Complexity analysis of inexact cubic-regularized primal-dual methods for finding second-order stationary points*

Xiao Wang†

Abstract

Motivated by recent developments of using cubic regularization to escape saddle points of unconstrained optimization, in this paper we explore its potential in pursuing second-order stationary points of nonconvex constrained optimization whose exact objective function information may be hard to obtain. We first propose an algorithmic framework, named as ICPD, of inexact cubic-regularized primal-dual methods for equality constrained optimization. To update the primal variable at each iteration, we construct a cubic regularized model relying on inexact first- and second-order derivatives of the objective function together with information of constraint functions. By allowing an inexact solution to each subproblem under certain conditions, we establish the iteration complexity of ICPD to find an ϵ-approximate first- and second-order stationary point, respectively. We then consider a stochastic variant of algorithm, SCPD for equality constrained optimization whose objective takes an expectation form. Through a proper sampling strategy to calculate stochastic gradients and Hessians, we address the oracle complexity of SCPD to reach approximate stationary points with high probability. We also investigate the behavior of the standard gradient descent when solving each subproblem with a random perturbation. We provide a detailed analysis on how to fulfill the required conditions on an inexact subproblem solution with high probability at each iteration. Additionally, we present an analysis of an adaptive variant of ICPD which updates penalty parameters dynamically and discuss the applicability of adaptive cubic regularization parameters. Finally, preliminary numerical results are reported to showcase the performances of our proposed algorithms.

Keywords: Constrained optimization, augmented Lagrangian function, cubic regularization, inexactness, second-order stationarity, stochastic approximation, complexity

Mathematics Subject Classification 2020: 65K05, 90C30

1 Introduction

In this paper, we consider the following equality constrained optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c(x) := (c_1(x), c_2(x), \ldots, c_m(x))^T = 0, \end{aligned} \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$ are twice continuously differentiable and possibly nonlinear and nonconvex. Nonlinear constrained optimization has been an important research field in optimization community and has been studied comprehensively for decades. Very recently, along with developments of deep learning (DL), it has revealed advantages of incorporating various constraints when training deep neural networks [38]. Related models include physics-constrained DL model [56], manifold regularized DL model [39] and constraint-aware DNN compression [15], and so on. In those models large

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

†wangx07@pcl.ac.cn, Pengcheng Laboratory, Shenzhen, 518066, China.

data sets are possibly involved. Thus under many circumstances it will be expensive sometimes even prohibitive to calculate exact information of functions.

Due to the development of complexity theories in the past ten years, it has witnessed great progress of research on nonlinear constrained optimization. Various algorithms have been studied with complexity analysis provided. Among those, penalty methods have attracted much attention. In [52] an inexact augmented Lagrangian method (ALM) is proposed for convex programs with both equality and inequality constraints. Each subproblem is solved inexactly up to a certain accuracy. Under different parameter settings, the author studies the global convergence rate and gradient evaluation complexity to produce a primal and/or primal-dual solution. With Nesterov's optimal first-order method as the subproblem solver, the number of gradient evaluations to reach a primal ϵ-solution is in order $\mathcal{O}(\epsilon^{-1})$ if the objective is convex and $(\epsilon^{-1/2}|\log\epsilon|)$ if it is strongly convex. Complexity to produce a primal-dual ϵ-solution is also established for convex and strongly convex case, respectively. It studies in [3] the complexity of an inexact ALM corresponding to Algencan [1] for general nonconvex constrained optimization. It is shown that the outer iteration complexity bound is $\mathcal{O}(|\log(\epsilon)|)$ to find an ϵ-approximate KKT point with KKT measure less than ϵ in l_2-norm, or an approximate stationary point of an infeasibility measure when the penalty parameter is unbounded. Paper [20] studies the worst-case complexity of an inexact ALM for inequality constrained optimization to find an ϵ-approximate KKT point. It shows the outer iteration complexity bound $\mathcal{O}(\epsilon^{-2/(\nu-1)})$ when the penalty parameter is unbounded with $\nu > 1$ used to control the increase rate. Oracle complexity for linearly constrained optimization is estimated in [20]. Recent paper [30] studies an inexact proximal-point penalty method for nonconvex constrained optimization. Objective of each subproblem is formed by adding a quadratic penalty function and a proximal term to the original objective function. Under the weak-convexity assumption, computational complexity in terms of the number of proximal gradient steps to find an ϵ-stationary point are analyzed for cases with convex constraints and with nonconvex constraints. In particular, when constraints are nonconvex, under a non-singularity condition the complexity is $\tilde{O}(\epsilon^{-3})$. It is shown in [9] that finding an ϵ-approximate first-order critical point of general smooth constrained optimizaiton needs $\mathcal{O}(\epsilon^{-2})$ function and constraint evaluations. But it requires each linearized trust-region subproblem be exactly solved.

To achieve second-order stationarity, [21] considers a class of linearly constrained optimization problems where the objective function may not be differentiable or twice differentiable. The paper focuses on interior trust-region point algorithms, establishing computational complexity for finding approximate second-order stationary points. In [31], two algorithms based on negative-curvature gradient projection are proposed to pursue second-order stationarity for smooth linearly constrained nonconvex optimization. The paper analyzes the per-iteration complexity and global sublinear rate of these algorithms. In addition, [23] by Razaviyayn et al. studies two first-order primal-dual based algorithms for a class of linearly constrained non-convex optimization problems and characterizes their global convergence to second-order stationary solutions. Furthermore, [37] investigates a trust-region algorithm and provides an analysis of the iteration complexity for finding an approximate second-order stationary point. In [8] a trust-region algorithm is studied for unconstrained optimization and then extended to convexly constrained problems. Evaluations of objective function values and its derivatives are at most $\mathcal{O}(\epsilon^{-(2p+1)})$ to reach an ϵ-approximate p-th order critical point, where $p \geq 1$. Nevertheless, it requires a global minimizer of a convexly constrained pth-order Taylor model. The problem of escaping saddle points in convexly constrained optimization is also studied in [32] and the iteration complexity to reach an approximate second-order stationary point is established. However, the proposed algorithm framework in [32] requires the convex feasible set be simple for a quadratic objective function such that an approximate solution of a convexly constrained quadratic program can be efficiently computed. For problems with possibly nonlinear and nonconvex constraints, authors of [49] study a proximal AL method for nonconvex equality constrained optimization and apply the Newton-CG algorithm [40] to solve each subproblem. The outer iteration complexity bound to find an approximate ϵ-first- and -second-order solution is $\mathcal{O}(\epsilon^{\eta-2})$ when the penalty parameter is in order $\mathcal{O}(\epsilon^{-\eta})$, where $\eta \in [0, 2]$ is user-defined. The authors also analyze the total iteration complexity regarding the number of inner-loop iterations as well as the operation complexity on matrix-vector products to reach approximate solutions. In the subsequent work [22] a new Newton-CG based ALM with improved complexities to achieve a second-order stationary point is proposed and studied. A two-phase minimization algorithm for nonconvex constrained optimization is studied in [10]

with the evaluation complexity analyzed to achieve first-, second- and third-order criticality. But a global minimizer of a convexly constrained high-order Taylor model is also required. As is noted, for general nonlinear constrained optimization, all aforementioned algorithms aiming for high-order stationary points has stronger conditions on subproblem solutions, such as high-order stationary points of an associated (possibly highly) nonlinear penalty function or a global minimizer of a relatively complicated subproblem. However, they are normally expensive or even impossible to realize in practical computations. Motivated by this, in this paper we hope to design algorithms based on simpler subproblems which can only be solved inexactly, towards second-order stationarity for nonlinear constrained optimization.

For unconstrained optimization, cubic regularization (CR) has been well studied due to its advantages in promoting second-order stationary. It originates from the pioneering work by Nesterov and Polyak [35]. Consider the unconstrained smooth optimization problem $\min_{x\in\mathbb{R}^n}\phi(x)$. At kth iteration the CR approach minimizes a local upper bound on ϕ, obtained by using its second-order Taylor expansion at current iterate x_k plus a cubic regularizer:

$$\phi(x_k)+\langle\nabla\phi(x_k),x-x_k\rangle+\frac{1}{2}\langle x-x_k,\nabla^2\phi(x_k)(x-x_k)\rangle+\frac{\sigma_k}{6}\|x-x_k\|^3, \tag{1.2}$$

where $\sigma_k>0$. Variants of algorithms have been proposed with complexity analysis to find an (ϵ_g,ϵ_H)-point x satisfying

$$\|\nabla\phi(x)\|\le\epsilon_g,\quad \lambda_{\min}(\nabla^2\phi(x))\ge-\epsilon_H,$$

where $\epsilon_g,\epsilon_H>0$. It is shown in [35] that the iteration complexity to find an $(\epsilon_g,\epsilon_g^{1/2})$-point is $\mathcal{O}(\epsilon_g^{-3/2})$, if each subproblem is solved exactly. Cartis et al. [11] propose an adaptive regularization with cubics and establish the $\mathcal{O}(\max(\epsilon_g^{-3/2},\epsilon_H^{-3}))$ iteration complexity to find an (ϵ_g,ϵ_H)-point, where each subproblem can be solved inexactly. Subsequent works focus on operation complexity which characterizes the total number of operations including function information evaluation and matrix-vector products. It has been shown that the operation complexity achieved by CR or its variants is in order $\tilde{O}(\epsilon_g^{-7/4})$ with high probability to find an $(\epsilon_g,\epsilon_g^{1/2})$-point of ϕ. Related works include, but not limited to, [7, 18, 24, 40, 41]. In recent years, there has been growing interest in the pursuit of second-order stationarity for stochastic CR methods in the context of unconstrained optimization [26,44,54,55]. In the work [26], the authors propose a sub-sampled CR approach for unconstrained optimization in the finite-sum form. This technique incorporates sub-sampling to compute gradient and Hessian estimates, ensuring that the required conditions are satisfied with high probability. This approach aims to promote desirable theoretical properties of the algorithm. Another two relevant works are [54] and [55], where stochastic cubic regularized algorithms based on variance reductions are studied for unconstrained finite-sum optimization. Additionally, [44] focuses on the unconstrained minimization of a ρ-Hessian Lipschitz function $f(x)=\mathbb{E}_{\xi\sim\mathcal{D}}[F(x;\xi)]$ and proposes a stochastic optimization method that utilizes stochastic gradients and Hessian-vector products to construct the cubic regularized approximation models. By using sampling minibatches for gradient and Hessian evaluations, the method can find an ϵ-second-order stationary point satisfying

$$\|\nabla f(x)\|\le\epsilon,\quad \lambda_{\min}(\nabla^2 f(x))\ge-\sqrt{\rho\epsilon}, \tag{1.3}$$

with high probability and within $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ oracle evaluations.

In the context of constrained optimization, CR-type methods have also been studied by works such as [12], [13], and [14]. The key of these methods is to employ an auxiliary function to construct a least-square subproblem, which is solved using a CR approach. Specifically, in the case of the equality constrained optimization problem (1.1), [12] introduces the function $r(x,t)$ as $r(x,t):=(c(x)^T,f(x)-t)^T$, with t serving as a target value for the objective function f. Then a CR approach is called to solve the least-square problem $\min\frac{1}{2}\|r(x,t_k)\|^2$, where t_k is adaptively varied. In [13], when solving the general nonlinear optimization, i.e. (1.1) together with a convex set constraint $x\in X$, a two-phase Short-Step ARC algorithm is proposed and in each phase a CR approach is called to solve a convex set constrained least square problem. To find an approximate first-order stationary point satisfying

$$\Big(\chi_l(x_k,y_k)\le\epsilon^{2/3}\|(y_k,1)\|_2 \text{ and } \|c(x_k)\|_2\le\delta\epsilon\Big) \text{ or } \Big(\chi_{\|c\|_2}(x_k)\le\epsilon^{2/3} \text{ and } \|c(x_k)\|_2>\delta\epsilon\Big) \tag{1.4}$$

for some $\delta > 0$, where $\chi_\omega(x) = |\min_{x+d\in X,\|d\|\le 1}\langle\nabla\omega(x), d\rangle|$ for a mapping $\omega : \mathbb{R}^n \to \mathbb{R}$ and $l(x) = f(x) + \langle y, c(x)\rangle$ is Lagrange function associated with (1.1), the proposed algorithm achieves an improved bound, $\mathcal{O}(\epsilon^{-3/2})$, on the evaluations of functions. Nevertheless it is worth noting that the works by [12–14] focus on CR-type methods for approximating first-order criticality, without providing a discussion on second-order criticality. Moreover, algorithms studied in these works are all designed for problems in deterministic settings and rely on exact function information including exact function values, gradients and/or Hessians. They cannot apply under circumstances when exact function information is costly sometimes even prohibitive to calculate. Currently, study on nonlinear constrained stochastic optimization is still limited. Wang et al. [46] study penalty methods based on first/zeroth-order stochastic approximations for stochastic optimization with deterministic equality constraints. Curtis et al. [17] consider the same kind of problems and propose stochastic SQP methods with complexity analysis provided. Xu studies in [50] a stochastic primal-dual algorithm for convex programs with nonlinear constraints. Recently, Jin and Wang [25] extend this algorithm to nonconvex settings. Boob et al. [4] study algorithms for inequality constrained optimization in different scenarios, including deterministic setting, semi-stochastic setting (constraints are deterministic) and fully-stochastic setting. But due to a strong feasibility assumption, the algorithms proposed in [4] cannot be applied to equality constrained optimization. Shi et al. [43] propose a linearized ALM based on momentum for general constrained stochastic optimization which may contain both equality and inequality constraints. There are also some work on nonlinear stochastic optimization with expectation constraints. Here we will not give more details while interested readers are referred to [28, 53]. Note that all previous works consentrate on the first-order stationarity of algorithms. The study on the second-order stationary for nonlinear constrained stochastic optimization is still rare. We note that in recent paper [32] a stochastic extension of the proposed algorithm framework escaping saddle points is studied. But it focuses on convexly constrained optimization and requires the convex feasible set be relatively easy to cope with. In literature there has not been any study on second-order stationarity for general stochastic optimization with possibly nonconvex constraints. Motivated by this, we will propose a stochastic approximation method for (nonconvexly) equality constrained stochastic optimization and investigate its oracle complexity for finding second-order stationary points.

1.1 Contributions

Main contributions of this paper lie in the following aspects.

- We propose an algorithm framework ICPD of inexact cubic-reguarized primal-dual methods applied for equality constrained optimization. To update the primal variable, at each iteration we construct a cubic model that combines a cubic regularizer with a quadratic approximation to the augmented Lagrangian (AL) function around the current iterate. One advantage of our approach is that the subproblems involved in our algorithm framework are much simpler to solve in comparison to existing methods.

- In the algorithm design we do not need to evaluate any objective function value, require exactness of gradients/Hessians or solve each subproblem exactly. Instead, we allow to use approximate first- and second-order derivatives of the objective function and solve each subproblem inexactly. Under certain conditions we can derive the iteration complexity bounds of ICPD to reach an ϵ-FSP and ϵ-SSP, respectively. Those complexity orders obtained for deterministic nonlinear constrained optimization are compatitative with and in certain scenarios even lower than existing algorithms that rely on exact function information.

- Due to the regime of ICPD that allows to use inexact derivatives of the objective function, we can easily extend it to cope with problems in stochastic settings, when only stochastic oracles can be available. We thus propose a stochastic variant of algorithm, SCPD for equality constrained optimization with the objective function in an expectation form. We investigate numbers of stochastic gradient and Hessian evaluations to achieve the required conditions on inexact oracles at each iteration, then establish corresponding oracle complexity bounds of SCPD to reach approximate

stationary points with high probability. To the best of our knowledge, the analysis on second-order stationarity for general nonlinear constrained stochastic optimization is new in the literature.

- In order to satisfy conditions imposed on inexact solution of the subproblem at each iteration, we investigate the behavior of the standard gradient descent with a random perturbation. Under appropriate parameter settings, we can make sure that required conditions are achieved with high probability, provided that the iteration number of the subsolver is sufficiently large.
- We present a theoretical analysis of an adaptive ICPD algorithm, which incorporates a scheme to dynamically update penalty parameters. We investigate the theoretical behavior of the algorithm under scenarios of bounded and unbounded penalty parameters respectively. In addition, we discuss the potential applicability of an adaptive update for the cubic regularization parameter.
- We conduct some preliminary numerical implementations on proposed algorithms for solving a quadratically constrained nonconvex program and a multi-class Neyman-Pearson classification problem. We report their performance profiles in different scenarios comparing with several existing algorithms in the literature.

1.2 Notations and preliminaries

The following notations are used throughout the remainder of the paper. Without any specification we use $\|\cdot\|$ to denote the Euclidean norm. Given two vectors $x, y \in \mathbb{R}^n$, $\langle x, y\rangle = x^T y$ refers to the inner product of x and y. We use $\mathbf{I}_n$ to denote the n-dimensional identity matrix. Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, $A \succeq 0$ means that A is positive semidefinite. We define the Jacobian matrix $\nabla c(x) = (\nabla c_1(x), \ldots, \nabla c_m(x))$ and the null space $\mathrm{Null}(\nabla c(x)^T) = \{d \in \mathbb{R}^n : \nabla c(x)^T d = 0\}$. Notation $\tilde{\mathcal{O}}(\cdot)$ is used to hide the dependence on logarithmic factor. The notation $\zeta \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ represents that ζ is uniformly distributed on the unit sphere in $\mathbb{R}^n$ and $\mathbb{N}$ represents the set of non-negative integers.

For nonlinear constrained optimization, in general it is NP-hard to reach the global minimizer or even a local minimizer. As is well-known, under certain constraint qualifications local minimizers satisfy first- and second-order necessary conditions [36]. Points satisfying those necessary conditions are usually called stationary points. Currently the main research interest has been put on seeking the more trackable stationary points of (1.1).

Definition **1.1.** *We call x a first-order stationary point of* (1.1)*, if there exists $\lambda \in \mathbb{R}^m$ such that*

$$\nabla f(x) + \nabla c(x)\lambda = 0, \quad c(x) = 0. \tag{1.5}$$

Definition **1.2.** *We call x a second-order stationary point of* (1.1)*, if there exists $\lambda \in \mathbb{R}^m$ such that* (1.5) *holds and*

$$d^T(\nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 c_i(x))d \geq 0 \quad \textit{for any } d \in \mathrm{Null}(\nabla c(x)^T). \tag{1.6}$$

In this paper we study algorithms for finding approximate stationary points of (1.1) which are defined as below.

Definition **1.3.** *We call x an ϵ-approximate first-order stationary point (ϵ-FSP) of* (1.1)*, if there exists $\lambda \in \mathbb{R}^m$ such that*

$$\|\nabla f(x) + \nabla c(x)\lambda\| \leq \epsilon \ \textit{ and } \ \|c(x)\| \leq \epsilon. \tag{1.7}$$

Definition **1.4.** *We call x an ϵ-approximate second-order stationary point (ϵ-SSP) of* (1.1)*, if there exists $\lambda \in \mathbb{R}^m$ such that* (1.7) *holds and*

$$d^T(\nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 c_i(x))d \geq -\sqrt{\epsilon}\|d\|^2 \textit{ for any } d \in \mathrm{Null}(\nabla c(x)^T). \tag{1.8}$$

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we present the detailed description of the algorithm framework, ICPD, for cubic-regularized primal-dual methods for solving problem (1.1). In Section 3, we establish the iteration complexity of ICPD to find an ϵ-FSP and an ϵ-SSP of (1.1), respectively. In Section 4 we propose a stochastic variant of ICPD for equality constrained optimization with the objective function in an expectation form. We analyze its oracle complexity to find approximate stationary points with high probability. In Section 5 we study the subproblem solver based on standard gradient descent approach for each subproblem with a random perturbation. We show that under certain conditions the required conditions on inexact subproblem solutions can be satisfied with high probability. In Section 6 we provide discussions on adaptive ICPD with penalty parameters and regularization parameters updated dynamically. We also report numerical experiment results on proposed algorithms in Section 7. Finally, we draw conclusions in Section 8.

2 Algorithm framework

The augmented Lagrangian (AL) function associated with (1.1) is defined as

$$\mathcal{L}_\beta(x,\lambda) = f(x) + \Psi_\beta(x,\lambda), \quad \text{where } \Psi_\beta(x,\lambda) := \lambda^T c(x) + \frac{\beta}{2}\|c(x)\|^2, \tag{2.1}$$

x is the primal variable, $\lambda \in \mathbb{R}^m$ refers to the dual variable, and $\beta > 0$ is a penalty parameter. In classic augmented Lagrangian methods (ALMs) for (1.1), to update the primal variable the AL function with fixed λ and β needs to be minimized approximately. However, the related computational burden is normally high. Recently, linearized ALMs have attracted much attention. In those methods simpler subproblems are solved. Each subproblem is built on a quadratic approximation to the AL function. See, for example, [51] for reference. Specifically, at current iterate x and for fixed β and λ, consider to approximate $\mathcal{L}_\beta(x+d,\lambda)$ by a quadratic model:

$$\mathcal{L}_\beta(x+d,\lambda) \approx \mathcal{L}_\beta(x,\lambda) + \langle \nabla \mathcal{L}_\beta(x,\lambda), d\rangle + \frac{1}{2\eta}\|d\|^2$$

and use this model to generate next primal iterate. It is worthy to note that linearized ALMs in the literature merely aim for first-order stationarity. A natural question is whether we can achieve higher stationarity based on AL function while constructing relatively simpler subproblems. Fortunately, the answer is affirmative. This is inspired by recent development on cubic regularization (CR) for unconstrained optimization. As discussed in Introduction, CR has shown great potential in helping find approximate second-order stationary points for unconstrained optimization. In CR approaches each subproblem is built on a cubic model by incorporating second-order information of the original problem. We can apply similar idea to nonlinear constrained optimization (1.1) in order to pursue second-order stationarity. But due to the existence of constraints, the cubic model we try to build should be able to merge the information of both objective function and constraints. This is easy to realize by means of the AL function. The key idea here is that at each iteration we use a cubic regularized model to approximate the AL function around current iterate:

$$\mathcal{L}_\beta(x+d,\lambda) \approx \mathcal{L}_\beta(x,\lambda) + \langle \nabla \mathcal{L}_\beta(x,\lambda), d\rangle + \frac{1}{2}\langle d, \nabla^2_{xx}\mathcal{L}_\beta(x,\lambda)d\rangle + \frac{\sigma}{6}\|d\|^3, \tag{2.2}$$

where $\sigma > 0$ is a regularization parameter.

The cubic model in (2.2) involves calculation of exact gradient and Hessian of f at current iterate. But under many circumstances it is expensive to compute exact derivatives of f, so we can only get access to an approximate gradient g_k^0 and Hessian H_k^0 of f at an inquiry point x_k. Then we compute

$$g_k = g_k^0 + \nabla_x \Psi_{\beta_k}(x_k,\lambda_k), \quad H_k = H_k^0 + \nabla^2_{xx}\Psi_{\beta_k}(x_k,\lambda_k), \tag{2.3}$$

which obviously are approximations to $\nabla_x \mathcal{L}_\beta(x_k, \lambda_k)$ and $\nabla^2_{xx} \mathcal{L}_\beta(x_k, \lambda_k)$. In order to update the primal variable at kth iteration we now define the subproblem:

$$\min_{d \in \mathbb{R}^n} \quad q_k(d) := \langle g_k, d\rangle + \frac{1}{2}\langle d, H_k d\rangle + \frac{\sigma_k}{6}\|d\|^3. \tag{2.4}$$

The lemma below characterizes properties of the optimal solution of (2.4).

LEMMA **2.1.** *The optimal solution of* (2.4)*, denoted by* s_k^**, satisfies following optimality conditions:*

$$H_k + \frac{\sigma_k}{2}\|s_k^*\|\mathbf{I}_n \succeq 0, \tag{2.5}$$

$$g_k + H_k s_k^* + \frac{\sigma_k}{2}\|s_k^*\| s_k^* = 0,$$

$$\langle g_k, s_k^*\rangle + \frac{1}{2}\langle s_k^*, H_k s_k^*\rangle + \frac{\sigma_k}{6}\|s_k^*\|^3 \leq -\frac{\sigma_k}{12}\|s_k^*\|^3. \tag{2.6}$$

In practical computations, however, the optimal solution of (2.4) is normally out of reach due to the possible nonconvexity of q_k. We may only obtain an approximate minimizer of q_k, denoted by s_k. In order to derive desirable theoretical properties we impose the following conditions on the inexact solution of (2.4). Here, the positive parameter ω is to control the tolerance.

Condition A The inexact solution s_k of subproblem (2.4) satisfies the following conditions:

$$|\|s_k^*\| - \|s_k\|| \leq \omega^{1/3}, \tag{2.7a}$$

$$q_k(s_k) \leq q_k(s_k^*) + \frac{1}{18}\sigma_k \omega, \tag{2.7b}$$

$$\left\| g_k + H_k s_k + \frac{\sigma_k}{2}\|s_k\| s_k \right\| \leq \sigma_k \omega^{2/3}, \tag{2.7c}$$

where $\omega > 0$.

Remark 2.1. *Methods, including Krylov subspace approach and gradient descent approach, for solving cubic regularized subproblems have been studied in the literature such as [6, 11, 26], which have different requirements on inexact subproblem solutions from ours. In Section 5 we will specify the subproblem solver used in our algorithm and analyze how Condition A can be guaranteed with high probability.*

With the new primal iterate $x_{k+1} := x_k + s_k$, we come to update the dual variable. A popular way in ALMs for (1.1) is to compute

$$\lambda_{k+1} = \bar{\lambda}_{k+1} := \lambda_k + \beta_k c(x_{k+1}).$$

However, to control the distance between λ_k and λ_{k+1}, we apply a convex combination of λ_k and $\bar{\lambda}^{k+1}$ by introducing parameter $\rho_k \in (0, \beta_k)$ to determine λ_{k+1}:

$$\lambda_{k+1} = (1 - \frac{\rho_k}{\beta_k})\lambda_k + \frac{\rho_k}{\beta_k}\bar{\lambda}_{k+1},$$

which is exactly

$$\lambda_{k+1} = \lambda_k + \rho_k c(x_{k+1}). \tag{2.8}$$

Same strategy has also been adopted in [25, 43].

We are now ready to present the algorithm framework, ICPD, for inexact cubic-regularized primal-dual methods for solving (1.1).

Algorithm 2.1 ICPD

Input: $x_1 \in \mathbb{R}^n, \lambda_1 \in \mathbb{R}^m, \beta_1 > 0, \sigma_1 > 0, \bar{\epsilon} \in (0,1), \omega \in (0,1), \rho_1 \in (0, \beta_1)$

1: **for** $k = 1 \ldots$ **do**
2: Generate approximate gradient g_k^0 and approximate Hessian H_k^0 of f at x_k and compute g_k and H_k through (2.3).
3: Solve subproblem (2.4) obtaining an inexact solution s_k satisfying Condition A and set $x_{k+1} := x_k + s_k$.
4: If $\max\{\|x_{k+1} - x_k\|, \|x_k - x_{k-1}\|\} < \bar{\epsilon}$, terminate the algorithm and return x_{k+1}.
5: Compute β_{k+1} satisfying $\beta_{k+1} \geq \beta_k$
6: Compute $\rho_{k+1} \in (0, \beta_{k+1})$ and σ_{k+1}
7: Compute λ_{k+1} through (2.8).
8: $k := k + 1$.
9: **end for**

Note that the termination criterion in Step 4 of ICPD relies on parameter $\bar{\epsilon}$. Obviously $\bar{\epsilon}$ depends on the type of approximate stationary point we are seeking. To reach ϵ-approximate stationary points, we will specify the value of $\bar{\epsilon}$ in next section. In addition, for simplicity we let $\lambda_1 = \mathbf{0}$ in the following analysis. We will specify in next section the condition that g_k^0 and H_k^0 should satisfy in theoretical analysis. In practical computations, there are various ways to select the penalty parameters β_k and regularization parameters σ_k. For instance, one can update them adaptively depending on the algorithm's progress or assign them predetermined values.

3 Iteration complexity

In this section, we will establish the iteration complexity of ICPD to find ϵ-approximate stationary points of (1.1). As each iteration of ICPD only involves a single inexact gradient and Hessian evaluation, its oracle complexity, in terms of the total number of inexact gradient and Hessian evaluations, is in the same order as the iteration complexity. To proceed, we first lay out several assumptions used throughout the remainder of the paper.

Assumption 1. *Function f is lower bounded by f_{low}, L_f-Lipschitz continuous and twice continuously differentiable with L_g^f-Lipschitz continuous gradient and L_H^f-Lipschitz continuous Hessian, i.e., for any $x, y \in \mathbb{R}^n$,*

$$\|\nabla f(x)\| \leq L_f, \quad \|\nabla f(x) - \nabla f(y)\| \leq L_g^f \|x - y\|, \quad \|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_H^f \|x - y\|.$$

Assumption 2. *Functions $c_i, i = 1, \ldots, m$ are L_c-Lipschitz continuous and twice continuously differentiable with L_g^c-Lipschitz continuous gradient and L_H^c-Lipschitz continuous Hessian, i.e., for any $x, y \in \mathbb{R}^n$,*

$$\|\nabla c_i(x)\| \leq L_c, \quad \|\nabla c_i(x) - \nabla c_i(y)\| \leq L_g^c \|x - y\|, \quad \|\nabla^2 c_i(x) - \nabla^2 c_i(y)\| \leq L_H^c \|x - y\|$$

for $i = 1, \ldots, m$.

Assumption 3. *Let $\mathcal{X}$ be an open convex set that contains $\{x_k\}$ generated by the associated algorithm, and f, $c_i, i = 1, \ldots, m$ are bounded over $\mathcal{X}$, namely there exists $C > 0$ such that $|f(x)| \leq C$ and $\|c(x)\| \leq C$ for any $x \in \mathcal{X}$.*

Remark 3.1. *In the study of stochastic constrained optimization, the boundedness condition plays a crucial role in ensuring reliable properties of the iteration sequence. Due to the inherent randomness of the algorithmic process, it becomes challenging to guarantee that all iterates remain within a specific level set. Several related works, such as [2, 19, 33, 34], have also recognized the necessity of the boundedness assumption in achieving desirable properties. Assumption 5 in the work [48], which focuses on second-order stationarity for deterministic constrained optimization, also imposes a boundedness condition. Specifically, all iterates generated within each subproblem are assumed to be contained in a bounded*

and convex set. Moreover, both the objective function f and constraint functions are assumed to be twice uniformly Lipschitz continuously differentiable on this set. The assumptions in [13] are also relevant to our work. These assumptions differ slightly, as they consider the weak Lipschitz continuity of $\nabla^2 c_i$ and $\nabla^2 f$ along the segments between x_k and a trial point x_k^+. It is important to note that our analysis in this paper remains valid when we relax the uniform Lipschitz continuity of $\nabla^2 c_i$ and $\nabla^2 f$ to weak Lipschitz continuity along the segments $[x_k, x_{k+1}]$ for $k \geq 1$.

The following lemma provides upper bounds on λ_k, $k \geq 1$.

LEMMA **3.1.** *Under Assumption 3, it holds that*

$$\|\lambda_{k+1} - \lambda_k\|_1 \leq \rho_k \sqrt{m} C \quad \text{and} \quad \|\lambda_k\|_1 \leq \sqrt{m} C \sum_{t=1}^{k-1} \rho_t, \quad \forall k \geq 1,$$

where $\sum_{t=1}^{0} \rho_t := 0$.

Proof. By the update strategy of λ_k as in (2.8), we obtain

$$\|\lambda_{k+1} - \lambda_k\|_1 = \rho_k \|c(x_{k+1})\|_1 \leq \rho_k \sqrt{m} \|c(x_{k+1})\| \leq \rho_k \sqrt{m} C,$$

which yields the conclusion from $\|\lambda_k\|_1 \leq \sum_{t=1}^{k-1} \|\lambda_{t+1} - \lambda_t\|_1$. □

The lemma below characterizes the smoothness of Ψ_{β_k} with respect to $x \in \mathcal{X}$ for fixed λ_k.

LEMMA **3.2.** *Under Assumptions 2-3, it holds that for any $x, y \in \mathcal{X}$ and $k \geq 1$,*

$$\|\nabla_x \Psi_{\beta_k}(x, \lambda_k) - \nabla_x \Psi_{\beta_k}(y, \lambda_k)\| \leq L_k^g \|x - y\|, \tag{3.1}$$

$$\|\nabla_{xx}^2 \Psi_{\beta_k}(x, \lambda_k) - \nabla_{xx}^2 \Psi_{\beta_k}(y, \lambda_k)\| \leq L_k^H \|x - y\|, \tag{3.2}$$

where $L_k^g = m\beta_k L_c^2 + \sqrt{m}\beta_k C L_g^c + \sqrt{m} L_g^c C \sum_{t=1}^{k-1} \rho_t$ *and* $L_k^H = \sqrt{m} C L_H^c \sum_{t=1}^{k-1} \rho_t + \sqrt{m} C \beta_k L_H^c + 3m\beta_k L_c L_g^c$.

Proof. It follows from the definition of Ψ_{β_k} in (2.1) that for any $x, y \in \mathcal{X}$,

$$\begin{aligned}
\|\nabla_x \Psi_{\beta_k}(x, \lambda) - \nabla_x \Psi_{\beta_k}(y, \lambda)\| &= \|\sum_{i=1}^{m} [(\lambda_i + \beta_k c_i(x))\nabla c_i(x) - (\lambda_i + \beta_k c_i(y))\nabla c_i(y)]\| \\
&\leq \sum_{i=1}^{m} \|(\lambda_i + \beta_k c_i(x))\nabla c_i(x) - (\lambda_i + \beta_k c_i(y))\nabla c_i(y)\| \\
&= \sum_{i=1}^{m} \|\beta_k (c_i(x) - c_i(y))\nabla c_i(x) + (\lambda_i + \beta_k c_i(y)) [\nabla c_i(x) - \nabla c_i(y)]\| \\
&\leq \sum_{i=1}^{m} \left[\beta_k |c_i(x) - c_i(y)| \|\nabla c_i(x)\| + |\lambda_i + \beta_k c_i(y)| L_g^c \|x - y\|\right] \\
&\leq \sum_{i=1}^{m} \left[\beta_k L_c^2 \|x - y\| + (|\lambda_i| + \beta_k |c_i(y)|) L_g^c \|x - y\|\right] \\
&= (m\beta_k L_c^2 + L_g^c \|\lambda\|_1 + \sqrt{m} C \beta_k L_g^c) \|x - y\|
\end{aligned}$$

which together with $\lambda = \lambda_k$ derives (3.1) by Lemma 3.1.

We next prove (3.2). It follows from Assumption 2 that

$$\begin{aligned}
&\|\nabla_{xx}^2 \Psi_{\beta_k}(x, \lambda) - \nabla_{xx}^2 \Psi_{\beta_k}(y, \lambda)\| \\
&= \|\sum_{i=1}^{m} (\lambda_i + \beta_k c_i(x))\nabla^2 c_i(x) + \beta_k \sum_{i=1}^{m} \nabla c_i(x) \nabla c_i(x)^T
\end{aligned} \tag{3.3}$$

$$
\begin{aligned}
&- \sum_{i=1}^{m}(\lambda_i + \beta_k c_i(y))\nabla^2 c_i(y) - \beta_k \sum_{i=1}^{m} \nabla c_i(y)\nabla c_i(y)^T\| \\
&\le \|\sum_{i=1}^{m} \lambda_i(\nabla^2 c_i(x) - \nabla^2 c_i(y))\| + \beta_k \|\sum_{i=1}^{m}[c_i(x)\nabla^2 c_i(x) - c_i(y)\nabla^2 c_i(y)]\| \\
&\quad + \beta_k \|\sum_{i=1}^{m}[\nabla c_i(x)\nabla c_i(x)^T - \nabla c_i(y)\nabla c_i(y)^T]\| \\
&\le L_H^c \|\lambda\|_1 \|x-y\| + \beta_k \|\sum_{i=1}^{m} c_i(x)(\nabla^2 c_i(x) - \nabla^2 c_i(y))\| \\
&\quad + \beta_k \|\sum_{i=1}^{m}(c_i(x) - c_i(y))\nabla^2 c_i(y)\| + \beta_k \|\sum_{i=1}^{m} \nabla c_i(x)(\nabla c_i(x) - \nabla c_i(y))^T\| \\
&\quad + \beta_k \|\sum_{i=1}^{m}(\nabla c_i(x) - \nabla c_i(y))\nabla c_i(y)^T\| \\
&\le (L_H^c \|\lambda\|_1 + \sqrt{m} C \beta_k L_H^c + 3m\beta_k L_c L_g^c)\|x-y\|, \qquad (3.4)
\end{aligned}
$$

which indicates (3.2) by Lemma 3.1. □

In our algorithm we do not require the exactness of derivatives of f, while only approximate first- and second-order derivatives are used. But to obtain desirable theoretical properties we impose the following condition on errors of those approximations.

Condition B For any $k \ge 1$, the gradient and Hessian approximations: g_k^0 and H_k^0 satisfy

$$\|g_k^0 - \nabla f(x_k)\| \le \theta\beta_k \max\{\|x_k - x_{k-1}\|^2, \hat{\epsilon}^2\}, \qquad (3.5a)$$

$$\|H_k^0 - \nabla^2 f(x_k)\| \le \theta\beta_k \max\{\|x_k - x_{k-1}\|, \hat{\epsilon}\}, \qquad (3.5b)$$

where $x_0 := x_1$, $\theta \in (0,1)$ and $\hat{\epsilon} \in [0, \bar{\epsilon}]$.

Remark 3.2. *In above condition, the coefficients in right hand sides of* (3.5a) *and* (3.5b) *can be different. Here we choose the same value $\theta\beta_k$ only for simplicity.*

Without loss of generality we assume that

$$\sigma_{k-1} \ge 12(\frac{1}{18} + \mu)\sigma_k \text{ with } \mu \in (0, \tfrac{1}{36}], \quad \forall k \ge 1. \qquad (3.6)$$

In the lemma below we provide an estimate on the upper bound of accumulated distances between iterates.

LEMMA **3.3.** *Under Assumptions 1-3, Conditions A and B, assume that ICPD does not terminate before Kth iteration. If $\omega = \bar{\epsilon}^3$ and* (3.6) *holds, then*

$$\sum_{k=1}^{K-1} \hat{\sigma}_k \max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} \le \mathcal{L}_{\beta_1}(x_1, \lambda_1) - f_{low} + 2\sqrt{m}C^2 \sum_{k=1}^{K-1} \rho_k + \frac{1}{2}C^2 \sum_{k=1}^{K-1}(\beta_{k+1} - \beta_k), \qquad (3.7)$$

where $\hat{\sigma}_k = \mu\sigma_k - \frac{1}{12}(L_H^f + \sqrt{m}CL_H^c \sum_{t=1}^{k-1} \rho_t + \beta_k(\sqrt{m}CL_H^c + 3mL_cL_g^c + 54\theta))$.

Proof. When $K = 1$, (3.7) holds obviously. Thus without loss of generality, we assume that $K > 1$. For any $k = 1, \ldots, K-1$, it holds that

$$\max\{\|x_{k+1} - x_k\|, \|x_k - x_{k-1}\|\} \ge \bar{\epsilon}. \qquad (3.8)$$

By (3.2), we attain that for any $x, y \in \mathcal{X}$,

$$\|\nabla_{xx}^2 \mathcal{L}_{\beta_k}(x, \lambda_k) - \nabla_{xx}^2 \mathcal{L}_{\beta_k}(y, \lambda_k)\| \le \|\nabla^2 f(x) - \nabla^2 f(y)\| + \|\nabla_{xx}^2 \Psi_{\beta_k}(x, \lambda_k) - \nabla_{xx}^2 \Psi_{\beta_k}(y, \lambda_k)\|$$

$$\leq (L_H^f + L_k^H)\|x - y\|,$$

where L_k^H is defined in Lemma 3.2. Then from (2.6) and (2.7b) it follows that

$$\begin{aligned}
&\mathcal{L}_{\beta_k}(x_{k+1}, \lambda_k) \\
&\leq \mathcal{L}_{\beta_k}(x_k, \lambda_k) + \langle \nabla_x \mathcal{L}_{\beta_k}(x_k, \lambda_k), x_{k+1} - x_k \rangle + \frac{1}{2}\langle x_{k+1} - x_k, \nabla_{xx}^2 \mathcal{L}_{\beta_k}(x_k, \lambda_k)(x_{k+1} - x_k)\rangle \\
&\quad + \frac{L_H^f + L_k^H}{6}\|x_{k+1} - x_k\|^3 \\
&\leq \mathcal{L}_{\beta_k}(x_k, \lambda_k) + \langle g_k, x_{k+1} - x_k\rangle - \langle g_k - \nabla_x \mathcal{L}_{\beta_k}(x_k, \lambda_k), x_{k+1} - x_k\rangle + \frac{1}{2}\langle x_{k+1} - x_k, H_k(x_{k+1} - x_k)\rangle \\
&\quad + \frac{1}{2}\langle x_{k+1} - x_k, (\nabla_{xx}^2 \mathcal{L}_{\beta_k}(x_k, \lambda_k) - H_k)(x_{k+1} - x_k)\rangle + \frac{L_H^f + L_k^H}{6}\|x_{k+1} - x_k\|^3 \\
&\leq \mathcal{L}_{\beta_k}(x_k, \lambda_k) - \frac{1}{6}(\sigma_k - (L_H^f + L_k^H))\|x_{k+1} - x_k\|^3 - \langle g_k - \nabla_x \mathcal{L}_{\beta_k}(x_k, \lambda_k), x_{k+1} - x_k\rangle \\
&\quad + \frac{1}{2}\langle x_{k+1} - x_k, (\nabla_{xx}^2 \mathcal{L}_{\beta_k}(x_k, \lambda_k) - H_k)(x_{k+1} - x_k)\rangle + q_k(x_{k+1} - x_k) \\
&\leq \mathcal{L}_{\beta_k}(x_k, \lambda_k) - \frac{1}{6}(\sigma_k - (L_H^f + L_k^H))\|x_{k+1} - x_k\|^3 - \langle g_k^0 - \nabla f(x_k), x_{k+1} - x_k\rangle \\
&\quad + \frac{1}{2}\langle x_{k+1} - x_k, (\nabla^2 f(x_k) - H_k^0)(x_{k+1} - x_k)\rangle + \frac{1}{18}\sigma_k \omega. \qquad (3.9)
\end{aligned}$$

Note that by Condition B and $a^2 b \leq a^3 + b^3$ for $a, b \geq 0$,

$$\begin{aligned}
|\langle g_k^0 - \nabla f(x_k), x_{k+1} - x_k\rangle| &\leq \|g_k^0 - \nabla f(x_k)\|\|x_{k+1} - x_k\| \\
&\leq \theta\beta_k(\max\{\|x_k - x_{k-1}\|, \hat{\epsilon}\})^2\|x_{k+1} - x_k\| \\
&\leq \theta\beta_k(\max\{\|x_k - x_{k-1}\|, \hat{\epsilon}\})^3 + \theta\beta_k\|x_{k+1} - x_k\|^3 \\
&\leq \theta\beta_k\|x_k - x_{k-1}\|^3 + \theta\beta_k\|x_{k+1} - x_k\|^3 + \theta\beta_k\hat{\epsilon}^3 \\
&\leq 2\theta\beta_k \max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} + \theta\beta_k\hat{\epsilon}^3 \qquad (3.10)
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{2}|\langle x_{k+1} - x_k, (\nabla^2 f(x_k) - H_k^0)(x_{k+1} - x_k)\rangle| &\leq \frac{1}{2}\|x_{k+1} - x_k\|^2\|\nabla^2 f(x_k) - H_k^0\| \\
&\leq \frac{1}{2}\theta\beta_k\|x_{k+1} - x_k\|^2 \max\{\|x_k - x_{k-1}\|, \hat{\epsilon}\} \\
&\leq \frac{1}{2}\theta\beta_k\|x_k - x_{k-1}\|^3 + \frac{1}{2}\theta\beta_k\|x_{k+1} - x_k\|^3 + \frac{1}{2}\theta\beta_k\hat{\epsilon}^3 \\
&\leq \theta\beta_k \max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} + \frac{1}{2}\theta\beta_k\hat{\epsilon}^3. \qquad (3.11)
\end{aligned}$$

With a slight abuse of notations, we define $\bar{\sigma}_k := \frac{1}{6}\sigma_k - \frac{1}{6}(L_H^f + L_k^H)$ and $\bar{\sigma}_0 := 0$. Then plugging (3.10) and (3.11) into (3.9) indicates that

$$\begin{aligned}
\mathcal{L}_{\beta_k}(x_{k+1}, \lambda_k) \leq &\mathcal{L}_{\beta_k}(x_k, \lambda_k) - \bar{\sigma}_k\|x_{k+1} - x_k\|^3 + 3\theta\beta_k \max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} \\
&+ \frac{3}{2}\theta\beta_k\hat{\epsilon}^3 + \frac{1}{18}\sigma_k\omega. \qquad (3.12)
\end{aligned}$$

Note that

$$\begin{aligned}
\mathcal{L}_{\beta_k}(x_{k+1}, \lambda_k) &= f(x_{k+1}) + \langle \lambda_k, c(x_{k+1})\rangle + \frac{\beta_k}{2}\|c(x_{k+1})\|^2 \\
&= f(x_{k+1}) + \langle \lambda_{k+1}, c(x_{k+1})\rangle + \frac{\beta_k}{2}\|c(x_{k+1})\|^2 - \langle \lambda_{k+1} - \lambda_k, c(x_{k+1})\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{L}_{\beta_{k+1}}(x_{k+1}, \lambda_{k+1}) - \langle \lambda_{k+1} - \lambda_k, c(x_{k+1})\rangle - \frac{1}{2}(\beta_{k+1} - \beta_k)\|c(x_{k+1}\|^2 \\
&\geq \mathcal{L}_{\beta_{k+1}}(x_{k+1}, \lambda_{k+1}) - C\|\lambda_{k+1} - \lambda_k\|_1 - \frac{1}{2}C^2(\beta_{k+1} - \beta_k), \qquad (3.13)
\end{aligned}
$$

where the last inequality is due to Assumption 3. Therefore, combining (3.12) and (3.13) yields

$$
\begin{aligned}
\mathcal{L}_{\beta_{k+1}}(x_{k+1}, \lambda_{k+1}) \leq \mathcal{L}_{\beta_k}(x_k, \lambda_k) - \bar{\sigma}_k\|x_{k+1} - x_k\|^3 + 3\theta\beta_k \max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} + \frac{3}{2}\theta\beta_k\hat{\epsilon}^3 \\
+ \frac{1}{18}\sigma_k\omega + C\|\lambda_{k+1} - \lambda_k\|_1 + \frac{1}{2}C^2(\beta_{k+1} - \beta_k).
\end{aligned}
$$

Summing up the above inequality over $k = 1, \ldots, K-1$ with terms rearranged and by

$$
\omega = \bar{\epsilon}^3 \quad \text{and} \quad \hat{\epsilon} \leq \bar{\epsilon} \leq \max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\},
$$

we obtain

$$
\begin{aligned}
\mathcal{L}_{\beta_K}(x_K, \lambda_K) \leq \mathcal{L}_{\beta_1}(x_1, \lambda_1) - \sum_{k=1}^{K-1} \bar{\sigma}_k\|x_{k+1} - x_k\|^3 + \sum_{k=1}^{K-1}(\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k)\max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} \\
+ C\sum_{k=1}^{K-1}\|\lambda_{k+1} - \lambda_k\|_1 + \frac{1}{2}C^2\sum_{k=1}^{K-1}(\beta_{k+1} - \beta_k). \qquad (3.14)
\end{aligned}
$$

Note that it implies from $\bar{\sigma}_0 = 0$ that

$$
\begin{aligned}
\sum_{k=1}^{K-1} \bar{\sigma}_k\|x_{k+1} - x_k\|^3 &= \frac{1}{2}(\sum_{k=1}^{K-1} \bar{\sigma}_k\|x_{k+1} - x_k\|^3 + \sum_{k=1}^{K-1} \bar{\sigma}_k\|x_{k+1} - x_k\|^3) \\
&= \frac{1}{2}(\sum_{k=1}^{K-1} \bar{\sigma}_k\|x_{k+1} - x_k\|^3 + \sum_{k=2}^{K} \bar{\sigma}_{k-1}\|x_k - x_{k-1}\|^3) \\
&= \frac{1}{2}(\sum_{k=1}^{K-1} \bar{\sigma}_k\|x_{k+1} - x_k\|^3 + \sum_{k=1}^{K} \bar{\sigma}_{k-1}\|x_k - x_{k-1}\|^3) \\
&\geq \frac{1}{2}\sum_{k=1}^{K-1}(\bar{\sigma}_k\|x_{k+1} - x_k\|^3 + \bar{\sigma}_{k-1}\|x_k - x_{k-1}\|^3) \\
&\geq \frac{1}{2}\sum_{k=1}^{K-1}\min\{\bar{\sigma}_k, \bar{\sigma}_{k-1}\}\max\{\|x_{k+1} - x_k\|^3, \|x_k - x_{k-1}\|^3\}.
\end{aligned}
$$

Then it together with (3.14) yields that

$$
\begin{aligned}
\mathcal{L}_{\beta_K}(x_K, \lambda_K) \leq \mathcal{L}_{\beta_1}(x_1, \lambda_1) - \sum_{k=1}^{K-1}(\frac{1}{2}\min\{\bar{\sigma}_k, \bar{\sigma}_{k-1}\} - (\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k))\max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} \\
+ \sqrt{m}C^2\sum_{k=1}^{K-1}\rho_k + \frac{1}{2}C^2\sum_{k=1}^{K-1}(\beta_{k+1} - \beta_k). \qquad (3.15)
\end{aligned}
$$

On the one hand, by $\bar{\sigma}_k = \frac{1}{6}\sigma_k - \frac{1}{6}(L_H^f + L_k^H)$, it holds that

$$
\frac{1}{2}\bar{\sigma}_k - (\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k) = \frac{1}{36}\sigma_k - \frac{1}{12}(L_H^f + \sqrt{m}CL_H^c\sum_{t=1}^{k-1}\rho_t + \beta_k(\sqrt{m}CL_H^c + 3mL_cL_g^c + 54\theta)), \qquad (3.16)
$$

while on the other hand,

$$\begin{aligned}
&\frac{1}{2}\bar{\sigma}_{k-1} - (\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k) \\
&= \frac{1}{12}\sigma_{k-1} - \frac{1}{12}(L_H^f + \sqrt{m}CL_H^c \sum_{t=1}^{k-2}\rho_t + \beta_{k-1}(\sqrt{m}CL_H^c + 3mL_cL_g^c)) - (\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k) \\
&\geq \frac{1}{12}\sigma_{k-1} - \frac{1}{12}(L_H^f + \sqrt{m}CL_H^c \sum_{t=1}^{k-1}\rho_t + \beta_k(\sqrt{m}CL_H^c + 3mL_cL_g^c)) - (\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k) \\
&\geq \mu\sigma_k - \frac{1}{12}(L_H^f + \sqrt{m}CL_H^c \sum_{t=1}^{k-1}\rho_t + \beta_k(\sqrt{m}CL_H^c + 3mL_cL_g^c + 54\theta)) \qquad (3.17)
\end{aligned}$$

due to (3.6). Thus it yields from the setting of $\hat{\sigma}_k$ that

$$\frac{1}{2}\min\{\bar{\sigma}_k, \bar{\sigma}_{k-1}\} - (\frac{9}{2}\theta\beta_k + \frac{1}{18}\sigma_k) \geq \hat{\sigma}_k.$$

Moreover, Lemma 3.1 indicates

$$\begin{aligned}
\mathcal{L}_{\beta_K}(x_K, \lambda_K) = f(x_K) + \langle \lambda_K, c(x_K)\rangle + \frac{\beta_K}{2}\|c(x_K)\|^2 &\geq f(x_K) + \langle \lambda_K, c(x_K)\rangle \\
&\geq f(x_K) - C\|\lambda_K\|_1 \\
&\geq f_{low} - \sqrt{m}C^2 \sum_{k=1}^{K-1}\rho_k \qquad (3.18)
\end{aligned}$$

which together with (3.15) gives (3.7). □

Next we will establish the finite termination of ICPD under certain conditions, and then characterize properties of the algorithm output. Before proceeding, we need another assumption ensuring the nonsingularity of constraints.

Assumption 4. *There exists a positive constant ν such that for any $x \in \mathcal{X}$,*

$$\nu\|c(x)\| \leq \|\nabla c(x)c(x)\|. \qquad (3.19)$$

Remark 3.3. *Assumption 4 holds naturally at feasible points. Finding a feasible solution for nonconvex constrained optimization can be challenging in general. To address this issue, it is common to impose a constraint qualification condition or a nonsingularity condition on the constraint functions. In the context of infeasible methods, specifically stochastic approximation methods for nonconvex constrained optimization, a constraint qualification or nonsingularity condition is often imposed on infeasible iterates. This condition helps analyze the complexity of algorithms, and its necessity is evident in various works, including references such as [2, 17, 29, 30, 42]. In our work, Assumption 4 plays a crucial role in ensuring finding an approximately feasible point, which, in turn, promotes approximate second-order stationarity. This assumption is closely related to the strong linear independence constraint qualification (LICQ) assumed in works such as [2] and [17]. Strong LICQ requires that singular values of the Jacobian matrix of constraints over the region containing all iterates and trial points be lower bounded away from zero. It is evident that this condition can imply Assumption 4.*

We next assume that

$$\lim_{K\to+\infty} \frac{\sum_{k=1}^{K-1}\alpha_k}{\alpha_K} = +\infty, \quad \text{where } \alpha_k := \sum_{t=1}^{k-1}\rho_t + \beta_k. \qquad (3.20)$$

Note that as $\alpha_k \geq \beta_k$ and $\beta_k \geq \beta_{k-1}$ for any $k \geq 1$, it implies $\lim_{K\to+\infty}\sum_{k=1}^{K-1}\alpha_k = +\infty$. Note that (3.20) holds if, for instance, $\beta_k = k^\tau$ with $\tau \in (0, +\infty)$ and $\rho_k = k^{-\iota}$ with $\iota \in (1, +\infty)$.

THEOREM **3.1.** *Under Assumptions 1-4, Conditions A and B, suppose that* $\omega = \bar{\epsilon}^3$, (3.6) *and* (3.20) *hold, and*

$$\sigma_k = \frac{1}{6\mu}(L_H^f + \sqrt{m}CL_H^c\alpha_k + \beta_k(3mL_cL_g^c + 54\theta)), \quad k \geq 1. \tag{3.21}$$

Then ICPD terminates finitely. Moreover, the output, if denoted by x_{K+1}, *satisfies*

$$\begin{cases} \|\nabla f(x_{K+1}) + \nabla c(x_{K+1})\hat{\lambda}_{K+1}\| = \mathcal{O}(\alpha_K\bar{\epsilon}^2 + \rho_K), \\ \|c(x_{K+1})\| = \mathcal{O}(\bar{\epsilon}^2 + {\beta_K}^{-1}(\sum_{k=1}^K \rho_k + 1)), \\ d^T(\nabla^2 f(x_{K+1}) + \sum_{i=1}^m \hat{\lambda}_i\nabla^2 c_i(x_{K+1}))d \geq -\upsilon_K\|d\|^2 \ \ \textit{for all } d \in \mathrm{Null}(\nabla c(x_{K+1})^T), \end{cases} \tag{3.22}$$

where $\hat{\lambda}_{K+1} := \lambda_{K+1} + \beta_K c(x_{K+1})$ *and* $\upsilon_K = \mathcal{O}(\alpha_K\bar{\epsilon} + \rho_K\bar{\epsilon}^2 + \rho_K{\beta_K}^{-1}(\sum_{t=1}^{K-1}\rho_t + 1))$.

Proof. Assume that ICPD does not terminate before Kth iteration with $K > 1$. Then it implies

$$\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} \geq \bar{\epsilon} \quad \text{for } k = 1, \ldots, K-1,$$

thus yields from (3.7) that

$$\sum_{k=1}^{K-1}\hat{\sigma}_k\bar{\epsilon}^3 \leq \mathcal{L}_{\beta_1}(x_1, \lambda_1) - f_{low} + 2\sqrt{m}C^2\sum_{k=1}^{K-1}\rho_k + \frac{1}{2}C^2(\beta_K - \beta_1). \tag{3.23}$$

By the setting of σ_k and $\hat{\sigma}_k$ in Lemma 3.3 as well as L_k^H in Lemma 3.2, it is easy to have

$$\hat{\sigma}_k = \frac{1}{12}(L_H^f + \sqrt{m}CL_H^c\alpha_k + \beta_k(3mL_cL_g^c + 54\theta)), \tag{3.24}$$

which together with (3.20) indicates that K must be finite, thus proves the finite termination of ICPD.

With a slight abuse of notation, we still use K as the iteration number when ICPD terminates with the output x_{K+1}. With $\hat{\lambda} = \lambda_{K+1} + \beta_K c(x_{K+1})$, it follows from (2.7c) that

$$\begin{aligned}
&\|\nabla f(x_{K+1}) + \nabla c(x_{K+1})\hat{\lambda}_{K+1}\| \\
&= \|\nabla f(x_{K+1}) + \nabla_x\Psi_{\beta_K}(x_{K+1}, \lambda_{K+1})\| \\
&\leq \|\nabla f(x_{K+1}) + \nabla_x\Psi_{\beta_K}(x_{K+1}, \lambda_{K+1}) - g_K - H_K(x_{K+1} - x_K) - \frac{\sigma_K}{2}\|x_{K+1} - x_K\|(x_{K+1} - x_K)\| + \sigma_K\omega^{2/3} \\
&\leq \|\nabla f(x_{K+1}) - \nabla f(x_K) - \nabla^2 f(x_K)(x_{K+1} - x_K)\| + \|\nabla f(x_K) - g_K^0\| \\
&\quad + \|\nabla_x\Psi_{\beta_K}(x_{K+1}, \lambda_K) - \nabla_x\Psi_{\beta_K}(x_K, \lambda_K) - \nabla_{xx}^2\Psi_{\beta_K}(x_K, \lambda_K)(x_{K+1} - x_K)\| + \|(\nabla^2 f(x_K) - H_K^0)(x_{K+1} - x_K)\| \\
&\quad + \|\nabla_x\Psi_{\beta_K}(x_{K+1}, \lambda_{K+1}) - \nabla_x\Psi_{\beta_K}(x_{K+1}, \lambda_K)\| + \frac{\sigma_K}{2}\|x_{K+1} - x_K\|^2 + \sigma_K\bar{\epsilon}^2 \\
&\leq \frac{L_H^f}{2}\|x_{K+1} - x_K\|^2 + \theta\beta_K\max\{\|x_K - x_{K-1}\|^2, \bar{\epsilon}^2\} + \frac{L_K^H}{2}\|x_{K+1} - x_K\|^2 \\
&\quad + \theta\beta_K\max\{\|x_K - x_{K-1}\|, \bar{\epsilon}\}\|x_{K+1} - x_K\| + L_c\|\lambda_{K+1} - \lambda_K\|_1 + \frac{\sigma_K}{2}\|x_{K+1} - x_K\|^2 + \sigma_K\bar{\epsilon}^2 \\
&< \frac{1}{2}(L_H^f + L_K^H + 4\theta\beta_K + 3\sigma_K)\bar{\epsilon}^2 + L_c\rho_K\|c(x_{K+1})\|_1 \qquad (3.25) \\
&= \mathcal{O}(\alpha_K\bar{\epsilon}^2 + \rho_K), \qquad (3.26)
\end{aligned}$$

where the second inequality is indicated by

$$\begin{aligned}
\|\nabla_x\Psi_{\beta_k}(x_{K+1}, \lambda_{K+1}) - \nabla_x\Psi_{\beta_k}(x_{K+1}, \lambda_K)\| &= \|\sum_{i=1}^m(\lambda_{K+1} - \lambda_K)_i\nabla c_i(x_{K+1})\| \\
&\leq \|\sum_{i=1}^m|(\lambda_{K+1} - \lambda_K)_i|\|\nabla c_i(x_{K+1})\|
\end{aligned}$$

$$\leq L_c\|\lambda_{K+1}-\lambda_K\|_1,$$

and the last inequality is due to Lemma 3.1 and the termination condition of ICPD that $\max\{\|x_K - x_{K-1}\|, \|x_{K+1}-x_K\|\} < \bar{\epsilon}$.

Regarding the feasibility of x_{K+1}, we can derive from Assumption 4 and $\bar{\epsilon}\in(0,1)$ that

$$\begin{aligned}
\|c(x_{K+1})\| &\leq \frac{1}{\nu\beta_K}\|\beta_K\nabla c(x_{K+1})c(x_{K+1})\| \\
&\leq \frac{1}{\nu\beta_K}\|\nabla f(x_{K+1})+\nabla c(x_{K+1})(\beta_K c(x_{K+1})+\lambda_{K+1})-\nabla f(x_{K+1})-\nabla c(x_{K+1})\lambda_{K+1}\| \\
&\leq \frac{1}{\nu\beta_K}\|\nabla f(x_{K+1})+\nabla_x\Psi_{\beta_K}(x_{K+1},\lambda_{K+1})\|+\frac{1}{\nu\beta_K}\|\nabla f(x_{K+1})\|+\frac{1}{\nu\beta_K}\|\nabla c(x_{K+1})\lambda_{K+1}\| \\
&\leq \frac{1}{\nu\beta_K}\|\nabla f(x_{K+1})+\nabla_x\Psi_{\beta_K}(x_{K+1},\lambda_{K+1})\|+\frac{1}{\nu\beta_K}(L_f+L_c\|\lambda_{K+1}\|_1) \\
&\leq \frac{1}{\nu\beta_K}\|\nabla f(x_{K+1})+\nabla_x\Psi_{\beta_K}(x_{K+1},\lambda_{K+1})\|+\frac{1}{\nu\beta_K}(L_f+mL_cC\sum_{k=1}^{K}\rho_k) \\
&= \mathcal{O}(\bar{\epsilon}^2+{\beta_K}^{-1}(\sum_{k=1}^{K}\rho_k+1)). \qquad (3.27)
\end{aligned}$$

Note that it follows from (2.5) and (2.7a) that

$$H_K+\frac{\sigma_K}{2}\|x_{K+1}-x_K\|\mathbf{I}_n \succeq H_K+\frac{\sigma_K}{2}\|s_K^*\|\mathbf{I}_n-\frac{\sigma_K}{2}(\|s_K^*\|-\|s_K\|)\mathbf{I}_n \succeq -\frac{\sigma_K}{2}\omega^{1/3}\mathbf{I}_n,$$

which indicates from $\omega^{1/3}=\bar{\epsilon}\geq\max\{\|x_{K+1}-x_K\|,\|x_K-x_{K-1}\|\}$ that

$$\begin{aligned}
\nabla^2 f(x_{K+1})+\sum_{i=1}^{m}\hat{\lambda}_i\nabla^2 c_i(x_{K+1}) &= \nabla^2 f(x_{K+1})+\sum_{i=1}^{m}((\lambda_{K+1})_i+\beta_K c_i(x_{K+1}))\nabla^2 c_i(x_{K+1}) \\
&\succeq \nabla^2 f(x_{K+1})+\nabla^2_{xx}\Psi_{\beta_K}(x_{K+1},\lambda_{K+1})-\beta_K\sum_{i=1}^{m}\nabla c_i(x_{K+1})\nabla c_i(x_{K+1})^T \\
&\quad -H_K^0-\nabla^2_{xx}\Psi_{\beta_K}(x_K,\lambda_K)-\frac{\sigma_K}{2}\|x_{K+1}-x_K\|\mathbf{I}_n-\frac{\sigma_K}{2}\omega^{1/3}\mathbf{I}_n \\
&\succeq (\nabla^2 f(x_{K+1})-H_K^0)+(\nabla^2_{xx}\Psi_{\beta_K}(x_{K+1},\lambda_{K+1})-\nabla^2_{xx}\Psi_{\beta_K}(x_K,\lambda_K)) \\
&\quad -\beta_K\sum_{i=1}^{m}\nabla c_i(x_{K+1})\nabla c_i(x_{K+1})^T-\sigma_K\bar{\epsilon}\mathbf{I}_n.
\end{aligned}$$

To estimate above lower bound, on the one hand, Assumption 1 together with Condition B implies

$$\begin{aligned}
\nabla^2 f(x_{K+1})-H_K^0 &\succeq -\|\nabla^2 f(x_{K+1})-\nabla^2 f(x_K)\|\mathbf{I}_n-\|\nabla^2 f(x_K)-H_K^0\|\mathbf{I}_n \\
&\succeq -L_H^f\|x_{K+1}-x_K\|\mathbf{I}_n-\|\nabla^2 f(x_K)-H_K^0\|\mathbf{I}_n \\
&\succeq -(L_H^f+\theta\beta_K)\bar{\epsilon}\mathbf{I}_n.
\end{aligned}$$

On the other hand, (3.2) implies

$$\begin{aligned}
&\nabla^2_{xx}\Psi_{\beta_k}(x_{K+1},\lambda_{K+1})-\nabla^2_{xx}\Psi_{\beta_k}(x_K,\lambda_K) \\
&=\nabla^2_{xx}\Psi_{\beta_k}(x_{K+1},\lambda_{K+1})-\nabla^2_{xx}\Psi_{\beta_k}(x_{K+1},\lambda_K)+\nabla^2_{xx}\Psi_{\beta_k}(x_{K+1},\lambda_K)-\nabla^2_{xx}\Psi_{\beta_k}(x_K,\lambda_K) \\
&=\sum_{i=1}^{m}((\lambda_{K+1})_i-(\lambda_K)_i)\nabla^2 c_i(x_{K+1})+\nabla^2\Psi_{\beta_k}(x_{K+1},\lambda_K)-\nabla^2\Psi_{\beta_k}(x_K,\lambda_K) \\
&\succeq -L_g^c\|\lambda_{K+1}-\lambda_K\|_1\mathbf{I}_n-L_K^H\|x_{K+1}-x_K\|\mathbf{I}_n.
\end{aligned}$$

Then the following relation holds:

$$\nabla^2 f(x_{K+1}) + \sum_{i=1}^{m} (\hat{\lambda}_{K+1})_i \nabla^2 c_i(x_{K+1}) \succeq -(L_H^f + \theta\beta_K + L_K^H + \sigma_K)\bar{\epsilon}\mathbf{I}_n - L_g^c \|\lambda_{K+1} - \lambda_K\|_1 \mathbf{I}_n - \beta_K \nabla c_i(x_{K+1}) \nabla c_i(x_{K+1})^T. \tag{3.28}$$

Thus for any $d \in \text{Null}(\nabla c(x_{K+1})^T)$, by $\|\lambda_{K+1} - \lambda_K\| = \rho_K \|c(x_{K+1})\|$, settings on σ_K, ρ_K and $\|x_{K+1} - x_K\| \leq \bar{\epsilon}$, we have

$$d^T(\nabla^2 f(x_{K+1}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla^2 c_i(x_{K+1}))d \geq -\upsilon_K \|d\|^2,$$

where $\upsilon_K = (L_H^f + \theta\beta_K + L_K^H + \sigma_K)\bar{\epsilon} + L_g^c \|\lambda_{K+1} - \lambda_K\|_1$. It is easy to derive from (3.27) that

$$\upsilon_K = \mathcal{O}(\alpha_K \bar{\epsilon} + \rho_K \bar{\epsilon}^2 + \rho_K {\beta_K}^{-1} (\sum_{t=1}^{K} \rho_t + 1)).$$

The proof is completed. □

Remark 3.4. *As indicated in* (3.27)*, Assumption 4 plays a crucial role in ensuring the near feasibility of x_{K+1}. Without this assumption, following the analysis to* (3.27) *and* (3.28) *and applying Assumption 3 we can obtain*

$$\|\nabla c(x)c(x)\| = \mathcal{O}(\bar{\epsilon}^2 + \beta_K^{-1}(\sum_{k=1}^{K} \rho_k + 1))$$

and

$$d^T(\nabla^2 f(x_{K+1}) + \sum_{i=1}^{m} (\hat{\lambda}_{K+1})_i \nabla^2 c_i(x_{K+1}))d \geq -\bar{\upsilon}_K \|d\|^2, \text{ for all } d \in \text{Null}(\nabla c(x_{K+1})^T),$$

where $\bar{\upsilon}_K = (L_H^f + \theta\beta_K + L_K^H + \sigma_K)\bar{\epsilon} + L_g^c C \rho_K$.

We next present the iteration complexity of ICPD to find an ϵ-FSP and an ϵ-SSP of (1.1), respectively, under a non-adaptive presetting of penalty parameters β_k.

THEOREM **3.2.** *Under Assumptions 1-4, Conditions A and B, suppose that $\omega = \bar{\epsilon}^3$, $\hat{\epsilon} \leq \bar{\epsilon}$,* (3.6)*,* (3.20)*,* (3.21)*. The following statements hold true.*

(i) The iteration complexity of ICPD to find an ϵ-FSP is in order $\mathcal{O}(\epsilon^{-3})$, if

$$\bar{\epsilon} = \epsilon,\ \beta_k = k^\tau,\ \tau = \frac{1}{3}, \text{ and } \rho_k = k^{-\iota},\ \iota \in (1, +\infty). \tag{3.29}$$

(ii) The iteration complexity of ICPD to find an ϵ-SSP is in order $\mathcal{O}(\epsilon^{-4.5})$, if

$$\bar{\epsilon} = \epsilon^{1.5},\ \beta_k = k^\tau,\ \tau = \frac{2}{9}, \text{ and } \rho_k = k^{-\iota},\ \iota \in (1, +\infty). \tag{3.30}$$

Proof. Let K be the index of iteration when ICPD terminates. By Theorem 3.1 we know that K is finite. Under setting (3.29), it holds that

$$\alpha_k = \sum_{t=1}^{k-1} \rho_t + \beta_k \geq k^\tau$$

which derives from (3.23) that

$$\frac{\sum_{k=1}^{K-1} k^\tau \bar{\epsilon}^3}{K^\tau} = \mathcal{O}(1), \tag{3.31}$$

thus there exists a positive constant $\hat{C} = \mathcal{O}(1)$ such that $K \leq \hat{C}\bar{\epsilon}^{-3}$. Consequently, under parameter setting (3.29), K is in order $\mathcal{O}(\epsilon^{-3})$, then it is straightforward to obtain conclusions from Theorem 3.1. For (ii), the conclusion can be derived through similar analysis. □

Remark 3.5.

(a) *The paper [49] presents an analysis of Proximal AL method designed for equality constrained optimization. This work characterizes the total iteration complexity of the proposed algorithm, in terms of total number of iterations of Newton-CG. Since at each iteration of Newton-CG, it needs to compute the gradient and Hessian once, the total iteration complexity in [49] aligns with the oracle complexity in our case, counting the number of gradient and Hessian evaluations. It is shown in [49, Theorem 4] that the total iteration complexity is in order* $\mathcal{O}(\epsilon^{-5.5})$ *to find* ϵ*-FSP satisfying* (1.7)*. And the oracle complexity is in order* $\mathcal{O}(\epsilon^{-7})$ *to find an approximate second-order stationary point with high probability satisfying* (1.7) *and*

$$d^T(\nabla^2 f(x) + \sum_{i=1}^{m} \lambda_i \nabla^2 c_i(x))d \geq -\epsilon\|d\|^2 \text{ for any } d \in \text{Null}(\nabla c(x)^T). \tag{3.32}$$

By Theorem 3.2 the oracle complexity of ICPD to find an ϵ*-FSP is* $\mathcal{O}(\epsilon^{-3})$*. To reach an approximate second-order stationary point satisfying* (1.7) *and* (3.32)*, we set*

$$\bar{\epsilon} = \epsilon^2, \quad \beta_k = k^\tau, \tau = \frac{1}{6}, \text{ and } \rho_k = k^{-\iota}, \iota \in (1, +\infty).$$

Then the corresponding oracle complexity of ICPD is in order $\mathcal{O}(\epsilon^{-6})$*. Meanwhile, as ICPD relies on inexact derivatives of the objective, we can easily extend it to solve problems in stochastic settings, which however is not straightforward for Proximal AL.*

(b) *A deterministic SQP method is studied for* (1.1) *in [17]. It is assumed in [17, Assumption 1] that* $\nabla f, c, \nabla c^T$ *are bounded,* $\nabla f, \nabla c^T$ *are Lipschitz continuous, and a strong LICQ holds, i.e., singular values of* ∇c^T *are uniformly lower bounded away from zero for all* k*. For Hessian approximations* $\{B_k\}$*, it only requires their boundedness in norm and* $\{\frac{u^T B_k u}{\|u\|^2} : u \neq 0, J_k u = 0\}$ *are uniformly lower bounded from zero. It is shown in [17] that the iteration complexity of deterministic SQP is in order* $\mathcal{O}(\epsilon^{-2})$ *to find a point satisfying*

$$\|\nabla f(x) + \nabla c(x)\lambda\| \leq \epsilon \text{ and } \sqrt{\|c(x)\|_1} \leq \epsilon. \tag{3.33}$$

Since many trivial approximations, including the identity matrix, satisfy the required assumption on Hessian approximations. Thus we consider the oracle complexity of deterministic SQP only regarding gradient approximation evaluations. Moreover, as at each iteration only a single gradient is computed, the iteration complexity of deterministic SQP is same as the oracle complexity regarding the gradient evaluations.

(c) *Another work closely related to ours is [13], where the two-phase Short-Step ARC algorithm is proposed for* $\min_{x \in X} f(x)$ *subject to* $c(x) = 0$*. It is based on a CR method that applies for (convex set constrained) least-square problems. Under the assumptions that the closed convex hull of all iterates and trial iterates is bounded, the Hessian is weakly uniformly Lipschitz continuous and is well approximated by a Hessian approximation (see AS2-AS3 in [13]), to find an approximate first-order critical point of accuracy* ϵ*, satisfying* (1.3)*, for general constrained optimization, Short-Step ARC algorithm owns the oracle complexity in order* $\mathcal{O}(\epsilon^{-1.5})$ *regarding problem-functions evaluations to find an approximate first-order stationary point satisfying* (1.4)*. But this algorithm requires a convexly constrained high-order Taylor model be solved exactly at each iteration.*

(d) *In [10] a two-phase minimization algorithm,* OUTER*, is studied for the same problem as [13], with an inner algorithm called to solve a convex set constrained least-square problem at each iteration. The related evaluation complexity is analyzed to achieve higher-order critical points and dependent on the choice of the inner algorithm that arrives at* ϵ*-approximate* q*th order critical point using the regularized Taylor series of degree* p*. Under a constraint qualification (Assumption AS.0) that*

ensures the existence of a feasible path, it achieves $\mathcal{O}(\epsilon_p^{-1}, \epsilon_p^{1-\pi}\epsilon_D^{-\pi})$ *evaluations of* f, c *and their derivatives up to order* p *to find* x_ϵ *and* $t_\epsilon \le f(x_\epsilon)$ *such that*

$$\begin{cases} \|c(x_\epsilon)\| > \delta\epsilon_p, \text{ and } \phi_{\nu,j}^{\Delta_k}(x_\epsilon) \le \epsilon_D \Delta_k^j \|c(x_\epsilon)\| \text{ for } j \in [q], & \text{if } t_\epsilon = f(x_\epsilon), \\ \|c(x_\epsilon)\| \le \delta\epsilon_p, \text{ and } \phi_{\mu,j}^{\Delta_k}(x_\epsilon, t_\epsilon) \le \epsilon_D \Delta_k^j \|r(x_\epsilon, t_\epsilon)\| \text{ for } j \in [q], & \text{if } t_\epsilon < f(x_\epsilon). \end{cases} \tag{3.34}$$

Here, $\nu(x) := \frac{1}{2}\|c(x)\|^2$, $\mu(x,t) := \frac{1}{2}\|r(x,t)\|^2$ *with* $r(x,t) := (c(x)^T, f(x)-t)^T$, *and* $\phi_{\psi,j}^{\Delta}(x) := \psi(x) - \min\{T_{\psi,j}(x,d) | x+d \in X, \|d\| \le \Delta\}$ *with the* j*th order Taylor model* $T_{\psi,j}(x,d)$. *And the parameter* $\pi \ge 1$ *indicates the lower bound of function decrease at successful iterations of inner algorithm is in* $\mathcal{O}(\epsilon^\pi)$. *Consequently, the evaluation complexity of* OUTER *to achieve approximate first- and second-order stationary point as defined by* (3.34) *is in* $\mathcal{O}(\epsilon^{-2})$ *and* $\mathcal{O}(\epsilon^{-3})$, *respectively.*

(e) We present a brief summary on aforementioned algorithms and ICPD in Table 1. Here, "1-o" refers to the first-order stationarity, while "2-o" refers to the second-order stationarity. And "StaMea" represents the stationarity measure, while "OraCom" represents the related oracle complexity. As all algorithms are deterministic, without specification the oracle complexity are in terms of the number of gradient and Hessian evaluations, if applicable. And "Stochasticity" is to represent if the original algorithm has been extended to a stochastic variant.

Algorithm	**ProbType**	**1-o**		**2-o**		**Stochasiticity**
		StaMea	**OraCom**	**StaMea**	**OraCom**	
Short-Step ARC [13]	$\min_{x\in X} f(x)$ s.t. $c(x)=0$	(1.4)	$\mathcal{O}(\epsilon^{-1.5})$	-	-	-
Deterministic SQP [17]	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	(3.33)	$\mathcal{O}(\epsilon^{-2})$	-	-	✓
OUTER [10]	$\min_{x\in X} f(x)$ s.t. $c(x)=0$	(3.34) ($p=q=1$)	$\mathcal{O}(\epsilon^{-2})$	(3.34) ($p=q=2$)	$\mathcal{O}(\epsilon^{-3})$	-
Proximal AL [49]	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	(1.7)	$\mathcal{O}(\epsilon^{-5.5})$	(1.7) & (3.32)	$\mathcal{O}(\epsilon^{-7})$	-
ICPD (this paper)	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	(1.7)	$\mathcal{O}(\epsilon^{-3})$	(1.7) & (3.32)	$\mathcal{O}(\epsilon^{-6})$	✓

Table 1: A brief summary on deterministic algorithms for (1.1).

4 Stochastic variant

In this section, we consider problem (1.1) with objective function f in an expectation form, i.e.

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) := \mathbb{E}[F(x;\xi)] \\ \text{s.t.} \quad & c_i(x) = 0, \quad i = 1, \dots, m. \end{aligned} \tag{4.1}$$

Here $\xi : \Omega \to \mathbb{W}$ is a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{W}$ is a measurable space and $\mathbb{E}$ represents the expectation with respect to ξ. Additionally, $F : \mathbb{R}^n \times \mathbb{W} \to \mathbb{R}$ is twice continuously differentiable with respect to x almost surely for ξ. We assume that exact derivatives of f cannot be accessed, while only stochastic oracles of f at an inquiry point x_k can be obtained. We now consider a stochastic variant of ICPD, a stochastic cubic-regularized primal-dual algorithm (shortened as SCPD), in which mini-batch stochastic approximations g_k^0 and H_k^0 are calculated:

$$g_k^0 = \frac{1}{I_k}\sum_{i\in\mathcal{I}_k} \nabla_x F(x_k;\xi_k^i), \quad H_k^0 = \frac{1}{J_k}\sum_{j\in\mathcal{J}_k} \nabla_{xx}^2 F(x_k;\xi_k^j), \tag{4.2}$$

where $\{\xi_k^i, i \in \mathcal{I}_k\}$ and $\{\xi_k^j, j \in \mathcal{J}_k\}$ are two randomly and independently generated sample sets with $I_k = |\mathcal{I}_k|$ and $J_k = |\mathcal{J}_k|$. As can be seen in previous section, although inexact derivatives of the objective function can be allowed, certain condition i.e., Condition B should always be satisfied at each iteration. It plays a crucial rule in establishing desirable properties of the algorithm. We will see that in stochastic settings this condition can be easily satisfied with high probability through a proper sampling strategy. But first we need the following assumption.

Assumption 5. *For any x,*

- $\mathbb{E}[\nabla_x F(x;\xi)] = \nabla f(x)$ *and* $\|\nabla_x F(x;\xi) - \nabla f(x)\| \leq M_1$ *almost surely;*
- $\mathbb{E}[\nabla^2_{xx} F(x;\xi)] = \nabla^2 f(x)$ *and* $\|\nabla^2_{xx} F(x;\xi) - \nabla^2 f(x)\| \leq M_2$ *almost surely.*

Remark 4.1. *Matrix concentration inequalities are commonly used in modern random matrix theory to characterize the deviations of random matrices. In the context of bounded random matrices, Assumption 5 is often imposed to establish the matrix Bernstein inequality from [45]. By using the sub-sampling scheme and leveraging the concentration inequality one can derive the oracle complexity regarding the number of random matrices. In the study of stochastic cubic regularization for unconstrained expectation minimization problems [44], a similar assumption to Assumption 5 is also assumed. This assumption enables the application of matrix concentration inequalities for analyzing the behavior of stochastic cubic regularization methods. If instead assuming that $F(x;\xi)$ has L_F-Lipschitz continuous gradients, the boundedness on the error of Hessian estimates can be removed, as indicated in [44]. Matrix concentration inequalities have also been used to analyze the gradient and Hessian oracle complexity in the work for stochastic cubic regularzation methods including [27, 48, 54].*

The lemma below characterizes the sample size per-iteration to ensure Condition B with high probability. Detailed proof can be referred to [44, Lemma 4].

LEMMA **4.1.** *Let $\mathcal{I}_k$ and $\mathcal{J}_k$ be sampled independently and randomly, and Assumption 5 hold. Then for any given $\delta' \in (0,1)$, Condition B with $\hat{\epsilon} = \bar{\epsilon}$ is satisfied at kth iteration with probability no less than $1-\delta'$, provided that*

$$I_k \geq \mathcal{O}\left(\max\left\{\frac{M_1}{\theta\beta_k \max\{\|x_k - x_{k-1}\|^2, \bar{\epsilon}^2\}}, \frac{M_1^2}{\theta^2 {\beta_k}^2 \max\{\|x_k - x_{k-1}\|^4, \bar{\epsilon}^4\}}\right\}\log(\frac{1}{\delta'})\right),$$
$$J_k \geq \mathcal{O}\left(\max\left\{\frac{M_2}{\theta\beta_k \max\{\|x_k - x_{k-1}\|, \bar{\epsilon}\}}, \frac{M_2^2}{\theta^2 {\beta_k}^2 \max\{\|x_k - x_{k-1}\|^2, \bar{\epsilon}^2\}}\right\}\log(\frac{1}{\delta'})\right).$$

Now we are ready to give the oracle complexity of SCPD in terms of stochastic gradients and Hessian evaluations calculated through (4.2).

THEOREM **4.1.** *Under Assumptions 1-5 and Condition A, $\omega = \bar{\epsilon}^3$, $\hat{\epsilon} = \bar{\epsilon}$, (3.6), (3.20) and (3.21). For any given $\delta \in (0,1)$, the following statements hold true.*

(i) If (3.29) holds, then to find an ϵ-FSP of (4.1) with probability no less than $1-\delta$, the oracle complexity regarding stochastic gradient and Hessian evaluations is in order $\tilde{\mathcal{O}}(\epsilon^{-5})$ and $\tilde{\mathcal{O}}(\epsilon^{-3})$, respectively.

(ii) If $\bar{\epsilon} = \epsilon, \beta_k = k^{1/6}$ and $\rho_k = k^{-\iota}$ with $\iota \in (1, +\infty)$, then to find a point x satisfying (1.8) and

$$\|\nabla f(x) + \nabla c(x)\lambda\| \leq \epsilon^{1.5} \text{ and } \|c(x)\| \leq \sqrt{\epsilon} \tag{4.3}$$

with probability no less than $1-\delta$, the oracle complexity regarding stochastic gradient and Hessian evaluations is in order $\tilde{\mathcal{O}}(\epsilon^{-6})$ and $\tilde{\mathcal{O}}(\epsilon^{-4})$, respectively.

Proof. Assume that the algorithm terminates in a finite number of iterations, with maximum iteration number denoted by K. We assume that Condition B holds with probability no less than $1-\delta'$ at each iteration, where $\delta' \in (0,1)$. To guarantee that Condition B holds for all iterations with probability at

least $1-\delta$, it is sufficient to require $1-K\delta' \geq 1-\delta$. So we can simply set $\delta' = \delta/K$. Accordingly, by Lemma 4.1, to ensure Condition B hold at kth iteration, the sample size associated with stochastic gradient and Hessian evaluations at kth iteration can be set as

$$I_k = \mathcal{O}\left(\left\lceil \max\left\{\frac{1}{\theta\beta_k\bar{\epsilon}^2}, \frac{1}{\theta^2\beta_k^2\bar{\epsilon}^4}\right\}\log(\frac{K}{\delta})\right\rceil\right) \text{ and } J_k = \mathcal{O}\left(\left\lceil \max\left\{\frac{1}{\theta\beta_k\bar{\epsilon}}, \frac{1}{\theta^2\beta_k^2\bar{\epsilon}^2}\right\}\log(\frac{K}{\delta})\right\rceil\right),$$

respectively. Thus the total number of stochastic gradient and Hessian evaluations are

$$\sum_{k=1}^{K} I_k = \mathcal{O}\left(\max\left\{\sum_k \frac{1}{\beta_k\bar{\epsilon}^2}, \sum_k \frac{1}{\beta_k^2\bar{\epsilon}^4}\right\}|\log(\epsilon\delta)| + \bar{\epsilon}^{-3}\right), \tag{4.4}$$

and

$$\sum_{k=1}^{K} J_k = \mathcal{O}\left(\max\left\{\sum_k \frac{1}{\beta_k\bar{\epsilon}}, \sum_k \frac{1}{\beta_k^2\bar{\epsilon}^2}\right\}|\log(\epsilon\delta)| + \bar{\epsilon}^{-3}\right), \tag{4.5}$$

respectively, where the term $\bar{\epsilon}^{-3}$ in (4.4)-(4.5) is due to the fact that $I_k \geq 1$ and $J_k \geq 1$ for $k = 1, \ldots, K$. (i) We obtain from Corollary 3.2 that with probability no less than $1-\delta$, $K = \epsilon^{-3}$ and ICPD can find an ϵ-FSP of (1.1) under parameter setting (3.29). Note that

$$\sum_{k=1}^{K} \frac{1}{\beta_k^2} = \mathcal{O}(K^{1/3}), \text{ and } \sum_{k=1}^{K} \frac{1}{\beta_k} = \mathcal{O}(K^{2/3}) \quad \text{for } \beta_k = k^{1/3}.$$

Thus when $\bar{\epsilon} = \epsilon$, to reach an ϵ-FSP of (4.1) with probability at least $1-\delta$, the oracle complexity of SCPD in terms of stochastic gradient and Hessian evaluations is in order $\mathcal{O}(\epsilon^{-5}|\log(\epsilon\delta)|)$ and $\mathcal{O}(\epsilon^{-3}|\log(\epsilon\delta)|)$, respectively. (ii) When $\bar{\epsilon} = \epsilon$, $\beta_k = k^{1/6}$ and $\rho_k = k^{-\iota}$ with $\iota \in (1, +\infty)$, from (3.31) we obtain that with probability no less than $1-\delta$, $K = \mathcal{O}(\epsilon^{-3})$ and ICPD reaches a point satisfying (4.3) and (1.8). Then by (4.4), (4.5) and

$$\sum_{k=1}^{K} \frac{1}{\beta_k^2} = \mathcal{O}(K^{2/3}), \quad \sum_{k=1}^{K} \frac{1}{\beta_k} = \mathcal{O}(K^{5/6}), \quad \text{for } \beta_k = k^{1/6},$$

related oracle complexity is in order $\mathcal{O}(\epsilon^{-6}|\log(\epsilon\delta)|)$ and $\mathcal{O}(\epsilon^{-4}|\log(\epsilon\delta)|)$, respectively. □

Remark 4.2. *The complexity bounds presented in* (4.4) *and* (4.5) *are affected by the choice of the parameter τ, which controls the rate of increase of the penalty parameter in the algorithm.*

Remark 4.3. *We now list several closely related works that either focus on nonconvex constrained stochastic optimization or study stochastic approximation method based on cubic regularization.*

(a) Penalty methods based on first- and zeroth-order stochastic approximations are studied in [46] for solving problem (4.1)*. By assuming that the iterate sequence is bounded and under standard assumptions that stochastic gradients are unbiased estimates and have bounded stochastic variances, i.e. $\mathbb{E}_\xi[\nabla F_x(x;\xi)] = \nabla f(x)$ and $\mathbb{E}_\xi[\|\nabla F_x(x;\xi) - \nabla f(x)\|^2] \leq M_1^2$ for some M_1, the penalty method based on first-order stochastic approximation, shorted as PFSA, can achieve $\mathcal{O}(\epsilon^{-3.5})$ oracle complexity, to find an ϵ-approximate critical point of* (4.1) *satisfying $\mathbb{E}[\|\nabla f(x) + \nabla c(x)\lambda\|^2] \leq \epsilon$ and $\mathbb{E}[\theta(x)] \leq \sqrt{\epsilon}$, where $\theta(x) := \|c(x)\| - \min_{\|d\|\leq 1}\|c(x) + \nabla c(x)d\|$. If the nonsingularity condition as Assumption 4 is assumed, the oracle complexity of PFSA in terms of stochastic gradients evaluations is in order $\mathcal{O}(\epsilon^{-7})$ to reach*

$$\mathbb{E}[\|\nabla f(x) + \nabla c(x)\lambda\|] \leq \epsilon, \quad \mathbb{E}[\|c(x)\|] \leq \epsilon. \tag{4.6}$$

(b) A stochastic primal-dual algorithm (SPD) was originally proposed in [25] for nonconvex optimization with a large number of possibly nonconvex constraints. This method can be applied to solve (4.1)*, with trivial modifications to the algorihtm framework. When starting from an arbitrary initial iterate and assuming the nonsigularity condition as shown in Assumption 4, one can find an ϵ-stationary point x of* (4.1) *satisfying* (4.6) *with oracle complexity in order $\mathcal{O}(\epsilon^{-6})$, under standard assumptions on stochastic gradients.*

(c) *Recently, a momemtum-based primal-dual algorithm for nonconvex programming with general deterministic equality and inequality constraints is studied in [43]. Besides the standard assumptions on stochastic gradients and the nonsigularity condition, if further assuming the mean-squared smoothness condition, i.e. there exists* $L > 0$ *such that* $\mathbb{E}_\xi[\|\nabla_x F(x;\xi) - \nabla_x F(y;\xi)\|^2] \leq L^2\|x-y\|^2$, *MLALM can find a point satisfying* (4.6) *within* $\mathcal{O}(\epsilon^{-4})$ *stochastic gradient evaluations. Moreover, when the initial iterate is (nearly) feasible, the corresponding complexity is* $\mathcal{O}(\epsilon^{-3})$. *Comparatively, starting from an arbitrary initial point and without assuming the mean-squared smoothness condition, SCPD can achieve a complexity order of* $\tilde{\mathcal{O}}(\epsilon^{-5})$ *to obtain an* ϵ*-FSP with high probability.*

(d) *An adaptive stochastic SQP algorithm (SSQP) is studied in [17]. Under standard assumptions on stochastic gradients and the strong LICQ condition, when the threshold value of the associated penalty parameter is unknown, with probability* $1-\delta \in (0,1)$, *SSQP can find an approximate solution* x *satisfying*

$$\mathbb{E}\left[\|\nabla f(x) + \nabla c(x)\lambda\| \,|E\right] \leq \epsilon, \quad \mathbb{E}[\sqrt{\|c(x)\|_1}|E] \leq \epsilon \tag{4.7}$$

after $\tilde{\mathcal{O}}(\epsilon^{-4}\log\frac{1}{\delta})$ *iterations, where* E *is the event satisfying certain conditions and* λ *is the associated Lagrange multiplier. It is noteworthy that the theories presented in [17] remain valid when simply using the identity matrix as the Hessian approximation at each iteration.*

(e) *In addition to aforementioned works on stochastic approximation algorithms for nonconvex constrained optimization, the work [44] is also closely related to ours. In [44], a stochastic cubic regularized algorithm SCR is studied for unconstrained optimization, with complexity analysis being provided. We present in Table 2 a brief summary on SCR and stochastic approximation algorithms that can be applied to solve* (4.1). *We list the problem type the corresponding algorithm is originally designed for, where* $f(x) = \mathbb{E}[F(x;\xi)]$. *Here "standard assumptions" mean that the stochastic oracles are unbiased and have bounded variances, "almost sure boundedness" assumption refers to the one on the error of stochastic gradients and Hessians, if applicable, as presented in Assumption 5, and "mean-squared smoothness" is same as given in (c). The stationarity measure "StaMea" that defines the approximate solution sought by each algorithm is also presented. We report the oracle complexity in terms of stochastic gradient and Hessian (or Hessian-vector) evaluations, shortened as "GraOra" and "HesOra", respectively. Notation "-" means that the related algorithm does not involve any computation of Hessian or its approximation. For SCR, SSQP and SCPD the corresponding complexity is in high probability.*

Algorithm	**ProbType**	**Assumption**	**StaMea**	**GraOra**	**HesOra**
SCR [44]	$\min_{x\in\mathbb{R}^n} f(x)$	almost sure boundedness	(1.3)	$\tilde{\mathcal{O}}(\epsilon^{-3.5})$	$\tilde{\mathcal{O}}(\epsilon^{-3.5})$
PFSA [46]	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	standard assumptions nonsigularity condition	(4.6)	$\mathcal{O}(\epsilon^{-7})$	-
SPD [25]	$\min_{x\in X} f(x)+h(x)$ s.t. $c_i(x)\leq 0, i\in\mathcal{I}$	standard assumptions nonsigularity condition	(4.6)	$\mathcal{O}(\epsilon^{-6})$	-
MLALM [43]	$\min_{x\in X} f(x)+h(x)$ s.t. $c_i(x)=0, i\in\mathcal{E}$ $c_i(x)\leq 0, i\in\mathcal{I}$	standard assumptions nonsigularity condition mean-squared smoothness	(4.6)	$\mathcal{O}\left(\epsilon^{-4}\right)$	-
SSQP [17]	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	standard assumptions strong LICQ	(4.7)	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$	-
SCPD	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	almost sure boundedness nonsigularity condition	(1.7) (4.3) &(1.8)	$\tilde{\mathcal{O}}(\epsilon^{-5})$ $\tilde{\mathcal{O}}(\epsilon^{-6})$	$\tilde{\mathcal{O}}(\epsilon^{-3})$ $\tilde{\mathcal{O}}(\epsilon^{-4})$

Table 2: A brief summary on SCR [44] and stochastic approximation algorithms that can solve (4.1).

5 Subproblem solver

As can be seen in previous sections, Condition A plays a crucial role in theoretical analysis for deriving the iteration complexity of the proposed algorithm. But we did not specify how to realize this condition back then. The goal of this section is to investigate how we can make sure it is satisfied at an inexact solution of each subproblem. For brievity, in this section we will omit the subscripts in the kth subproblem by simply considering

$$\min_{d\in\mathbb{R}^n} \quad q(d) := \langle g, d\rangle + \frac{1}{2}\langle d, Hd\rangle + \frac{\sigma}{6}\|d\|^3. \tag{5.1}$$

We denote its optimal solution as s^*. Algorithms for solving the cubic regularized problem (5.1) have been studied in recent work [5] and [6]. But there is a so-called "hard case" when $g^{(1)} = 0$. Here $g^{(1)}$ refers to the first coordinate of g in the eigenbasis of H [5]. In this case it may appear that $(s^*)^{(1)} \neq 0$ but the gradient descent remains in a subspace orthogonal to the first eigenvector of H [16]. To cope with this issue a randomization scheme is proposed by slightly perturbing the vector g to $\tilde{g}$ with $\tilde{g}^{(1)} \neq 0$. Then solve the resulting problem

$$\min_{d\in\mathbb{R}^n} \quad \tilde{q}(d) := \langle \tilde{g}, d\rangle + \frac{1}{2}\langle d, Hd\rangle + \frac{\sigma}{6}\|d\|^3 \tag{5.2}$$

by means of the standard gradient descent approach. We present the approach as the cubic-subsolver in the following algorithm. Recalling that we are seeking an inexact solution satisfying Condition A, we will analyze in details how this condition can be accomplished at the output of the solver with high probability.

Algorithm 5.1 Cubic-Subsolver via Gradient descent

Input: g, η, γ, H and σ
Output: s^t

1: $s^1 = \mathbf{0}$
2: $\tilde{g} = g + \gamma\zeta$ with $\zeta \sim \text{Unif}(\mathbb{S}^{n-1})$
3: **for** $t = 1, 2, \ldots$ **do**
4: $\quad s^{t+1} = s^t - \eta(\tilde{g} + Hs^t + \frac{\sigma}{2}\|s^t\|s^t)$
5: **end for**

By the optimality of s^*, it holds that

$$g + Hs^* + \frac{\sigma}{2}\|s^*\|s^* = 0,$$

from which it implies

$$\frac{\sigma}{2}\|s^*\|^2 \leq \|g\| + \|Hs^*\| \leq \|g\| + \|H\|\|s^*\| \leq \|g\| + \frac{\|H\|^2}{\sigma} + \frac{\sigma}{4}\|s^*\|^2,$$

thus leading to

$$\|s^*\| \leq S := 2\sqrt{\frac{\|g\|}{\sigma}} + \frac{2\|H\|}{\sigma}.$$

We next define

$$\psi = \|H\|, \quad \lambda = -\lambda_{\min}(H), \quad \lambda_+ = \max\{\lambda, 0\} \tag{5.3}$$

and set

$$\gamma = \frac{\sigma\bar{\sigma}\varepsilon}{25(\psi + \sigma S)}, \quad R = \frac{\psi}{\sigma} + \sqrt{(\frac{\psi}{\sigma})^2 + \frac{2\|g\|}{\sigma}}, \quad \eta = \frac{1}{2\max\{\psi + \sigma S + 2.5\sqrt{\gamma\sigma}, 4(\psi + \frac{\sigma}{2}R)\}}, \tag{5.4}$$

where $\bar{\sigma} \in (0,1)$ and $\varepsilon > 0$. Note that $q(0) \le q(s) + \varepsilon$ for any $\varepsilon \ge (\frac{1}{2}\psi + \frac{\sigma}{2}\|s^*\|)\|s^*\|^2$ as q is $(\psi + \sigma\|s^*\|)$-smooth on the set $\{d \in \mathbb{R}^n : \|d\| \le \|s^*\|\}$. In the following, without loss of generality, we assume

$$\varepsilon \le (\frac{1}{2}\psi + \frac{\sigma}{2}S)\|s^*\|^2.$$

Then it yields from the setting of γ that

$$\gamma \le \frac{1}{48}\bar{\sigma}\sigma\|s^*\|^2. \tag{5.5}$$

From the setting of η and $s^1 = 0$, it is easy to check that Assumptions A and B in [6] are satisfied. Then as $\tilde{g}^{(1)} \neq 0$, [6, Proposition 3.3] implies

$$s^t \to \tilde{s}^*, \quad \text{where } \tilde{s}^* \text{ is the optimal solution of (5.2).} \tag{5.6}$$

And by [6, Lemma 3.1] we have $\|s^t\| \le \|\tilde{s}^*\|$ for all $t \ge 1$. The lemma below provides an upper bound on the error between $\|s^t\|^2$ and $\|s^*\|^2$.

LEMMA **5.1.** *Under parameter settings as* (5.4)*, it holds that for sufficiently large* t*,*

$$\big|\|s^t\|^2 - \|s^*\|^2\big| \le \frac{\bar{\sigma}}{6(\psi + \sigma S)}\varepsilon. \tag{5.7}$$

Proof. By the gradient perturbation $\tilde{g} = g + \gamma\zeta$, applying [5, Lemma 4.6 (iii)] we obtain

$$|\|\tilde{s}^*\|^2 - \|s^*\|^2| \le \frac{4\gamma}{\sigma} \tag{5.8}$$

which derives from (5.4) and (5.6) that (5.7) holds for sufficiently large t. □

Motivated by Lemma 5.1 together with (5.6) and $|\|s^t\| - \|s^*\|| \le \sqrt{|\|s^t\|^2 - \|s^*\|^2|}$, by setting

$$\varepsilon \le \frac{6(\psi + \sigma S)}{\bar{\sigma}}\omega^{2/3}, \tag{5.9}$$

we can obtain (2.7a) at kth iteration with $s_k := s^t$ and $s_k^* = s^*$ when t is sufficiently large. We next consider when (2.7b) can be reached at the output of the cubic solver. First, note that by (5.5) and (5.8),

$$|\|s^*\|^2 - \|\tilde{s}^*\|^2| \le \frac{4\gamma}{\sigma} \le \frac{1}{12}\bar{\sigma}\|s^*\|^2,$$

which indicates

$$\|\tilde{s}^*\| \in (\sqrt{1 - \bar{\sigma}/12}, \sqrt{1 + \bar{\sigma}/12})\|s^*\|. \tag{5.10}$$

Then it is easy to obtain

$$|\|s^*\| - \|\tilde{s}^*\|| \le \frac{4\gamma}{\sigma(\|s^*\| + \|\tilde{s}^*\|)} \le \frac{4\gamma}{\sigma(1 + \sqrt{11/12})\|s^*\|} \le \frac{\bar{\sigma}\varepsilon}{6(1 + \sqrt{11/12})\|s^*\|(\psi + \sigma S)} \le \frac{\bar{\sigma}\varepsilon}{10\sigma\|s^*\|^2}. \tag{5.11}$$

Under parameter settings (5.3)-(5.4), the following lemma can be derived.

LEMMA **5.2.** *Under parameter settings* (5.3)-(5.4)*, it holds that* $\tilde{q}(s^t) \le \tilde{q}(\tilde{s}^*) + \varepsilon$ *for all*

$$t \ge \frac{6}{\eta}\left(\log\left(1 + \frac{\lambda_+^2}{2\sigma|\tilde{g}^{(1)}|}\right) + \log\frac{(\|H\| + \sigma\|\tilde{s}^*\|)\|\tilde{s}^*\|^2}{\varepsilon}\right)\min\left\{\frac{1}{\frac{\sigma}{2}\|\tilde{s}^*\| - \lambda}, \frac{10\|\tilde{s}^*\|^2}{\varepsilon}\right\}. \tag{5.12}$$

If, in particular, $\frac{\sigma}{2}\|s^*\| - \lambda \le (1 - \frac{2\bar{\sigma}}{3})\frac{\varepsilon}{10\|s^*\|^2}$*, then* $\tilde{q}(s^t) \le \tilde{q}(\tilde{s}^*) + \varepsilon$ *for all*

$$t \ge \frac{6}{\eta}\left(\log\left(1 + \frac{\lambda_+^2}{2\sigma|\tilde{g}^{(1)}|}\right) + \log\frac{(\|H\| + \sigma\|\tilde{s}^*\|)\|\tilde{s}^*\|^2}{\varepsilon}\right)\sqrt{\frac{10\|\tilde{s}^*\|^2}{\varepsilon} \cdot \frac{1}{\text{gap} \wedge \frac{\sigma}{2}\|\tilde{s}^*\|}}, \tag{5.13}$$

where " gap*" represents the distance between* $\lambda_{\min}$ *and the first eigenvalue of* H *that is larger than* $\lambda_{\min}$*.*

Proof. By [5, Theorem 3.1], it is straightforward to obtain (5.12). As for (5.13), it can be derived by [5, Theorem 3.1] when $\frac{\sigma}{2}\|\tilde{s}^*\| - \lambda \le \frac{\varepsilon}{10\|\tilde{s}^*\|^2}$, which is implied by $\frac{\sigma}{2}\|s^*\| - \lambda \le (1 - \frac{2\bar{\sigma}}{3})\frac{\varepsilon}{10\|s^*\|^2}$, since

$$\frac{\sigma}{2}\|\tilde{s}^*\| - \lambda \le \frac{\sigma}{2}\|s^*\| - \lambda + \frac{\bar{\sigma}\varepsilon}{20\|s^*\|^2} \le \frac{\varepsilon}{10\|s^*\|^2}(1 - \frac{2\bar{\sigma}}{3} + \frac{\bar{\sigma}}{2}) \le \frac{\varepsilon}{10\|s^*\|^2}(1 + \bar{\sigma}/12)(1 - \bar{\sigma}/6) \le \frac{\varepsilon}{10\|\tilde{s}^*\|^2}$$

by (5.11). □

It follows from (5.10) that

$$\frac{10\|\tilde{s}^*\|^2}{\varepsilon} \le (1 + \bar{\sigma}) \cdot \frac{10\|s^*\|^2}{\varepsilon}, \tag{5.14}$$

$$\log \frac{(\|H\| + \sigma\|\tilde{s}^*\|)\|\tilde{s}^*\|^2}{\varepsilon} \le \log \frac{(1 + \bar{\sigma})(\|H\| + \sqrt{(1 + \bar{\sigma}/12)}\sigma\|s^*\|)\|s^*\|^2}{\varepsilon} \tag{5.15}$$

and

$$\sqrt{\frac{10\|\tilde{s}^*\|^2}{\varepsilon} \cdot \frac{1}{\mathrm{gap} \wedge \frac{\sigma}{2}\|\tilde{s}^*\|}} \le (1 + \bar{\sigma})\sqrt{\frac{10\|s^*\|^2}{\varepsilon} \cdot \frac{1}{\mathrm{gap} \wedge \frac{\sigma}{2}\|s^*\|}}. \tag{5.16}$$

In particular, if $\frac{\varepsilon}{10\|s^*\|^2} \le \frac{\sigma}{2}\|s^*\| - \lambda$, then by (5.11),

$$|\|s^*\| - \|\tilde{s}^*\|| \le \frac{\bar{\sigma}}{10\sigma\|s^*\|^2} \cdot 10\|s^*\|^2(\frac{\sigma}{2}\|s^*\| - \lambda) = \frac{\bar{\sigma}}{\sigma}(\frac{\sigma}{2}\|s^*\| - \lambda)$$

which yields

$$\frac{\sigma}{2}\|\tilde{s}^*\| - \lambda \ge (1 - \frac{\bar{\sigma}}{2})(\frac{\sigma}{2}\|s^*\| - \lambda) \ge \frac{\frac{\sigma}{2}\|s^*\| - \lambda}{1 + \bar{\sigma}}.$$

We thus obtain from (5.14) that

$$\min\left\{\frac{1}{\frac{\sigma}{2}\|\tilde{s}^*\| - \lambda}, \frac{10\|\tilde{s}^*\|^2}{\varepsilon}\right\} \le (1 + \bar{\sigma})\min\left\{\frac{1}{\frac{\sigma}{2}\|s^*\| - \gamma}, \frac{10\|s^*\|^2}{\varepsilon}\right\}. \tag{5.17}$$

In addition, whenever $\tilde{q}(s^t) \le \tilde{q}(\tilde{s}^*) + \varepsilon$, the following relation holds

$$\begin{aligned} q(s^t) \le \tilde{q}(s^t) + \gamma\|s^t\| \le \tilde{q}(\tilde{s}^*) + \varepsilon + \gamma\|s^t\| &\le \tilde{q}(\tilde{s}^*) + \varepsilon + \gamma\|\tilde{s}^*\| \\ &\le \tilde{q}(s^*) + \varepsilon + \gamma\|\tilde{s}^*\| \\ &\le q(s^*) + \varepsilon + \gamma(\|s^*\| + \|\tilde{s}^*\|) \\ &\le q(s^*) + \varepsilon + \gamma(S + \sqrt{1 + \bar{\sigma}/12}\|s^*\|) \\ &\le q(s^*) + \varepsilon + 3\gamma S \\ &\le q(s^*) + (1 + \bar{\sigma}/8)\varepsilon. \end{aligned} \tag{5.18}$$

It is also noteworthy that by Lemma 4.6 (i) in [5] and the setting of γ in (5.4), with probability at least $1 - \delta \in (0, 1)$,

$$\mathbb{P}\left(|\tilde{g}^{(1)}| \le \frac{\sqrt{\pi}\gamma\delta}{\sqrt{2n}}\right) \le \mathbb{P}\left(|\tilde{g}^{(1)}| \le \frac{\sqrt{\pi}\sigma\bar{\sigma}\varepsilon\delta}{24(\psi + \sigma\|s^*\|)\sqrt{2n}}\right) \le \delta. \tag{5.19}$$

Then it indicates that

$$\mathbb{P}(|\tilde{g}^{(1)}| > b) = \mathbb{P}\left(|\tilde{g}^{(1)}| > \sqrt{\pi}\gamma\delta/\sqrt{2n}\right) \ge 1 - \delta,$$

where $b = \frac{\sqrt{\pi}\delta\sigma\bar{\sigma}\varepsilon}{24\sqrt{2n}(\psi + \sigma S)}$. Hence, as a result of (5.15)-(5.19), the following lemma can be obtained.

LEMMA **5.3.** *Under parameter settings* (5.3)-(5.4) *and with probability at least* $1 - \delta$, $q(s^t) \le q(s^*) + \varepsilon$ *if*

$$t \ge \frac{6(1 + \bar{\sigma})}{\eta}\left(\log\left(1 + \frac{\lambda_+^2}{2\sigma b}\right) + \log\frac{(1 + \bar{\sigma})(\|H\| + \sqrt{(1 + \bar{\sigma}/12)}\sigma\|s^*\|)\|s^*\|^2}{\varepsilon/(1 + \bar{\sigma}/8)}\right)\min\left\{\frac{1}{\frac{\sigma}{2}\|s^*\| - \lambda}, \frac{10\|s^*\|^2}{\varepsilon/(1 + \bar{\sigma}/8)}\right\}.$$

If, in particular, $\frac{\sigma}{2}\|s^*\| - \lambda \le (1 - \frac{2\bar{\sigma}}{3})\frac{\varepsilon}{10\|s^*\|^2}$, *then* $\tilde{q}(s^t) \le \tilde{q}(\tilde{s}^*) + \varepsilon$ *for all*

$$t \ge \frac{6(1+\bar{\sigma})}{\eta}\left(\log\left(1+\frac{\lambda_+^2}{2\sigma b}\right) + \log\frac{(1+\bar{\sigma})(\|H\| + \sqrt{(1+\bar{\sigma}/12)}\sigma\|s^*\|)\|s^*\|^2}{\varepsilon/(1+\bar{\sigma}/8)}\right)\sqrt{\frac{10\|s^*\|^2}{\varepsilon/(1+\bar{\sigma}/8)}\cdot\frac{1}{\mathrm{gap}\wedge\frac{\sigma}{2}\|s^*\|}}.$$

Based on Theorem 5.3, by setting ε properly we can estimate the lower bound of iteration number with high probability in order to find an output, s^t for some t, at kth iteration satisfying (2.7b) with high probability. Suppose now that (2.7b) holds at kth iteration, namely, with subscripts omitted,

$$q(s^t) - q(s^*) \le \varepsilon, \text{ where } \varepsilon \le \tfrac{1}{18}\sigma\omega.$$

It is easy to check from (5.7) and $\|s^t\| \le \sqrt{|\|s^t\|^2 - \|s^*\|^2|} + \|s^*\|$ that

$$\begin{aligned}\tilde{q}(s^t) \le q(s^t) + \gamma\|s^t\| \le q(s^*) + \varepsilon + \gamma\|s^t\| &\le q(\tilde{s}^*) + \varepsilon + \gamma\|s^*\| + 2.5\gamma\sqrt{\frac{\gamma}{\sigma}}\\ &\le \tilde{q}(\tilde{s}^*) + \gamma\|\tilde{s}^*\| + \varepsilon + \gamma\|s^*\| + 2.5\gamma\sqrt{\frac{\gamma}{\sigma}}\\ &\le \tilde{q}(\tilde{s}^*) + \varepsilon + 2\gamma\|s^*\| + 4.5\gamma\sqrt{\frac{\gamma}{\sigma}}. \qquad (5.20)\end{aligned}$$

Since $\tilde{q}$ is smooth, for any t, by Taylor's theorem the following equality holds:

$$\tilde{q}(s^{t+1}) = \tilde{q}(s^t) + \langle\nabla\tilde{q}(s^t), s^{t+1} - s^t\rangle + \frac{1}{2}\langle s^{t+1} - s^t, \nabla^2\tilde{q}(s)(s^{t+1} - s^t)\rangle$$

for some $s \in [s^t, s^{t+1}]$. Note that for all sufficiently large t,

$$\|\nabla^2\tilde{q}(s)\| = \left\|H + \frac{\sigma}{2}\|s\|\mathrm{I} + \frac{\sigma}{2}\frac{ss^T}{\|s\|}\right\| \le \|H\| + \sigma\max\{\|s^t\|, \|s^{t+1}\|\} \le \|H\| + \sigma(S + 2.5\sqrt{\frac{\gamma}{\sigma}}) =: L_{\tilde{q}}.$$

Hence, we obtain

$$\tilde{q}(\tilde{s}^*) \le \tilde{q}(s^{t+1}) \le \tilde{q}(s^t) + \langle\nabla\tilde{q}(s^t), s^{t+1} - s^t\rangle + \frac{L_{\tilde{q}}}{2}\|s^{t+1} - s^t\|^2,$$

which indicates from (5.20) that

$$(\eta - \frac{\eta^2 L_{\tilde{q}}}{2})\|\nabla\tilde{q}(s^t)\|^2 \le \tilde{q}(s^t) - \tilde{q}(\tilde{s}^*) \le \varepsilon + 2\gamma\|s^*\| + 4.5\gamma\sqrt{\frac{\gamma}{\sigma}}.$$

As $\tilde{g} - g = \gamma\zeta$ and $\|\nabla q(s^t)\|^2 \le 2\|\nabla\tilde{q}(s^t)\|^2 + 2\gamma^2$, we derive the following inequality:

$$(\eta - \frac{\eta^2 L_{\tilde{q}}}{2})\|\nabla q(s^t)\|^2 \le 2\varepsilon + 4\gamma\|s^*\| + 9\gamma\sqrt{\frac{\gamma}{\sigma}} + 2\gamma^2(\eta - \frac{\eta^2 L_{\tilde{q}}}{2})$$

which implies

$$\|\nabla q(s^t)\|^2 \le \frac{2\varepsilon + 4\gamma\|s^*\| + 9\gamma\sqrt{\gamma/\sigma}}{\eta - \eta^2 L_{\tilde{q}}/2} + 2\gamma^2.$$

Recall that $\eta < 1/L_{\tilde{q}}$ where $L_{\tilde{q}} = \|H\| + \sigma S + 2.5\sqrt{\gamma\sigma}$. Besides, by the definition of g and H at each iteration and with the parameter setting (3.21), we can set g, $\|H\|$ and σ in the same order as β. Then $\|s^*\| = \mathcal{O}(1)$, $R = \mathcal{O}(1)$ and $\gamma = \mathcal{O}(\varepsilon)$ indicating

$$\|\nabla q(s^t)\| = \mathcal{O}(\sqrt{L_{\tilde{q}}\varepsilon} + \varepsilon) = \mathcal{O}(L_{\tilde{q}}\omega^{2/3}), \text{ if } \varepsilon = \mathcal{O}(L_{\tilde{q}}\,\omega^{4/3}).$$

Consequently, (2.7c) can be satisfied at the output of the cubic-solver at kth iteration.

To summarize above analysis, we can finally obtain that under proper parameter settings Condition A can be achieved with high probability at each iteration, provided that the iteration number of the cubic solver is sufficiently large.

6 Discussions on adaptive parameter settings

In the previous sections, non-adaptive penalty parameters and regularization parameters have been considered. It is worth noting that our algorithms can be adapted to variants with adaptive parameter settings. In this section, we will delve into discussions on the behavior of ICPD when incorporating adaptive parameter settings.

6.1 Adaptive penalty parameters

One popular adaptive approach for updating penalty parameters in penalty methods, as discussed in the literature [3], is based on the improvement of constraint violation. If the constraint violation is reduced, it indicates that the penalty parameter is sufficiently large and there is no need to increase it. Conversely, if the constraint violation does not improve, the penalty parameter should be increased. In the case of equality-constrained optimization, Wang and Yuan [47] studied an augmented Lagrangian trust region method based on exact function information. They modeled the subproblem at the kth iteration in a similar way to (2.4), but incorporated a trust region strategy instead of cubic regularization. Both penalty parameter and trust region radius are adaptively updated, and the global convergence of the proposed method is presented in [47]. In our proposed algorithms for solving (1.1), we can also adopt a scheme to update the penalty parameter adaptively. We will consider an adaptive version of ICPD with the penalty parameter β_k updated following the rule:

$$\beta_{k+1} = \begin{cases} \beta_k, & \text{if } \|c(x_{k+1})\| \leq \varpi \|c(x_k)\|, \\ \beta_k + \Gamma, & \text{if } \|c(x_{k+1})\| > \varpi \|c(x_k)\| \end{cases} \tag{6.1}$$

for $k \geq 1$. Here, $\varpi \in (0,1)$ and $\Gamma > 0$. As shown in Theorem 3.1, ICPD can terminate in a finite number of iterations. However, to better understand the theoretical properties of adaptive ICPD, we will skip the finite termination test and investigate the behavior of $\{\beta_k\}_{k\geq 1}$ generated during the algorithm. Since $\{\beta_k\}$ is a non-decreasing sequence following the update scheme (6.1), either $\{\beta_k\}$ is upper bounded, or β_k approaches infinity as k increases to infinity. In the subsequent analysis, we will address each case separately. For completeness, we present the adaptive ICPD algorithm in Algorithm 6.1.

Algorithm 6.1 Adaptive ICPD

Input: $x_1, \lambda_1 = \mathbf{0}, \beta_1 > 0, \sigma_1 > 0, \bar{\epsilon} \in (0,1)$, $\omega \in (0,1), \rho_1 \in (0, \beta_1)$

1: **for** $k = 1 \ldots$ **do**
2: Generate approximate gradient g_k^0 and Hessian H_k^0 of f at x_k, and compute g_k and H_k by (2.3).
3: Solve subproblem (2.4) obtaining an inexact solution s_k satisfying Condition A and set $x_{k+1} := x_k + s_k$.
4: Compute β_{k+1} through (6.1).
5: Compute $\rho_{k+1} \in (0, \beta_{k+1})$ and σ_{k+1}.
6: Compute λ_{k+1} through (2.8).
7: $k := k + 1$.
8: **end for**

6.1.1 Bounded penalty parameters

In this case we assume that there exists $\beta_{\text{bnd}} > 0$ such that $\beta_k \leq \beta_{\text{bnd}}$ for all $k \geq 1$. The lemma below indicates that the number of consecutive iterations in which $\max\{\|x_{k+1} - x_k\|, \|x_k - x_{k-1}\|\} \geq \bar{\epsilon}$ is upper bounded.

LEMMA **6.1.** *Under Assumptions 1-3, Conditions A and B, assume that* $\omega = \bar{\epsilon}^3$ *and* $\sum_{k=1}^{\infty} \rho_k \leq \rho$ *with* $\rho > 0$, (3.6) *and* (3.21) *hold. Further assume that there exist two positive integers* K_1, K_2 *such that* $\max\{\|x_{k+1} - x_k\|, \|x_k - x_{k-1}\|\} \geq \bar{\epsilon}$ *for any* $k \in [K_1, K_2)$. *Then it holds that*

$$K_2 - K_1 \leq l_1 := \left\lceil \frac{C + (3\sqrt{m}\rho + \beta_{\text{bnd}})C^2 - f_{low}}{\hat{\sigma}\bar{\epsilon}^3} \right\rceil, \tag{6.2}$$

where $\hat{\sigma} := \frac{1}{12} L_H^f$.

Proof. By applying Lemma 3.3, we obtain that

$$\sum_{k=K_1}^{K_2-1} \hat{\sigma}_k \max\{\|x_k - x_{k-1}\|^3, \|x_{k+1} - x_k\|^3\} \leq \mathcal{L}_{\beta_{K_1}}(x_{K_1}, \lambda_{K_1}) - f_{low} + 2\sqrt{m}C^2 \sum_{k=K_1}^{K_2-1} \rho_k + \frac{1}{2}C^2\beta_{\rm bnd}. \quad (6.3)$$

Note that under the condition (3.21) and by (3.24), $\hat{\sigma}_k \geq \hat{\sigma}$. Additionally, it follows from Assumption 3 and Lemma 3.1 together with $\sum_{k=1}^{\infty} \rho_t \leq \rho$ that

$$\mathcal{L}_{\beta_{K_1}}(x_{K_1}, \lambda_{K_1}) \leq C + \sqrt{m}\rho C^2 + \frac{1}{2}C^2\beta_{\rm bnd}.$$

Then we derive the conclusion from (6.3). □

The update rule (6.1) indicates that

$$\frac{\beta_k - \beta_1}{\Gamma} \leq J := \frac{\beta_{\rm bnd} - \beta_1}{\Gamma}, \quad \forall k \geq 1.$$

For simplicity, we next denote k_j as the index of iteration when the penalty parameter increases to $\beta_1 + j\Gamma$, where $j \in \{0, 1, \ldots, J\}$, i.e.,

$$k_j = \min_{k \in \mathbb{N}} \{k : \beta_k = \beta_1 + j\Gamma\}, \quad j \in \{0, 1, \ldots, J\}.$$

Obviously, $k_0 = 1$. Our next theorem shows that for any $j \in \{0, 1, \ldots, J\}$, either $k_j - k_{j-1}$ is upper bounded or there exists $K \in [k_{j-1}, k_j]$ such that

$$\|c(x_{k+1})\| < \epsilon \text{ and } \max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon} \quad (6.4)$$

hold at $k = K$. Then we can identify the iteration index K such that x_{K+1} satisfies the approximate second-order stationarity.

THEOREM **6.1.** *Under the conditions of Lemma 6.1 and Assumption 4, for any $j \in \{0, 1, \ldots, J\}$, either $k_j - k_{j-1} \leq l_1 + l_2 + 1$ with $l_2 = \lceil \frac{\log(C\epsilon^{-1})}{\log \varpi^{-1}} + 1 \rceil$ and l_1 defined in (6.2), or there exists $K \in [k_{j-1} + l_2, k_{j-1} + l_1 + l_2]$ such that (6.4) holds at $k = K$. Consequently, letting $\bar{\epsilon} = \sqrt{\epsilon}$, there exists $K \in [1, (l_1 + l_2 + 1)(J + 1)]$ such that*

$$\begin{cases} \|\nabla f(x_{K+1}) + \nabla c(x_{K+1})\hat{\lambda}_{K+1}\| = \mathcal{O}((\rho + \beta_{\rm bnd})\epsilon + \rho_K), \\ \|c(x_{K+1})\| < \epsilon, \\ d^T(\nabla^2 f(x_{K+1}) + \sum_{i=1}^m \hat{\lambda}_i \nabla^2 c_i(x_{K+1}))d \geq -\upsilon_K \|d\|^2 \text{ for all } d \in \mathrm{Null}(\nabla c(x_{K+1})^T), \end{cases}$$

where $\hat{\lambda} := \lambda_{K+1} + \beta_k c(x_{K+1})$ and $\upsilon_K = \mathcal{O}((\rho + \beta_{\rm bnd})\sqrt{\epsilon} + \rho_K\epsilon + \rho_K{\beta_1}^{-1}(\rho + 1))$. Moreover, K is in the order

$$\mathcal{O}\left(\frac{\beta_{\rm bnd} - \beta_1}{\Gamma} \cdot \left(\frac{\log(C\epsilon^{-1})}{\log(\varpi^{-1})} + \frac{C + (3\sqrt{m}\rho + \beta_{\rm bnd})C^2 - f_{low}}{\hat{\sigma}\epsilon^{1.5}}\right)\right).$$

Proof. Following the update scheme (6.1), for any given $\epsilon > 0$ and under Assumption 4, to achieve $\|c(x_{k+1})\| \leq \epsilon$ the number of consecutive iterations when the penalty parameters keep the same does not exceed l_2. Thus when $k_j - k_{j-1} > l_1 + l_2 + 1$, by Lemma 6.1 we obtain that there exists $K \in [k_{j-1} + l_2, k_{j-1} + l_1 + l_2]$ such that (6.4) holds at $k = K$. Therefore, due to $j \leq J$, it occurs that either $k_j - k_{j-1} \leq l_1 + l_2 + 1$ for any $j \leq J$, or there exists some $j \in [1, J]$ suth that $k_j - k_{j-1} > l_1 + l_2 + 1$. No matter which case occurs, there always exists some $K \in [1, (l_1 + l_2 + 1)(J + 1)]$ such that (6.4) holds at $k = K$. Then in analogy to Theorem 3.1 and by $\bar{\epsilon} = \sqrt{\epsilon}$ we obtain the approximate second-order stationarity and the order of K. □

6.1.2 Unbounded penalty parameters

We now assume that penalty parameters are unbounded and will eventually approach the infinity. The following lemma shows that there are infinite number of iterations, denoted by $\{K_i\}_{i\geq 1}$, when $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon}$, and the times of increase of the penalty parameter between K_i and K_{i+1} is upper bounded for any $i \geq 1$.

LEMMA **6.2.** *Under Assumptions 1-4, Conditions A and B, assume that $\omega = \bar{\epsilon}^3$ and $\sum_{k=1}^{\infty} \rho_k \leq \rho$ with $\rho > 0$, (3.6) and (3.21) hold. Then there exists an infinite sequence of iteration indices, denoted by $\{K_i\}_{i\geq 1}$, such that for any $i \geq 1$,*

$$\begin{cases} \max\{\|x_{k+1} - x_k\|, \|x_k - x_{k-1}\|\} < \bar{\epsilon}, & k = K_i, \\ \max\{\|x_{k+1} - x_k\|, \|x_k - x_{k-1}\|\} \geq \bar{\epsilon}, & K_i < k < K_{i+1}, \end{cases} \tag{6.5}$$

and for any $k \in [K_i, K_{i+1}]$,

$$\frac{\beta_k - \beta_{K_i}}{\Gamma} \leq J_1 := \left\lceil \frac{16((C + 3\sqrt{m}C^2\rho)/\beta_1 + C^2)}{mL_cL_g^c + 18\theta} \bar{\epsilon}^{-3} \right\rceil. \tag{6.6}$$

Additionally, (3.22) holds true for any $K \in \{K_i\}_{i\geq 1}$.

Proof. For any $k \geq 1$, let $j_k = (\beta_k - \beta_1)/\Gamma$. It must hold that $j_k \to \infty$ as $k \to \infty$ since the penalty parameters are unbounded. For a given $K > 2$, without loss of generality we assume that $j_K \geq 2$. Then it is easy to obtain from $j_K \leq j_{K-1} + 1$ that

$$\begin{aligned} \frac{\sum_{k=1}^{K-1} \beta_k}{\beta_K} &\geq \frac{\beta_1 + (\beta_1 + 1 \cdot \Gamma) + \cdots + (\beta_1 + j_{K-1} \cdot \Gamma)}{\beta_1 + j_K\Gamma} \\ &\geq \frac{1}{2} \cdot \frac{(j_{K-1} + 1)(\beta_1 + j_{K-1}\Gamma)}{\beta_1 + j_K\Gamma} \\ &\geq \frac{j_{K-1} + 1}{4} \\ &\to \infty \text{ as } K \to \infty. \end{aligned} \tag{6.7}$$

We denote K_1 as the smallest iteration index such that $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon}$. Without loss of generality, we assume $K_1 > 1$, thus obviously

$$\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} \geq \bar{\epsilon} \tag{6.8}$$

for $k \in [1, K_1 - 1]$. Then (3.23) holds with $\hat{\sigma}_k$ satisfying (3.24), i.e.

$$\sum_{k=1}^{K_1-1} \hat{\sigma}_k \bar{\epsilon}^3 \leq \mathcal{L}_{\beta_1}(x_1, \lambda_1) - f_{low} + 2\sqrt{m}C^2\rho + \frac{1}{2}C^2\beta_{K_1} \leq C + 3\sqrt{m}C^2\rho + C^2\beta_{K_1}$$

with

$$\hat{\sigma}_k \geq \frac{1}{4}\beta_k(mL_cL_g^c + 18\theta), \quad \forall k \in [1, K_1 - 1].$$

It, together with (6.7), indicates that K_1 is finite and

$$\frac{\beta_{K_1} - \beta_1}{\Gamma} = j_{K_1} \leq J_1 = \frac{16((C + 3\sqrt{m}C^2\rho)/\beta_1 + C^2)}{mL_cL_g^c + 18\theta} \bar{\epsilon}^{-3}.$$

Furthermore, in analogy to Theorem 3.1 we can show that (3.22) holds for $K = K_1$. We now assume that $K_{i+1} > K_i$ is the smallest iteration index such that $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon}$ for $k > K_i$. Without loss of generality we assume $K_{i+1} - K_i > 1$. Then by replicating the above analysis with related information initialized at $(K_i + 1)$th iteration in (3.23) and (3.24), we obtain (6.6) and (3.22) holds at $K = K_{i+1}$. Hence, due to the unboundedness of penalty parameters and by induction, we ensure the existence of the infinite sequence $\{K_i\}$ as desired. □

With a slight abuse of notation, we still denote k_j, $j \geq 0$, as the index of iteration when the penalty parameter increases to $\beta_1 + j\Gamma$ for $j \in \mathbb{N}$, i.e.,

$$k_j = \min_{k\in\mathbb{N}}\{k : \beta_k = \beta_1 + j\Gamma\}, \quad j \in \mathbb{N}.$$

Obviously, we have $k_0 = 1$ and the sequence $\{k_j\}_{j\in\mathbb{N}}$ is infinite as β_k is unbounded. We next show that for any $j \in \mathbb{N}_+$, either $k_j - k_{j-1}$ is upper bounded or there must exist K_i for some i such that (6.4) holds at $k = K_i$. We thus arrive at an iteration index K with x_{K+1} satisfying the approximate second-order stationarity.

THEOREM **6.2.** *Under the conditions of Lemma 6.2, for any $j \in \mathbb{N}_+$, either $k_j - k_{j-1} \leq \bar{K}_1 + \bar{K}_2 + 1$ with*

$$\bar{K}_1 = \left\lceil \frac{\log(C\epsilon^{-1})}{\log \varpi^{-1}} + 1 \right\rceil \quad \text{and} \quad \bar{K}_2 = \left\lceil \frac{1 + 4((C + 3\sqrt{m}C^2\rho + C^2\Gamma)/\beta_1 + 1.5C^2)}{mL_cL_g^c + 18\theta}\bar{\epsilon}^{-3} \right\rceil, \tag{6.9}$$

or there exists $K_i \in (k_{j-1} + \bar{K}_1, k_{j-1} + \bar{K}_1 + \bar{K}_2]$ such that (6.4) holds. Furthermore, with $J_2 := \lceil(\epsilon^{-1} - \beta_1)/\Gamma\rceil$, one of the following cases holds true.

(i) There exists integer $K \in [1, (\bar{K}_1 + \bar{K}_2 + 1)J_2 - 1]$ such that

$$\begin{aligned} &\|c(x_{K+1})\| < \epsilon, \\ &\|\nabla f(x_{K+1}) + \nabla c(x_{K+1})\hat{\lambda}_{K+1}\| = \mathcal{O}(\epsilon^{-1}\bar{\epsilon}^2 + \rho_K), \qquad (6.10) \\ &d^T(\nabla^2 f(x_{K+1}) + \sum_{i=1}^m \hat{\lambda}_i \nabla^2 c_i(x_{K+1}))d \geq -\upsilon_K\|d\|^2 \quad \text{for all } d \in \mathrm{Null}(\nabla c(x_{K+1})^T), \qquad (6.11) \end{aligned}$$

where $\hat{\lambda}_{K+1} := \lambda_{K+1} + \beta_K c(x_{K+1})$ and $\upsilon_K = \mathcal{O}(\epsilon^{-1}\bar{\epsilon} + \rho_K\bar{\epsilon}^2 + \rho_K\epsilon)$.

(ii) There exists integer $K \in [(\bar{K}_1+\bar{K}_2+1)J_2, (\bar{K}_1+\bar{K}_2+1)(J_2+J_1+1))$ such that $\|c(x_{K+1})\| = \mathcal{O}(\bar{\epsilon}^2+\epsilon)$, and (6.10)-(6.11) hold.

Proof. For any fixed $j \in \mathbb{N}_+$, if $k_j - k_{j-1} > \bar{K}_1 + \bar{K}_2 + 1$, it is easy to obtain from Assumption 3 and (6.1) that $\|c(x_{k+1})\| < \epsilon$ for any $k \in (k_{j-1} + \bar{K}_1, k_j)$. We next prove that there exists $K_i \in (k_{j-1} + \bar{K}_1, k_{j-1} + \bar{K}_1 + \bar{K}_2]$ such that $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon}$ for $k = K_i$, when $k_j - k_{j-1} > \bar{K}_1 + \bar{K}_2 + 1$. By the way of contradiction, we now assume that for any $k \in (k_{j-1} + \bar{K}_1, k_{j-1} + \bar{K}_1 + \bar{K}_2]$, $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} \geq \bar{\epsilon}$. Then similar to (3.23) and due to $\beta_k = \beta_{k_{j-1}+\bar{K}_1+1}$ for any $k \in (k_{j-1} + \bar{K}_1, k_{j-1} + \bar{K}_1 + \bar{K}_2]$ and $\beta_{k_{j-1}+\bar{K}_1+\bar{K}_2+1} \leq \beta_{k_{j-1}+\bar{K}_1+1} + \Gamma$, we have

$$\begin{aligned} \sum_{k=k_{j-1}+\bar{K}_1+1}^{k_{j-1}+\bar{K}_1+\bar{K}_2} \hat{\sigma}_k\bar{\epsilon}^3 &\leq \mathcal{L}_{\beta_{k_{j-1}+\bar{K}_1+1}}(x_{k_{j-1}+\bar{K}_1+1}, \lambda_{k_{j-1}+\bar{K}_1+1}) - f_{low} + 2\sqrt{m}C^2\rho + \frac{1}{2}C^2\beta^2_{k_{j-1}+\bar{K}_1+\bar{K}_2+1} \\ &\leq C + 3\sqrt{m}C^2\rho + C^2\Gamma + 1.5C^2\beta_{k_{j-1}+\bar{K}_1+1}, \end{aligned}$$

where

$$\hat{\sigma}_k \geq \frac{1}{4}\beta_{k_{j-1}+\bar{K}_1+1}(mL_cL_g^c + 18\theta).$$

It thus leads to

$$\bar{K}_2 \leq \frac{4(C + 3\sqrt{m}C^2\rho + C^2\Gamma + 1.5C^2\beta_{k_{j-1}+\bar{K}_1+1})}{\beta_{k_{j-1}+\bar{K}_1+1}(mL_cL_g^c + 18\theta)}\bar{\epsilon}^{-3},$$

which however contradicts the setting of $\bar{K}_2$ in (6.9). Hence, there must exist integer $K_i \in (k_{j-1} + \bar{K}_1, k_{j-1} + \bar{K}_1 + \bar{K}_2]$ such that $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon}$, thus (6.4) holds at $k = K_i$.

We now prove the second part of the theorem. Note that for any $j \geq J_2 := \lceil(\epsilon^{-1} - \beta_1)/\Gamma\rceil$, the penalty parameter satisfies $\beta_{k_j} \geq \beta_{k_{J_2}} \geq \epsilon^{-1}$. Moreover, starting from (k_{J_2})th iteration, by Lemma 6.2 we obtain that after at most $\mathcal{O}(\bar{\epsilon}^{-3})$ times of increase of the penalty parameter we come to some K such that $\max\{\|x_K - x_{K-1}\|, \|x_{K+1} - x_K\|\} < \bar{\epsilon}$. Therefore, one of the following cases occurs.

(i) **Case 1**: there exists some $j \in [1, J_2]$ such that $k_j - k_{j-1} > \bar{K}_1 + \bar{K}_2 + 1$. In this case, by the first part of this theorem there exists some $K \in [1, (\bar{K}_1 + \bar{K}_2 + 1)(J_2 - 1) + \bar{K}_1 + \bar{K}_2]$ such that

$$\beta_K \leq \epsilon^{-1}, \quad \|c(x_{K+1})\| < \epsilon, \quad \max\{\|x_K - x_{K-1}\|, \|x_{K+1} - x_K\|\} < \bar{\epsilon}.$$

Then through similar analysis to Theorem 3.1, we further derive the desired conclusions.

(ii) **Case 2:** $k_j - k_{j-1} \leq \bar{K}_1 + \bar{K}_2 + 1$ for any $j \in [1, J_2]$. In this case, for any $k \geq (\bar{K}_1 + \bar{K}_2 + 1)J_2$, it must have $\beta_k^{-1} = \mathcal{O}(\epsilon)$. Moreover, by the existence of $\{K_i\}$ as demonstrated in Lemma 6.2, from $((\bar{K}_1 + \bar{K}_2 + 1)J_2)$th iteration until (K_i)th iteration where K_i is the smallest among $k > (\bar{K}_1 + \bar{K}_2 + 1)J_2$, the times of increase of the penalty parameter is bounded by J_1 according to (6.6). Between any two successive increase, with iteration indices still denoted by k_{j-1} and k_j for example, we have proved previously that either $k_j - k_{j-1} \leq \bar{K}_1 + \bar{K}_2 + 1$ or $K_i \in (k_{j-1}, k_{j-1} + \bar{K}_1 + \bar{K}_2]$ for some j. It must hold that

$$K_i - (\bar{K}_1 + \bar{K}_2 + 1)J_2 \leq (\bar{K}_1 + \bar{K}_2 + 1)J_1 + \bar{K}_1 + \bar{K}_2.$$

Then following (3.22) and $\beta_{K_i}^{-1} = \mathcal{O}(\epsilon)$ we can further obtain (6.10) and (6.11) with $K = K_i$ and $\|c(x_{K+1})\| = \mathcal{O}(\bar{\epsilon}^2 + \epsilon)$.

The proof is completed. □

Remark 6.1. *For both cases presented in Theorem 6.2, by setting appropriate values of $\bar{\epsilon}$ and ρ_k, we can obtain x_{K+1} satisfying the approximate second-order stationarity. This allows us to derive an estimate on the iteration complexity of ICPD with unbounded adaptive penalty parameters. However, we have to admit that the complexity order appears to be not quite satisfying. One reason may lie in the estimate of* (6.7)*. As it is unclear how often β_k is changed, the lower bound in* (6.7) *seems very conservative. On the other hand, to derive the associated approximate criticality measure we currently simply pursue the criterion* $\max\{\|x_k - x_{k-1}\|, \|x_{k+1} - x_k\|\} < \bar{\epsilon}$ *for some k. This however leads to a large gap estimate between two successive K_i and K_{i+1}. In order to improve the results we may need to adopt a different criticality measure.*

6.2 Adaptive regularization parameter

Another issue related to Adaptive ICPD is about the adaptive update of cubic regularization parameter σ. For problem (1.1), if assuming that inexact function values of f can be evaluated at x_k and $x_k + s_k$, denoted as f_k and f_k^+, we can adopt a similar strategy to [11] to update σ_k adaptively. More specifically, we can compute the reduction ratio r_k, i.e. the reduction $f_k + \Psi_{\beta_k}(x_k, \lambda_k) - (f_k^+ - \Psi_{\beta_k}(x_k + s_k, \lambda_k))$ divided by the predicted model reduction $(-q_k(s_k))$. Provided that $s_k \neq 0$, r_k is well-defined. This ratio can be used to estimate the quality of the predicted model q_k. If the ratio is greater than a preset positive constant, we trust the predicted model and believe it is a good approximation to $\mathcal{L}_{\beta_k}(x_k + s_k, \lambda_k)$. In this case, there is no need to increase σ in the next iteration. If, on the other hand, r_k is too small or even negative, σ should be increased to control the step size. It is noteworthy that unlike CR methods for unconstrained optimization and due to the nonlinearity of constraint functions in general, the ratio r_k will also rely on the penalty parameter β_k. As in [11], it requires there exist some $\gamma_3 \in (0, 1]$ such that $\sigma_{k+1} \geq \gamma_3 \sigma_k$ for all successful iterations (see (2.10) in [11]), which also complies with our assumption on σ_k as given in (3.6). With this assumption and following our algorithm framework, the complexity analysis given in this paper can be applied to the case with adaptive σ.

7 Numerical experiments

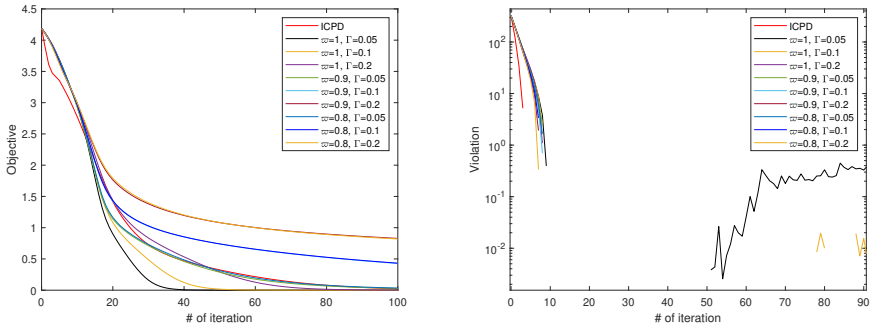
7.1 Quadratically constrained nonconvex program

In this subsection, we consider the following quadratically constrained nonconvex program:

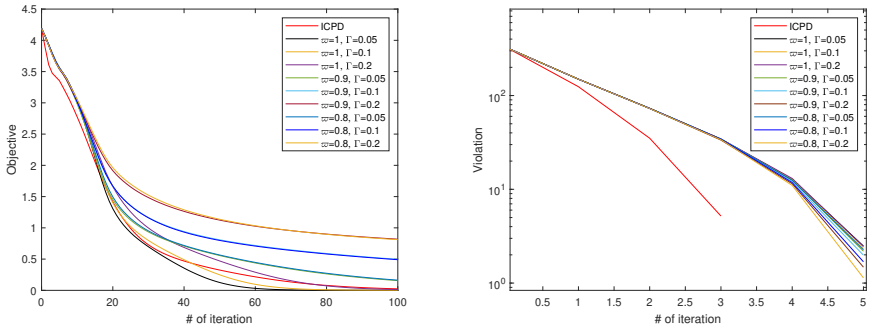
$$\begin{aligned} \min_{x\in X} \quad & f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1+\frac{1}{2}\|H_i x - c_i\|^2) \\ \text{s.t.} \quad & f_j(x) = \frac{1}{2}x^T Q_j x + a_j^T x \le b_j,\ j = 1, ..., M, \end{aligned} \tag{7.1}$$

where $X = [-10, 10]^n$, $H_i \in \mathbb{R}^{p\times n}$, $c_i \in \mathbb{R}^p$, $Q_j \in \mathbb{R}^{n\times n}$ and $a_i, b_i \in \mathbb{R}^n$. For each $i \in [N] := \{1, \ldots, N\}$, we randomly and independently generate H_i with elements following the standard Gaussian distribution. For each $j \in [M]$, we generate $Q_j \in \mathbb{R}^{n\times n}$ as the sum of a random matrix and a diagonal matrix with elements following the uniform distribution on $[-1, 1]$, i.e. U$[-1, 1]$, and a_j following U$[0.1, 1.1]^n$. With a randomly generated point $x_* \sim$ U$[0, 1]^n$, we let $c_i = H_i x_*$, $i \in [N]$ and $b_j = \frac{1}{2}x_*^T Q_j x_* + a_j^T x_*$, $j \in [M]$. Obviously, x_* is a feasible point of (7.1) and $f(x_*) = 0$, thus x_* is the optimal solution of (7.1). To obtain more stable results, we report the trend of average objective function values at all previous iterates, i.e., $\sum_{i=1}^{k} f(x_i)/k$ and the average of constraint violation $\sum_{i=1}^{M}[f_j(x) - b_j]_+$ over past iterates. All reported results are the average values obtained from 5 independent runs of each algorithm.

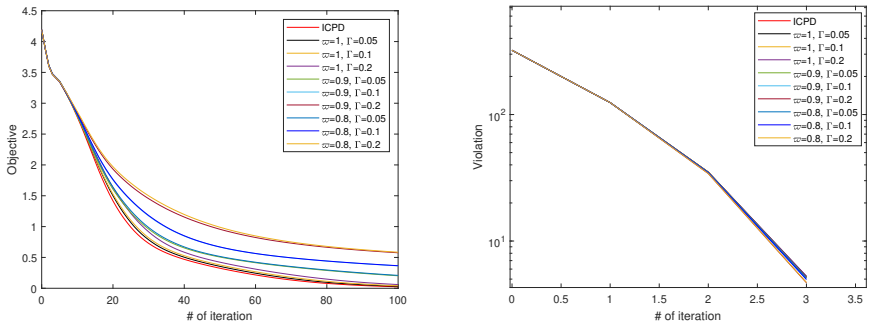
We begin by comparing the performances of ICPD with non-adaptive penalty parameters and adaptive penalty parameters. In this comparison, we set $\sigma = 10$ and $\gamma = 0.1$, where γ represents the perturbation parameter in the subproblem solver. For the non-adaptive case, we set $\beta_k = T^{1/4}$, where T is the maximum iteration number of the algorithm. In the adaptive setting, we follow the update rule in (6.1) with varying values for β_0, ϖ, and Γ. Figure 1 illustrates the results obtained from solving (7.1) with $M = 100$, $n = 100$, and $N = 2000$. From the figure, it can be observed that the decrease in objective function values achieved by ICPD with $\beta_k \equiv T^{1/4}$ is comparable to that achieved by the adaptive settings of penalty parameters, while the constraint violation decreases at a faster rate. Based on these observations, we adopt the non-adaptive setting of the penalty parameter for the subsequent tests.



(a) $\beta_0 = 0.1$



(b) $\beta_0 = 1$



(c) $\beta_0 = T^{1/4}$

Figure 1: Comparison between ICPD with non-adaptive penalty parameters and adaptive ones: $M = 100, n = 100, N = 2000$

To provide a detailed analysis of SCPD's performance, we first examine the impact of the penalty parameter β on its numerical performance. In this case we consider a scenario where $M = n = 50$, $T = 1000$, $\sigma = 50$, $I_k = J_k = 2$, and $\gamma = 0.1$. We report the results in Figure 2. As can be observed from this figure, when β decreases, the algorithm tends to achieve faster convergence concerning the objective function value, but this may result in a larger constraint violation. When β is set to $T^{1/4}$, the algorithm achieves relatively lower objective function values and constraint violations. Figure 3 showcases the impact of σ on the performance of SCPD. We use the same parameter settings as previously mentioned, with β set to $T^{1/4}$. From this figure, we observe that when $\sigma = 20$ or 50, SCPD achieves relatively better overall performance. This result is evident in both the objective function values and the constraint violations. It is important to note that the specific optimal value of σ may vary depending on the problem at hand and its characteristics.

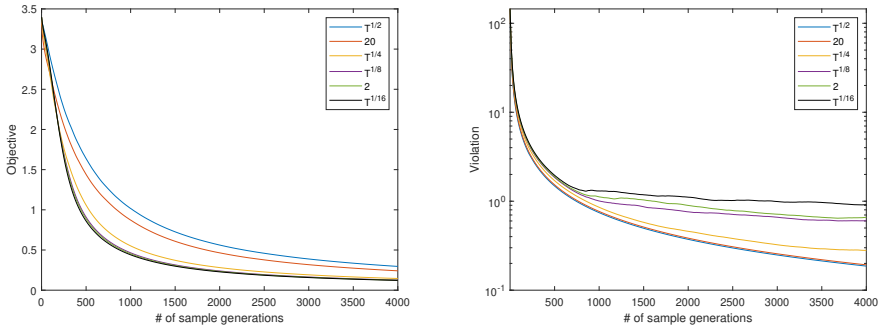


Figure 2: Impact of β on SCPD for (7.1)

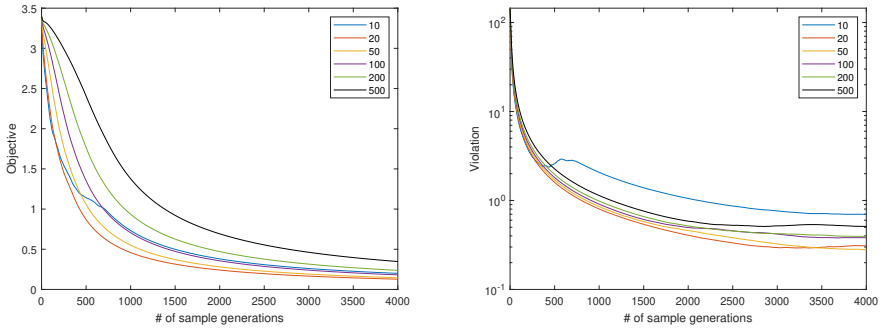


Figure 3: Impact of σ on SCPD for (7.1)

In Figures 4 and 5, we present a comparison between Matlab's built-in solver Fmincon with ICPD and SCPD, respectively. For SCPD, we utilize the parameters as mentioned earlier. In the case of ICPD, we set $\sigma = 10$, $\beta = 8$, and $\gamma = 0.1$. Regarding the settings of Fmincon, we employ the following configurations: Algorithm: 'interior-point'; ConstraintTolerance: '1e-6'; StepTolerance: '1e-10'; BarrierParamUpdate: 'monotone'; HessianApproximation: 'bfgs'; SubproblemAlgorithm: 'factorization'. When comparing ICPD and Fmincon, we draw the trend of sequences of objective function values and constraint violations, instead of taking the average over past iterates, for both algorithms. As can be seen, both ICPD and Fmincon approach similar levels in terms of objective function value and constraint violation. But within the same number of iterations, ICPD demonstrates slower convergence compared to Fmincon. On the other hand, when comparing the performances of SCPD and Fmincon, SCPD exhibits superior behavior within the given number of sample calls. This is reasonable since Fmincon calls a deterministic algorithm to solve the associated problem thus needs to compute the full information of functions and gradients. The results in Figures 4 and 5 highlight the potential advantage of SCPD in large-scale settings.

In Figure 6, we present a numerical comparison between SCPD, SPD [25], and ICPPC [4]. The experimental settings for SCPD are as follows: $T = 1000$, $\sigma = 50$, $\beta = T^{1/4}$, sampleSize $= 2$, and $\gamma = 0.1$. For SPD, the parameters are set as $T = 995$, $\eta_t = 0.03 * T^{-1/4}$, $\beta = T^{1/4}$, and $\rho = 10$, while for ICPPC we consider a scenario with $\mathcal{M} = 0.1M$ and an inner iteration of 2. We test multiple scenarios with varying values of M, n and N. Analyzing the objective function values, SCPD exhibits a significantly faster rate of decrease within the same number of sample generations compared to both SPD and ICPPC. Regarding the constraint violations, while SPD demonstrates a slightly faster reduction in the early stages, SCPD eventually achieves lower constraint violations. This indicates that the incorporation of second-order information in SCPD brings benefits to deliver improved performance when solving nonconvex constrained optimization problems.

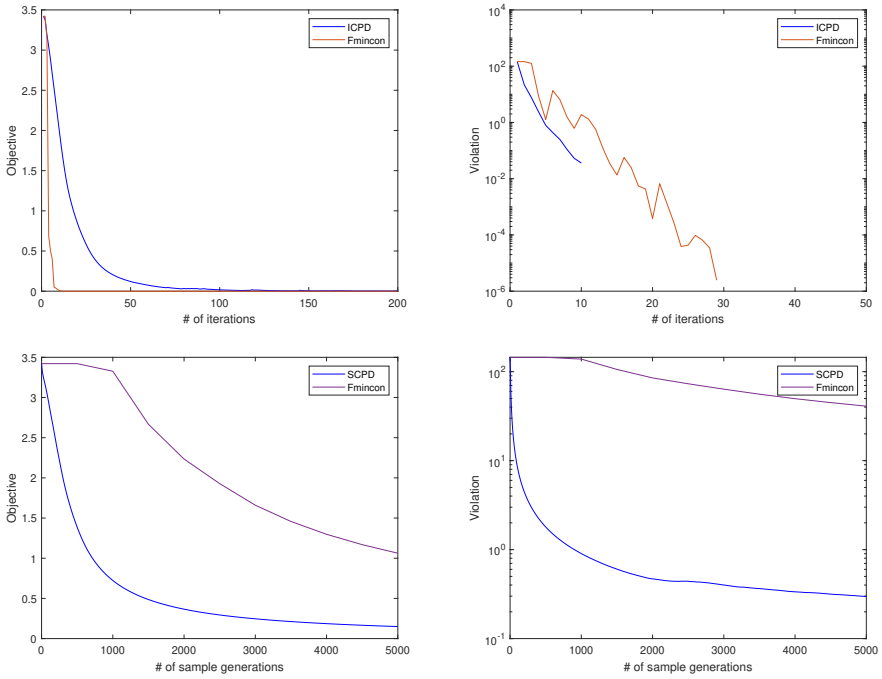


Figure 4: Comparison between ICPD, SCPD and Fmincon for (7.1): $M = 50, n = 50, N = 500$

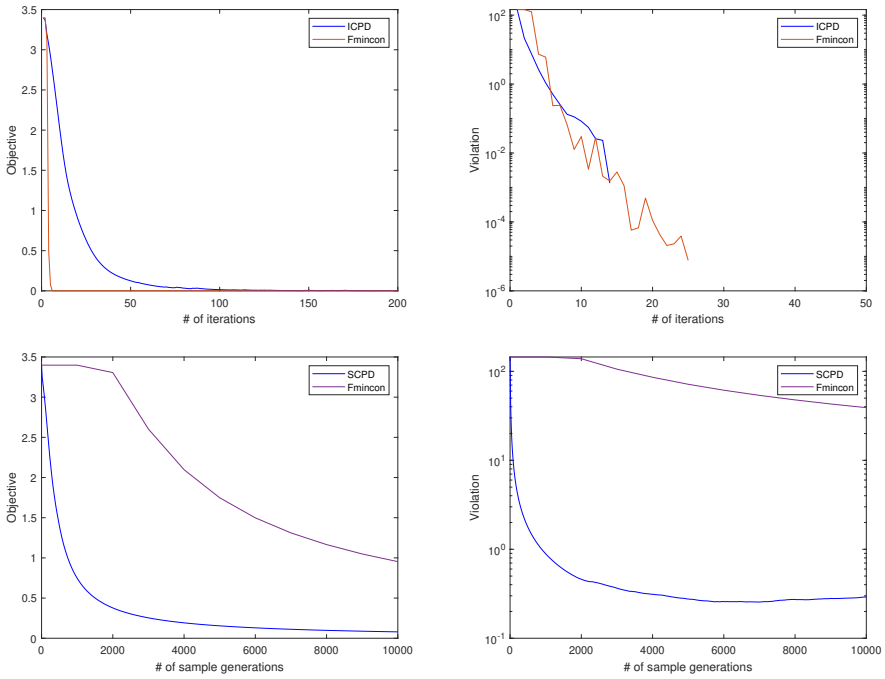
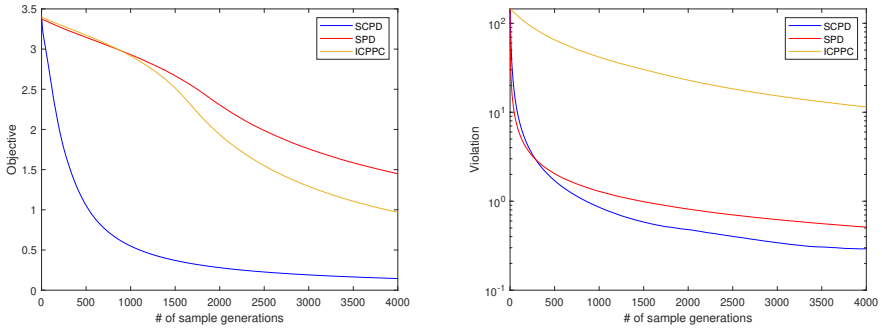
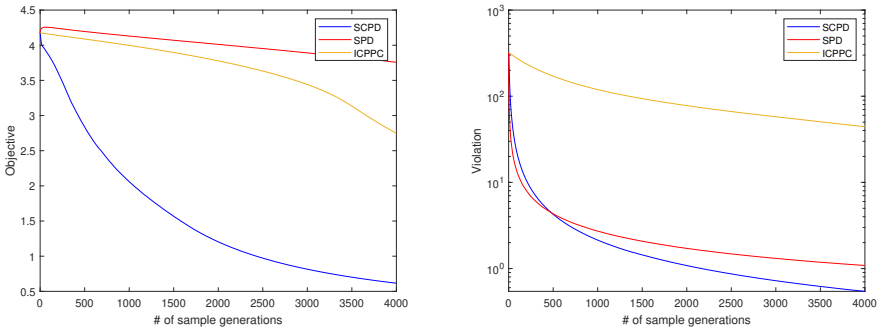


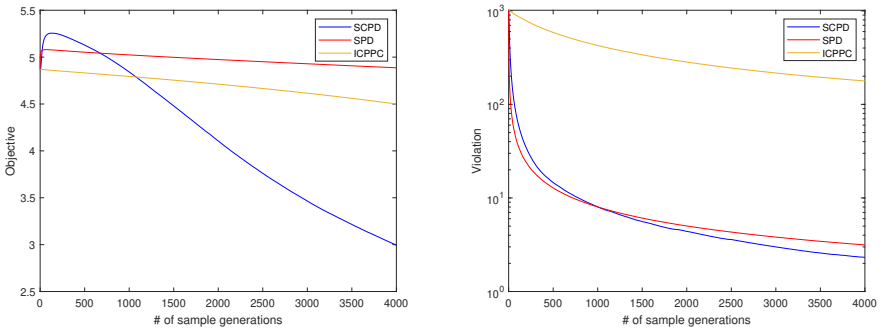
Figure 5: Comparison between ICPD, SCPD and Fmincon for (7.1): $M = 50, n = 50, N = 1000$



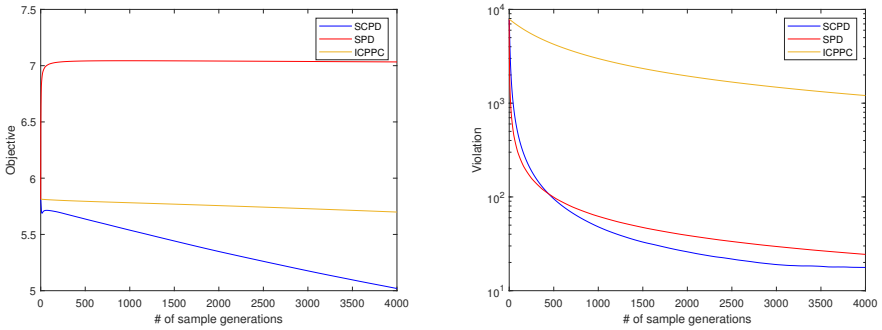
(a) $M = 50, n = 50, N = 1000$



(b) $M = 100, n = 100, N = 5000$



(c) $M = 200, n = 200, N = 10000$



(d) $M = 500, n = 500, N = 10000$

Figure 6: Comparison between three algorithms for (7.1)

7.2 Multi-class Neyman-Pearson classification

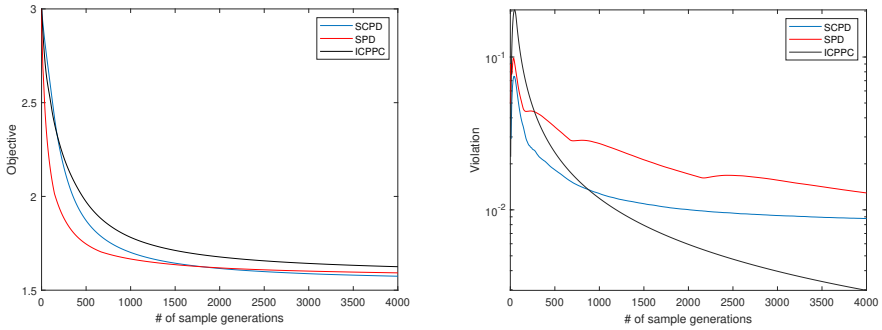
In this subsection we consider the multi-class Neyman-Pearson classification (mNPC) problem. The mNPC problem focuses on learning K models $x_1, \ldots, x_K$ in order to predict the class of a potential data point ξ. Specifically, the optimization problem is to minimize the loss on one class while controlling its value on others:

$$\begin{aligned} \min_{\|x_k\| \le r,\, k \in [K]} \quad & \frac{1}{D_1} \sum_{l>1} \sum_{\xi \in D_1} h(x_1^T \xi - x_l^T \xi) \\ \text{s.t.} \quad & \frac{1}{D_k} \sum_{l \neq k} \sum_{\xi \in D_k} h(x_1^T \xi - x_l^T \xi) \le \gamma_k, \quad k = 2, \ldots, K, \end{aligned} \tag{7.2}$$

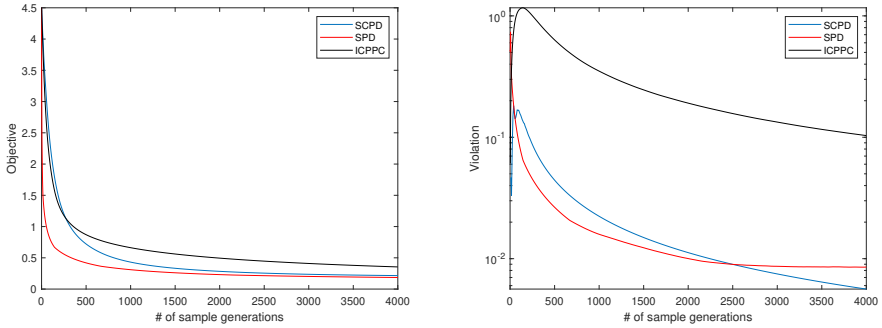
where $h(z) = (1 + e^z)^{-1}$ is the loss function and D_k represents the training data of the kth class. We use two datasets from LibSVM[1]: *covtype* ($K = 7$), and *mnist* ($K = 10$). In numerical tests, we set $r = 0.3$ and $\gamma_k = 0.5(K - 1)$, i.e. $\gamma_k = 4.5$ for *mnist* and $\gamma_k = 3$ for *covtype*. For numerical comparison, we report performance profiles on SCPD, SPD and ICPPC. For these three algorithms, we set the maximum number of stochastic gradient computations is 4000. All reported results are the average values obtained from 5 independent runs of each algorithm.

In Figure 7, we report the numerical results for the datasets *mnist* and *covtype*. For the data set *mnist*, we set the following parameters for SCPD: $T = 1000$, $\lambda = 0.3$, $\sigma = 30000$, $\beta = T^{1/4}$, $I_k = J_k = 2$ and $\gamma = 0.1$. For SPD, we set the parameters as follows: $T = 995$, $\eta_t = 0.005/t^{1/4}$, $\beta_t = T^{1/4}$ and $\rho = 0.0067$. For ICPPC, we set $T = 2000$, $\theta_t = 0.67$, $\tau_t = 2.5$ and $\eta_t = 2.6 \times 10^{-4}$. For the dataset *covtype*, the parameter settings for SCPD remain the same as above, except for σ which is chosen as 10000. For SPD, the parameters are set as: $T = 995$, $\eta_t = 0.01/t^{1/4}$, $\beta_t = 5T^{1/4}$, and $\rho = 0.67$. As for ICPPC, we use the following parameter values: $T = 2000$, $\theta_t = 0.67$, $\tau_t = 2.5$, and $\eta_t = 0.003$. We set $x_0 = \mathbf{0}$ for all three algorithms. In Figure 7, we observe that although SPD initially decreases faster, SCPD can eventually approach close or even lower objective function values and constraint violations. Furthermore, when evaluating the performance of ICPPC, we notice that its behavior with respect to constraint violations varies significantly across the two datasets. On the other hand, SCPD appears to be more robust, performing well regardless of the dataset.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/



(a) convtype



(b) minst

Figure 7: Comparison on dataset *convtype* and *mnist* for solving (7.2)

8 Conclusion

In this paper we propose an algorithm framework ICPD of cubic-regularized primal-dual methods for equality constrained optimization. To update the primal variable at each iteration, we construct a subproblem based on a cubic model obtained from a quadratic approximation to the AL function plus a cubic regularizer. Under certain conditions on approximate gradients and Hessians of the objective function, as well as inexact subproblem solutions, we establish the iteration complexity of ICPD to find an ϵ-FSP and ϵ-SSP, respectively. We then extend the algorithm to a stochastic variant, suitable for problems where the objective function is in an expectation form. We investigate the oracle complexity regarding the total number of stochastic gradient and Hessian evaluations to reach approximate stationary points with high probability. To solve each subproblem with a random perturbation, we apply the standard gradient descent approach. We show that, under proper parameter settings, the required conditions imposed on the inexact subproblem solution can be satisfied with high probability at each iteration. Moreover, we provide theoretical analysis on the behaviour of Adaptive ICPD which updates the penalty parameter dynamically, and also discuss the applicability of adaptive cubic regularization parameters. Additionally, we present preliminary numerical results on two test problems to showcase the performance of the proposed algorithms. As far as we know, the oracle complexity bounds established in this paper for deterministic constrained optimization without using objective function values and for stochastic problems to reach first-order stationarity are comparative with existing algorithms. Moreover, as study on second-order stationarity for general nonlinear constrained stochastic optimization is a relatively unexplored area in the literature, our analysis in this paper regarding the second-order stationarity are novel.

Acknowledgments

The author would like to thank Luwei Bai for his kind help with part of the numerical implementations in the revised version of the paper. The author also sincerely appreciates the editor and two anonymous reviewers for their valuable comments and suggestions.

References

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. On augmented Lagrangian methods with general lower-level constraints. *SIAM J. Optim.*, 18:1286–1309, 2008.

[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] E.G. Birgin and J.M. Martt́nez. Complexity and performance of an augmented Lagrangian algorithm. *Optim. Methods Softw.*, 35:885–920, 2020.

[4] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Math. Program.*, 197:215–279, 2023.

[5] Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM J. Optim.*, 29:2146–2178, 2019.

[6] Y. Carmon and J.C. Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Rev.*, 2020.

[7] Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM J. Optim.*, 28:1751–1772, 2018.

[8] C. Cartis, N. Gould, and P. Toint. Second-order optimality and beyond: characterization and evaluation complexity in convexly constrained nonlinear optimization. *Found. Comput. Math.*, page 1–35, 2017.

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

[10] C. Cartis, N.I. Gould, and P.L. Toint. Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. *J. Complex.*, 53:68–94, 2019.

[11] C. Cartis, N.I.M. Gould, and P.L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Math. Program.*, 127:245–295, 2011.

[12] C Cartis, NIM Gould, and PL Toint. On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization. *SIAM J. Optim.*, 23(3), 2013.

[13] C Cartis, NIM Gould, and PL Toint. On the evaluation complexity of constrained nonlinear least-squares and general constrained nonlinear optimization using second-order methods. *SIAM J. Numer. Anal.*, 53(2):836–851, 2015.

[14] C Cartis, NIM Gould, and PL Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. SIAM, 2022.

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

[16] A.R. Conn, N.I.M. Gould, and P.L. Toint. *Trust Region Methods*. MPS-SIAM Series on Optimization, SIAM, 2000.

[17] 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.*, 2023.

[18] F.E. Curtis, D.P. Robinson, C.W. Royer, and S.J. Wright. Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization. *SIAM J. Optim.*, 31(1):518–544, 2021.

[19] 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.

[20] G. N. Grapiglia and Y. Yuan. On the complexity of an augmented Lagrangian method for nonconvex optimization. *IMA J. Numer. Anal.*, 41(2):1546–1568, 2021.

[21] G. Haeser, H. Liu, and Y. Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Math. Program.*, 178:263–299, 2019.

[22] C. He, Z. Lu, and T.K. Pong. A newton-cg based augmented lagrangian method for finding a second-order stationary point of nonconvex equality constrained optimization with complexity guarantees. *SIAM J. Optim.*, 33(3):1734–1766, 2023.

[23] M. Hong, M. Razaviyayn, and J.D. Lee. Gradient primal-dual algorithm converges to second-order stationary solution for nonconvex distributed optimization over networks. In *International Conference on Machine Learning*, 2018.

[24] R.J. Jiang, Z.S. Zhou, and Z.R. Zhou. Cubic regularization methods with second-order complexity guarantee based on a new subproblem reformulation. *J. Oper. Res. Soc. China*, pages 1–36, 2022.

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

[26] J.M. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. *In Proc. 34th International Conference on Machine Learning (ICML)*, 70:1895–1904, 2017.

[27] J.M. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. *ICML*, pages 1895–1904, 2017.

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

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

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

[31] S. Lu, M. Razaviyayn, B. Yang, K. Huang, and M. Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 2811–2822. Curran Associates, Inc., 2020.

[32] A. Mokhtari, A. Ozdaglar, and A. Jadbabaie. Escaping saddle points in constrained optimization. *NeurIPS*, 2018.

[33] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic pro- gramming with differentiable exact augmented lagrangians. *Math. Program*, 199:721–791, 2023.

[34] Sen Na, Mihai Anitescu, and Mladen Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Math. Program.*, 202:279–353, 2023.

[35] Y. Nesterov and B. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108:177–205, 2006.

[36] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2006.

[37] M. Nouiehed and M. Razaviyayn. A trust region method for finding second-order stationarity in linearly constrained nonconvex optimization. *SIAM J. Optim.*, 30(3):2501–2529, 2020.

[38] 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. *AAAI*, 33:4772–4779, 2019.

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

[40] C.W. Royer, M. O'Neill, and S.J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Math. Program.*, 180:451–488, 2020.

[41] C.W. Royer and S.J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM J. Optim.*, 28:1448–1477, 2018.

[42] M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. *NeurIPS*, 32, 2019.

[43] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization. *Optimization Online*, 2024.

[44] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *NeurIPS*, pages 2899–2908, 2018.

[45] J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8:1–1230, 2015.

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

[47] Xiao Wang and Ya xiang Yuan. An augmented Lagrangian trust region method for equality constrained optimization. *Optim. Methods Softw.*, 30:559–582, 2015.

[48] Z. Wang, Y. Zhou, Y. Liang, and G. Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR*, 89:2731–2740, 2019.

[49] Y. Xie and S.J. Wright. Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. *J. Sci. Comput.*, 86(38), 2021.

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

[51] Y. Xu. First-order methods for constrained convex programming based on linearized augmented Lagrangian function. *Informs J. Optim.*, 3:89–117, 2021.

[52] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Math. Program.*, 185:199–244, 2021.

[53] L. Zhang, Y. Zhang, X. Xiao, and J. Wu. Stochastic approximation proximal method of multipliers for convex stochastic programming. *Math. Oper. Res.*, 48(1):177–193, 2022.

[54] D. Zhou, P. Xu, and Q. Gu. Stochastic variance-reduced cubic regularization methods. *J. Machine Learn. Res.*, 20:1–47, 2019.

[55] D. Zhou, P. Xu, and Q. Gu. Stochastic variance-reduced cubic regularization methods. *J. Machine Learn. Res.*, 2019.

[56] 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.