obtain

$$|\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\frac{(z^{t+1})^{\mathrm{T}}\varepsilon^t}{a_k}]| \leq \frac{ZC_{eps}}{M}.$$

Furthermore, the following relations hold:

$$\begin{aligned}
&|\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}(\frac{(\nabla c(x^t)-\nabla c(x^{t+1}))\widehat{\lambda}^t}{a_k}+\frac{x^{t+1}-x^t}{\eta_t a_k})]| \\
\leq&\frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\|z^{t+1}\|\|\nabla c(x^t)-\nabla c(x^{t+1})\|_{\mathrm{F}}\|\widehat{\lambda}^t\|+\frac{\|z^{t+1}\|\|x^{t+1}-x^t\|}{\eta_t})] \\
\leq&\frac{Z}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(C_\lambda L_{\beta_t}\|\nabla c(x^t)-\nabla c(x^{t+1})\|_{\mathrm{F}}+\frac{\|x^{t+1}-x^t\|}{\eta_t})] \\
\leq&\frac{Z}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\sqrt{m}LC_\lambda L_{\beta_t}\|x^{t+1}-x^t\|+\frac{\|x^{t+1}-x^t\|}{\eta_t})] \\
\leq&\frac{Z}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{\sqrt{m}LC_\lambda+1}{k}\sum_{t=1}^{k}(L_{\beta_t}+\frac{1}{\eta_t})\|x^{t+1}-x^t\|] \\
\leq&\frac{Z}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[(\sqrt{m}LC_\lambda+1)\sqrt{\frac{1}{k}\sum_{t=1}^{k}(L_{\beta_t}+\frac{1}{\eta_t})^2\|x^{t+1}-x^t\|^2}] \\
\leq&\frac{Z}{a_k}\cdot\sqrt{2}(\sqrt{m}LC_\lambda+1)\sqrt{\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(L_{\beta_t}^2+\frac{1}{\eta_t^2})\|x^{t+1}-x^t\|^2]} \\
\leq&\frac{Z}{a_k}\cdot\sqrt{2}(\sqrt{m}LC_\lambda+1)C_{sum}\leq\frac{\sqrt{2}Z(\sqrt{m}LC_\lambda+1)C_{sum}}{M},
\end{aligned}$$

where the second inequality is due to the fact that there exists C_λ independent of β_t such that $\|\hat{\lambda}^t\|\leq C_\lambda L_{\beta_t}$ and C_{sum} denotes the upper bound derived by (32) and (76), i.e., there is a $C_{sum}>0$ such that

$$\begin{aligned}
&\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(L_{\beta_t}^2+\frac{1}{\eta_t^2})\|x^{t+1}-x^t\|^2] \\
\leq&\frac{5}{\eta_k k}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+\frac{m\beta_{k+1}C^2}{2}+2m\tilde{C}^2\sum_{t=1}^{k}\rho_t)+\mathbb{E}_{\xi^{[k]}}[\frac{5}{2k}\sum_{t=1}^{k}\|\varepsilon^t\|^2] \\
\leq&\frac{5}{\eta_k k}(\mathcal{L}_{\beta_1}(x^1,\lambda^1)-C^*+\frac{m\beta_{k+1}C^2}{2}+2m\tilde{C}^2\sum_{t=1}^{k}\rho_t)+\frac{5}{2}C_{eps}^2\leq C_{sum}^2,
\end{aligned}$$

owing to the setting of η_t, β_t, ρ_t in (33) and $\iota<\frac{1}{2}$. Therefore, it follows from (75) that

$$\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}(\frac{\nabla c(x^{t+1})\hat{\lambda}^t}{a_k}+v^{t+1})]\geq-\frac{Z(2G+C_{eps}+\sqrt{2}(\sqrt{m}LC_\lambda+1)C_{sum})}{M}=:\frac{C_{lower}}{M}.$$

We next introduce notations:

$$\mathcal{E}_1 := \{i \in \mathcal{E} : c_i(x^{t+1}) \neq 0\}, \quad \mathcal{E}_2 := \{i \in \mathcal{E} : c_i(x^{t+1}) = 0\}, \quad \mathcal{I}_1 := \{i \in \mathcal{I} : c_i(x^{t+1}) > 0\},$$
$$\mathcal{I}_2 := \{i \in \mathcal{I} : -\frac{\lambda_i^t}{\beta_t} \leq c_i(x^{t+1}) \leq 0\}, \quad \mathcal{I}_3 := \{i \in \mathcal{I} : c_i(x^{t+1}) < -\frac{\lambda_i^t}{\beta_t}\}. \tag{77}$$

For the sake of simplicity, and without causing any potential confusion, we omit the index t in the notations as defined in (77). For any $i \in \mathcal{E}_1$, it follows from $|\lambda_i^t| \geq |\widehat{\lambda}_i^t| - \beta_t|c_i(x^t)|$ and $\beta_t|c_i(x^t)| - \beta_t|c_i(x^{t+1})| \leq \beta_t|c_i(x^t) - c_i(x^{t+1})|$ that

$$\begin{aligned}
(z^{t+1})^{\mathrm{T}}\nabla c_i(x^{t+1})\hat{\lambda}_i^t &= -\delta \cdot \mathrm{sgn}(c_i(x^{t+1}))\widehat{\lambda}_i^t \\
&= -\delta \cdot \mathrm{sgn}(c_i(x^{t+1}))(\lambda_i^t + \beta_t c_i(x^{t+1}) - \beta_t c_i(x^{t+1}) + \beta_t c(x^t)) \\
&\leq -\delta|\lambda_i^t| + 2\delta|\lambda_i^t| - \delta\beta_t|c_i(x^{t+1})| + \delta\beta_t|c_i(x^{t+1}) - c_i(x^t)| \\
&\leq -\delta|\widehat{\lambda}_i^t| + \delta\beta_t|c_i(x^t)| + 2\delta|\lambda_i^t| - \delta\beta_t|c_i(x^{t+1})| + \delta\beta_t|c_i(x^{t+1}) - c_i(x^t)| \\
&\leq -\delta|\widehat{\lambda}_i^t| + 2\delta|\lambda_i^t| + 2\delta\beta_t|c_i(x^{t+1}) - c_i(x^t)| \\
&\leq -\delta|\widehat{\lambda}_i^t| + 2\delta|\lambda_i^t| + 2\delta G\beta_t\|x^{t+1} - x^t\|.
\end{aligned}$$

In addition, for any $i \in \mathcal{I}_1$, it follows from (74) and $\hat{\lambda}_i^t \geq 0$ that $(z^{t+1})^T\nabla c_i(x^{t+1})\hat{\lambda}_i^t \leq -\delta\hat{\lambda}_i^t$. We then obtain from (74) together with $\tilde{L} \geq mCL + mG^2$ and $(z^{t+1})^{\mathrm{T}}v^{t+1} \leq 0$ that

$$\begin{aligned}
&\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}(\frac{\nabla c(x^{t+1})\widehat{\lambda}^t}{a_k} + v^{t+1})] \leq \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(z^{t+1})^{\mathrm{T}}\nabla c(x^{t+1})\widehat{\lambda}^t] \\
\leq &\frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(-\delta\|\widehat{\lambda}_{\mathcal{E}_1}^t\|_1 + 2\delta\|\lambda_{\mathcal{E}_1}^t\|_1 + 2\delta|\mathcal{E}_1|G\beta_t\|x^{t+1} - x^t\|)] \\
&+ \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(-\delta\|\widehat{\lambda}_{\mathcal{I}_1}^t\|_1 + (z^{t+1})^{\mathrm{T}}\nabla c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1})\widehat{\lambda}_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}^t)] \\
\leq &\frac{1}{a_k} \cdot (-\delta\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|\widehat{\lambda}^t\|_1] + \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(2\delta\|\lambda_{\mathcal{E}_1}^t\|_1 + \delta\|\widehat{\lambda}_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}^t\|_1 + (z^{t+1})^{\mathrm{T}}\nabla c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1})\widehat{\lambda}_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}^t)]) \\
&+ \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}2\delta|\mathcal{E}_1|G\beta_t\|x^{t+1} - x^t\|] \\
\leq &-\delta + \frac{2\delta|\mathcal{E}|\Lambda}{a_k} + \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\delta + ZG)\|\widehat{\lambda}_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}^t\|_1 + 2\delta|\mathcal{E}_1|G\beta_t\|x^{t+1} - x^t\|] \\
\leq &-\delta + \frac{2\delta|\mathcal{E}|\Lambda}{a_k} + \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\delta + ZG)(\|\lambda_{\mathcal{E}_2}^t + \beta_t c_{\mathcal{E}_2}(x^t)\|_1 + \|[\lambda_{\mathcal{I}_2\cup\mathcal{I}_3}^t + \beta_t c_{\mathcal{I}_2\cup\mathcal{I}_3}(x^t)]_+\|_1)] \\
&+ \frac{1}{a_k} \cdot \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}2\delta G^{-1}\tilde{L}\beta_t\|x^{t+1} - x^t\|]
\end{aligned}$$

$$\leq -\delta + \frac{2\delta|\mathcal{E}|\Lambda}{a_k} + \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\delta+ZG)\|\lambda_{\mathcal{E}_2}^t + \beta_t c_{\mathcal{E}_2}(x^t) - \beta_t c_{\mathcal{E}_2}(x^{t+1})\|_1]$$

$$+ \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\delta+ZG)\|[\lambda_{\mathcal{I}_2\cup\mathcal{I}_3}^t + \beta_t c_{\mathcal{I}_2\cup\mathcal{I}_3}(x^t) - \beta_t c_{\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1})]_+\|_1]$$

$$+ \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}2\delta G^{-1}\tilde{L}\beta_t\|x^{t+1}-x^t\|]$$

$$\leq -\delta + \frac{2\delta|\mathcal{E}|\Lambda}{a_k} + \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\delta+ZG)\|\lambda_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}^t + \beta_t c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^t) - \beta_t c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1})\|_1]$$

$$+ \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}2\delta G^{-1}\tilde{L}\beta_t\|x^{t+1}-x^t\|]$$

$$\leq -\delta + \frac{2\delta|\mathcal{E}|\Lambda}{a_k} + \frac{\delta+ZG}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}(\|\lambda_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}^t\|_1 + \beta_t\|c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^t) - c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1})\|_1)$$

$$+ \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}2\delta G^{-1}\tilde{L}\beta_t\|x^{t+1}-x^t\|]$$

$$\leq -\delta + \frac{m(3\delta+ZG)\Lambda}{a_k} + \frac{\delta+ZG}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\beta_t\|c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^t) - c_{\mathcal{E}_2\cup\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1})\|_1]$$

$$+ \frac{1}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}2\delta G^{-1}\tilde{L}\beta_t\|x^{t+1}-x^t\|]$$

$$\leq -\delta + \frac{m(3\delta+ZG)\Lambda}{a_k} + \frac{(3\delta+ZG)G^{-1}}{a_k}\cdot\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\tilde{L}\beta_t\|x^{t+1}-x^t\|]$$

$$\leq -\delta + \frac{m(3\delta+ZG)\Lambda}{a_k} + \frac{(3\delta+ZG)G^{-1}}{a_k}\cdot(\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}L_{\beta_t}^2\|x^{t+1}-x^t\|^2])^{\frac{1}{2}}$$

$$\leq -\delta + \frac{(3\delta+ZG)(m\Lambda+G^{-1}C_{sum})}{a_k} \leq -\delta + \frac{(3\delta+ZG)(m\Lambda+G^{-1}C_{sum})}{M} =: -\delta + \frac{C_{upper}}{M},$$

where $\Lambda > 0$ is a constant such that $\|\lambda^t\| \leq \Lambda$ for any $t \geq 1$ by Lemma 2 and the setting of ρ_t, and the 5-th inequality is due to $c_{\mathcal{E}_2}(x^{t+1}) = 0$ and $c_{\mathcal{I}_2\cup\mathcal{I}_3}(x^{t+1}) \leq 0$. Summarizing above analysis we obtain $M \leq \delta^{-1}(C_{upper} - C_{lower})$. However, this contradicts the arbitrariness of M. Thus we derive that $\limsup_{k\to\infty} a_k < +\infty$.

We now bound $\mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|\tilde{\lambda}^t\|]$. According to the definitions of $\tilde{\lambda}^t$ and $\widehat{\lambda}^t$, we have

$$\begin{aligned}\|\tilde{\lambda}^t\| &\leq \|\lambda_{\mathcal{E}}^t + \beta_{t-1}c_{\mathcal{E}}(x^t)\| + \|[\lambda_{\mathcal{I}}^t + \beta_{t-1}c_{\mathcal{I}}(x^t)]_+\| \\ &= \|\lambda_{\mathcal{E}}^t + \beta_{t-1}c_{\mathcal{E}}(x^t) - \beta_t c_{\mathcal{E}}(x^t) + \beta_t c_{\mathcal{E}}(x^t)\| + \|[\lambda_{\mathcal{I}}^t + \beta_{t-1}c_{\mathcal{I}}(x^t) - \beta_t c_{\mathcal{I}}(x^t) + \beta_t c_{\mathcal{I}}(x^t)]_+\| \\ &\leq \|\widehat{\lambda}_{\mathcal{E}}^t\| + \|\beta_{t-1}c_{\mathcal{E}}(x^t) - \beta_t c_{\mathcal{E}}(x^t)\| + \|\widehat{\lambda}_{\mathcal{I}}^t\| + \|[\beta_{t-1}c_{\mathcal{I}}(x^t) - \beta_t c_{\mathcal{I}}(x^t)]_+\|\end{aligned}$$

$$\leq \sqrt{2}\|\widehat{\lambda}^t\| + mC(\beta_t - \beta_{t-1}).$$

Therefore, it yields from the setting of β_t that

$$\limsup_{k\to\infty} \mathbb{E}_{\xi^{[k]}}[\frac{1}{k}\sum_{t=1}^{k}\|\tilde{\lambda}^t\|] \leq \sqrt{2}\limsup_{k\to\infty} a_k + mC\limsup_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k}(\beta_t - \beta_{t-1}) < +\infty.$$

The proof is completed.

C. Verification of Assumption 5.

We apply MLALM to solve the following problem with both equality and inequality constraints:

$$\begin{aligned}
\min_{x\in X} \quad & f(x) = \frac{1}{N}\sum_{i=1}^{N}\log(1+\frac{1}{2}\|H_i x - c_i\|^2) \\
\text{s.t.} \quad & f_j(x) = \frac{1}{2}x^T Q_j x + a_j^T x \leq b_j, \quad j = 1,\ldots,M_1, \\
& f_k(x) = \frac{1}{2}x^T Q_k x + a_k^T x = b_k, \quad k = 1,\ldots,M_2,
\end{aligned}$$

where $M_1, M_2 \geq 1$ and $X = [-10,10]^n$ is a convex set. Each matrix $H_i \in \mathbb{R}^{p\times n}$, $i \in [N]$, is generated randomly, with elements independently selected following a standard Gaussian distribution. Elements of $Q_j, Q_k \in \mathbb{R}^{n\times n}$ are also randomly and independently selected from $\mathrm{U}[-0.5, 0.5]$. Here, $a_j, j \in [M_1]$ and $a_k, k \in [M_2]$ are randomly generated following $\mathrm{U}[0.1, 1.1]^n$. Then we generate a random point $x_* \sim \mathrm{U}(0,1)^n$ and set $c_i = H_i x_*$, $i \in [N]$, $b_j = \frac{1}{2}x_*^T Q_j x_* + a_j^T x_*$, $j \in [M_1]$, $b_k = \frac{1}{2}x_*^T Q_k x_* + a_k^T x_*$, $k \in [M_2]$.

To observe how often Assumption 5 holds during the iteration process, we solve the auxiliary problem

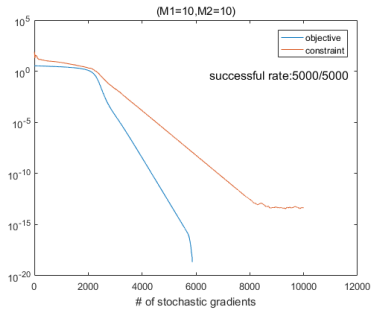
$$\begin{aligned}
\min_{z} \quad & 0 \\
\text{s.t.} \quad & \delta\cdot\mathrm{sgn}(c_i(x^t)) + \nabla c_i(x^t)^T z = 0, \qquad i \in \mathcal{E} : c_i(x^t) \neq 0\,; \\
& \delta + \nabla c_i(x^t)^{\mathrm{T}} z \leq 0, \qquad\qquad\quad i \in \mathcal{I} : c_i(x^t) > 0
\end{aligned} \tag{78}$$

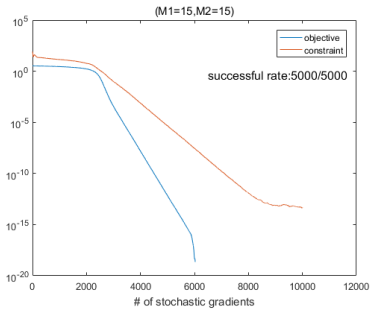
by calling *linprog* function in *matlab*. Here we set $\delta = 1$. The *linprog* function will return an *exitflag*. When *exitflag* $= 1$ the solution of (78) is reliable. Then Assumption 5 holds, if the solution $z \in -\mathcal{N}_X^*(x^t)$ with

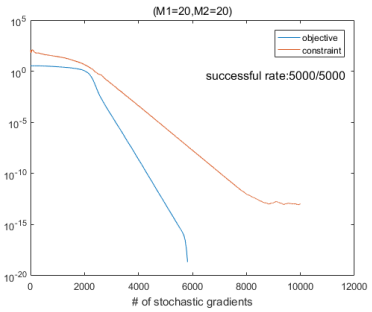
$$-\mathcal{N}_X^*(x^t) = \left\{ z \in \mathbb{R}^n \mid \begin{array}{ll} z_i \geq 0, & \text{if } x_i^t = -10 \\ z_i \leq 0, & \text{if } x_i^t = 10 \\ z_i \in \mathbb{R}, & \text{o.w.} \end{array} \right\}.$$

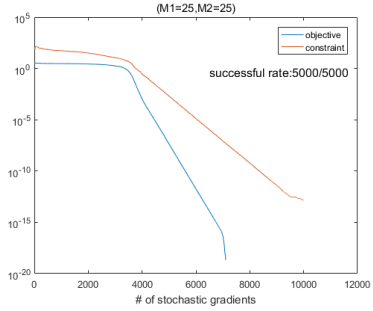
We consider the case with $p = 5, n = 50$ and $N = 1000$. Initial point is chosen as $x^1 = \mathbf{0}$, and the maximum number of iterations is set to $T = 5000$. With varying numbers of constraints, we report the successful rate during all iterations when Assumption 5 holds at iterates.

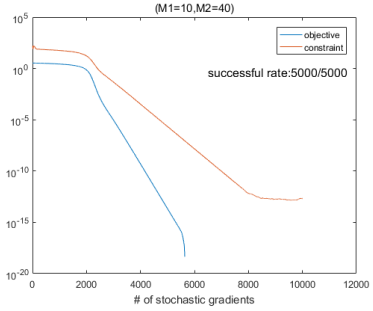
The following figures report the performances of MLALM in terms of objective value $f(x^t)$ and constraint violation $\sum_{j\in[M_1]}[f_j(x^t) - b_j, 0]_+ \sum_{k\in[M_2]}|f_k(x^t) - b_k|$, the successful rate of Assumption 5 and the KKT residual at iteration t_0 when Assumption 5 holds for all $t \geq t_0$. We can see from Figures 6 and 7 that when the number of constraints is relatively small ($M_1 + M_2 \leq 60$), Assumption 5 always holds, while it holds with a high percentage when the number of constraints is slightly more ($70 \leq M_1 + M_2 \leq 90$).

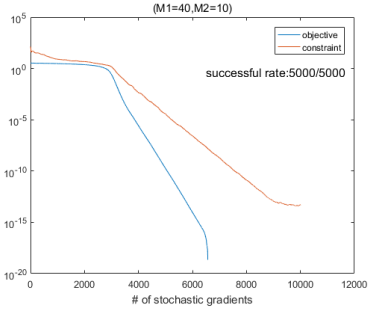


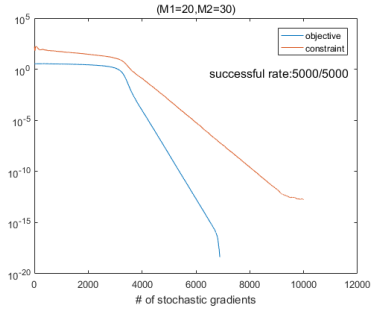


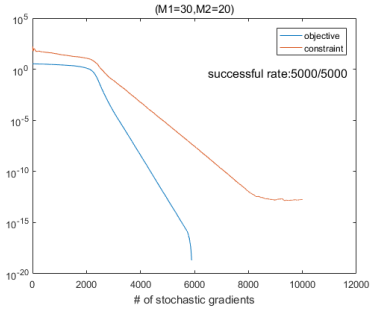












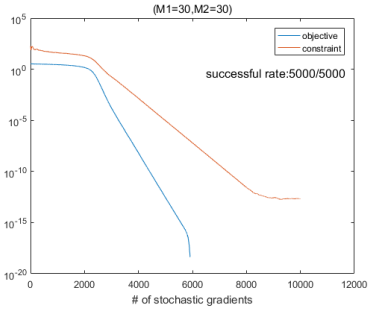


Figure 6: $M_1 + M_2 \leq 60$

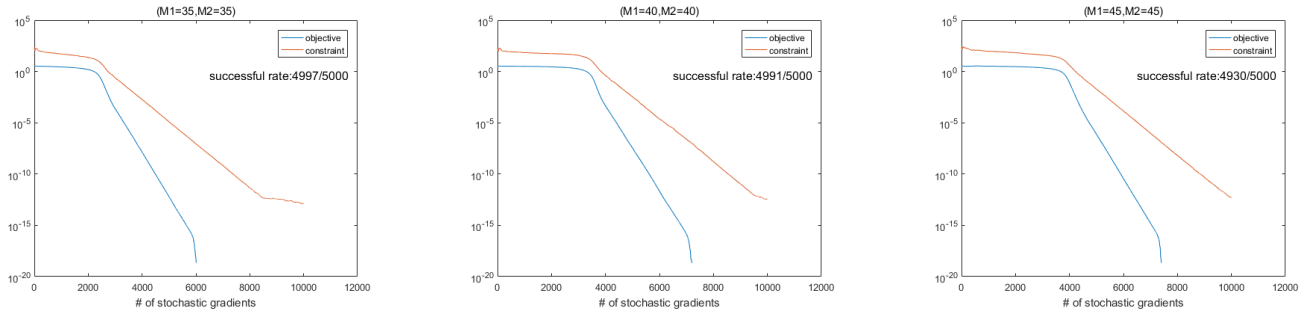
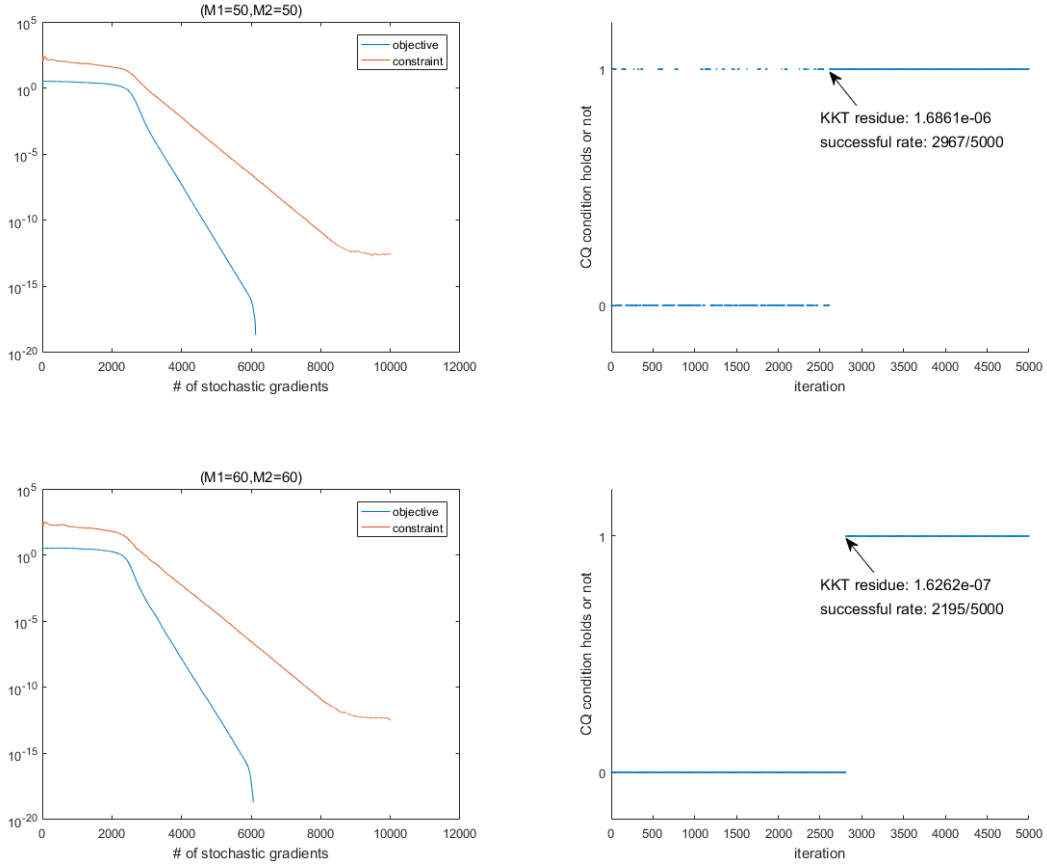
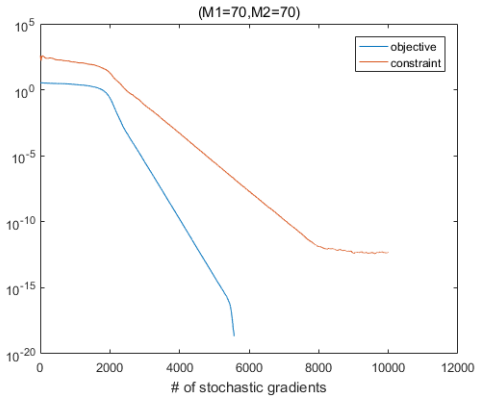
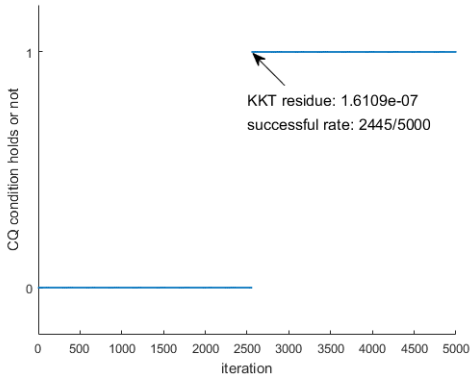


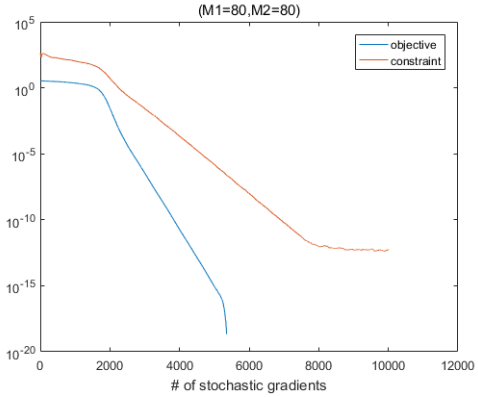
Figure 7: $70 \leq M_1 + M_2 \leq 90$

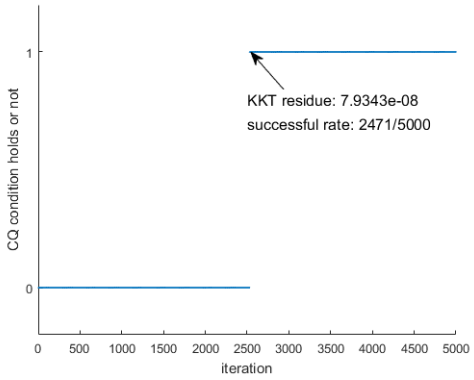
As the number of constraints increases, Assumption 5 gradually transitions from being satisfied by all iterates to being satisfied only at iterates in later stage of the algorithmic process. This occurs because at the earlier stage of the algorithm, since the number of infeasible constraints exceeds the dimension of the variable x, the system of equations is normally overdetermined, causing Assumption 5 hard to hold. However, as the iteration proceeds, less constraints are violated, leading to Assumption 5 more likely to hold.

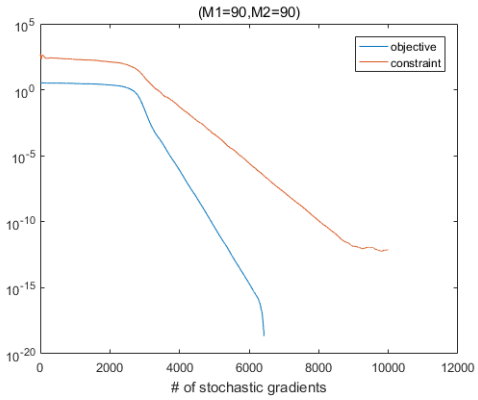


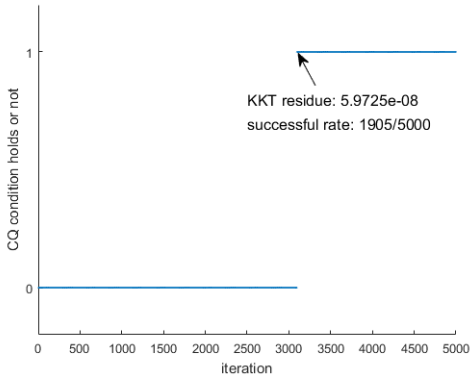

(M1=70,M2=70)
objective
constraint
# of stochastic gradients

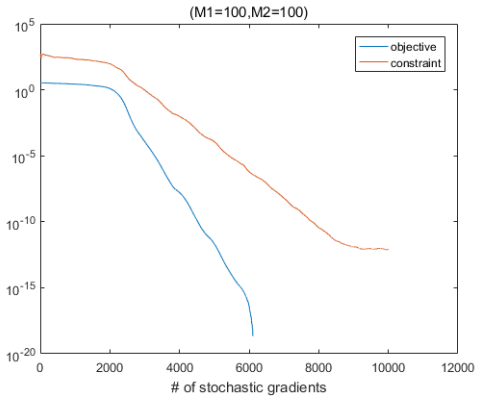

KKT residue: 1.6109e-07
successful rate: 2445/5000
CQ condition holds or not
iteration

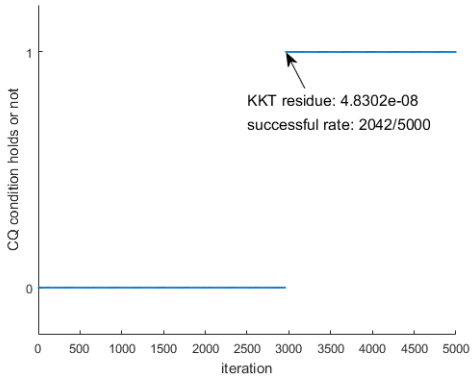

(M1=80,M2=80)
objective
constraint
# of stochastic gradients


KKT residue: 7.9343e-08
successful rate: 2471/5000
CQ condition holds or not
iteration


(M1=90,M2=90)
objective
constraint
# of stochastic gradients


KKT residue: 5.9725e-08
successful rate: 1905/5000
CQ condition holds or not
iteration


(M1=100,M2=100)
objective
constraint
# of stochastic gradients


KKT residue: 4.8302e-08
successful rate: 2042/5000
CQ condition holds or not
iteration

D. More discussions and extensions.

In this part we will discuss more about CQ conditions and explore potential extensions. For the sake of simplicity, we consider (1) with $X = \mathbb{R}^n$ and $h \equiv 0$, i.e.

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \{f(x) \equiv \mathbb{E}_\xi[\mathbf{F}(x;\xi)]\} \\ \text{s.t.} \quad & c_i(x) = 0, \quad i \in \mathcal{E}, \\ & c_i(x) \le 0, \quad i \in \mathcal{I}. \end{aligned} \tag{79}$$

Here, $n \ge |\mathcal{E}|+|\mathcal{I}|$. We assume that Assumptions 1-4 hold. Apply MLALM to solve (79) and let $\{x^t\}$ be the generated sequence of iterates.

The following proposition provides some insight on the relationship between LICQ and Assumption 5.

Proposition D.1 *Let LICQ hold at a feasible point $\tilde{x}$ of (79) and $\delta > 0$. Then there exist $Z > 0$ and a neighborhood of $\tilde{x}$, denoted by $\mathcal{N}(\tilde{x})$, such that for any $x \in \mathcal{N}(\tilde{x})$ the linear system*

$$\begin{aligned} \delta \cdot \operatorname{sgn}(c_i(x)) + \nabla c_i(x)^T z = 0, \qquad & i \in \mathcal{E} : c_i(x) \neq 0\,; \\ \delta + \nabla c_i(x)^T z \le 0, \qquad & i \in \mathcal{I} : c_i(x) > 0 \end{aligned} \tag{80}$$

has a solution located in the ball $\{z : \|z\| \le Z\}$.

Proof. Recall that LICQ holds at $\tilde{x}$, if the gradients

$$\nabla c_i(\tilde{x}) \in \mathbb{R}^n, \quad i \in \tilde{\mathcal{A}} := \mathcal{E} \cup \{i \mid c_i(\tilde{x}) = 0, i \in \mathcal{I}\}$$

are linearly independent. Introduce $\nabla c_{\tilde{\mathcal{A}}}(x) \in \mathbb{R}^{n\times|\tilde{\mathcal{A}}|}$ with columns consisting of $\nabla c_i(x), i \in \tilde{\mathcal{A}}$, i.e.

$$\nabla c_{\tilde{\mathcal{A}}}(x) = (\nabla c_i(x), i \in \mathcal{E}, \nabla c_i(x), i \in \tilde{\mathcal{A}} \cap \mathcal{I}).$$

By the linear independence and continuity, there exist positive constants κ and a neighborhood $\mathcal{N}(\tilde{x})$ of $\tilde{x}$ such that for all $x \in \mathcal{N}(\tilde{x})$, $\nabla c_i(x) \in \mathbb{R}^n, i \in \tilde{\mathcal{A}}$ are linearly independent,

$$c_i(x) < 0, \ \forall i \notin \tilde{\mathcal{A}}, \text{ and singular values of } \nabla c_{\tilde{\mathcal{A}}}(x) \text{ are lower bounded by } \kappa.$$

Obviously, there exists an index set $\mathcal{B}(x) \subseteq [n]$ with $|\mathcal{B}(x)| = |\tilde{\mathcal{A}}|$ such that

$$\{\nabla c_i(x), i \in \tilde{\mathcal{A}}\} \subseteq \operatorname{span}\{q_i, i \in \mathcal{B}(x)\},$$

where $q_i, i \in [n]$ denote n orthonormal vectors spanning $\mathbb{R}^n$. Then we can construct $Q_0(x) \in \mathbb{R}^{n\times(n-|\tilde{\mathcal{A}}|)}$ with columns being $q_i, i \in [n]\backslash\mathcal{B}(x)$. It is straightforward to attain

$$Q_0(x)^T Q_0(x) = I_{n-|\tilde{A}|} \text{ and } Q_0(x)^T \nabla c_{\tilde{\mathcal{A}}}(x) = 0.$$

Here, $I_{n-|\tilde{\mathcal{A}}|}$ is the identity matrix with dimension $(n - |\tilde{\mathcal{A}}|)$. Hence, the matrix $Q(x) := [\nabla c_{\tilde{\mathcal{A}}}(\tilde{x}), Q_0]$ is non-singular and

$$Q(x)^T Q(x) = \begin{bmatrix} \nabla c_{\tilde{\mathcal{A}}}(x)^T \nabla c_{\tilde{\mathcal{A}}}(x) & 0 \\ 0 & I_{n-|\tilde{\mathcal{A}}|} \end{bmatrix}.$$

And the vector of singular values of $Q(x)$, denoted by $s(Q(x))$, is

$$s(Q(x)) = (\text{singular values of } \nabla c_{\tilde{\mathcal{A}}}(x), 1, \ldots, 1)^T,$$

and thus the SVD of $Q(x)$ can be denoted as $Q(x) = U\Sigma V^T$ with $U, V \in \mathbb{R}^{n\times n}$ being orthogonal matrices and $\Sigma = \operatorname{Diag}(s(Q(x))) \in \mathbb{R}^{n\times n}$. Therefore, it is easy to obtain

$$\|(Q(x)^T)^{-1}\| \le \max(\kappa^{-1}, 1), \quad \forall x \in \mathcal{N}(\tilde{x}).$$

Moreover, for any given $x \in \mathcal{N}(\tilde{x})$, there exists a nonzero vector $z(\omega) \in \mathbb{R}^n$ such that the matrix

$$Q(x)^T z(\omega) = -\delta \cdot \omega,$$

where

$$\omega = (\mathrm{sgn}(c_i(x)), i \in \mathcal{E}, \mathrm{sgn}([c_i(x)]_+), i \in \tilde{\mathcal{A}} \cap \mathcal{I}, 0, \ldots, 0)^T \in \mathbb{R}^n.$$

Then it holds that

$$\|z(\omega)\| \leq Z := \delta n \cdot \max(\kappa^{-1}, 1).$$

And due to $\{i \in \mathcal{I} : c_i(x) > 0\} \subseteq \tilde{\mathcal{A}} \cap \mathcal{I}$, we obtain

$$\delta \cdot \mathrm{sgn}(c_i(x)) + \nabla c_i(x)^T z(\omega) = 0, \qquad i \in \mathcal{E} : c_i(x) \neq 0\,;$$

$$\delta + \nabla c_i(x)^T z(\omega) \leq 0, \qquad i \in \mathcal{I} : c_i(x) > 0\,,$$

which implies the conclusion. □

By Proposition D.1, if iterates are sufficiently close to a feasible point at which LICQ holds, then Assumption 5 is satisfied at those iterates. As can be seen, Assumption 5 is required on all infeasible iterates, depending on a specific algorithm's behavior. A natural question that arises is whether we can relax such requirement and assume a CQ condition on limit points of iterates. We next explore the answer to this question.

Although we are looking for a feasible solution of (79), in general due to the nonconvexity of this problem it is difficult to obtain a global minimizer of the feasibility minimization problem

$$\min_x \frac{1}{2}\|c_{\mathcal{E}}(x)\|^2 + \frac{1}{2}\|[c_{\mathcal{I}}(x)]_+\|^2. \tag{81}$$

Instead, we may only reach a stationary point $\tilde{x}^\star$ satisfying

$$\nabla c_{\mathcal{E}}(\tilde{x}^\star) c_{\mathcal{E}}(\tilde{x}^\star) + \nabla c_{\mathcal{I}}(\tilde{x}^\star)[c_{\mathcal{I}}(\tilde{x}^\star)]_+ = 0. \tag{82}$$

Define

$$\mathcal{Y}^\star := \{y \in \mathbb{R}^n : \|\nabla c_{\mathcal{E}}(y) c_{\mathcal{E}}(y) + \nabla c_{\mathcal{I}}(y)[c_{\mathcal{I}}(y)]_+\| = 0\}\,. \tag{83}$$

Obviously, $\mathcal{Y}^\star$ is closed due to the continuity of ∇c and c. The lemma below characterizes a property for the points in $\mathcal{Y}^\star$.

Lemma D.1 *For any $\tilde{x}^\star \in \mathcal{Y}^\star$, either $\tilde{x}^\star$ is feasible, or the linear system (80) at $\tilde{x}^\star$ has no solution for any $\delta > 0$.*

Proof. Suppose that $\tilde{x}^\star$ is infeasible. If there exist δ and z such that (80) holds at $\tilde{x}^\star$, then by (82) we obtain the following relations:

$$0 = c_{\mathcal{E}}(\tilde{x}^\star)^T \nabla c_{\mathcal{E}}(\tilde{x}^\star)^T z + [c_{\mathcal{I}}(\tilde{x}^\star)]_+^T \nabla c_{\mathcal{I}}(\tilde{x}^\star)^T z \leq -\delta \|c_{\mathcal{E}}(\tilde{x}^\star)\|_1 - \delta \|[c_{\mathcal{I}}(\tilde{x}^\star)]_+\|_1,$$

which however contradicts the infeasibility of $\tilde{x}^*$. Hence, the conclusion is derived. □

Lemma D.2 *Let $\mathcal{C}$ be a compact subset of $\mathbb{R}^n$ such that $\mathcal{Y}^\star \cap \mathcal{C} = \emptyset$. Then $\min_{y \in \mathcal{C}} \|\nabla c_{\mathcal{E}}(y) c_{\mathcal{E}}(y) + \nabla c_{\mathcal{I}}(y)[c_{\mathcal{I}}(y)]_+\| > 0$.*

Proof. Since $\mathcal{C}$ is compact, the infimum of $\|\nabla c_{\mathcal{E}}(y) c_{\mathcal{E}}(y) + \nabla c_{\mathcal{I}}(y)[c_{\mathcal{I}}(y)]_+\|$ over $\mathcal{C}$ is attainable. Arguing by contradiction, we assume that this minimum value is equal to 0. Then there exists a subsequence $\{y_k\} \subseteq \mathcal{C}$ such that $y_k \to \hat{y}$ and $\|\nabla c_{\mathcal{E}}(y_k) c_{\mathcal{E}}(y_k) + \nabla c_{\mathcal{I}}(y_k)[c_{\mathcal{I}}(y_k)]_+\| \to 0$ as $k \to \infty$. It obviously holds that $\hat{y} \in \mathcal{C}$ by the compactness of $\mathcal{C}$ and $\|\nabla c_{\mathcal{E}}(\hat{y}) c_{\mathcal{E}}(\hat{y}) + \nabla c_{\mathcal{I}}(\hat{y})[c_{\mathcal{I}}(\hat{y})]_+\| = 0$, thus $\hat{y} \in \mathcal{Y}^\star$. However, it contradicts the assumption that $\mathcal{C}$ is disjoint from $\mathcal{Y}^\star$. Therefore, the conclusion is derived. □

We now introduce the sequence $\{y^t\}$, where for each $t \geq 1$, $y^t = x^{k_t}$ with k_t being randomly and uniformly chosen from $\{1, \ldots, t\}$. Then by Theorem 1 we have

$$\lim_{t \to \infty} \mathbb{E}[\mathbf{d}^2(\nabla f(y^t) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_i^t \nabla c_i(y^t), 0)] = 0,$$

$$\lim_{t\to\infty} \mathbb{E}[\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\|^2] = 0.$$

Here $\bar{\lambda}^t$ is defined as

$$\bar{\lambda}_i^t = \begin{cases} \beta_{k_t-1}c_i(y^t) + \lambda_i^{k_t}, & i \in \mathcal{E}, \\ [\beta_{k_t-1}c_i(y^t) + \lambda_i^{k_t}]_+, & i \in \mathcal{I}. \end{cases}$$

It then implies the convergence in probability[1]:

$$\|\nabla f(y^t) + \sum_{i\in\mathcal{E}\cup\mathcal{I}} \bar{\lambda}_i^t \nabla c_i(y^t)\| \xrightarrow{p} 0, \tag{84}$$

$$\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\| \xrightarrow{p} 0. \tag{85}$$

To proceed, we lay out another assumption.

Assumption D.1 *$\{y^t\}$ is contained in a compact set $\mathcal{Y} \subset \mathbb{R}^n$.*

The next proposition establishes that $\{y^t\}$ converges in probability to $\mathcal{Y}^\star$ under Assumption D.1.

Proposition D.2 *Let Assumption D.1 hold, then we have*

$$\mathbf{d}\left(y^t, \mathcal{Y}^\star\right) \xrightarrow{p} 0, \quad \text{i.e.,} \quad \lim_{t\to\infty} \mathrm{P}(\mathbf{d}\left(y^t, \mathcal{Y}^\star\right) < \epsilon) = 1, \quad \forall \epsilon > 0. \tag{86}$$

Proof. Given $\bar{\epsilon} > 0$, define $\mathcal{C} := \mathcal{Y} \cap \{y \mid \mathbf{d}\left(y, \mathcal{Y}^\star\right) \geq \bar{\epsilon}\}$. It holds that $\mathcal{C}$ is compact and $\mathcal{Y}^\star \cap \mathcal{C} = \emptyset$. Then it follows from Lemma D.2 that

$$0 < \hat{\epsilon} := \min_{y\in\mathcal{C}} \|\nabla c_{\mathcal{E}}(y)c_{\mathcal{E}}(y) + \nabla c_{\mathcal{I}}(y)[c_{\mathcal{I}}(y)]_+\| \leq \|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\|, \quad \forall y^t \in \mathcal{C}.$$

Since $\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\| \xrightarrow{p} 0$, we have

$$\lim_{t\to\infty} \mathrm{P}(\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\| < \hat{\epsilon}) = 1,$$

which together with

$$\mathrm{P}(\mathbf{d}\left(y^t, \mathcal{Y}^\star\right) < \bar{\epsilon}) \geq \mathrm{P}(\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\| < \hat{\epsilon})$$

indicates

$$\lim_{t\to\infty} \mathrm{P}(\mathbf{d}\left(y^t, \mathcal{Y}^\star\right) < \bar{\epsilon}) = 1.$$

Therefore, due to the arbitrariness of $\bar{\epsilon}$, we derive the conclusion. □

By Lemma D.1, any point in $\mathcal{Y}^\star$ is either feasible or the linear system (80) at this point has no solution for any $\delta > 0$. To make sure that $\{y^t\}$ is approaching a KKT point of (79) with high probability, we need to impose the following assumption on $\mathcal{Y}^\star$.

Assumption D.2 *There exists $\nu > 0$ such that singular values of $\nabla c(y), y \in \mathcal{Y}^\star$ are uniformly lower bounded away from ν, where $\nabla c(y) = (\nabla c_{\mathcal{E}}(y), \nabla c_{\mathcal{I}}(y))$.*

Assumption D.2 guarantees the feasibility of any point in $\mathcal{Y}^\star$. Moreover, there exists $\tilde{\epsilon} > 0$ such that singular values of $\nabla c(y)$ are uniformly lower bounded away from $\nu/\sqrt{2}$ for any y satisfying $d(y, \mathcal{Y}^\star) < \tilde{\epsilon}$. Then for any y^t satisfying $d(y^t, \mathcal{Y}^\star) < \tilde{\epsilon}$, we have

$$\|c_{\mathcal{E}}(y^t)\| + \|[c_{\mathcal{I}}(y^t)]_+\| \leq \frac{2}{\nu}\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\|$$

which implies that

$$\mathrm{P}(\|c_{\mathcal{E}}(y^t)\| + \|[c_{\mathcal{I}}(y^t)]_+\| < \epsilon) \geq \mathrm{P}(\|\nabla c_{\mathcal{E}}(y^t)c_{\mathcal{E}}(y^t) + \nabla c_{\mathcal{I}}(y^t)[c_{\mathcal{I}}(y^t)]_+\| < \frac{\nu}{2}\epsilon) \quad \forall \epsilon > 0.$$

[1] A sequence of random variables $\{x_t\}_{t\geq 1} \subseteq \mathbb{R}^n$ is called to converge in probability to a random variable x, denoted by $x_t \xrightarrow{p} x$, if $\lim_{n\to\infty} \mathrm{P}(\|x_t - x\| \geq \epsilon) = 0$ for any $\epsilon > 0$.

Hence, it indicates from (85) that $\lim_{t\to\infty} \mathrm{P}(\|c_{\mathcal{E}}(y^t)\| + \|[c_{\mathcal{I}}(y^t)]_+\| < \epsilon) = 1$ for any $\epsilon > 0$, i.e.,

$$\|c_{\mathcal{E}}(y^t)\| + \|[c_{\mathcal{I}}(y^t)]_+\| \xrightarrow{p} 0. \tag{87}$$

Based on above analysis, for any $\epsilon \in (0, \tilde{\epsilon})$ and $\gamma \in (0, 1)$, we obtain the following statements.

(i) By (86), there exists $T_1(\gamma, \epsilon)$ such that

$$\mathrm{P}(\mathbf{d}(y^t, \mathcal{Y}^*) < \epsilon) \geq 1 - \gamma, \quad \forall t \geq T_1(\gamma, \epsilon).$$

Then it implies that the probability where singular values of $\nabla c(y^t), y^t \in \mathcal{Y}^\star$ for all $t \geq T_1(\gamma, \epsilon)$ are lower bounded away from $\nu/\sqrt{2}$ is at least $1 - \gamma$.

(ii) By (84), there exists $T_2(\gamma, \epsilon)$ such that

$$\mathrm{P}(\|\nabla f(y^t) + \nabla c(y^t)\bar{\lambda}^t\| < \epsilon) \geq 1 - \gamma, \quad \forall t \geq T_2(\gamma, \epsilon). \tag{88}$$

Then combining (i) we obtain that with probability at least $(1 - \gamma)^2$, (88) holds and meanwhile $\bar{\lambda}^t$ is upper bounded for all $t \geq \max\{T_1(\gamma, \epsilon), T_2(\gamma, \epsilon)\}$. Let us denote this upper bound by $\bar{\Lambda}$ for simplicity. Then we have $\mathrm{P}(\|\lambda^t\| < \bar{\Lambda}) \geq (1 - \gamma)^2$ for all $t \geq \max\{T_1(\gamma, \epsilon), T_2(\gamma, \epsilon)\}$.

(iii) By (87), there exists $T_3(\gamma, \epsilon)$ such that,

$$\mathrm{P}(\|c_{\mathcal{E}}(y^t)\| + \|[c_{\mathcal{I}}(y^t)]_+\| < \frac{\epsilon}{\max\{1, \bar{\Lambda}\}}) \geq 1 - \gamma, \quad \forall t \geq T_3(\gamma, \epsilon).$$

Then together with (ii) we derive that

$$\mathrm{P}(\sum_{i\in\mathcal{I}} \bar{\lambda}_i^t [c_i(y^t)]_+ < \epsilon) \geq \mathrm{P}(\|\bar{\lambda}^t\| \|[c_{\mathcal{I}}(y^t)]_+\| < \epsilon) \geq (1 - \gamma)^3, \quad \forall t \geq \max\{T_1(\gamma, \epsilon), T_2(\gamma, \epsilon), T_3(\gamma, \epsilon)\}.$$

We thus arrive at the main proposition. Here we define

$$U^t := \max\{\|c_{\mathcal{E}}(y^t)\| + \|[c_{\mathcal{I}}(y^t)]_+\|, \|\nabla f(y^t) + \nabla c(y^t)\bar{\lambda}^t\|, \sum_{i\in\mathcal{I}} \bar{\lambda}_i^t [c_i(y^t)]_+\}.$$

Proposition D.3 *Under Assumptions D.1 and D.2, for any sufficiently small $\epsilon > 0$ and any $\gamma \in (0, 1)$, there exists $T(\gamma, \epsilon)$ such that $\mathrm{P}(U^t < \epsilon) \geq 1 - \gamma$ for all $t \geq T(\gamma, \epsilon)$.*