Computational Optimization and Applications
https://doi.org/10.1007/s10589-022-00358-y

Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization

Qihang Lin[1] · **Runchao Ma**[1] · **Yangyang Xu**[2]

Received: 30 November 2020 / Accepted: 4 February 2022

Abstract

In this paper, an inexact proximal-point penalty method is studied for constrained optimization problems, where the objective function is non-convex, and the constraint functions can also be non-convex. This method approximately solves a sequence of subproblems, each of which is formed by adding to the original objective function a proximal term and quadratic penalty terms associated to the constraint functions. Under a weak-convexity assumption, each subproblem is made strongly convex and can be solved effectively to a required accuracy by an optimal gradient-based method. The computational complexity of this approach is analyzed separately for the cases of convex constraint and non-convex constraint. For both cases, the complexity results are established in terms of the number of proximal gradient steps needed to find an ϵ-stationary point. When the constraint functions are convex, we show a complexity result of $\tilde{O}(\epsilon^{-5/2})$ to produce an ϵ-stationary point under the Slater's condition. When the constraint functions are non-convex, the complexity becomes $\tilde{O}(\epsilon^{-3})$ if a non-singularity condition holds on constraints and otherwise $\tilde{O}(\epsilon^{-4})$ if a feasible initial solution is available.

Keywords Constrained optimization · Nonconvex optimization · Proximal-point method · Penalty method

✉ Qihang Lin
qihang-lin@uiowa.edu

Runchao Ma
runchao-ma@uiowa.edu

Yangyang Xu
xuy21@rpi.edu

[1] Department of Business Analytics, University of Iowa, Iowa City, IA 52242, USA

[2] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Published online: 03 March 2022

1 Introduction

We consider the nonconvex optimization problem with inequality and equality constraints:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f_0(\mathbf{x}) + g(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{f}(\mathbf{x}) \le \mathbf{0}, \quad \mathbf{c}(\mathbf{x}) = \mathbf{0}, \tag{1.1}$$

where $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $\mathbf{f} = [f_1, \ldots, f_m]$ with $f_i : \mathbb{R}^d \to \mathbb{R}$ for each $i = 0, \ldots, m$, and $\mathbf{c} = [c_1, \ldots, c_n]$ with $c_j : \mathbb{R}^d \to \mathbb{R}$ for each $j = 1, \ldots, n$. We assume that g is a proper lower-semicontinuous convex function with a compact domain and all other functions are continuously differentiable.

For a general non-convex function, finding its global minimizer is intractable, and it becomes even more difficult, when there are (non-convex) constraints. Therefore, instead of finding a global minimizer of (1.1), we focus on finding a stationary point. We call a point $\mathbf{x}^* \in \mathrm{dom}(g)$ a *stationary* point of (1.1), if there are $\boldsymbol{\lambda}^* \in \mathbb{R}^m_+$ and $\mathbf{y}^* \in \mathbb{R}^n$, which exist if some constraint qualification is assumed, such that the Karush-Kuhn-Tucker (KKT) conditions hold:

$$\mathbf{0} \in \nabla f_0(\mathbf{x}^*) + J_{\mathbf{f}}(\mathbf{x}^*)^\top \boldsymbol{\lambda}^* + J_{\mathbf{c}}(\mathbf{x}^*)^\top \mathbf{y}^* + \partial \quad g(\mathbf{x}^*), \tag{1.2a}$$

$$f_i(\mathbf{x}^*) \le 0,\ i = 1, \ldots, m, \quad c_j(\mathbf{x}^*) = 0,\ j = 1, \ldots, n, \tag{1.2b}$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0,\ i = 1, \ldots, m, \tag{1.2c}$$

where $\partial g(\mathbf{x}^*)$ denotes the subdifferential of g at $\mathbf{x}^*$, $J_{\mathbf{f}}(\mathbf{x}^*)$ denotes the Jacobian matrix of $\mathbf{f}$ at $\mathbf{x}^*$, and $J_{\mathbf{c}}(\mathbf{x}^*)$ denotes the Jacobian matrix of $\mathbf{c}$ at $\mathbf{x}^*$. The vectors $\boldsymbol{\lambda}^*$ and $\mathbf{y}^*$ are called Lagrangian multipliers. Due to the inevitable truncation error, it is hard to compute a solution that satisfies the above conditions exactly. Numerically, it is more reasonable to pursue an approximate stationary point defined as follow. Here, $\|\cdot\|$ stands for the Euclidean norm.

Definition 1 (ε-stationary point and its weak version) Given $\varepsilon > 0$, a point $\bar{\mathbf{x}}$ is an ε -stationary point of (1.1) if there are $\bar{\boldsymbol{\xi}} \in \partial g(\bar{\mathbf{x}})$, $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^m_+$, and $\bar{\mathbf{y}} \in \mathbb{R}^n$ such that $\bar{\lambda}_i = 0$ if $f_i(\bar{\mathbf{x}}) < 0$ for $i = 1, \ldots, m$ and

$$\left\| \nabla f_0(\bar{\mathbf{x}}) + J_{\mathbf{f}}(\bar{\mathbf{x}})^\top \bar{\boldsymbol{\lambda}} + J_{\mathbf{c}}(\bar{\mathbf{x}})^\top \bar{\mathbf{y}} + \bar{\boldsymbol{\xi}} \right\| \le \varepsilon, \tag{1.3a}$$

$$\sqrt{\|\mathbf{c}(\bar{\mathbf{x}})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\|^2} \le \varepsilon, \tag{1.3b}$$

$$\sum_{i=1}^{m} |\bar{\lambda}_i f_i(\bar{\mathbf{x}})| \le \varepsilon. \tag{1.3c}$$

If only (1.3a) and (1.3b) hold, $\bar{\mathbf{x}}$ is called a *weak ε-stationary point* of (1.1).

Here, the three conditions in (1.3) are ϵ-approximation of the three conditions in (1.2) while the condition that $\bar{\lambda}_i = 0$ if $f_i(\bar{\mathbf{x}}) < 0$ essentially requires the complementary slackness condition in (1.2) holds exactly when $f_i(\bar{\mathbf{x}}) < 0$. When there are only equality conditions, Definition 1 is the same as that for the ϵ-approximate first-order solution considered in several existing papers, e.g., [60, 73]. When there are inequality constraints, the ϵ-stationary solution in Definition 1 is stronger than the solutions guaranteed by [5, 29], which only requires $\bar{\lambda}_i = 0$ if $f_i(\bar{\mathbf{x}}) < -\epsilon$. A different definition of approximate stationary point is considered in [8, 47] where the objective and constraint functions can be non-smooth. The differences between their definitions and Definition 1 are discussed in Appendix 3.1.

Our goal is to establish the theoretical complexity of finding an ϵ-stationary point or a weak ϵ-stationary point of (1.1). To achieve this goal, we consider an *inexact proximal-point penalty* (iPPP) method (see Algorithm 1 below). Our method solves a sequence of strongly-convex unconstrained subproblems that are constructed by combining two classical techniques: the proximal-point method and the quadratic penalty method; see (4.1) below. The *adaptive accelerated proximal gradient* (AdapAPG) method by [45, 54] (see Algorithm 4) is applied to approximately solve each subproblem. To show the complexity results, we consider two cases of (1.1) separately and assume different regularity conditions for them. In the first case, the problem has a weakly-convex objective (see Definition 5) but convex constraint functions, and we assume Slater's condition. In the second case, the objective and constraint functions are all weakly convex, and we assume either a non-singularity condition (see Assumption 4) or the feasibility of the initial solution (see Assumption 5).

1.1 Contributions

We make contributions to understanding the theoretical complexity of finding an ϵ -stationary point of a non-convex constrained problem in the form of (1.1). Three scenarios are studied and the computational complexity of the iPPP method, measured by the number of proximal gradient steps, is established in each scenario.

They are summarized as follows.

- For the case where f_0 is weakly convex, f_i is convex for $i = 1, \ldots, m$, and c_j is affine for $j = 1, \ldots, n$, we show that, when Slater's condition holds, the proposed iPPP method can find an ϵ-stationary point within $\tilde{O}(\epsilon^{-5/2})$ proximal gradient steps.[1] *This complexity is first achieved by this paper and remains by far the best complexity for* (1.1) *under these assumptions.*
- When $\{f_i\}_{i=0}^m$ and $\{c_j\}_{j=1}^n$ are all weakly convex, we show that, if a non-singularity condition (see Assumption 4) is satisfied by the constraint functions, the iPPP method can find an ϵ-stationary point within $\tilde{O}(\epsilon^{-3})$ proximal gradient steps. This

[1] Here and in the rest of paper, we suppress all logarithmic terms in $\tilde{O}$.

complexity improves the one $\tilde{O}(\varepsilon^{-4})$ achieved in [60] that uses an inexact augmented Lagrangian method under the same assumptions.[2]

– When $\{f_i\}_{i=0}^m$ and $\{c_j\}_{j=1}^n$ are all weakly convex, we show that, if an initial feasible solution is available (but the aforementioned non-singularity condition is not needed), the iPPP method can find a *weak* ε-stationary point within $\tilde{O}(\varepsilon^{-4})$ proximal gradient steps. In Sect. 2, we will discuss how this result is compared with other works that also consider non-convex constraints without the non-singularity condition.

1.2 Organization of the paper

The rest of the paper is organized as follows. In Sect. 2, we discuss related works on convex and non-convex constrained optimization. In Sect. 3, we introduce some definitions, notations, and a subproblem to solve in the proposed algorithm. Details of the proposed algorithm are described in Sect. 4. The complexity analysis is conducted in Sect. 5 for the convex constrained case and in Sect. 6 for the non-convex constrained case. Numerical results are presented in Sects. 7 and 8 concludes the paper.

2 Related works

There has been growing interest in first-order algorithms for non-convex minimization problems with no constraints or simple constraints [3] in both stochastic and deterministic settings. See, e.g., [1, 16–18, 21, 25, 26, 39, 58, 82]. However, for (1.1) with constraints that are not simple, these methods are not applicable. There is a long history of studies on continuous optimization with constraints. The recent works on first-order methods for convex optimization with convex constraints include [4, 44, 62, 71, 74–78] for deterministic constraints and [3, 40, 79] for stochastic constraints. Different from these works, this paper studies the problems with a non-convex objective function and with potentially non-convex constraints.

When all constraint functions in (1.1) are affine, a primal-dual Frank-Wolfe method is proposed in [70], and it finds an ε-stationary point with a complexity of $O(\varepsilon^{-3})$ in general and $O(\varepsilon^{-2})$ when there exists a strictly feasible solution. We adopt a notion of ε-stationary point different from that in [70], and our constraint functions can be nonlinear and non-convex.

As a classical approach for solving problems in the form of (1.1), a penalty method finds an approximate solution by solving a sequence of unconstrained

[2] A complexity of $\tilde{O}(\varepsilon^{-3})$ is claimed in Corollary 4.2 in [60]. However, there is an error in its proof. The authors claimed the complexity of solving their subproblem is $O(\frac{\lambda_{\beta_k}^2 \rho^2}{\epsilon_{k+1}})$ but it should be $O(\frac{\lambda_{\beta_k}^2 \rho^2}{\epsilon_{k+1}^2})$. (See [60] for the definitions of λ_{β_k}, ρ and ϵ_{k+1}.) After correcting this error, following the same proof they used gives a total complexity of $\tilde{O}(\epsilon^{-4})$.

[3] Here, simple constraints mean the constraints allow a closed-form projection onto the feasible set.

subproblems, where the violation of constraints is penalized by the positively weighted penalty terms in the objective function of the subproblems. Unconstrained optimization techniques are then applied to the subproblems along with an updating scheme for the weighting parameters. The computational complexity of penalty methods for convex problems has been well established [38, 51, 61]. For non-convex problems, most existing studies focus on the asymptotic convergence to a stationary point. See, e.g., [9, 10, 19, 20, 23, 28, 57]. On the contrary, we analyze the finite complexity of penalty methods for finding an ϵ-stationary point.

An exact penalty method has been studied in [11] as an application of a trust region method for a composite non-smooth problem. When applied to (1.1) with $g \equiv 0$, the method in [11] either finds an ϵ-infeasible and ϵ-critical point of (1.1) (see [11] for the definition) or finds a solution that is infeasible to (1.1) but ϵ-critical to the infeasibility measure $\sum_{j=1}^{n} \left|c_j(\mathbf{x})\right| + \sum_{i=1}^{m} \max\{f_i(\mathbf{x}), 0\}$. It needs to exactly solve $O(\epsilon^{-2})$ linearized trust-region subproblems if the penalty parameter is bounded above and solve $O(\epsilon^{-5})$ subproblems otherwise. In a subsequent study [12] and its corrigendum [13], a target-following algorithm is developed. It can find an approximate Fritz-John (instead of KKT) solution with similar guarantee as [11] by solving $O(\epsilon^{-2})$ subproblems regardless of the boundedness of penalty parameter. This method has been extended to the case when f_0 is the expectation of a stochastic function in [68]. We want to emphasize that the complexity result of [11, 12] is given in terms of the number of exactly-solved trust-region subproblems, and thus it is not exactly computational (time) complexity, especially when the subproblem is not trivially solvable. On the contrary, we directly analyze the total computational (time) complexity of the proposed method. When the constraints are non-convex, our method has complexity $\tilde{O}(\epsilon^{-3})$ and $\tilde{O}(\epsilon^{-4})$, respectively, when a non-singularity condition (Assumption 4) is assumed and when an feasible initial solution (Assumption 5) is assumed. Neither assumption is needed in [11]. Suppose the time complexity of solving a trust-region subproblem in [11] is the same as a proximal gradient step in our method. The complexity of [11] is lower than ours if their penalty parameters are bounded and higher than ours, otherwise. Moreover, the method by [11] did not always guarantee an ϵ-feasible solution while our method does, which is mainly because of Assumptions 4 or 5 we make.

On solving a problem with a non-convex objective and linear constraint, [36] has developed a quadratic-penalty accelerated inexact proximal point method. That method can generate an ϵ-stationary point in the sense of (1.3) with a complexity of $O(\epsilon^{-3})$. Our method is similar to that in [36] by utilizing the techniques from both the proximal point method and the quadratic penalty method. Although we make a little stronger assumption than [36] by requiring the boundedness of $\mathrm{dom}(g)$, our method and analysis apply to the problems with non-convex objectives and convex/non-convex nonlinear constraint functions. When the constraints are convex (but possibly nonlinear), our method can find an ϵ-stationary point with a complexity of $\tilde{O}(\epsilon^{-5/2})$ that is a nearly $O(\epsilon^{-1/2})$ improvement over the complexity in [36].

Barrier methods are another traditional class of algorithms for constrained optimization. Similar to penalty methods, they also solve a sequence of unconstrained subproblems with barrier functions added to the objective function. The barrier

functions will increase to infinity as the iterates approach the boundary of the feasible set, and thus enforce the iterates to stay in the interior of the feasible set. The convergence rate of barrier methods has been studied by [53, 63–65] for convex problem. For a general non-convex problems, most studies only focus on asymptotic convergence analysis. Recent works [30, 56] proposed algorithms based on logarithmic barrier function for non-convex problems with only non-negative and linear constraints. They established the complexity of their algorithms for finding first-order and second-order ε-KKT point (whose definitions are slightly different in [30, 56] and different from our definition). However, they do not consider nonlinear constraints as we do.

The augmented Lagrangian method (ALM) is another effective approach for constrained optimization. At each iteration, ALM updates the primal variable by minimizing the augmented Lagrangian function and then performs a dual gradient ascent step to update the dual variable. The iteration complexity of ALM has been established for convex problems [38, 51, 74–76]. For non-convex problems, asymptotic convergence or local convergence rate of ALM has been studied by [6, 7, 15, 22, 24, 66, 69]. The computational complexity of ALM and its variants (e.g. ADMM) for finding an ε-stationary point for linearly constrained non-convex problems has been studied by [27, 31, 32, 34, 48–50, 80, 81]. For example, the proximal inexact ALM method by [50] achieves complexity of $O(\varepsilon^{-5/2})$ and a related but different ALM method by [80, 81] achieves complexity of $O(\varepsilon^{-2})$, the latter of which is by far the best result for nonconvex problems with linear constraints.

ALM and its proximal variant are analyzed by [5, 29, 37, 41, 42, 60, 73] for non-convex problems with nonlinear constraints. In each main iteration of those methods, an approximate stationary point of the (proximal) augmented Lagrangian function is computed by first- or second-order methods. Utilizing the Hessian information, the methods in [60, 73] can find a second-order ε-stationary point while our method cannot. Without Hessian information, the methods by [60, 73] can still find a first-order ε-stationary point. In [60], under a *non-singularity assumption* that the smallest singular value of the Jacobian matrix of the constraint functions is uniformly bounded away from zero, it is showed that ALM finds an ε-stationary point with complexity of $\tilde{O}(\varepsilon^{-4})$.

The complexity of [73] is also $O(\varepsilon^{-4})$ if we set the parameter η in their algorithm to the optimal value, i.e., zero; see Theorem 2 in [73].

On the contrary, our method has a complexity of $\tilde{O}(\varepsilon^{-3})$ for problems with non-convex constraints under the assumptions similar to [60, 73] and only has a complexity of $\tilde{O}(\varepsilon^{-2.5})$ for convex constrained problems. Moreover, we consider both inequalities and equality constraints while [60, 73] only consider equality constraints.

In addition, even if the non-singularity assumption does not hold, our method can still find an ε-stationary point as long as an initial feasible solution is available. This result benefits the applications where the constraints are non-convex but have some special structure that allows finding a feasible solution easily (e.g. [67] and [35]). After the release of the first draft [43] of this paper, [42] gave a hybrid of the quadratic penalty method and ALM, which also achieves an $\tilde{O}(\varepsilon^{-2.5})$ complexity for non-convex problems with convex constraints under the same assumptions as we make in Assumptions 1 and 2. However, [42] shows that the complexity of the pure-ALM-based

first-order method is $\tilde{O}(\varepsilon^{-3})$. Thus the usage of the quadratic penalty method in the hybrid method of [42] is the key to obtain the $\tilde{O}(\varepsilon^{-2.5})$ complexity. The $\tilde{O}(\varepsilon^{-3})$ complexity result has also been established in [37] for a proximal ALM on solving nonconvex problems with nonlinear convex constraints. [41] adopts a proximal-point based subroutine and improves to $\tilde{O}(\varepsilon^{-3})$ the complexity result of the first-order ALM in [60] for equality-constrained nonconvex problems.

In [5], the authors assume neither the non-singularity assumption nor a feasible initial solution while are still able to achieve $O(\varepsilon^{-3})$ complexity for ALM. However, in their setting, ALM does not necessarily guarantee an ε-stationary point of (1.1) but may only return a point that is infeasible to (1.1) and ε-stationary to the infeasibility measure (similar to the guarantee by [11]). The totaly number of iterations needed by ALM is also analyzed by [29] when the constraints are linear or quadratic. However, they solve the ALM subproblems by a second-order or high-order method so their complexity per iteration can be much higher than ours when the problem's dimension is high. Finally, we want to emphasize again that the ε-stationary point we consider in Definition 1 is stronger than the solutions guaranteed by [5, 29] which do not satisfy (1.3c) and only satisfy $\bar{\lambda}_i = 0$ when $f_i(\bar{\mathbf{x}}) < -\varepsilon$.

In addition to [60, 73], the algorithms by [33, 46, 55] also utilize Hessian information to find a second-order ε-stationary point for linearly constrained non-convex optimization. Different from these works, we focus on finding an approximate first-order stationary point for nonlinear constrained non-convex optimization using only gradient information.

Two recent works [8, 47] proposed similar algorithms for non-convex constrained optimization based on the proximal-point technique. In their approaches, a strongly convex constrained subproblem is constructed in each main iteration by adding proximal terms to the objective and constraints. When applied to non-convex smooth constrained optimization, both methods find an ε-stationary point in complexity of $\tilde{O}(\varepsilon^{-3})$. The definitions of an ε-stationary point in [8, 47] are different from ours, and the differences are discussed in Appendix 3.1. Their analysis requires a (nearly) feasible initial solution and uniform boundedness of the dual solutions of all subproblems. To satisfy the latter requirement, [47] assumes that a uniform Slater's condition holds and [8] assumes that the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at the limiting points of the generated iterates. On the contrary, when the non-singularity assumption (Assumption 4) holds, our method also has complexity $\tilde{O}(\varepsilon^{-3})$ but does not require a (nearly) feasible initial solution. The uniform Slater's condition in [47] and MFCQ in [8] do not imply the non-singularity assumption, and our non-singularity assumption does not imply their assumptions either. See Appendix 3.2 for the related examples. When an initial feasible solution is indeed available, our method can be analyzed in an alternative way without the non-singularity assumption or the constraint qualification conditions required by [8, 47], although the complexity becomes $\tilde{O}(\varepsilon^{-4})$.

3 Preliminary

In this section, we provide some basic definitions and discuss about a subproblem solved in each main iteration of our algorithm.

3.1 Definitions and Assumptions

We denote $\|\cdot\|$ as the ℓ_2-norm and $\|\cdot\|_1$ as the ℓ_1-norm. Let

$$\mathcal{X} = \mathrm{dom}\,(g) := \{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) < +\infty\} \tag{3.1}$$

be the domain of a function g. The interior and boundary of $\mathcal{X}$ are respectively denoted by $\mathrm{int}(\mathcal{X})$ and $\partial\mathcal{X}$. We use $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ for the normal cone of $\mathcal{X}$ at $\mathbf{x}$. Given $a > 0$, we use $\mathcal{B}_a$ to represent the ball $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \le a\}$. We denote $\mathbf{0}$ as an all-zero vector whose dimension is clear from the context, and $[\mathbf{a}]_+ = \max\{\mathbf{0}, \mathbf{a}\}$ is the vector of component-wise maximum between $\mathbf{0}$ and $\mathbf{a}$. For a convex set $\mathcal{S}$, we use $\mathrm{dist}(\mathbf{x}, \mathcal{S}) = \min_{\mathbf{y}\in\mathcal{S}} \|\mathbf{y} - \mathbf{x}\|$ for the distance of $\mathbf{x}$ to $\mathcal{S}$. For any $\mathbf{x} \in \mathbb{R}^d$, $J_{\mathbf{f}}(\mathbf{x}) = [\nabla f_1(\mathbf{x}), \ldots, \nabla f_m(\mathbf{x})]^\top \in \mathbb{R}^{m\times d}$ and $J_{\mathbf{c}}(\mathbf{x}) = [\nabla c_1(\mathbf{x}), \ldots, \nabla c_n(\mathbf{x})]^\top \in \mathbb{R}^{n\times d}$ denote the Jacobian matrices of $\mathbf{f}$ and $\mathbf{c}$ at $\mathbf{x}$, respectively.

We adopt the following definitions.

Definition 2 (subdifferential) Given a proper lower-semicontinuous convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its subdifferential at any $\mathbf{x}$ in the domain is defined as

$$\partial h(\mathbf{x}) = \left\{\boldsymbol{\zeta} \in \mathbb{R}^d \,\middle|\, h(\mathbf{x}') \ge h(\mathbf{x}) + \boldsymbol{\zeta}^\top(\mathbf{x}' - \mathbf{x}), \forall \mathbf{x}' \in \mathbb{R}^d\right\},$$

and each $\boldsymbol{\zeta} \in \partial h(\mathbf{x})$ is called a subgradient of h at $\mathbf{x}$.

Definition 3 (L-smoothness) A function $h : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if it is differentiable on $\mathbb{R}^d$ and satisfies

$$h(\mathbf{x}) \le h(\mathbf{x}') + \langle \nabla h(\mathbf{x}'), \mathbf{x} - \mathbf{x}'\rangle + \frac{L}{2}\|\mathbf{x}' - \mathbf{x}\|^2, \ \forall\, \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d. \tag{3.2}$$

Definition 4 (μ-strong convexity) A function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is μ-strongly convex for $\mu \ge 0$ if $h - \frac{\mu}{2}\|\cdot\|^2$ is convex. When $\mu = 0$, μ-strong convexity is reduced to convexity.

Definition 5 (ρ-weak convexity) A function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is ρ-weakly convex for $\rho \ge 0$ if $h + \frac{\rho}{2}\|\cdot\|^2$ is convex. When $\rho = 0$, ρ-weak convexity is reduced to convexity.

When h is smooth and ρ-weakly convex, it holds that for any $\mathbf{x}$ and $\mathbf{x}'$,

$$h(\mathbf{x}) \ge h(\mathbf{x}') + \left\langle \nabla h(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \right\rangle - \frac{\rho}{2}\|\mathbf{x}' - \mathbf{x}\|^2. \tag{3.3}$$

Definition 6 (proximal mapping) Given a proper lower-semicontinuous convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its *proximal mapping* at $\mathbf{x}$ is defined as

$$\mathbf{prox}_h(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \left\{ h(\mathbf{z}) + \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 \right\}.$$

The following assumption on problem (1.1) is made throughout the paper.

Assumption 1 The following statements hold:

A. f_i is L_{f_i}-smooth with $L_{f_i} \geq 0$ for $i = 0, 1, \ldots, m$; c_j is L_{c_j}-smooth with $L_{c_j} \geq 0$ for $j = 1, \ldots, n$;
B. $\mathcal{X}$ is compact, and its diameter is denoted by $D = \max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x}-\mathbf{x}'\|$;
C. There exist constants G and M such that $|g(\mathbf{x})| \leq G$, $\partial g(\mathbf{x}) \neq \emptyset$, and $\partial g(\mathbf{x}) \subseteq \mathcal{N}_{\mathcal{X}}(\mathbf{x}) + \mathcal{B}_M, \forall \mathbf{x} \in \mathcal{X}$.
D. $\mathbf{prox}_g(\mathbf{x})$ can be computed easily, e.g., in a closed form.

With Assumption 1A and 1B, there must exist constants $\{B_{f_i}\}_{i=0}^m$ and $\{B_{c_j}\}_{j=1}^n$ such that

$$\max\left\{|f_i(\mathbf{x})|, \|\nabla f_i(\mathbf{x})\|\right\} \leq B_{f_i}, \ \forall \mathbf{x} \in \mathcal{X}, \ \forall\, i = 0, 1, \ldots, m, \tag{3.4a}$$

$$\max\left\{|c_j(\mathbf{x})|, \|\nabla c_j(\mathbf{x})\|\right\} \leq B_{c_j}, \ \forall \mathbf{x} \in \mathcal{X}, \ \forall\, j = 1, \ldots, n. \tag{3.4b}$$

Assumption 1C holds, for example, if $g(\mathbf{x}) = r(\mathbf{x}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x})$, where $\mathbf{1}_{\mathcal{X}}$ is the indicator function on $\mathcal{X}$, and r is a real-valued function with the norm of every subgradient bounded by M. In addition to Assumption 1, we will make more assumptions on the (weak) convexity of the constraint functions. Details will be given in Sects. 5 and 6, where we conduct the complexity analysis.

3.2 Strongly convex composite subproblem

In each main iteration of the algorithm we propose for (1.1), a *strongly convex composite optimization* subproblem of the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ F(\mathbf{x}) := \phi(\mathbf{x}) + r(\mathbf{x}) \right\} \tag{3.5}$$

will be approximately solved, where $\phi : \mathbb{R}^d \to \mathbb{R}$ is μ_ϕ-strongly convex and L_ϕ-smooth, and $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower-semicontinuous convex function. Function r will be g in (1.1) but function ϕ will vary with the main iteration. For each subproblem, we need to find a solution $\bar{\mathbf{x}}$ satisfying

$$\operatorname{dist}\big(-\nabla\phi(\bar{\mathbf{x}}), \partial r(\bar{\mathbf{x}})\big) := \min_{\boldsymbol{\xi}'\in\partial r(\bar{\mathbf{x}})} \|\nabla\phi(\bar{\mathbf{x}}) + \boldsymbol{\xi}'\| \leq \hat{\epsilon}, \tag{3.6}$$

where the value of the left-hand side measures the suboptimality of $\bar{\mathbf{x}}$ to (3.5) and $\hat{\epsilon}$ is the targeted suboptimality that decreases with the main iteration. To find $\bar{\mathbf{x}}$, we

can solve (3.5) by an *accelerated proximal gradient* (APG) method which is a first-order method and whose main step per iteration is a *proximal gradient step*, namely, computing

$$T_L(\mathbf{w}) := \mathbf{prox}_{L^{-1}r}(\mathbf{w} - L^{-1}\nabla\phi(\mathbf{w})) \tag{3.7}$$

for some $\mathbf{w} \in \mathbb{R}^d$ and $L > 0$. In this paper, the *complexity* of the APG method and our method is measured by the total number of proximal gradient steps they perform.

The standard APG method (e.g., [52]) requires knowing the exact values of μ_ϕ and L_ϕ which may be unknown. To address this issue, *adaptive accelerated proximal gradient* (AdapAPG) methods [45, 54] have been developed to dynamically estimate μ_ϕ and L_ϕ during the algorithm at the cost of a little higher complexity than APG. We will apply the AdapAPG method in [45] to our subproblems and the total complexity of our algorithm will be derived using the complexity of AdapAPG for solving each subproblem with a specific level of suboptimality (i.e., $\hat{\varepsilon}$). That said, describing the AdapAPG method in details requires introducing additional notations and technical results, which are not necessary for readers to understand the proposed algorithm and main theories in this paper. Hence, to avoid interrupting the flow of the presentation, the details of the AdapAPG method are postponed to Appendix 1, and here we only present its complexity in terms of μ_ϕ, L_ϕ and $\hat{\varepsilon}$ in the following theorem, which is sufficient to derive the complexity of our algorithm. The exact complexity of AdapAPG in terms of all related parameters is presented in Theorem 5 in Appendix 1.

Theorem 1 (Complexity of AdapAPG) *When applied to* (3.5), *the AdapAPG method by* [45] *terminates in* $\tilde{O}\left(\sqrt{\frac{L_\phi}{\mu_\phi}}\log\left(\frac{1}{\hat{\varepsilon}}\right)\right)$ *proximal gradient steps with an ouput* $\bar{\mathbf{x}}$ *satisfying* (3.6).

4 Inexact proximal-point penalty methods

In this section, we describe the inexact proximal-point penalty (iPPP) method for (1.1) in details. This method incorporates the ideas of the proximal point method and the quadratic penalty method by iteratively updating the estimated solution $\bar{\mathbf{x}}^{(k)}$ as follows

$$\bar{\mathbf{x}}^{(k+1)} \approx \widetilde{\mathbf{x}}^{(k)} := \arg\min_{\mathbf{x}\in\mathbb{R}^d} f_0(\mathbf{x}) + g(\mathbf{x}) + \frac{\gamma_k}{2}\|\mathbf{x} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{\beta_k}{2}\left(\|\mathbf{c}(\mathbf{x})\|^2 + \left\|[\mathbf{f}(\mathbf{x})]_+\right\|^2\right), \tag{4.1}$$

where $\beta_k > 0$ is the penalty parameter, and $\gamma_k > 0$ is the proximal parameter. We formally describe our method in Algorithm 1, where ϕ_k in (4.5) is the smooth part of the objective function in (4.1). In iteration k, we only need to guarantee (4.6) through solving (4.1). When all f_i's and c_j's are weakly convex, a sufficiently large γ_k can be chosen such that the minimization problem in (4.1) becomes strongly convex.

Then the AdapAPG method can be applied to (4.1) to obtain $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6); see Theorem 1. In addition to $\bar{\mathbf{x}}^{(k+1)}$, Algorithm 1 also computes the Lagrangian multipliers $\bar{\mathbf{y}}^{(k+1)}$ and $\bar{\boldsymbol{\lambda}}^{(k+1)}$, as well as the following three quantities

$$\mathbf{S}_{k+1} = \operatorname{dist}\left(\nabla f_0(\bar{\mathbf{x}}^{(k+1)}) + J_{\mathbf{f}}(\bar{\mathbf{x}}^{(k+1)})^\top \bar{\boldsymbol{\lambda}}^{(k+1)} + J_{\mathbf{c}}(\bar{\mathbf{x}}^{(k+1)})^\top \bar{\mathbf{y}}^{(k+1)},\ -\partial g(\bar{\mathbf{x}}^{(k+1)})\right) \tag{4.2a}$$

$$\mathbf{F}_{k+1} = \sqrt{\|\mathbf{c}(\bar{\mathbf{x}}^{(k+1)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2} \tag{4.2b}$$

$$\mathbf{C}_{k+1} = \sum_{i=1}^{m} |\bar{\lambda}_i^{(k+1)} f_i(\bar{\mathbf{x}}^{(k+1)})|, \tag{4.2c}$$

which correspond to the three inequalities in (1.3) and will be used to select the output solution from $\{\bar{\mathbf{x}}^{(l)}\}_{l=1}^{k+1}$. In particular, depending on if the goal is to find an ϵ -stationary point or its weak version, Algorithm 1 will return $\bar{\mathbf{x}}^{(R_{k+1})}$ with

$$R_{k+1} = \arg\min_{1\le l\le k+1} \max\left\{\mathbf{S}_l, \mathbf{F}_l, \mathbf{C}_l\right\} \tag{4.3}$$

$$R_{k+1} = \arg\min_{1\le l\le k+1} \max\left\{\mathbf{S}_l, \mathbf{F}_l\right\}. \tag{4.4}$$

Algorithm 1 Inexact Proximal-Point Penalty (iPPP) Method for (1.1)

1: **Input:** Initial solution $\bar{\mathbf{x}}^{(0)}$, proximal parameters $\{\gamma_k\}_{k\ge 0}$, penalty parameters $\{\beta_k\}_{k\ge 0}$, and the targeted optimality measure for subproblems $\{\hat{\varepsilon}_k\}_{k\ge 0}$.
2: **for** $k = 0, 1, \ldots,$ **do**
3: Let

$$\phi_k(\mathbf{x}) := f_0(\mathbf{x}) + \frac{\gamma_k}{2}\|\mathbf{x} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{\beta_k}{2}\left(\|\mathbf{c}(\mathbf{x})\|^2 + \|[\mathbf{f}(\mathbf{x})]_+\|^2\right), \tag{4.5}$$

4: Fine $\bar{\mathbf{x}}^{(k+1)}$ such that

$$\min_{\boldsymbol{\xi}'\in\partial g(\bar{\mathbf{x}}^{(k+1)})} \|\nabla\phi_k(\bar{\mathbf{x}}^{(k+1)}) + \boldsymbol{\xi}'\| \le \hat{\varepsilon}_k. \tag{4.6}$$

5: Set $\bar{\mathbf{y}}^{(k+1)} \leftarrow \beta_k \mathbf{c}(\bar{\mathbf{x}}^{(k+1)})$ and $\bar{\boldsymbol{\lambda}}^{(k+1)} \leftarrow \beta_k [\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+$
6: Compute $\mathbf{S}_{k+1}$, $\mathbf{F}_{k+1}$, and $\mathbf{C}_{k+1}$ according to (4.2)
7: **if** a stopping condition is satisfied **then**
8: Return $\bar{\mathbf{x}}^{(R_{k+1})}$ and stop, where R_{k+1} is defined in (4.3) or (4.4)
9: **end if**
10: **end for**

Remark 1 Computing $\mathbf{S}_{k+1}$ in (4.2a) requires projection onto $-\partial g(\bar{\mathbf{x}}^{(k+1)})$, which is also needed when evaluating the stopping condition $\omega(\mathbf{x}^{(t+1)})$ in Algorithm 4. We have assumed g to be simple enough to allow a closed-form solution for the proximal gradient step. For such a g, this projection is usually no harder than a proximal gradient step, e.g., when $g(\mathbf{x}) = \mathbf{1}_{\mathcal{X}}(\mathbf{x})$ where $\mathcal{X}$ is a Euclidean ball or a box.

In the rest of the paper, we analyze the theoretical properties of Algorithm 1. For technical reasons, we consider two different cases. In the first case, we assume that the objective function is weakly convex while the constraint functions are convex. In the second case, we assume that the objective function and the constraint functions are all weakly convex. The parameters $\{\gamma_k\}$ and $\{\beta_k\}$ will be chosen differently for the two cases. We will show that the output of Algorithm 1 with appropriate settings is an ϵ-stationary point or a weak ϵ-stationary point of (1.1) in each case. We will also analyze the computational complexity of Algorithm 1, measured by the total number of proximal gradient steps it performs.

Lemma 1 *Suppose ϕ_k in (4.5) is convex. Let $\{\bar{\mathbf{x}}^{(k)}\}$ be generated from Algorithm 1. Then for any $\mathbf{x} \in \mathcal{X}$, it holds that*

$$\phi_k(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \phi_k(\mathbf{x}) - g(\mathbf{x}) \le \hat{\epsilon}_k D, \ \forall k \ge 0 \tag{4.7}$$

and that

$$\begin{aligned}
&\sum_{k=0}^{K-1} \frac{\gamma_k}{2} \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{\beta_{K-1}}{2} \left(\|\mathbf{c}(\bar{\mathbf{x}}^{(K)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(K)})]_+\right\|^2 \right) \\
&\le 2B_{f_0} + 2G + \frac{\beta_0}{2} \left(\|\mathbf{c}(\mathbf{x}^{(0)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 \right) \\
&+ \frac{1}{2} \sum_{k=1}^{K-1} (\beta_k - \beta_{k-1}) \left(\|\mathbf{c}(\bar{\mathbf{x}}^{(k)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2 \right) + \left(\sum_{k=0}^{K-1} \hat{\epsilon}_k \right) D, \ \forall K \ge 1.
\end{aligned} \tag{4.8}$$

Proof According to Line 4 of Algorithm 1, there exists $\bar{\boldsymbol{\xi}}^{(k+1)} \in \partial g(\bar{\mathbf{x}}^{(k+1)})$ such that $\|\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)}\| \le \hat{\epsilon}_k$. Since ϕ_k is convex, so is $\phi_k + g$. Hence, we obtain (4.7) by noting

$$\begin{aligned}
&\phi_k(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \phi_k(\mathbf{x}) - g(\mathbf{x}) \\
\le &\left(\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)} \right)^\top (\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}) \le \hat{\epsilon}_k \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}\| \le \hat{\epsilon}_k D,
\end{aligned}$$

which gives (4.7). Now let $\mathbf{x} = \bar{\mathbf{x}}^{(k)}$ in (4.7) and sum it over $k = 0$ through $K-1$ to obtain (4.8) by Assumption 1 and the equation (3.4a). □

By the definitions of $\mathbf{S}_{k+1}$, $\mathbf{F}_{k+1}$, and $\mathbf{C}_{k+1}$ in (4.2), we have for any $K \ge 1$ that if $\{R_k\}$ is chosen as (4.3) in Algorithm 1, then

$$\begin{aligned}
\max\left\{\mathbf{S}_{R_K}, \mathbf{F}_{R_K}, \mathbf{C}_{R_K}\right\} &\le \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \max\left\{\mathbf{S}_{k+1}, \mathbf{F}_{k+1}, \mathbf{C}_{k+1}\right\} \\
&\le \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{S}_{k+1} + \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{F}_{k+1} + \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{C}_{k+1},
\end{aligned} \tag{4.9}$$

and if $\{R_k\}$ is chosen as (4.4) in Algorithm 1, then

$$\max\left\{\mathbf{S}_{R_K}, \mathbf{F}_{R_K}\right\} \le \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \max\left\{\mathbf{S}_{k+1}, \mathbf{F}_{k+1}\right\} \le \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{S}_{k+1} + \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \mathbf{F}_{k+1}. \tag{4.10}$$

Lemma 2 *Let* $\{\mathbf{S}_{k+1}\}$ *be defined in* (4.2*a*). *Then for any* $K \geq 1$,

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbf{S}_{k+1} \leq \frac{1}{K}\sum_{k=0}^{K-1} \hat{\varepsilon_k} + \frac{1}{K}\sum_{k=0}^{K-1} \gamma_k \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|. \tag{4.11}$$

Proof First, note

$$\begin{aligned}\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) = \nabla f_0(\bar{\mathbf{x}}^{(k+1)}) + \gamma_k(\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}) + \beta_k J_{\mathbf{c}}(\bar{\mathbf{x}}^{(k+1)})^\top \mathbf{c}(\bar{\mathbf{x}}^{(k+1)}) \\ + \beta_k J_{\mathbf{f}}(\bar{\mathbf{x}}^{(k+1)})^\top [\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+.\end{aligned} \tag{4.12}$$

Second, according to Line 4 of Algorithm 1, there exists $\bar{\boldsymbol{\xi}}^{(k+1)} \in \partial g(\bar{\mathbf{x}}^{(k+1)})$ such that $\|\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)}\| \leq \hat{\varepsilon_k}$, which, by the definition of $\bar{\mathbf{y}}^{(k+1)}$ and $\bar{\boldsymbol{\lambda}}^{(k+1)}$ in Algorithm 1, implies

$$\begin{aligned}&\left\|\nabla f_0(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)} + J_{\mathbf{c}}(\bar{\mathbf{x}}^{(k+1)})^\top \bar{\mathbf{y}}^{(k+1)} + J_{\mathbf{f}}(\bar{\mathbf{x}}^{(k+1)})^\top \bar{\boldsymbol{\lambda}}^{(k+1)}\right\| \\ &\leq \hat{\varepsilon_k} + \gamma_k \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|.\end{aligned}$$

Hence, by the definition of $\mathbf{S}_{k+1}$, we have $\mathbf{S}_{k+1} \leq \hat{\varepsilon_k} + \gamma_k \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|$, which, after taking average over $k = 0, \ldots, K-1$, implies the desired result. □

5 Complexity with convex constraints

Throughout this section, we assume that f_i is convex for each $i = 1, \ldots, m$ and c_j is affine for each $j = 1, \ldots, n$, namely, we consider the following problem with only convex constraints:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f_0(\mathbf{x}) + g(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{f}(\mathbf{x}) \leq \mathbf{0}, \tag{5.1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$ are given. In addition to Assumption 1, we make the following assumption.

Assumption 2 The following statements hold:

A. f_0 is ρ_0-weakly convex for $\rho_0 \geq 0$, f_i is convex for $i = 1, \ldots, m$ and $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$.

B. There exists $\mathbf{x}_{\text{feas}} \in \text{int}(\mathcal{X})$ satisfying $\mathbf{A}\mathbf{x}_{\text{feas}} = \mathbf{b}$ and $\mathbf{f}(\mathbf{x}_{\text{feas}}) < \mathbf{0}$.

Here, we only require the existence of $\mathbf{x}_{\text{feas}}$ but not its availability to our algorithm. In addition, the assumption on $\text{int}(\mathcal{X}) \neq \emptyset$ does not lose generality. If $\mathcal{X}$ does not have a full dimension, it can be written as $\mathcal{X}' \cap \{\mathbf{x} \in \mathbb{R}^d : \mathbf{C}\mathbf{x} = \mathbf{d}\}$ for some full-dimensional convex compact set $\mathcal{X}' \subset \mathbb{R}^d$. Then, we can put $\mathbf{C}\mathbf{x} = \mathbf{d}$ as a part of the affine constraints and replace $\mathcal{X}$ with $\mathcal{X}'$.

Under Assumptions 1 and 2, the function ϕ_k in (4.5) is L_{ϕ_k}-smooth with

$$L_{\phi_k} = L_{f_0} + \gamma_k + \beta_k\big(\|\mathbf{A}^\top\mathbf{A}\| + \textstyle\sum_{i=1}^m B_{f_i}(B_{f_i} + L_{f_i})\big) \tag{5.2}$$

and is $(\gamma_k - \rho_0)$-strongly convex if $\gamma_k > \rho_0$. To facilitate our analysis in this section, we define

$$\widehat{\mathbf{x}}^{(k)} \equiv \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^d} \left\{ f_0(\mathbf{x}) + g(\mathbf{x}) + \frac{\gamma_k}{2}\|\mathbf{x} - \bar{\mathbf{x}}^{(k)}\|^2, \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{f}(\mathbf{x}) \le \mathbf{0} \right\}. \tag{5.3}$$

5.1 Technical Lemmas

From Assumption 2B, we have Slater's condition, and thus $\widehat{\mathbf{x}}^{(k)}$ must be a KKT point of (5.3), i.e., there are $\widehat{\boldsymbol{\xi}}^{(k)} \in \partial g(\widehat{\mathbf{x}}^{(k)})$, Lagrangian multipliers $\widehat{\mathbf{y}}^{(k)} \in \mathbb{R}^n$, and $\widehat{\boldsymbol{\lambda}}^{(k)} \in \mathbb{R}^m$ associated to $\widehat{\mathbf{x}}^{(k)}$ such that (c.f. [59, Theorem 28.2]):

$$\nabla f_0(\widehat{\mathbf{x}}^{(k)}) + \widehat{\boldsymbol{\xi}}^{(k)} + \gamma_k(\widehat{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k)}) + \mathbf{A}^\top\widehat{\mathbf{y}}^{(k)} + J_{\mathbf{f}}(\widehat{\mathbf{x}}^{(k)})^\top\widehat{\boldsymbol{\lambda}}^{(k)} = \mathbf{0}, \tag{5.4a}$$

$$\widehat{\boldsymbol{\lambda}}^{(k)} \ge \mathbf{0}, \quad \mathbf{A}\widehat{\mathbf{x}}^{(k)} = \mathbf{b}, \quad \mathbf{f}(\widehat{\mathbf{x}}^{(k)}) \le \mathbf{0}, \tag{5.4b}$$

$$\widehat{\lambda}_i^{(k)} f_i(\widehat{\mathbf{x}}^{(k)}) = 0,\ i = 1, \dots, m. \tag{5.4c}$$

Note that the direct sum of $\mathrm{Range}(\mathbf{A}\mathbf{A}^\top)$ and $\mathrm{Null}(\mathbf{A}\mathbf{A}^\top)$ forms the whole space $\mathbb{R}^d$, and also $\mathbf{A}^\top\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{A}\mathbf{A}^\top\mathbf{y} = \mathbf{0}$. Hence, we can choose $\widehat{\mathbf{y}}^{(k)} \in \mathrm{Range}(\mathbf{A}\mathbf{A}^\top)$. With this choice, we next prove the boundedness of $(\widehat{\mathbf{y}}^{(k)}, \widehat{\boldsymbol{\lambda}}^{(k)})$.

Lemma 3 *Suppose Assumptions 1 and 2 hold. Let $(\widehat{\mathbf{x}}^{(k)}, \widehat{\mathbf{y}}^{(k)}, \widehat{\boldsymbol{\lambda}}^{(k)})$ be the solution satisfying the conditions in (5.4) and $\widehat{\mathbf{y}}^{(k)} \in \mathrm{Range}(\mathbf{A}\mathbf{A}^\top)$ for $k \ge 0$. Then*

$$\|\widehat{\boldsymbol{\lambda}}^{(k)}\| \le M_\lambda(\gamma_k) := \frac{Q_k}{\min_i |f_i(\mathbf{x}_{\mathrm{feas}})|} \tag{5.5}$$

$$\|\widehat{\mathbf{y}}^{(k)}\| \le M_y(\gamma_k) := Q_k\|(\mathbf{A}\mathbf{A}^\top)^\dagger\mathbf{A}\|\left(\frac{1}{D} + \frac{1}{\mathrm{dist}(\mathbf{x}_{\mathrm{feas}}, \partial\mathcal{X})} + \frac{\max_i B_{f_i}}{\min_i |f_i(\mathbf{x}_{\mathrm{feas}})|}\right), \tag{5.6}$$

where $Q_k = D(B_{f_0} + \gamma_k D + M)$, and $(\mathbf{A}\mathbf{A}^\top)^\dagger$ denotes the pseudoinverse of $\mathbf{A}\mathbf{A}^\top$.

Proof Let $\mathbf{x}_{\mathrm{feas}}$ be the point in Assumption 2. Then from the convexity of $\{f_i\}_{i=1}^m$ and the fact $\widehat{\boldsymbol{\lambda}}^{(k)} \ge \mathbf{0}$, it follows that

$$\textstyle\sum_{i=1}^m \widehat{\lambda}_i^{(k)} f_i(\mathbf{x}_{\mathrm{feas}}) \ge \sum_{i=1}^m \widehat{\lambda}_i^{(k)}\big[f_i(\widehat{\mathbf{x}}^{(k)}) + (\mathbf{x}_{\mathrm{feas}} - \widehat{\mathbf{x}}^{(k)})^\top\nabla f_i(\widehat{\mathbf{x}}^{(k)})\big].$$

The above inequality together with (5.4a) and (5.4c) yields

$$\begin{aligned}\sum_{i=1}^{m}\widehat{\lambda}_i^{(k)} f_i(\mathbf{x}_{\text{feas}}) \geq & -(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})^{\top}\Big[\nabla f_0(\widehat{\mathbf{x}}^{(k)})+\widehat{\boldsymbol{\xi}}^{(k)}+\mathbf{A}^{\top}\widehat{\mathbf{y}}^{(k)}+\gamma_k(\widehat{\mathbf{x}}^{(k)}-\bar{\mathbf{x}}^{(k)})\Big]\\ = & -(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})^{\top}\Big[\nabla f_0(\widehat{\mathbf{x}}^{(k)})+\widehat{\boldsymbol{\xi}}^{(k)}+\gamma_k(\widehat{\mathbf{x}}^{(k)}-\bar{\mathbf{x}}^{(k)})\Big]\\ \geq & -(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})^{\top}\widehat{\boldsymbol{\xi}}^{(k)}-DB_{f_0}-\gamma_k D^2,\end{aligned} \tag{5.7}$$

where the equality follows from $\mathbf{A}(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})=\mathbf{b}-\mathbf{b}=\mathbf{0}$ and the last inequality is by Assumption 1B and (3.4a).

By Assumption 1C, we have $\widehat{\boldsymbol{\xi}}^{(k)}=\widehat{\boldsymbol{\xi}}_1+\widehat{\boldsymbol{\xi}}_2$ with $\widehat{\boldsymbol{\xi}}_1\in\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}^{(k)})$ and $\|\widehat{\boldsymbol{\xi}}_2\|\leq M$, and thus

$$(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})^{\top}\widehat{\boldsymbol{\xi}}^{(k)}\leq(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})^{\top}\widehat{\boldsymbol{\xi}}_1+DM. \tag{5.8}$$

Next we bound the term $(\mathbf{x}_{\text{feas}}-\widehat{\mathbf{x}}^{(k)})^{\top}\widehat{\boldsymbol{\xi}}_1$. If $\widehat{\mathbf{x}}^{(k)}\in\operatorname{int}(\mathcal{X})$, then $\widehat{\boldsymbol{\xi}}_1=\mathbf{0}$. Hence, we only need to consider the case when $\widehat{\mathbf{x}}^{(k)}\in\partial\mathcal{X}$ and $\widehat{\boldsymbol{\xi}}_1\neq\mathbf{0}$. In this case, $\mathcal{H}=\{\mathbf{x}\in\mathbb{R}^d\,|\,(\mathbf{x}-\widehat{\mathbf{x}}^{(k)})^{\top}\widehat{\boldsymbol{\xi}}_1=0\}$ is a supporting hyperplane of $\mathcal{X}$ at $\widehat{\mathbf{x}}^{(k)}$. Hence, $\operatorname{dist}(\mathbf{x}_{\text{feas}},\mathcal{H})\geq\operatorname{dist}(\mathbf{x}_{\text{feas}},\partial\mathcal{X})>0$. By the distance formula of a point to a hyperplane, we have $\operatorname{dist}(\mathbf{x}_{\text{feas}},\mathcal{H})=\frac{|(\widehat{\mathbf{x}}^{(k)}-\mathbf{x}_{\text{feas}})^{\top}\widehat{\boldsymbol{\xi}}_1|}{\|\widehat{\boldsymbol{\xi}}_1\|}$, and thus

$$(\widehat{\mathbf{x}}^{(k)}-\mathbf{x}_{\text{feas}})^{\top}\widehat{\boldsymbol{\xi}}_1=|(\widehat{\mathbf{x}}^{(k)}-\mathbf{x}_{\text{feas}})^{\top}\widehat{\boldsymbol{\xi}}_1|=\operatorname{dist}(\mathbf{x}_{\text{feas}},\mathcal{H})\|\widehat{\boldsymbol{\xi}}_1\|\geq\operatorname{dist}(\mathbf{x}_{\text{feas}},\partial\mathcal{X})\|\widehat{\boldsymbol{\xi}}_1\|,$$

where the first equality follows from $\widehat{\boldsymbol{\xi}}_1\in\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}^{(k)})$ so that $(\widehat{\mathbf{x}}^{(k)}-\mathbf{x}_{\text{feas}})^{\top}\widehat{\boldsymbol{\xi}}_1\geq 0$. Applying this inequality to (5.8) and using (5.7), and also noting $f_i(\mathbf{x}_{\text{feas}})<0$, we have

$$\sum_{i=1}^{m}\widehat{\lambda}_i^{(k)}|f_i(\mathbf{x}_{\text{feas}})|+\operatorname{dist}(\mathbf{x}_{\text{feas}},\partial\mathcal{X})\|\widehat{\boldsymbol{\xi}}_1\|\leq DB_{f_0}+\gamma_k D^2+DM=Q_k.$$

The above inequality implies

$$\|\widehat{\lambda}^{(k)}\|\leq\|\widehat{\lambda}^{(k)}\|_1\leq\frac{\sum_{i=1}^{m}\widehat{\lambda}_i^{(k)}|f_i(\mathbf{x}_{\text{feas}})|}{\min_i|f_i(\mathbf{x}_{\text{feas}})|}\leq\frac{Q_k}{\min_i|f_i(\mathbf{x}_{\text{feas}})|}, \tag{5.9}$$

and

$$\|\widehat{\boldsymbol{\xi}}^{(k)}\|\leq\|\widehat{\boldsymbol{\xi}}_1\|+\|\widehat{\boldsymbol{\xi}}_2\|\leq\frac{Q_k}{\operatorname{dist}(\mathbf{x}_{\text{feas}},\partial\mathcal{X})}+M. \tag{5.10}$$

Furthermore, since $\widehat{\mathbf{y}}^{(k)}\in\operatorname{Range}(\mathbf{A}\mathbf{A}^{\top})$, we have from (5.4a) that

$$\widehat{\mathbf{y}}^{(k)}=-(\mathbf{A}\mathbf{A}^{\top})^{\dagger}\mathbf{A}\Big(\nabla f_0(\widehat{\mathbf{x}}^{(k)})+\gamma_k(\widehat{\mathbf{x}}^{(k)}-\bar{\mathbf{x}}^{(k)})+J_{\mathbf{f}}(\widehat{\mathbf{x}}^{(k)})^{\top}\widehat{\lambda}^{(k)}+\widehat{\boldsymbol{\xi}}^{(k)}\Big).$$

Therefore,

$$\begin{aligned}\|\widehat{\mathbf{y}}^{(k)}\| &\le \|(\mathbf{A}\mathbf{A}^\top)^\dagger\mathbf{A}\|\left(\left\|\nabla f_0(\widehat{\mathbf{x}}^{(k)}) + \gamma_k(\widehat{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k)})\right\| + \|\widehat{\boldsymbol{\xi}}^{(k)}\| + \left\|J_{\mathbf{f}}(\widehat{\mathbf{x}}^{(k)})^\top\widehat{\boldsymbol{\lambda}}^{(k)}\right\|\right)\\ &\le \|(\mathbf{A}\mathbf{A}^\top)^\dagger\mathbf{A}\|\left(B_{f_0} + \gamma_k D + M + \frac{Q_k}{\operatorname{dist}(\mathbf{x}_{\text{feas}}, \partial\mathcal{X})} + \|\widehat{\lambda}^{(k)}\|_1 \max_i \|\nabla f_i(\widehat{\mathbf{x}}^{(k)})\|\right)\\ &\le M_y(\gamma_k),\end{aligned}$$

where the second inequality is from (5.10), Assumption 1B, and (3.4a), and the third inequality is from (5.9), Assumption 1B, and the definition of $M_y(\gamma_k)$. □

The next lemma bounds the feasibility violation of iterate $\bar{\mathbf{x}}^{(k+1)}$.

Lemma 4 *Suppose Assumptions 1 and 2 hold. Given $\gamma_k > \rho_0$ and $\beta_k > 0$ for $k \ge 0$, let ϕ_k be defined in (4.5) with $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $\widehat{\mathbf{x}}^{(k)}$ be defined in (5.3), and $\bar{\mathbf{x}}^{(k+1)}$ be generated as in Algorithm 1. Then for any $k \ge 0$,*

$$\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \le \frac{4\hat{\varepsilon}_k D}{\beta_k} + \frac{4\|\widehat{\mathbf{y}}^{(k)}\|^2}{\beta_k^2} + \frac{4\|\widehat{\lambda}^{(k)}\|^2}{\beta_k^2}. \tag{5.11}$$

Proof Notice that when $\gamma_k > \rho_0$, ϕ_k in (4.5) is convex. Hence, letting $\mathbf{x} = \widehat{\mathbf{x}}^{(k)}$ in (4.7), we have from the feasibility of $\widehat{\mathbf{x}}^{(k)}$ for (5.3) that

$$\begin{aligned}&f_0(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) + \frac{\gamma_k}{2}\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{\beta_k}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2\right)\\ &\le f_0(\widehat{\mathbf{x}}^{(k)}) + g(\widehat{\mathbf{x}}^{(k)}) + \frac{\gamma_k}{2}\|\widehat{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 + \hat{\varepsilon}_k D.\end{aligned} \tag{5.12}$$

Recall that $\widehat{\mathbf{y}}^{(k)}$ and $\widehat{\lambda}^{(k)}$ are the Lagrangian multipliers satisfying (5.4). Hence, from the convexity of the objective and constraint functions of (5.3), we have

$$\begin{aligned}f_0(\widehat{\mathbf{x}}^{(k)}) + g(\widehat{\mathbf{x}}^{(k)}) + \tfrac{\gamma_k}{2}\|\widehat{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 \le f_0(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) + \frac{\gamma_k}{2}\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2\\ + (\widehat{\mathbf{y}}^{(k)})^\top(\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}) + \textstyle\sum_{i=1}^m \widehat{\lambda}_i^{(k)} f_i(\bar{\mathbf{x}}^{(k+1)}).\end{aligned}$$

The above inequality, together with (5.12), implies

$$\begin{aligned}\hat{\varepsilon}_k D \ge \tfrac{\beta_k}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2\right) - (\widehat{\mathbf{y}}^{(k)})^\top(\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b})\\ - \textstyle\sum_{i=1}^m \widehat{\lambda}_i^{(k)} f_i(\bar{\mathbf{x}}^{(k+1)}).\end{aligned} \tag{5.13}$$

By the Young's inequality, it holds

$$\begin{aligned}&-(\widehat{\mathbf{y}}^{(k)})^\top(\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}) \ge -\tfrac{\|\widehat{\mathbf{y}}^{(k)}\|^2}{\beta_k} - \tfrac{\beta_k}{4}\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2,\\ &-\textstyle\sum_{i=1}^m \widehat{\lambda}_i^{(k)} f_i(\bar{\mathbf{x}}^{(k+1)}) \ge -\sum_{i=1}^m \widehat{\lambda}_i^{(k)}[f_i(\bar{\mathbf{x}}^{(k+1)})]_+\\ &\ge -\sum_{i=1}^m \frac{(\widehat{\lambda}_i^{(k)})^2}{\beta_k} - \sum_{i=1}^m \frac{\beta_k}{4}[f_i(\bar{\mathbf{x}}^{(k+1)})]_+^2.\end{aligned}$$

Plugging the above two inequalities into (5.13) gives the desired result.

□

5.2 The complexity of the iPPP method

In this subsection, we specify the parameters in Algorithm 1 and estimate its complexity in order to find an ε-stationary point of (5.1).

Theorem 2 *Suppose that Assumptions 1 and 2 hold and the parameters* $\{\gamma_k\}$, $\{\beta_k\}$ *and* $\{\hat{\epsilon}_k\}$ *in Algorithm* 1 *are chosen as*

$$\gamma_k = \gamma > \rho_0, \quad \beta_k = \beta\sqrt{k+1}, \quad \text{and} \quad \hat{\epsilon}_k = \frac{1}{\beta_k(k+1)}, \tag{5.14}$$

where $\beta > 0$ *is a constant. If* R_k *is defined as* (4.3), *it holds for any* $K \geq 1$ *that*

$$\max\left\{\mathbf{S}_{R_K}, \mathbf{F}_{R_K}, \mathbf{C}_{R_K}\right\} \leq \frac{3}{\beta K} + \sqrt{\frac{2\gamma \mathcal{C}_1}{K}} + \frac{4\sqrt{D+M_y^2+M_\lambda^2}}{\beta\sqrt{K}} + \frac{8(D+M_y^2+M_\lambda^2)}{\beta\sqrt{K}}, \tag{5.15}$$

where $\{(\mathbf{S}_k, \mathbf{F}_k, \mathbf{C}_k)\}_{k\geq 1}$ *is defined in* (4.2), $M_y = M_y(\gamma)$, $M_\lambda = M_\lambda(\gamma)$ *defined in* (5.5), *and*

$$\mathcal{C}_1 = 2B_{f_0} + 2G + \frac{\beta}{2}\|\mathbf{A}\bar{\mathbf{x}}^{(0)} - \mathbf{b}\|^2 + \frac{\beta}{2}\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 + \frac{3}{\beta}\left(2D + M_y^2 + M_\lambda^2\right). \tag{5.16}$$

Proof Notice that ϕ_k is convex when $\gamma_k > \rho_0$. Hence, (4.8) holds.

Since $\gamma_k = \gamma$ for all k, we have from Lemma 3 that $\|\widehat{\lambda}^{(k)}\| \leq M_\lambda$ and $\|\widehat{\mathbf{y}}^{(k)}\| \leq M_y$ for all k. Hence, it follows from (5.11) and (5.14) that

$$\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \leq \frac{4\hat{\epsilon}_k D}{\beta_k} + \frac{4(M_y^2+M_\lambda^2)}{\beta_k^2} \leq \frac{4(D+M_y^2+M_\lambda^2)}{\beta_k^2}, \tag{5.17}$$

for any $k \geq 0$. Since $\beta_k = \beta\sqrt{k+1}$, we have from the above inequality that

$$(\beta_k - \beta_{k-1})\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2\right) \leq \frac{(\sqrt{k+1}-\sqrt{k})}{\beta k}\left(4D + 4M_y^2 + 4M_\lambda^2\right). \tag{5.18}$$

Noting $\sqrt{k+1} - \sqrt{k} = \frac{1}{\sqrt{k}+\sqrt{k+1}} \leq \frac{1}{2\sqrt{k}}$ and $\sum_{k=1}^{K-1} k^{-\frac{3}{2}} \leq 1 + \int_1^{K-1} x^{-\frac{3}{2}} dx \leq 3$, we sum up (5.18) to have

$$\sum_{k=1}^{K-1}(\beta_k - \beta_{k-1})\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2\right) \leq \frac{6}{\beta}\left(D + M_y^2 + M_\lambda^2\right). \tag{5.19}$$

In addition, by the setting of $\hat{\epsilon}_k$ in (5.14), it holds

$$\sum_{k=0}^{K-1} \hat{\varepsilon}_k = \frac{1}{\beta} \sum_{k=0}^{K-1} (k+1)^{-\frac{3}{2}} \leq \frac{1}{\beta}\left(1 + \int_1^K x^{-\frac{3}{2}} dx\right) \leq \frac{3}{\beta}. \tag{5.20}$$

Now plugging (5.19) and (5.20) into (4.8) with $\mathbf{c}(\mathbf{x}) = \mathbf{A}\bar{\mathbf{x}} - \mathbf{b}$, we obtain

$$\sum_{k=0}^{K-1} \frac{\gamma_k}{2} \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 \leq \mathcal{C}_1, \tag{5.21}$$

where we have used the definition of $\mathcal{C}_1$ in (5.16).

From (5.21), the Cauchy–Schwarz inequality, and the setting $\gamma_k = \gamma, \forall k$, it follows

$$\frac{1}{K} \sum_{k=0}^{K-1} \gamma_k \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\| \leq \frac{1}{K} \sqrt{\sum_{k=0}^{K-1} \gamma_k} \sqrt{\sum_{k=0}^{K-1} \gamma_k \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2} \tag{5.22}$$

$$\leq \sqrt{\frac{2\gamma \mathcal{C}_1}{K}}. \tag{5.23}$$

Applying (5.20) and (5.23) to (4.11) gives

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{S}_{k+1} \leq \frac{3}{\beta K} + \sqrt{\frac{2\gamma \mathcal{C}_1}{K}}. \tag{5.24}$$

Futhermore, by the definition of $\mathbf{F}_{k+1}$ in (4.2b), we have from (5.17) that

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{F}_{k+1} \leq \frac{1}{K} \sum_{k=0}^{K-1} \frac{2\sqrt{D + M_y^2 + M_\lambda^2}}{\beta_k} \leq \frac{4\sqrt{D + M_y^2 + M_\lambda^2}}{\beta \sqrt{K}}, \tag{5.25}$$

where we have used the following arguments:

$$\frac{1}{K} \sum_{k=0}^{K-1} \frac{1}{\beta_k} = \frac{1}{\beta K} \sum_{k=0}^{K-1} \frac{1}{\sqrt{k+1}} \leq \frac{1}{\beta K} \int_0^K x^{-\frac{1}{2}} dx \leq \frac{2}{\beta \sqrt{K}}. \tag{5.26}$$

Finally, by $\bar{\lambda}_i^{(k+1)} = \beta_k [f_i(\bar{\mathbf{x}}^{(k+1)})]_+$ and also using (5.17), we have

$$\sum_{i=1}^m |\bar{\lambda}_i^{(k+1)} f_i(\bar{\mathbf{x}}^{(k+1)})| = \beta_k \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \leq \frac{4(D + M_y^2 + M_\lambda^2)}{\beta_k}, \tag{5.27}$$

Hence, from the definition of $\mathbf{C}_{k+1}$ in (4.2c), we average the above inequality to have

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbf{C}_{k+1} \leq \frac{1}{K} \sum_{k=0}^{K-1} \frac{4(D + M_y^2 + M_\lambda^2)}{\beta_k} \leq \frac{8(D + M_y^2 + M_\lambda^2)}{\beta \sqrt{K}}, \tag{5.28}$$

where we have used (5.26) again.

Now the result in (5.15) follows by plugging (5.24), (5.25), and (5.28) into (4.9). □

Remark 2 In the parameter setting (5.14), we require the knowledge of the weak-convexity constant ρ_0. In case it is unknown but the smoothness constant L_{f_0} is known, we can set $\gamma > L_{f_0}$. Without knowledge of ρ_0 or L_{f_0}, we cannot guarantee strong convexity of the function ϕ_k given in (4.5). To the best of our knowledge,

smoothness constants are assumed in all existing works on the complexity analysis of first-order methods for non-convex problems, e.g., [11, 26, 36].

According to Theorem 2, the convergence rate of Algorithm 1 is $O(\frac{1}{\sqrt{K}})$, in terms of the number of outer iterations. Suppose (4.6) is guaranteed through the AdapAPG method in Algorithm 4 in Appendix 1. By the complexity result of the AdapAPG method in Theorem 1, we below give the overall computational complexity of Algorithm 1 for finding an ϵ-stationary point of (5.1).

Corollary 1 (complexity result) *Under the assumptions of Theorem 2, let*

$$K = \left\lceil \max\left\{ \frac{6}{\beta\epsilon}, \frac{4}{\epsilon^2}\left[\sqrt{2\gamma \mathcal{C}_1} + \frac{4\sqrt{D + M_y^2 + M_\lambda^2}}{\beta} + \frac{8(D + M_y^2 + M_\lambda^2)}{\beta} \right]^2 \right\} \right\rceil$$
$$= O\left(\left(\gamma \left(\beta + \frac{1+\gamma^2}{\beta} \right) + \frac{1+\gamma^4}{\beta^2} \right) \frac{1}{\epsilon^2} \right),$$

where $\mathcal{C}_1$ is defined as in (5.16), β and γ are the algorithmic parameters in (5.14). Then $\bar{\mathbf{x}}^{(R_K)}$ is an ϵ-stationary point of (5.1). In addition, if $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6) is found by the AdapAPG method, the total complexity for Algorithm 1 to produce $\bar{\mathbf{x}}^{(R_K)}$ is

$$\tilde{O}\left(\sqrt{\frac{\beta}{\gamma - \rho_0}} K^{\frac{5}{4}} \right) = \tilde{O}\left(\sqrt{\frac{\beta}{\gamma - \rho_0}} \left(\gamma \left(\beta + \frac{1+\gamma^2}{\beta} \right) + \frac{1+\gamma^4}{\beta^2} \right)^{\frac{5}{4}} \frac{1}{\epsilon^{\frac{5}{2}}} \right).$$

Proof With the given K, the right hand side of (5.15) is upper bounded by ϵ. Hence, $\bar{\mathbf{x}}^{(R_K)}$ is an ϵ-stationary point of (5.1). The order of magnitude of K in terms of ϵ, β and γ is then obtained by the fact that $\mathcal{C}_1 = O\big(\beta + (1+\gamma^2)/\beta\big)$ and $D + M_y^2 + M_\lambda^2 = O\big(1+\gamma^2\big)$ according to the definitions of $\mathcal{C}_1$, M_y and M_λ.

Let T_k be the number of proximal gradient steps performed by the AdapAPG method (Algorithm 4) to find $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6). Then according to Theorem 1, the settings of γ_k and β_k in (5.14), and the formula of L_{ϕ_k} in (5.2), we have

$$T_k = \tilde{O}\left(\sqrt{\frac{L_{\phi_k}}{\gamma_k - \rho_0}} \right) = \tilde{O}\left(\sqrt{\frac{\beta_k}{\gamma - \rho_0}} \right) = \tilde{O}\left(\frac{\sqrt{\beta}(k+1)^{\frac{1}{4}}}{\sqrt{\gamma - \rho_0}} \right),$$

for $k = 0, 1, \ldots, K-1$. Therefore, the total complexity is

$$T_{\text{total}} = \sum_{k=0}^{K-1} T_k = \sum_{k=0}^{K-1} \tilde{O}\left(\frac{\sqrt{\beta}(k+1)^{\frac{1}{4}}}{\sqrt{\gamma - \rho_0}} \right) = \tilde{O}\left(\sqrt{\frac{\beta}{\gamma - \rho_0}} K^{\frac{5}{4}} \right),$$

which completes the proof after plugging in the order of K.

□

6 Complexity of the iPPP method with non-convex constraints

In this section, we consider the problem in (1.1) with a non-convex objective and non-convex constraints. Instead of Assumption 2, we make the following assumption.

Assumption 3 f_i is ρ_i-weakly convex for $\rho_i \geq 0$ for $i = 0, 1, \ldots, m$. c_j is σ_j-weakly convex for $\sigma_j \geq 0$ for $j = 1, \ldots, n$.

The non-convexity of the constraints further increases the difficulty of finding a stationary point of (1.1). Fortunately, with a sufficiently large γ_k, the proximal-point penalty subproblem (4.1) is strongly convex under Assumption 3 and thus can be effectively solved by Algorithm 4. By this observation, we show that Algorithm 1 can still guarantee an approximate stationary solution of (1.1) within a polynomial time.

6.1 Technical Lemmas

To show the complexity result, we first establish a few technical lemmas. A proof of the following lemma has been given in [21, Lemma 2]. We present it here for the readers' convenience.

Lemma 5 *Suppose Assumptions 1 and 3 hold. For any $\beta > 0$, the function $\frac{\beta}{2}[f_i(\mathbf{x})]_+^2$ is $(\beta\rho_i B_{f_i})$-weakly convex for $i = 1, \ldots, m$, and $\frac{\beta}{2}[c_j(\mathbf{x})]^2$ is $(\beta\sigma_j B_{c_j})$-weakly convex for $j = 1, \ldots, n$.*

Proof Since $f_i(\mathbf{x})$ is ρ_i-weakly convex, we have

$$f_i(\mathbf{x}) - f_i(\mathbf{x}') \geq \left\langle \nabla f_i(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \right\rangle - \frac{\rho_i}{2}\|\mathbf{x}' - \mathbf{x}\|^2, \ \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Using this inequality, the fact $|f_i(\mathbf{x})| \leq B_{f_i}$, and also the convexity of $[t]_+^2$ about t, we have

$$\begin{aligned}\frac{\beta}{2}[f_i(\mathbf{x})]_+^2 \geq& \frac{\beta}{2}[f_i(\mathbf{x}')]_+^2 + \beta[f_i(\mathbf{x}')]_+\big(f_i(\mathbf{x}) - f_i(\mathbf{x}')\big)\\ \geq& \frac{\beta}{2}[f_i(\mathbf{x}')]_+^2 + \beta[f_i(\mathbf{x}')]_+\left\langle \nabla f_i(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \right\rangle - \frac{\beta\rho_i B_{f_i}}{2}\|\mathbf{x}' - \mathbf{x}\|^2,\end{aligned}$$

which implies the $(\beta\rho_i B_{f_i})$-weak convexity of $\frac{\beta}{2}[f_i(\mathbf{x})]_+^2$. Similarly, we can show the $(\beta\sigma_j B_{c_j})$-weak convexity of $\frac{\beta}{2}[c_j(\mathbf{x})]^2$ for each j and thus complete the proof. □

With a little abuse of notation, under Assumption 3, ϕ_k defined in (4.5) is L_{ϕ_k}-smooth with

$$L_{\phi_k} = L_{f_0} + \gamma_k + \beta_k\Big(\textstyle\sum_{i=1}^m B_{f_i}(B_{f_i} + L_{f_i}) + \sum_{j=1}^n B_{c_j}(B_{c_j} + L_{c_j})\Big). \tag{6.1}$$

Note that the value of L_{ϕ_k} is different from that defined in (5.2). In addition, by Assumption 3 and Lemma 5, the function $f_0(\mathbf{x}) + \frac{\beta_k}{2}\left(\|\mathbf{c}(\mathbf{x})\|^2 + \left\|[\mathbf{f}(\mathbf{x})]_+\right\|^2\right)$ is Γ_k -weakly convex with

$$\Gamma_k := \rho_0 + \beta_k \rho_c, \quad \rho_c = \left(\sum_{i=1}^m \rho_i B_{f_i} + \sum_{j=1}^n \sigma_j B_{c_j}\right). \tag{6.2}$$

6.2 The complexity of the iPPP method under a non-singularity condition

In this subsection, we make the following assumption in addition to Assumptions 1 and 3.

Assumption 4 There exists a constant $\nu > 0$ such that for any $\mathbf{x} \in \mathcal{X}$, the following inequality holds

$$\nu\sqrt{\|[\mathbf{f}(\mathbf{x})]_+\|^2 + \|\mathbf{c}(\mathbf{x})\|^2} \le \operatorname{dist}\left(J_{\mathbf{c}}(\mathbf{x})^\top \mathbf{c}(\mathbf{x}) + J_{\mathbf{f}}(\mathbf{x})^\top [\mathbf{f}(\mathbf{x})]_+,\ -\mathcal{N}_{\mathcal{X}}(\mathbf{x})\right). \tag{6.3}$$

This assumption is inspired by a similar assumption made in [60, 73], where only equality constraints are considered. This assumption is closely related to the Kurdyka-Łojasiewicz inequality [2]. To see the connection, we consider the minimization problem

$$\min_{\mathbf{x}} \left\{ h(\mathbf{x}) := \frac{1}{2}\|[\mathbf{f}(\mathbf{x})]_+\|^2 + \frac{1}{2}\|\mathbf{c}(\mathbf{x})\|^2 + \mathbf{1}_{\mathcal{X}}(\mathbf{x}) \right\}.$$

Its optimal objective value is zero and its optimal set is identical to the feasible set of (1.1). Due to non-convexity, finding a feasible solution to (1.1) and solving this minimization are equally difficult in general. However, when $h(\mathbf{x})$ above globally satisfies a special case of the Kurdyka-Łojasiewicz inequality, namely,

$$2\nu^2 h(\mathbf{x}) \le [\operatorname{dist}(\mathbf{0}, \partial h(\mathbf{x}))]^2, \quad \forall \mathbf{x} \in \mathcal{X},$$

it is possible to minimize $h(\mathbf{x})$, or equivalently, to find a feasible solution to (1.1) by reducing $\operatorname{dist}(\mathbf{0}, \partial h(\mathbf{x}))$, which is a relatively easy task for a non-convex problem. Note that the global Kurdyka-Łojasiewicz inequality satisfied by h above is exactly (6.3). This explains Assumption 4 and why it helps to numerically find an ε-stationary point of (1.1). In Appendix 2, we show that Assumption 4 can hold for the application (7.1) we test in the numerical experiment under the additional assumption (B.2) that holds when the data is preprocessed appropriately (e.g., normalized and lifted). In Appendix 3, we further discuss how Assumption 4 compares to the assumptions made by [47] and [8].

Suppose (4.6) is guaranteed through the AdapAPG method in Algorithm 4 in Appendix 1. Under Assumptions 1, 3, and 4, we are able to show that our iPPP method can find an ε-stationary point of (1.1) in a complexity of $\tilde{O}(\frac{1}{\varepsilon^3})$. Similar to Theorem 2, we first show a convergence rate result, in terms of the number of outer iterations.

Theorem 3 *Suppose that Assumptions 1, 3 and 4 hold and the parameters* $\{\gamma_k\}$, $\{\beta_k\}$ *and* $\{\hat{\varepsilon}_k\}$ *in Algorithm* 1 *are taken as*

$$\beta_k = \beta(k+1)^{\frac{1}{3}}, \quad \gamma_k = 2\Gamma_k, \quad \text{and} \quad \hat{\varepsilon}_k = \frac{1}{\beta(k+1)^{\frac{4}{3}}}, \tag{6.4}$$

where $\beta > 0$ *is a constant, and* Γ_k *is defined in (6.2). If* R_k *is defined as (4.3), then for any* $K \geq 1$, *it holds*

$$\begin{aligned}\max\left\{\mathbf{S}_{R_K}, \mathbf{F}_{R_K}, \mathbf{C}_{R_K}\right\} \leq & \frac{1}{K}\left(\frac{4}{\beta} + \frac{4}{\nu\beta^2} + \frac{9}{2\nu^2\beta^3} + \frac{6\mathcal{C}_2(\rho_0/\beta+\rho_c)}{\nu^2}\right) + \frac{\sqrt{2\rho_0\mathcal{C}_2}}{\sqrt{K}}\left(1+\frac{1}{\nu\beta}\right) \\ & + \frac{1}{K^{1/3}}\left(\left(1+\frac{1}{\nu\beta}\right)\sqrt{2\beta\rho_c\mathcal{C}_2} + \frac{3B_{f_0}+3M}{2\nu\beta} + \frac{9(B_{f_0}+M)^2}{2\nu^2\beta}\right)\end{aligned} \tag{6.5}$$

where $\{(\mathbf{S}_k, \mathbf{F}_k, \mathbf{C}_k)\}_{k\geq 1}$ *is defined in (4.2),* ρ_c *is defined in (6.2), and*

$$\begin{aligned}\mathcal{C}_2 := & 4\left[2B_{f_0} + 2G + \frac{\beta}{2}\|\mathbf{c}(\bar{\mathbf{x}}^{(0)})\|^2 + \frac{\beta}{2}\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 + \frac{8}{3\nu^2\beta}(1/\beta + B_{f_0} + M)^2 + \frac{4D}{\beta}\right] \\ & + \frac{64\rho_0^2D^2}{3\nu^2\beta} + \frac{16\beta\rho_c^2D^2}{\nu^2}\left(\left\lceil \max\left\{\left(\frac{32\rho_0}{3\nu^2\beta}\right)^{\frac{3}{4}}, \frac{32\rho_c}{3\nu^2}\right\}\right\rceil - 1\right)^{\frac{1}{3}}.\end{aligned} \tag{6.6}$$

Proof Similar to Theorem 2, we first bound the three summations on the right-hand side of (4.9).

According to Line 4 of Algorithm 1, there must exist $\bar{\boldsymbol{\xi}}^{(k+1)} \in \partial g(\bar{\mathbf{x}}^{(k+1)})$ such that $\|\nabla\phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)}\| \leq \hat{\varepsilon_k}$.

From Assumption 1C, we have $\bar{\boldsymbol{\xi}}^{(k+1)} = \bar{\boldsymbol{\xi}}_1 + \bar{\boldsymbol{\xi}}_2$ where $\bar{\boldsymbol{\xi}}_1 \in \mathcal{N}_{\mathcal{X}}(\bar{\mathbf{x}}^{(k+1)})$ and $\|\bar{\boldsymbol{\xi}}_2\| \leq M$. Hence, it follows from $\|\nabla\phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)}\| \leq \hat{\varepsilon_k}$ and (4.12) that

$$\begin{aligned}&\left\|\bar{\boldsymbol{\xi}}_1 + \beta_k J_{\mathbf{c}}(\bar{\mathbf{x}}^{(k+1)})^\top \mathbf{c}(\bar{\mathbf{x}}^{(k+1)}) + \beta_k J_{\mathbf{f}}(\bar{\mathbf{x}}^{(k+1)})^\top [\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\| \\ &\leq \hat{\varepsilon_k} + \gamma_k\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\| + \|\nabla f_0(\bar{\mathbf{x}}^{(k+1)})\| + \|\bar{\boldsymbol{\xi}}_2\| \\ &\leq \hat{\varepsilon_k} + \gamma_k\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\| + B_{f_0} + M,\end{aligned} \tag{6.7}$$

where we have used (3.4a) in the last inequality. Now noting $\frac{\bar{\boldsymbol{\xi}}_1}{\beta_k} \in \mathcal{N}_{\mathcal{X}}(\bar{\mathbf{x}}^{(k+1)})$, we have from (6.7) and Assumption 3 that

$$\nu\sqrt{\|\mathbf{c}(\bar{\mathbf{x}}^{(k+1)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2} \leq \frac{\hat{\varepsilon_k} + \gamma_k\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\| + B_{f_0} + M}{\beta_k}, \ \forall\, k \geq 0. \tag{6.8}$$

Since $\hat{\varepsilon_k} \leq \frac{1}{\beta}$ for all k, (6.8) implies

$$\begin{aligned}\|\mathbf{c}(\mathbf{x}^{(k)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2 &\leq \frac{1}{\nu^2\beta_{k-1}^2}(1/\beta + \gamma_{k-1}\|\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k-1)}\| + B_{f_0} + M)^2 \\ &\leq \frac{2}{\nu^2\beta_{k-1}^2}\left[(1/\beta + B_{f_0} + M)^2 + \gamma_{k-1}^2\|\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k-1)}\|^2\right],\end{aligned} \tag{6.9}$$

for all $k \geq 1$. Notice $(k+1)^{\frac{1}{3}} - k^{\frac{1}{3}} = \frac{1}{k^{\frac{2}{3}} + k^{\frac{1}{3}}(k+1)^{\frac{1}{3}} + (k+1)^{\frac{2}{3}}} \leq \frac{1}{3k^{\frac{2}{3}}}$. Hence, by the setting of $\{\beta_k\}$ in (6.4), it holds $\frac{\beta_k - \beta_{k-1}}{\beta_{k-1}^2} \leq \frac{1}{3\beta k^{\frac{4}{3}}}$. Therefore, multiplying $\beta_k - \beta_{k-1}$ to both sides of (6.9) and summing it over $k = 1$ to $K-1$, we have

$$\begin{aligned} &\textstyle\sum_{k=1}^{K-1}(\beta_k - \beta_{k-1})\Big(\|\mathbf{c}(\mathbf{x}^{(k)})\|^2 + \big\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\big\|^2\Big) \\ &\leq \textstyle\sum_{k=1}^{K-1} \frac{2}{3\nu^2\beta k^{\frac{4}{3}}}\big[(1/\beta + B_{f_0} + M)^2 + \gamma_{k-1}^2\|\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k-1)}\|^2\big] \end{aligned} \tag{6.10}$$

$$\leq \tfrac{8}{3\nu^2\beta}(1/\beta + B_{f_0} + M)^2 + \textstyle\sum_{k=1}^{K-1} \frac{2\gamma_{k-1}^2\|\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(k-1)}\|^2}{3\nu^2\beta k^{\frac{4}{3}}}, \tag{6.11}$$

where we have used $\sum_{k=1}^{K-1} k^{-\frac{4}{3}} \leq 1 + \int_1^{K-1} x^{-\frac{4}{3}} dx \leq 4$ in the last inequality. In addition, it follows from (6.4) that

$$\textstyle\sum_{k=0}^{K-1} \hat{\varepsilon}_k = \frac{1}{\beta}\sum_{k=0}^{K-1}(k+1)^{-\frac{4}{3}} \leq \frac{1}{\beta}\Big(1 + \int_1^K x^{-\frac{4}{3}} dx\Big) \leq \tfrac{4}{\beta}. \tag{6.12}$$

Since $\gamma_k > \Gamma_k$, ϕ_k defined in (4.5) is convex, and thus (4.8) holds. Adding (6.10) and (6.12) to (4.8), we obtain

$$\begin{aligned} &\textstyle\sum_{k=0}^{K-2}\left(\frac{\gamma_k}{2} - \frac{2\gamma_k^2}{3\nu^2\beta(k+1)^{\frac{4}{3}}}\right)\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 + \tfrac{\gamma_{K-1}}{2}\|\bar{\mathbf{x}}^{(K)} - \bar{\mathbf{x}}^{(K-1)}\|^2 \\ \leq\ & 2B_{f_0} + 2G + \frac{\beta_0}{2}\Big(\|\mathbf{c}(\mathbf{x}^{(0)})\|^2 + \big\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\big\|^2\Big) + \frac{8}{3\nu^2\beta}(1/\beta + B_{f_0} + M)^2 + \frac{4D}{\beta}. \end{aligned} \tag{6.13}$$

By $\gamma_k = 2\Gamma_k$ with Γ_k defined in (6.2), we have

$$\tfrac{2\gamma_k}{3\nu^2\beta(k+1)^{\frac{4}{3}}} = \tfrac{4\rho_0}{3\nu^2\beta(k+1)^{\frac{4}{3}}} + \tfrac{4\rho_c}{3\nu^2(k+1)}.$$

Let

$$K' := \left\lceil \max\left\{\left(\tfrac{32\rho_0}{3\nu^2\beta}\right)^{\frac{3}{4}}, \tfrac{32\rho_c}{3\nu^2}\right\}\right\rceil - 1. \tag{6.14}$$

When $k \geq K'$, it holds that $\frac{4\rho_0}{3\nu^2\beta(k+1)^{\frac{4}{3}}} \leq \frac{1}{8}$ and $\frac{4\rho_c}{3\nu^2(k+1)} \leq \frac{1}{8}$,

and thus

$$\tfrac{2\gamma_k}{3\nu^2\beta(k+1)^{\frac{4}{3}}} \leq \tfrac{1}{4},\ \forall\, k \geq K'. \tag{6.15}$$

Applying (6.15) for $K' \leq k \leq K-2$ in (6.13) and also noting $\frac{\gamma_k}{2} \geq \frac{\gamma_k}{4}$, we obtain

$$\sum_{k=0}^{K-1}\frac{\gamma_k}{4}\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\|^2 \le 2B_{f_0}+2G+\frac{\beta_0}{2}\left(\|\mathbf{c}(\mathbf{x}^{(0)})\|^2+\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2\right)$$
$$+\frac{8}{3\nu^2\beta}(1/\beta+B_{f_0}+M)^2+\frac{4D}{\beta}+\sum_{k=0}^{K'-1}\frac{2\gamma_k^2}{3\nu^2\beta(k+1)^{\frac{4}{3}}}\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\|^2. \tag{6.16}$$

By Assumption 1B and the definitions of β_k and γ_k, we can show that

$$\begin{aligned}\sum_{k=0}^{K'-1}\frac{2\gamma_k^2}{3\nu^2\beta(k+1)^{\frac{4}{3}}}\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\|^2 &\le \sum_{k=0}^{K'-1}\frac{2(2\rho_0+2\beta_k\rho_c)^2D^2}{3\nu^2\beta(k+1)^{\frac{4}{3}}}\\ &\le \sum_{k=0}^{K'-1}\frac{16\rho_0^2D^2}{3\nu^2\beta(k+1)^{\frac{4}{3}}}+\sum_{k=0}^{K'-1}\frac{16\beta\rho_c^2D^2}{3\nu^2(k+1)^{\frac{2}{3}}}\\ &\le \frac{64\rho_0^2D^2}{3\nu^2\beta}+\frac{16\beta\rho_c^2D^2}{\nu^2}(K')^{\frac{1}{3}},\end{aligned} \tag{6.17}$$

where, in the last inequality, we use the facts that $\sum_{k=0}^{K'-1}(k+1)^{-\frac{4}{3}} \le 1+\int_1^{K'}x^{-\frac{4}{3}}dx \le 4$ and that $\sum_{k=0}^{K'-1}(k+1)^{-\frac{2}{3}} \le 1+\int_1^{K'}x^{-\frac{2}{3}}dx \le 3(K')^{\frac{1}{3}}$. Applying (6.17) and the definition of K' in (6.14) to (6.16) gives

$$\sum_{k=0}^{K-1}\frac{\gamma_k}{4}\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\|^2 \le \mathcal{C}_2, \tag{6.18}$$

with $\mathcal{C}_2$ defined in in (6.6).

Using (6.4) and recalling Γ_k in (6.2), we have $\sum_{k=0}^{K-1}\gamma_k = 2\rho_0K+2\rho_c\sum_{k=0}^{K-1}\beta_k$, and in addition, $\sum_{k=0}^{K-1}\beta_k = \beta\sum_{k=0}^{K-1}(k+1)^{\frac{1}{3}} \le \beta K^{\frac{4}{3}}$. Therefore, by (5.22) and (6.18), it holds that

$$\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\gamma_k\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\| \le \frac{\sqrt{\mathcal{C}_2}}{K}\sqrt{2\rho_0K+2\rho_c\beta K^{\frac{4}{3}}} \le \sqrt{\tfrac{2\rho_0\mathcal{C}_2}{K}}+\tfrac{\sqrt{2\rho_c\beta\mathcal{C}_2}}{K^{1/3}}. \tag{6.19}$$

Now apply (6.12) and (6.19) to (4.11) to have

$$\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\mathbf{S}_{k+1} \le \frac{4}{\beta K}+\sqrt{\tfrac{2\rho_0\mathcal{C}_2}{K}}+\tfrac{\sqrt{2\rho_c\beta\mathcal{C}_2}}{K^{1/3}}. \tag{6.20}$$

From the definition of $\mathbf{F}_{k+1}$ in (4.2b), we use (6.8) to have

$$\begin{aligned}\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\mathbf{F}_{k+1} &\le \tfrac{1}{K}\sum_{k=0}^{K-1}\frac{\hat{\epsilon_k}+B_{f_0}+\gamma_k\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\|+M}{\nu\beta_k}\\ &\le \tfrac{1}{\nu\beta K}\textstyle\sum_{k=0}^{K-1}\hat{\epsilon}_k+\tfrac{1}{\nu\beta K}\sum_{k=0}^{K-1}\gamma_k\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\|+\tfrac{B_{f_0}+M}{\nu K}\sum_{k=0}^{K-1}\tfrac{1}{\beta_k}\\ &\le \tfrac{4}{\nu\beta^2K}+\tfrac{\sqrt{2\rho_0\mathcal{C}_2}}{\nu\beta\sqrt{K}}+\tfrac{\sqrt{2\rho_c\beta\mathcal{C}_2}}{\nu\beta K^{1/3}}+\tfrac{3B_{f_0}+3M}{2\nu\beta K^{1/3}},\end{aligned} \tag{6.21}$$

where the second inequality follows from $\beta_k \geq \beta$, and the last inequality holds because of (6.12), (6.19), and the fact that

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{\beta_k} = \frac{1}{\beta K}\sum_{k=0}^{K-1}\frac{1}{(k+1)^{1/3}} \leq \frac{1}{\beta K}\int_0^K x^{-\frac{1}{3}}dx \leq \frac{3}{2\beta K^{1/3}}. \tag{6.22}$$

By (6.8) and the definition of $\lambda_i^{(k+1)}$ in Algorithm 1, it holds

$$\sum_{i=1}^{m}|\bar{\lambda}_i^{(k+1)}f_i(\bar{\mathbf{x}}^{(k+1)})| = \beta_k\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \leq \frac{(\hat{\varepsilon_k} + B_{f_0} + \gamma_k\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\| + M)^2}{\nu^2\beta_k}.$$

From the definition of $\mathbf{C}_{k+1}$ in (4.2c), we average both sides of the above inequality to have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{C}_{k+1} &\leq \frac{1}{K}\sum_{k=0}^{K-1}\frac{(\hat{\varepsilon_k} + B_{f_0} + \gamma_k\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\| + M)^2}{\nu^2\beta_k}\\ &\leq \frac{3}{\nu^2 K}\sum_{k=0}^{K-1}\frac{\hat{\varepsilon}_k^2}{\beta_k} + \frac{3}{K}\sum_{k=0}^{K-1}\frac{(B_{f_0}+M)^2}{\nu^2\beta_k} + \frac{3}{K}\sum_{k=0}^{K-1}\frac{\gamma_k^2\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2}{\nu^2\beta_k}\\ &\leq \frac{9}{2\nu^2\beta^3 K} + \frac{9(B_{f_0}+M)^2}{2\nu^2\beta K^{1/3}} + \frac{3}{K}\sum_{k=0}^{K-1}\frac{\gamma_k^2\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2}{\nu^2\beta_k},\end{aligned} \tag{6.23}$$

where the second inequality uses $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$, and the third inequality follows from (6.22) and

$$\sum_{k=0}^{K-1}\frac{\hat{\varepsilon}_k^2}{\beta_k} = \frac{1}{\beta^3}\sum_{k=0}^{K-1}\frac{1}{(k+1)^3} \leq \frac{1}{\beta^3}\left(1 + \int_1^K \frac{1}{x^3}dx\right) \leq \frac{3}{2\beta^3}.$$

Recall $\gamma_k = 2\rho_0 + 2\beta_k\rho_c$ for all $k \geq 0$ and also note $\beta_k \geq \beta$. We have $\frac{\gamma_k^2}{\beta_k} \leq \gamma_k(\frac{2\rho_0}{\beta} + 2\rho_c)$, and thus by (6.16), it holds

$$\frac{3}{K}\sum_{k=0}^{K-1}\frac{\gamma_k^2\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2}{\nu^2\beta_k} \leq \frac{6\mathcal{C}_2(\rho_0/\beta + \rho_c)}{\nu^2 K}, \tag{6.24}$$

Now apply (6.24) to (6.23) to have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{C}_{k+1} \leq \frac{9}{2\nu^2\beta^3 K} + \frac{9(B_{f_0}+M)^2}{2\nu^2\beta K^{1/3}} + \frac{6\mathcal{C}_2(\rho_0/\beta + \rho_c)}{\nu^2 K}. \tag{6.25}$$

Plugging (6.20), (6.21), and (6.25) into (4.11) gives (6.5). □

Corollary 2 (*Complexity result*) *Under the same assumptions of Theorem 3, let*

$$K_1 = \frac{3}{\varepsilon}\left(\frac{4}{\beta} + \frac{4}{\nu\beta^2} + \frac{9}{2\nu^2\beta^3} + \frac{6\mathcal{C}_2(\rho_0/\beta + \rho_c)}{\nu^2}\right),\ K_2 = \frac{18\rho_0\mathcal{C}_2}{\varepsilon^2}\left(1 + \frac{1}{\nu\beta}\right)^2,$$
$$K_3 = \frac{27}{\varepsilon^3}\left(\left(1 + \frac{1}{\nu\beta}\right)\sqrt{2\beta\rho_c\mathcal{C}_2} + \frac{3B_{f_0} + 3M}{2\nu\beta} + \frac{9(B_{f_0} + M)^2}{2\nu^2\beta}\right)^3$$

and

$$K = \lceil \max\{K_1, K_2, K_3\} \rceil = O\left(\left(\left(\beta + \frac{1}{\beta}\right)^{\frac{3}{2}}\left[\beta\rho + \frac{\rho}{\beta^3} + \frac{\rho^3}{\beta} + \beta\rho^3\left((\rho/\beta)^{\frac{1}{4}} + \rho^{\frac{1}{3}}\right)\right]^{\frac{3}{2}} + \frac{1}{\beta^3}\right)\frac{1}{\varepsilon^3}\right),$$

where $\mathcal{C}_2$ *is defined as in* (6.6), β *is the algorithmic parameter in* (6.4) *and* $\rho = \max\{\rho_0, \rho_c\}$. *Then* $\bar{\mathbf{x}}^{(R_K)}$ *is an* ε*-stationary point of* (1.1). *In addition, if* $\bar{\mathbf{x}}^{(k+1)}$ *satisfying* (4.6) *is found by the AdapAPG method,*

the total complexity for Algorithm 1 *to produce* $\bar{\mathbf{x}}^{(R_K)}$ *is*

$$\tilde{O}\left(\sqrt{\frac{1+\rho}{\rho}}K\right) = O\left(\sqrt{\frac{1+\rho}{\rho}}\left(\left(\beta + \frac{1}{\beta}\right)^{\frac{3}{2}}\left[\beta\rho + \frac{\rho}{\beta^3} + \frac{\rho^3}{\beta} + \beta\rho^3\left((\rho/\beta)^{\frac{1}{4}} + \rho^{\frac{1}{3}}\right)\right]^{\frac{3}{2}} + \frac{1}{\beta^3}\right)\frac{1}{\varepsilon^3}\right).$$

Proof With the given K, the right hand side of (6.5) is upper bounded by ε, so $\bar{\mathbf{x}}^{(R_K)}$ is an ε-stationary point of (1.1). The order of magnitude of K in terms of ε, β and ρ is then obtained by the fact that $\mathcal{C}_2 = O\left(\beta + \beta^{-3} + \rho^2/\beta + \beta\rho^2\left((\rho/\beta)^{\frac{1}{4}} + \rho^{\frac{1}{3}}\right)\right)$ according to the definitions of $\mathcal{C}_2$ and ρ.

Let T_k be the number of proximal gradient steps performed by the AdapAPG method (Algorithm 4) to find $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6). Then according to Theorem 1 and the definitions of Γ_k, γ_k, β_k, and L_{ϕ_k} in (6.1), (6.2) and (6.4), we have

$$T_k = \tilde{O}\left(\sqrt{\frac{L_{\phi_k}}{\gamma_k - \Gamma_k}}\right) = \tilde{O}\left(\sqrt{\frac{\Gamma_k + \beta_k}{\Gamma_k}}\right) = \tilde{O}\left(\sqrt{\frac{\beta_k\rho + \beta_k}{\beta_k\rho}}\right) = \tilde{O}\left(\sqrt{\frac{1+\rho}{\rho}}\right),$$

for $k = 0, 1, \ldots, K-1$. Therefore, the total complexity is $T_{\text{total}} = \sum_{k=0}^{K-1} T_k = \tilde{O}\left(\sqrt{\frac{1+\rho}{\rho}}K\right)$, which completes the proof after plugging in the order of K.

□

6.3 The complexity of the iPPP method under initial feasibility assumption

In this subsection, we drop Assumption 4 and analyze the complexity of the proposed iPPP method by starting from an initial feasible point, namely, in addition to Assumptions 1 and 3, we assume the follows.

Assumption 5 The initial solution $\bar{\mathbf{x}}^{(0)} \in \mathcal{X}$ in Algorithm 1 is feasible, i.e., $f_i(\bar{\mathbf{x}}^{(0)}) \le 0$ for each $i = 1, \dots, m$ and $c_j(\bar{\mathbf{x}}^{(0)}) = 0$ for each $j = 1, \dots, n$.

Remark 3 This feasibility assumption on $\bar{\mathbf{x}}^{(0)}$ can be weakened to near-feasibility depending on the required accuracy. Unless with certain regularity conditions like the one we assumed in the previous subsection, or with certain special structures, it is generally impossible to find a (near) feasible solution of a nonlinear system in a polynomial time. Existing works, such as [8, 11, 47], also need the (near)-feasibility assumption to guarantee a near-stationary point.

Below, we specify the parameters of Algorithm 1 and analyze its complexity with Option II to find a weak ϵ-stationary point of (1.1).

Theorem 4 *Suppose that Assumptions 1, 3, and 5 hold and the parameters* $\{\gamma_k\}, \{\beta_k\}$ *and* $\{\hat{\epsilon}_k\}$ *in Algorithm* 1 *are taken as*

$$\beta_k = \beta, \quad \gamma_k = 2(\rho_0 + \beta\rho_c), \quad \text{and} \quad \hat{\epsilon}_k = \tfrac{1}{(k+1)^2}, \forall\, k \ge 0, \tag{6.26}$$

where $\beta > 0$ *is a constant, and* ρ_c *is defined in* (6.2). *If* R_k *is defined as* (4.4), *then for any* $K \ge 1$, *it holds that*

$$\max\left\{\mathbf{S}_{R_K}, \mathbf{F}_{R_K}\right\} \le \tfrac{\pi^2}{6K} + \tfrac{\sqrt{\mathcal{C}_3}}{\sqrt{K}} + \sqrt{\tfrac{4(B_{f_0}+G)+\pi^2 D/3}{\beta}}, \tag{6.27}$$

where

$$\mathcal{C}_3 := \left(2\rho_0 + 2\beta\rho_c\right)\left(4(B_{f_0} + G) + \tfrac{\pi^2 D}{3}\right). \tag{6.28}$$

Proof By the setting $\beta_k = \beta$, $\forall\, k \ge 0$ and $\sum_{k=0}^{K-1} \hat{\epsilon}_k = \sum_{k=0}^{K-1} \frac{1}{(k+1)^2} \le \frac{\pi^2}{6}$, we obtain from (4.8) and also the feasibility of $\bar{\mathbf{x}}^{(0)}$ that

$$\textstyle\sum_{k=0}^{K-1} \frac{\gamma_k}{2}\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{\beta}{2}\left\|\mathbf{c}(\bar{\mathbf{x}}^{(K)})\right\|^2 + \frac{\beta}{2}\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(K)})]_+\right\|^2 \le 2(B_{f_0} + G) + \frac{\pi^2 D}{6}. \tag{6.29}$$

Hence, from (5.22) and (6.29) and also the setting of γ_k in (6.26), we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\gamma_k\|\bar{\mathbf{x}}^{(k+1)}-\bar{\mathbf{x}}^{(k)}\| \le \frac{1}{K}\sqrt{4(B_{f_0}+G)+\frac{\pi^2 D}{3}}\sqrt{2\rho_0 K+2K\beta\rho_c} = \frac{\sqrt{\mathcal{C}_3}}{\sqrt{K}}. \tag{6.30}$$

Applying (6.30) and the fact that $\sum_{k=0}^{K-1}\hat{\varepsilon}_k \le \frac{\pi^2}{6}$ to (4.11) leads to

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{S}_{k+1} \le \frac{\pi^2}{6K}+\frac{\sqrt{\mathcal{C}_3}}{\sqrt{K}}. \tag{6.31}$$

In addition, notice that (6.29) actually holds for any $K \ge 1$. Hence,

$$\frac{\beta}{2}\left\|\mathbf{c}(\bar{\mathbf{x}}^{(k+1)})\right\|^2+\frac{\beta}{2}\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \le 2(B_{f_0}+G)+\frac{\pi^2 D}{6},\ \forall\, k\ge 0,$$

which, together with the definition of $\mathbf{F}_{k+1}$ in (4.2b), implies

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{F}_{k+1} \le \sqrt{\frac{4(B_{f_0}+G)+\pi^2 D/3}{\beta}}. \tag{6.32}$$

Now plugging (6.31) and (6.32) into (4.10) gives the desired result.

□

Corollary 3 (*complexity result*) *Under the same assumptions of Theorem 4, let* $\beta=\frac{36(B_{f_0}+G)+3\pi^2 D}{\varepsilon^2}$ *and* $K=\left\lceil \max\left\{\frac{9\mathcal{C}_3}{\varepsilon^2},\frac{\pi^2}{2\varepsilon}\right\}\right\rceil = O\left(\frac{\rho}{\varepsilon^4}\right)$,

where $\mathcal{C}_3$ *is defined in* (6.28) *and* $\rho=\max\{\rho_0,\rho_c\}$. *Then*

$\bar{\mathbf{x}}^{(R_K)}$ *is a weak* ε*-stationary point of* (1.1).

In addition, if $\bar{\mathbf{x}}^{(k+1)}$ *satisfying* (4.6) *is found by the AdapAPG method, the total complexity for Algorithm* 1 *to produce* $\bar{\mathbf{x}}^{(R_K)}$ *is* $\tilde{O}\left(\sqrt{\frac{1+\rho}{\rho}}K\right)=\tilde{O}\left(\frac{\sqrt{(1+\rho)\rho}}{\varepsilon^4}\right)$.

Proof With the chosen β and K, it holds that $\sqrt{\frac{4(B_{f_0}+G)+\pi^2 D/3}{\beta}} \le \frac{\varepsilon}{3}$ and $\frac{\pi^2}{6K}+\frac{\sqrt{\mathcal{C}_3}}{\sqrt{K}} \le \frac{2\varepsilon}{3}$. Hence, by (6.27), $\bar{\mathbf{x}}^{(R_K)}$ is a weak ε-stationary point of (1.1). The order of magnitude of K in terms of ε and ρ is then obtained by the fact that $\mathcal{C}_3 = O(\beta\rho)=\frac{\rho}{\varepsilon^2}$ according to the definitions of β and ρ.

Let T_k be the number of proximal gradient steps performed by the AdapAPG method (Algorithm 4) to find $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6).

Notice that with the parameters set in (6.26), ϕ_k defined in (4.5) is $(\rho_0+\beta\rho_c)$-strongly convex, and in addition, its smoothness constant $L_{\phi_k}=\Theta(\gamma_k+\beta_k)=\Theta(\rho\beta+\beta)$. Hence, according to Theorem 1 and the choice of β, we have

$$T_k = \tilde{O}\left(\sqrt{\frac{L_{\phi_k}}{\rho_0 + \beta\rho_c}}\right) = \tilde{O}\left(\sqrt{\frac{\beta\rho + \beta}{\beta\rho}}\right) = \tilde{O}\left(\sqrt{\frac{1+\rho}{\rho}}\right)$$

for all $k \geq 0$. Therefore, the total complexity is $T_{\text{total}} = \sum_{k=0}^{K-1} T_k = \tilde{O}\left(\sqrt{\frac{1+\rho}{\rho}}K\right)$, which completes the proof after plugging in the order of K. □

Remark 4 Notice that in Corollary 3, we only guarantee a weak ε-stationary point because no constraint qualification (CQ) is assumed. Without a CQ, even a global optimal solution is not guaranteed to be a KKT point.

7 Numerical experiments

In spite of the theoretical focus of this paper, we evaluate the numerical performance of the iPPP method on a multi-class Neyman-Pearson classification (mNPC) problem in this section. Suppose there is a set of training data with K classes, denoted by $\mathcal{D}_k \subseteq \mathbb{R}^d$ for $k = 1, 2, \ldots, K$. The goal is to learn K linear models $\mathbf{x}_k$, $k = 1, 2, \ldots, K$ and predict the class of a data point ξ as $\arg\max_{k=1,2,\ldots,K} \mathbf{x}_k^\top \xi$. To achieve a high classification accuracy, $\{\mathbf{x}_k\}$ is found such that $\mathbf{x}_k^\top \xi - \mathbf{x}_l^\top \xi$ is positively large for any $k \neq l$ and any $\xi \in \mathcal{D}_k$ [14, 72]. This leads to minimizing the average loss $\frac{1}{|\mathcal{D}_k|}\sum_{l\neq k}\sum_{\xi\in\mathcal{D}_k}\phi(\mathbf{x}_k^\top\xi - \mathbf{x}_l^\top\xi)$, where ϕ is a non-increasing (potentially non-convex) loss function. Suppose misclassifying ξ has a cost depending on its true class label k. When training these K linear models, the mNPC prioritizes minimizing the loss on one class, say $\mathcal{D}_1$, and meanwhile controls the losses on other classes, namely,

$$\begin{aligned} \min_{\|\mathbf{x}_k\|\leq\lambda, k=1,\ldots,K} \ & \tfrac{1}{|\mathcal{D}_1|}\textstyle\sum_{l>1}\sum_{\xi\in\mathcal{D}_1}\phi(\mathbf{x}_1^\top\xi - \mathbf{x}_l^\top\xi), \\ \text{s.t.} \ & \tfrac{1}{|\mathcal{D}_k|}\textstyle\sum_{l\neq k}\sum_{\xi\in\mathcal{D}_k}\phi(\mathbf{x}_k^\top\xi - \mathbf{x}_l^\top\xi) \leq r_k, \quad k = 2, 3, \ldots, K. \end{aligned} \tag{7.1}$$

Here, r_k controls the loss for $\mathcal{D}_k$, and $\lambda > 0$ is a regularization parameter.

We created test instances of (7.1) using the LIBSVM multi-class classification datasets *covtype* and *mnist*, which have $K = 7$ and $K = 10$ classes, respectively.

The first class of each dataset is used to formulate the objective function in (7.1), and the other classes are used to formulate the constraints. The function ϕ in (7.1) is chosen as the sigmoid function $\phi(z) = 1/(1 + \exp(z))$. We set $r_k = 0.5(K - 1)$, $\forall\, k = 2, \ldots, K$ and set $\lambda = 0.3$ for both datasets.

We compare the proposed method to the exact penalty method proposed in [11] and the inexact augmented Lagrangian method (iALM) in [60]. We choose [11] because their theoretical complexity is given in terms of how many trust-region subproblems their algorithm needs to solve while the complexity we consider in this paper is measured by the total number of the gradients computed. Since it is not clear how to compare the theoretical complexity between our method and [11], we directly compare their empirical performances. We choose [60] because the augmented Lagrangian method typically has better performance than the penalty

method in practice although the theoretical complexity by [60] is higher than ours. We do not compare with other methods because they either have a higher theoretical complexity than ours or have no theoretical guarantee as (7.1) does not satisfy their assumptions (e.g. linear constraints).

All methods are implemented in Matlab on a 64-bit MacOS Catalina machine with a 4.20 Ghz Intel Core i7-7700K CPU and 16GB of memory. For all algorithms, the initial iterate is set to $\bar{\mathbf{x}}^{(0)} = \mathbf{0}$ and we verify that it is a feasible solution of (7.1) with r_k's chosen above. In Appendix 2, we discuss how Assumption 4 can hold for problem (7.1) when the data ξ satisfies a mild condition that can be ensured by a standard preprocessing. The values of algorithm-related parameters in all algorithms are selected from a discrete set of candidates based on the value of the objective function after 10,000 data passes.

On solving (1.1) with $g \equiv 0$, the method in [11] applies a non-smooth trust-region method to solve a sequence of unconstrained subproblems in the form of

$$\min_{\mathbf{x}} f_0(\mathbf{x}) + \rho \sum_{i=1}^{m} \left[f_i(\mathbf{x})\right]_+ + \rho \sum_{j=1}^{n} \left|c_j(\mathbf{x})\right|, \tag{7.2}$$

where $\rho > 0$ is a penalty parameter which will be increased sequentially. At iteration k of the non-smooth trust-region method for solving (7.2), an updating direction is computed as[4]

$$\mathbf{s}^{(k)} \in \arg\min_{\|\mathbf{s}\|_1 \le \Delta_k} \left\{ \begin{array}{l} f_0(\mathbf{x}^{(k)}) + \nabla f_0(\mathbf{x}^{(k)})^\top \mathbf{s} + \rho \sum_{i=1}^{m} \left[f_i(\mathbf{x}^{(k)}) + \nabla f_i(\mathbf{x}^{(k)})^\top \mathbf{s}\right]_+ \\ +\rho \sum_{j=1}^{n} \left| c_j(\mathbf{x}^{(k)}) + \nabla c_j(\mathbf{x}^{(k)})^\top \mathbf{s} \right| \end{array} \right\}, \tag{7.3}$$

where Δ_k is the radius of the trust region. Upon obtaining $\mathbf{s}^{(k)}$, the estimated solution is updated to $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}$ if this update significantly reduces the objective value of (7.2). Once an ε-critical point of (7.2) (see equation (2.2) in [11] for the definition) is found, a steering procedure [10] is utilized to increase the penalty parameter ρ in (7.2).

In our implementation, we formulate the problem in (7.3) as a linear program and then use Matlab built-in LP solver to obtain $\mathbf{s}^{(k)}$. The outer iterations in the method by [11] require a steering parameter ξ, an increase factor to update ρ, an initial value of ρ, and a tolerance for solving subproblem (7.2). Steering parameter ξ is set to be 0.3 for *covtype* and 0.1 for *mnist*. The initial value of ρ is set to be $1/\xi$ for both datasets. We choose the increase factor to be 10 and tolerance $\varepsilon = 0.001$ for both datasets. Moreover, the trust-region method for solving subproblem (7.2) requires five control parameters: Δ_0, η_1, η_2, γ_1, and γ_2. For both datasets, we choose $\Delta_0 = 1, \eta_1 = 0.3, \eta_2 = 0.7, \gamma_1 = 0.3$, and $\gamma_2 = 0.7$.

The iALM method in [60] is developed for (1.1) with only equality constraints (i.e., $\mathbf{f} \equiv \mathbf{0}$). At the kth outer iteration, it applies another optimization algorithm to the following subproblem

[4] The method in [11] allows using any norm in the ball constraint of (7.3). Here, we choose ℓ_1-norm so that (7.3) can be solved as a linear program.

$$\min_{\mathbf{x}\in\mathbb{R}^d} f_0(\mathbf{x}) + g(\mathbf{x}) + (\bar{\mathbf{y}}^{(k)})^\top \mathbf{c}(\mathbf{x}) + \frac{\beta_k^{\mathrm{ALM}}}{2}\|\mathbf{c}(\mathbf{x})\|^2, \tag{7.4}$$

where $\beta_k^{\mathrm{ALM}} \geq 0$ and $\bar{\mathbf{y}}^{(k)} \in \mathbb{R}^n$ is the dual variable, in order to find an $\hat{\varepsilon}_{k+1}$-statioanry point of (7.4), namely, a point $\bar{\mathbf{x}}^{(k+1)} \in \mathbb{R}^d$ such that

$$\min_{\boldsymbol{\xi}'\in\partial g(\bar{\mathbf{x}}^{(k+1)})} \|\nabla f_0(\bar{\mathbf{x}}^{(k+1)}) + J_{\mathbf{c}}(\bar{\mathbf{x}}^{(k+1)})^\top \bar{\mathbf{y}}^{(k)} + \boldsymbol{\xi}'\| \leq \hat{\varepsilon}_{k+1}.$$

Then it applies a dual ascent step to update $\bar{\mathbf{y}}^{(k+1)} = \bar{\mathbf{y}}^{(k)} + \sigma_{k+1}\mathbf{c}(\bar{\mathbf{x}}^{(k+1)})$ with a step size $\sigma_{k+1} = \sigma_0 \min\{\frac{\|\mathbf{c}(\bar{\mathbf{x}}^{(0)})\|\log^2 2}{\|\mathbf{c}(\bar{\mathbf{x}}^{(k+1)})\|(k+1)\log^2(k+2)}, 1\}$ for all $k \geq 0$. Here σ_0 is user-specified. Since (7.1) has inequality constraints, we apply iALM to the following equivalent problem

$$\begin{aligned} &\min_{\|\mathbf{x}_k\|\leq\lambda, k=1,\ldots,K, s_k\geq 0, k=2,\ldots,K} \frac{1}{|\mathcal{D}_1|}\sum_{l>1}\sum_{\xi\in\mathcal{D}_1}\phi(\mathbf{x}_1^\top\xi - \mathbf{x}_l^\top\xi), \\ &\text{s.t.}\,\frac{1}{|\mathcal{D}_k|}\sum_{l\neq k}\sum_{\xi\in\mathcal{D}_k}\phi(\mathbf{x}_k^\top\xi - \mathbf{x}_l^\top\xi) + s_k = r_k, \quad k = 2,3,\ldots,K, \end{aligned} \tag{7.5}$$

where $s_k \in \mathbb{R}_+$ for $k = 2,\ldots,K$ are slack variables.

When implementing the iALM method, we set $\bar{\mathbf{y}}^{(0)} = \mathbf{0}$, $\sigma_0 = 5$, $\beta_k^{\mathrm{ALM}} = 5^k$ and $\hat{\varepsilon}_{k+1} = 1/\beta_k^{\mathrm{ALM}}$ for both *covtype* and *mnist* datasets. The forms of β_k^{ALM} and $\hat{\varepsilon}_{k+1}$ are consistent with Algorithm 1 and Corollary 4.2 in [60]. We find an $\hat{\varepsilon}_{k+1}$-statioanry point of (7.4) using the accelerated proximal gradient method (APGM) in [26]. The APGM itself requires three control parameters. Despite a little abuse of notation, we denote the control parameters in iteration t of APGM by α_t, β_t and λ_t to be consistent with the notation in [26]. According to Corollary 2 in [26], we set $\alpha_t = \frac{2}{t+1}$ and $\lambda_t = \beta_t$ for *covtype* and *mnist* with $\beta_t = \frac{1}{10\beta_k^{\mathrm{ALM}}}$ for *covtype* and $\beta_t = \frac{1}{200\beta_k^{\mathrm{ALM}}}$ for *mnist*, where β_k^{ALM} is from subproblem (7.4) solved in the kth outer iteration of the iALM method.

For our iPPP method, we need to specify the parameters $\hat{\varepsilon}_k, \gamma_k$ and β_k for each k as well as constant M^{ini} and μ^{ini}. The inner algorithms also require parameters $\gamma_{\mathrm{inc}}, \gamma_{\mathrm{dec}}, \gamma_{\mathrm{sc}}$, and θ_{sc}. We set $M^{\mathrm{ini}} = 10$, $\mu^{\mathrm{ini}} = 1$, $\gamma_{\mathrm{inc}} = 1.5$, $\gamma_{\mathrm{dec}} = \gamma_{\mathrm{sc}} = 1.2$, and $\theta_{\mathrm{sc}} = 0.5$. For other parameters, we compare two different settings: one using $\hat{\varepsilon}_k = 1/(k+1)^2, \gamma_k = 0.1, \beta_k = 1000, \forall\, k$, and the other using $\hat{\varepsilon}_k = \frac{1}{\beta(k+1)^{\frac{4}{3}}}, \gamma_k = 0.1(k+1)^{\frac{1}{3}}, \beta_k = \beta(k+1)^{\frac{1}{3}}, \forall\, k$.

In the latter setting, we choose $\beta = 200$ for *mnist* and choose $\beta = 500$ for *covtype*.

The numerical results are presented in Figure 1. The x-axis represents the number of data passes each algorithm performs. The y-axis represents the objective value of iterates in the first column, the infeasibility of iterates (i.e., $\max_{i=1,\ldots,m}\{f_i(\mathbf{x}), 0\}$) in the second column, and the stationarity of iterates in the third column. Let $I(\mathbf{x}) = \{1 \leq i \leq m | f_i(\mathbf{x}) \geq 0\}$ and $\mathcal{X} = \{\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_K) \,|\, \|\mathbf{x}_k\| \leq \lambda, k = 1,\ldots,K\}$. We calculate the stationarity of a solution $\mathbf{x}$ as the optimal objective value of the following convex optimization

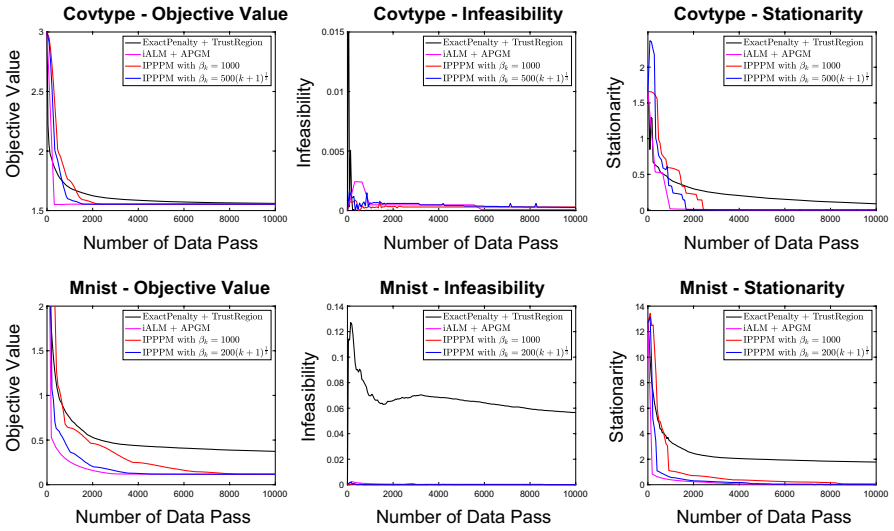

Fig. 1 Comparison between the iPPP method and the trust-region-based penalty method in [11] for solving multi-class Neyman-Pearson classification problem (7.1) on two datasets from LIBSVM

$$\min_{\lambda\in\mathbb{R}_+^{|I(\mathbf{x}^{(k)})|},\mathbf{y}\in\mathbb{R}^n} \operatorname{dist}\left(\nabla f_0(\mathbf{x}^{(k)}) + \sum_{i\in I(\mathbf{x})} \lambda_i \nabla f_i(\mathbf{x}^{(k)}) + \sum_{j=1}^{n} y_j \nabla c_j(\mathbf{x}^{(k)}), -\mathcal{N}_{\mathcal{X}}(\mathbf{x}^{(k)})\right),$$

which can be solved as a convex quadratic program and we solve using Matlab built-in QP solver. We observe from Figure 1 that, for these two instances, our iPPP method outperforms the trust-region-based penalty method by [11] in terms of its capability of improving objective value, feasibility, and stationarity of the iterates simutaneously. Moreover, these two instances also suggest that the iPPP method using growing penalty parameters performs better than using a fixed penalty parameter. However, the iALM method using APGM as a subroutine has better performance than our iPPP method for both instances although the former has a higher theoretical complexity ($\tilde{O}(\varepsilon^{-4})$) than the latter ($\tilde{O}(\varepsilon^{-3})$). It is possibly because the $\tilde{O}(\varepsilon^{-4})$ complexity proved by [60] is not tight and can be further reduced with more sophisticated analysis.

8 Conclusion

We proposed a gradient-based penalty method for a constrained non-convex optimization problem. The complexity of the proposed algorithm for finding an approximate stationary point is derived for two cases: (i) when the objective function is non-convex but the constraint functions are convex and, (ii) when the objective and constraint functions are all non-convex. For the first case, our method can produce an ε-stationary point with complexity of $\tilde{O}(\varepsilon^{-5/2})$ under Slater's condition. For the

second case, the complexity is $\tilde{O}(\varepsilon^{-3})$ if a non-singularity condition holds on the constraints and otherwise $\tilde{O}(\varepsilon^{-4})$ if an initial feasible solution is assumed.

Appendix 1: Adaptive accelareted proximal gradient method

In this section, we introduce the AdapAPG method by [45] for solving a strongly convex composite optimizaiton in the form of (3.5). It can be applied to (4.1), which is an instance of (3.5), in order to find $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6) as required in Line 4 of Algorithm 1.

Consider problem (3.5), where $\phi : \mathbb{R}^d \to \mathbb{R}$ is μ_ϕ-strongly convex and L_ϕ -smooth, and $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower-semicontinuous convex function. Given $\mathbf{w} \in \mathbb{R}^d$ and a constant $L > 0$, we define a local model of $\phi(\mathbf{x})$ as

$$\psi_L(\mathbf{w};\mathbf{x}) := \phi(\mathbf{w}) + \nabla\phi(\mathbf{w})^\top(\mathbf{x} - \mathbf{w}) + \frac{L}{2}\|\mathbf{x} - \mathbf{w}\|^2 + r(\mathbf{x}). \tag{A.1}$$

As defined in (3.7), the proximal gradient step of (3.5) at $\mathbf{w}$ is

$$T_L(\mathbf{w}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \psi_L(\mathbf{w};\mathbf{x}) = \mathbf{prox}_{L^{-1}r}(\mathbf{w} - L^{-1}\nabla\phi(\mathbf{w})), \tag{A.2}$$

and the *proximal gradient mapping* of (3.5) at $\mathbf{w}$ is

$$g_L(\mathbf{w}) := L(\mathbf{w} - T_L(\mathbf{w})). \tag{A.3}$$

By the optimality condition satisfied by $T_L(\mathbf{w})$ in (A.2), there exists $\boldsymbol{\xi} \in \partial r(T_L(\mathbf{w}))$ such that

$$\boldsymbol{\xi} + \nabla\phi(\mathbf{w}) + L(T_L(\mathbf{w}) - \mathbf{w}) = \mathbf{0},$$

which implies $\nabla\phi(T_L(\mathbf{w})) + \boldsymbol{\xi} = \nabla\phi(T_L(\mathbf{w})) - \nabla\phi(\mathbf{w}) - L(T_L(\mathbf{w}) - \mathbf{w})$ so that

$$\begin{aligned}\omega(T_L(\mathbf{w})) &\le \|\nabla\phi(T_L(\mathbf{w})) + \boldsymbol{\xi}\| \le \|\nabla\phi(T_L(\mathbf{w})) - \nabla\phi(\mathbf{w})\| + \|g_L(\mathbf{w})\| \\ &= \left(1 + \frac{S_L(\mathbf{w})}{L}\right)\|g_L(\mathbf{w})\|.\end{aligned} \tag{A.4}$$

Here,

$$\omega(\mathbf{x}) := \min_{\boldsymbol{\xi}'\in\partial r(\mathbf{x})} \|\nabla\phi(\mathbf{x}) + \boldsymbol{\xi}'\| \tag{A.5}$$

is a first-order suboptimality measure of $\mathbf{x}$ and

$$S_L(\mathbf{w}) := \frac{\|\nabla\phi(T_L(\mathbf{w})) - \nabla\phi(\mathbf{w})\|}{\|T_L(\mathbf{w}) - \mathbf{w}\|} \le L_\phi \tag{A.6}$$

is a local Lipschitz constant of ϕ. Inequality (A.4) means that, when $\|g_L(\mathbf{w})\|$ is small, the solution generated by a proximal gradient step from $\mathbf{w}$ has a small subgradient and thus is a high-quality solution of (3.5).

Algorithm 2 $\{\mathbf{x}^{(t+1)}, M_t, \mathbf{p}^{(t)}, S_t\} \leftarrow \texttt{LineSearch}(\phi, r, \mathbf{x}^{(t)}, L_t)$

Choose: $\gamma_{\text{inc}} > 1$
$L \leftarrow L_t/\gamma_{\text{inc}}$
repeat
  $L \leftarrow L\gamma_{\text{inc}}$
  $\mathbf{x}^{(t+1)} \leftarrow T_L(\mathbf{x}^{(t)})$
until $F(\mathbf{x}^{(t+1)}) \le \psi_L(\mathbf{x}^{(t)}; \mathbf{x}^{(t+1)})$
$M_t \leftarrow L$
$\mathbf{p}^{(t)} \leftarrow M_t(\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)})$
$S_t \leftarrow S_L(\mathbf{x}^{(t)})$

Algorithm 3
$\{\mathbf{x}^{(t+1)}, M_t, \alpha_t, \mathbf{p}^{(t)}, S_t\} \leftarrow \texttt{AccelLineSearch}(\phi, r, \mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, L_t, \mu, \alpha_{t-1})$

Choose: $\gamma_{\text{inc}} > 1$
$L \leftarrow L_t/\gamma_{\text{inc}}$
repeat
  $L \leftarrow L\gamma_{\text{inc}}$
  $\alpha_t \leftarrow \sqrt{\frac{\mu}{L}}$
  $\mathbf{w}^{(t)} \leftarrow \mathbf{x}^{(t)} + \frac{\alpha_t(1-\alpha_{t-1})}{\alpha_{t-1}(1+\alpha_t)}(\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)})$
  $\mathbf{x}^{(t+1)} \leftarrow T_L(\mathbf{w}^{(t)})$
until $F(\mathbf{x}^{(t+1)}) \le \psi_L(\mathbf{w}^{(t)}; \mathbf{x}^{(t+1)})$
$M_t \leftarrow L$
$\mathbf{p}^{(t)} \leftarrow M_t(\mathbf{w}^{(t)} - \mathbf{x}^{(t+1)})$
$S_t \leftarrow S_L(\mathbf{w}^{(t)})$

Algorithm 4 $\{\hat{\mathbf{x}}, \hat{M}, \hat{\mu}\} \leftarrow \texttt{AdapAPG}(\phi, r, \mathbf{x}^{\text{ini}}, L_{\text{ini}}, \mu_0, \hat{\varepsilon})$

Choose: $L_{\min} \in [\mu_0, L_{\text{ini}}]$, $\gamma_{\text{dec}} \ge 1$, $\gamma_{\text{sc}} > 1$, $\theta_{\text{sc}} \in (0,1)$
$\{\mathbf{x}^{(0)}, M_{-1}, \mathbf{p}^{(-1)}, S_{-1}\} \leftarrow \texttt{LineSearch}(\phi, r, \mathbf{x}^{\text{ini}}, L_{\text{ini}})$
$\mathbf{x}^{(-1)} \leftarrow \mathbf{x}^{(0)}$, $L_0 \leftarrow M_{-1}$, $\mu \leftarrow \mu_0$, $\alpha_{-1} \leftarrow 1$, $\tau_0 \leftarrow 1$, $t \leftarrow 0$
repeat
  $\{\mathbf{x}^{(t+1)}, M_t, \alpha_t, \mathbf{p}^{(t)}, S_t\} \leftarrow \texttt{AccelLineSearch}(\phi, r, \mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, L_t, \mu, \alpha_{t-1})$
  $\tau_{t+1} \leftarrow \tau_t(1-\alpha_t)$
  if $\|\mathbf{p}^{(t)}\| \le \theta_{\text{sc}}\|\mathbf{p}^{(-1)}\|$ **then**
    $\mathbf{x}^{(0)} \leftarrow \mathbf{x}^{(t+1)}$, $\mathbf{x}^{(-1)} \leftarrow \mathbf{x}^{(t+1)}$, $L_0 \leftarrow M_t$, $\mathbf{p}^{(-1)} \leftarrow \mathbf{p}^{(t)}$, $M_{-1} \leftarrow M_t$, $S_{-1} \leftarrow S_t$
    $t \leftarrow 0$
  else
    if $2\sqrt{2\tau_t}\,\frac{M_t}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right) \le \theta_{\text{sc}}$ **then**
      $\mu \leftarrow \mu/\gamma_{\text{sc}}$
      $t \leftarrow 0$
    else
      $L_{t+1} \leftarrow \max\{L_{\min}, M_t/\gamma_{\text{dec}}\}$
      $t \leftarrow t+1$
    end if
  end if
until $\omega(\mathbf{x}^{(t+1)}) \le \hat{\varepsilon}$
$\hat{\mathbf{x}} \leftarrow \mathbf{x}^{(t+1)}$, $\hat{M} \leftarrow M_t$, $\hat{\mu} \leftarrow \mu$

With these notations, we briefly describe the AdapAPG method in Algorithm 4, where we treat ϕ and r as the input because we need to apply it to instances of (3.5) with different ϕ's. We refer the interested readers to [45] for details. AdapAPG calls two different line-search schemes that are described in Algorithm 2 and Algorithm 3, respectively. Here, Algorithm 2 is only used for initialization while Algorithm 3 is the main subroutine in each iteration of Algorithm 4. AdapAPG maintains and updates estimations of μ_ϕ and L_ϕ.

In iteration t, the AdapAPG method calls Algorithm 3, which performs the updating steps of the APG method [52] by using an estimation of μ_ϕ, denoted as μ, in place of μ_ϕ and using a line search scheme to update the estimation of L_ϕ, denoted as M_t. After the tth call of Algorithm 3, the AdapAPG method stores $\mathbf{p}^{(t)} = g_{M_t}(\mathbf{w}^{(t)})$ and $\mathbf{x}^{(t+1)} = T_{M_t}(\mathbf{w}^{(t)})$. It can be shown that, if $\mu \le \mu_\phi$, the value of $\|\mathbf{p}^{(t)}\|$ should decrease geometrically to zero. Therefore, if such a decrease is not observed, it must happen that $\mu > \mu_\phi$. Then the algorithm is restarted with μ divided by $\gamma_{\rm sc} > 1$, which leads to the adaptivity to the unknown μ_ϕ. The following theorem shows the complexity of the AdapAPG method for solving (3.5).

Theorem 5 (Theorem 2 in [45]) *Assume $\mu_0 \ge \mu_\phi > 0$. Let $\mathbf{p}^{\rm ini}$ denote the first $\mathbf{p}^{(-1)}$ computed by Algorithm* 4. *Algorithm* 4 *terminates in at most*

$$T_{\rm APG} = \left(\left\lceil \log_{1/\theta_{sc}} \left(\left(1 + \frac{L_\phi}{L_{\min}} \right) \frac{\|\mathbf{p}^{\rm ini}\|}{\hat{\varepsilon}} \right) \right\rceil + \left\lceil \log_{\gamma_{\rm sc}} \left(\frac{\mu_0}{\mu_\phi} \right) \right\rceil \right) \times \sqrt{\frac{L_\phi \gamma_{\rm inc} \gamma_{sc}}{\mu_\phi}} \ln \left(8 \left(\frac{L_\phi \gamma_{\rm inc} \gamma_{sc}}{\mu_\phi \theta_{sc}} \right)^2 \left(1 + \frac{L_\phi}{L_{\min}} \right)^2 \right) \tag{A.7}$$

iterations with an output $\bar{\mathbf{x}}$ satisfying (3.6) and the total complexity is

$$\left(1 + \frac{\ln \gamma_{\rm dec}}{\ln \gamma_{\rm inc}} \right) (T_{\rm APG} + 1) + \frac{1}{\ln \gamma_{\rm inc}} \max \left\{ \ln \frac{\gamma_{\rm inc} L_\phi}{\gamma_{\rm dec} \mu_\phi}, 0 \right\} = O(T_{\rm APG}), \tag{A.8}$$

where $T_{\rm APG}$ is the number of iterations given by (A.7).

Remark 5 In this paper, we measure the complexity of an algorithm using the total number of proximal gradient steps it performs. The value $T_{\rm APG}$ in (A.7) is the total number of iterations by Algorithm 4, but multiple proximal gradient steps can be performed in one iteration of Algorithm 4 inside the subroutine `AccelLineSearch`. According to the inequality between inequalities 16 and 17 in [45], the total number of proximal gradient steps is given in (A.8) which differs from $T_{\rm APG}$ only by logarithmic factors. If $0 < \mu_0 < \mu_\phi$, following the same proof as for Theorem 2 in [45], we can show that the total number of iterations performed by Algorithm 4 is at most

$$\left\lceil \log_{1/\theta_{sc}} \left(\left(1 + \frac{L_\phi}{L_{\min}} \right) \frac{\|\mathbf{p}^{\rm ini}\|}{\hat{\varepsilon}} \right) \right\rceil \times \sqrt{\frac{L_\phi \gamma_{\rm inc}}{\mu_0}} \ln \left(8 \left(\frac{L_\phi \gamma_{\rm inc}}{\mu_0 \theta_{sc}} \right)^2 \left(1 + \frac{L_\phi}{L_{\min}} \right)^2 \right),$$

which is obtained by replacing μ_ϕ / γ_{sc} by μ_0 in (A.7).

By this theorem, the total complexity of the AdapAPG method to find a solution $\bar{\mathbf{x}}$ to (3.5) satisfying (3.6) is $O\left(\kappa_\phi^{1/2} \log(\kappa_\phi) \log\left(\frac{1}{\hat{\varepsilon}} \right) \right)$,

where $\kappa_\phi = \frac{L_\phi}{\mu_\phi}$ is the condition number of (3.5). Compared to APG whose complexity is $O\left(\kappa_\phi^{1/2} \log\left(\frac{1}{\hat{\varepsilon}} \right) \right)$, AdapAPG has an additional factor of $\log(\kappa_\phi)$ in the

complexity result, but the latter does not require knowing the exact values of μ_ϕ and L_ϕ. We present this complexity result in terms of μ_ϕ, L_ϕ and $\hat{\varepsilon}$ in Theorem 1.

By (A.5) and the stopping condition (Line 19) of Algorithm 4, we can use AdapAPG to find $\bar{\mathbf{x}}^{(k+1)}$ satisfying (4.6) at Line 4 of Algorithm 1. To do so, we only need to initialize $\hat{M}$ and $\hat{\mu}$ at the beginning (between Lines 1 and 2) of Algorithm 1 as

$$\hat{M} \leftarrow M^{\text{ini}} \quad \text{and} \quad \hat{\mu} \leftarrow \mu^{\text{ini}} \tag{A.9}$$

using any constants μ^{ini} and M^{ini} satisfying $0 < \mu^{\text{ini}} \le M^{\text{ini}}$ and then replace Line 4 of Algorithm 1 with

$$\text{Call Alg. 4} : \{\bar{\mathbf{x}}^{(k+1)}, \hat{M}, \hat{\mu}\} \leftarrow \texttt{AdapAPG}\,(\phi_k, g, \bar{\mathbf{x}}^{(k)}, \hat{M}, \hat{\mu}, \hat{\varepsilon}_k). \tag{A.10}$$

This is also what we implement in our numerical experiments in Sect. 7.

In Algorithms 2, 3 and 4, users need to provide parameters $L_{\text{ini}} \ge \mu_0 > 0$, $L_{\min} \in [\mu_0, L_{\text{ini}}]$, $\gamma_{\text{inc}} > 1$, $\gamma_{\text{dec}} > 1$, $\gamma_{\text{sc}} > 1$, and $\theta_{\text{sc}} \in (0, 1)$. Next, we will explain the roles of these parameters. Parameters $L_{\min}$, L_{ini}, γ_{inc} and γ_{dec} are introduced for updating the local estimate of L_ϕ, which is M_t in Algorithm 4. More specifically, L_{ini} is the initial guess of L_ϕ while $L_{\min}$ is an estimated lower bound of L_ϕ. According to equation 12 in [45], the stopping condition $F(\mathbf{x}^{(t+1)}) \le \psi_L(\mathbf{x}^{(t)}; \mathbf{x}^{(t+1)})$ in Algorithms 2 and 3 holds whenever $L \ge L_\phi$. However, the condition $L \ge L_\phi$ is not necessary for the above stopping condition to hold for a particular t, and an overly large L will slow down the convergence. Hence, the estiamte M_t is reduced by a factor $\gamma_{\text{dec}} > 1$ in Line 15 in Algorithm 4 to ensure Algorithm 2 begins with a relatively small L and, in each iteration of Algorithms 2 and 3 , L is increased by a factor $\gamma_{\text{inc}} > 1$ to ensure the aforementioned stopping condition will eventually hold.

Parameters μ_0, γ_{sc} and θ_{sc} are introduced for updating the estimate of μ_ϕ, which is μ in Algorithm 4. Parameter θ_{sc} is the desired shrinking factor, which is used in the condition $\|\mathbf{p}^{(t)}\|_2 \le \theta_{\text{sc}} \|\mathbf{p}^{(-1)}\|_2$ for restarting the iterate at $\mathbf{x}^{(t+1)}$ as shown in Lines 8 and 9 in Algorithm 4. Because of (A.4), Algorithm 4 will terminate if this condition holds sufficiently many times. Parameter μ is initialized with μ_0. By Lemma 11 in [45], if $\mu \le \mu_\phi$, it must hold that

$$\|\mathbf{p}^{(t)}\| \le 2\sqrt{2\tau_t}\,\frac{M_t}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right)\|\mathbf{p}^{(-1)}\|,$$

where τ_t is updated in Line 6 of Algorithm 4 and decreases geometrically to zero. Hence, for any $\theta_{\text{sc}} \in (0, 1)$, if $2\sqrt{2\tau_t}\,\frac{M_t}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right) \le \theta_{\text{sc}}$ happens earlier than $\|\mathbf{p}^{(t)}\|_2 \le \theta_{\text{sc}}\|\mathbf{p}^{(-1)}\|_2$, we must have that $\mu > \mu_\phi$ and need to decrease μ by a factor $\gamma_{\text{sc}} > 1$ and restart the iterate at $\mathbf{x}^{(0)}$ as shown in Lines 12 and 13 of Algorithm 4.

Although the choices for the aforementioned parameters must depend on specific problems, we can provide some guideline in general for users in practice. In our iPPP method, L_{ini} and μ_0 are set to $\hat{M}$ and $\hat{\mu}$, which are initialized in (A.9) and updated after each call of Algorithm 4 in (A.10). Since AdapAPG can adaptively update $\hat{M}$ and $\hat{\mu}$, the initial value M^{ini} does not need to be large and μ^{ini} can be set relatively close to M^{ini}. In practice, we suggest setting $M^{\text{ini}} = 10$ and $\mu^{\text{ini}} \in [1, 10]$. According to Theorem 5 and the discussion afterwards, setting γ_{inc}, γ_{dec}, γ_{sc}, and θ_{sc} too large or too small

will increase the complexity. According to the numerical experiments in [45, 54] and in this paper, where different classes of problems are solved, γ_{inc} and γ_{dec} can be selected from a grid in $(1, 2]$, γ_{sc} in $(1, 10]$ and θ_{sc} in $(0, 1)$. The selection can be maded based on the objective value achieved after a fixed number of iterations.

Appendix 2: Discussion on Assumption 4 for application (7.1)

We explain that Assumption 4 can hold for the tested problem (7.1). For simplicity, we consider the case of $K = 2$, and in this case, we have a single inequality constraint in the form of

$$f(\mathbf{x}) := \frac{1}{N_2}\sum_{i=1}^{N_2} \phi(\mathbf{x}_2^\top \xi_i - \mathbf{x}_1^\top \xi_i) - r_2,$$

where $\mathbf{x} = [\mathbf{x}_1;\mathbf{x}_2]$, $\phi(z) = 1/(1+\exp(z))$, N_2 denotes the number of data points in $\mathcal{D}_2$, and ξ_i is the i-th data point in $\mathcal{D}_2$. In addition, let $\mathcal{X} = \{(\mathbf{x}_1, \mathbf{x}_2) : \|\mathbf{x}_1\| \le \lambda, \|\mathbf{x}_2\| \le \lambda\}$. It is easy to have

$$\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \begin{cases} \{\mathbf{0}\}, & \text{if } \|\mathbf{x}_1\| < \lambda, \|\mathbf{x}_2\| < \lambda, \\ \{\mathbf{0}\} \times \{a_2\mathbf{x}_2 : a_2 \ge 0\}, & \text{if } \|\mathbf{x}_1\| < \lambda, \|\mathbf{x}_2\| = \lambda, \\ \{a_1\mathbf{x}_1 : a_1 \ge 0\} \times \{\mathbf{0}\}, & \text{if } \|\mathbf{x}_1\| = \lambda, \|\mathbf{x}_2\| < \lambda, \\ \{a_1\mathbf{x}_1 : a_1 \ge 0\} \times \{a_2\mathbf{x}_2 : a_2 \ge 0\}, & \text{if } \|\mathbf{x}_1\| = \lambda, \|\mathbf{x}_2\| = \lambda. \end{cases}$$

The condition in Assumption 4 reduces to

$$\exists\, \nu > 0 \text{ such that } \nu[f(\mathbf{x})]_+ \le \operatorname{dist}\big([f(\mathbf{x})]_+ \nabla f(\mathbf{x}), -\mathcal{N}_{\mathcal{X}}(\mathbf{x})\big),\ \forall\, \mathbf{x} \in \mathcal{X}. \tag{B.1}$$

Let

$$\mathbf{E} = \frac{1}{N_2}[\xi_1, \ldots, \xi_{N_2}], \quad u_i(\mathbf{x}) = \phi'(\mathbf{x}_2^\top \xi_i - \mathbf{x}_1^\top \xi_i), \text{ for } i = 1, \ldots, N_2.$$

Then

$$\nabla f(\mathbf{x}) = \frac{1}{N_2}\sum_{i=1}^{N_2} \phi'(\mathbf{x}_2^\top \xi_i - \mathbf{x}_1^\top \xi_i)[-\xi_i;\xi_i] = [-\mathbf{E}\mathbf{u}(\mathbf{x});\mathbf{E}\mathbf{u}(\mathbf{x})].$$

When $f(\mathbf{x}) \le 0$, the condition in (B.1) trivially holds for any $\nu > 0$. Below, we assume the feasibility of the origin, i.e., $f(\mathbf{0}) \le 0$ as in our numerical experiment and also assume

$$\xi_i \ne \mathbf{0},\ \forall\, i, \quad \xi_i^\top \xi_j \ge 0,\ \forall\, i, j. \tag{B.2}$$

Notice that the condition in (B.2) may not naturally hold but can be ensured by lifting all data points by one more dimension, i.e., $\xi_i \leftarrow [\xi_i;c], \forall\, i$ for some $c > 0$. If the original data points are normalized, it suffices to take $c = 1$. For linear classification

models like (7.1), lifting the data is essentially changing the intercept coefficient that does not change the separability of the data.

However, the optimization problem will be changed slightly. The goal here is not to provide an equivalent transformation of the original optimization model so that it satisfies (B.2). Instead, we just want to present a practical scenario where (B.2) is ensured at modeling stage, instead of optimization stage, by normalizing and lifting data.

We establish the condition in (B.1) through discussing three cases on $\mathbf{x}$ with $f(\mathbf{x}) > 0$.

Case I $\|\mathbf{x}_1\| < \lambda, \|\mathbf{x}_2\| < \lambda$. In this case, $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{\mathbf{0}\}$, and thus the inequality in (B.1) becomes $\nu \le \|\nabla f(\mathbf{x})\|$. Let

$$\nu_1 = \inf_{\|\mathbf{x}_1\| < \lambda, \|\mathbf{x}_2\| < \lambda} \|\nabla f(\mathbf{x})\| = \min_{\mathbf{x}\in\mathcal{X}} \sqrt{2}\|\mathbf{Eu}(\mathbf{x})\|. \tag{B.3}$$

Since $\phi'(\mathbf{x}) < 0$ for any $\mathbf{x}$, there exist $\eta_1 > 0$ and $\eta_2 > 0$ such that $-\eta_2 \le u_i(\mathbf{x}) \le -\eta_1$, $\forall\, i$ for all $\mathbf{x} \in \mathcal{X}$. Hence $\nu_1 > 0$ by (B.2).

Case II $\|\mathbf{x}_1\| < \lambda, \|\mathbf{x}_2\| = \lambda$ or $\|\mathbf{x}_1\| = \lambda, \|\mathbf{x}_2\| < \lambda$. We only consider the former because the latter can be discussed in the same way. In the former case, $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{\mathbf{0}\} \times \{a_2\mathbf{x}_2 : a_2 \ge 0\}$, and the inequality in (B.1) becomes

$$\big(\nu[f(\mathbf{x})]_+\big)^2 \le \big\|[f(\mathbf{x})]_+\mathbf{Eu}(\mathbf{x})\big\|^2 + \min_{a_2\ge 0} \big\|[f(\mathbf{x})]_+\mathbf{Eu}(\mathbf{x}) + a_2\mathbf{x}_2\big\|^2,$$

which is implied by the fact that $\nu \le \|\mathbf{Eu}(\mathbf{x})\|$ with $\nu = \nu_1/\sqrt{2}$ and ν_1 defined in (B.3).

Case III $\|\mathbf{x}_1\| = \lambda, \|\mathbf{x}_2\| = \lambda$. In this case, $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{a_1\mathbf{x}_1 : a_1 \ge 0\} \times \{a_2\mathbf{x}_2 : a_2 \ge 0\}$, and the inequality in (B.1) becomes

$$\big(\nu[f(\mathbf{x})]_+\big)^2 \le \min_{a_1\ge 0} \big\|[f(\mathbf{x})]_+\mathbf{Eu}(\mathbf{x}) - a_1\mathbf{x}_1\big\|^2 + \min_{a_2\ge 0} \big\|[f(\mathbf{x})]_+\mathbf{Eu}(\mathbf{x}) + a_2\mathbf{x}_2\big\|^2,$$

which, as $f(\mathbf{x}) > 0$, is equivalent to

$$\nu^2 \le \min_{a_1\ge 0} \big\|\mathbf{Eu}(\mathbf{x}) - a_1\mathbf{x}_1\big\|^2 + \min_{a_2\ge 0} \big\|\mathbf{Eu}(\mathbf{x}) + a_2\mathbf{x}_2\big\|^2.$$

Let $\nu_2 \ge 0$ be defined as

$$\nu_2^2 = \min_{\substack{\|\mathbf{x}_1\| = \lambda \\ \|\mathbf{x}_2\| = \lambda \\ f(\mathbf{x}) \ge 0}} \left\{ \min_{a_1\ge 0} \big\|\mathbf{Eu}(\mathbf{x}) - a_1\mathbf{x}_1\big\|^2 + \min_{a_2\ge 0} \big\|\mathbf{Eu}(\mathbf{x}) + a_2\mathbf{x}_2\big\|^2 \right\}. \tag{B.4}$$

Notice that the minimum of the above problem is reached at a point $\bar{\mathbf{x}}$ and numbers $\bar{a}_1$ and $\bar{a}_2$. Suppose $\nu_2 = 0$. It must hold that $\bar{\mathbf{x}}_1 = \lambda\frac{\mathbf{Eu}(\bar{\mathbf{x}})}{\|\mathbf{Eu}(\bar{\mathbf{x}})\|}$ and $\bar{\mathbf{x}}_2 = -\lambda\frac{\mathbf{Eu}(\bar{\mathbf{x}})}{\|\mathbf{Eu}(\bar{\mathbf{x}})\|}$ with the corresponding $\bar{a}_1 = \bar{a}_2 = \frac{\|\mathbf{Eu}(\bar{\mathbf{x}})\|}{\lambda}$. Since $u_i(\bar{\mathbf{x}}) < 0$, $\forall\, i$, we have from (B.2) that $\bar{\mathbf{x}}_2^\top\xi_i - \bar{\mathbf{x}}_1^\top\xi_i = -\frac{\lambda}{\|\mathbf{Eu}(\bar{\mathbf{x}})\|}\xi_i^\top\mathbf{Eu}(\bar{\mathbf{x}}) > 0$ for all i. This means $f(\bar{\mathbf{x}}) < f(\mathbf{0}) \le 0$ by the

monotonicity of ϕ, which contradicts with the fact that $f(\bar{\mathbf{x}}) \geq 0$. Therefore, we must have $\nu_2 > 0$.

By the above discussions, we can set $\nu = \min\{\nu_1/\sqrt{2}, \nu_2\} > 0$ to ensure condition (B.1), which gives the following conclusion.

Claim: Assumption 4 can hold for the tested problem (7.1).

Appendix 3: Comparison with [47] and [8]

In this section, we compare [47] and [8] with this paper in two aspects: the definition of approximate stationary point and the assumptions needed for processing non-convex constraints.

Since [47] and [8] only consider the problems with inequality constraints and their key assumptions are also stated only for the problems with inequality constraints, we also assume $\mathbf{c}(\mathbf{x}) \equiv \mathbf{0}$ in (1.1) in this section. In addition, we assume Assumptions 1 and 3 during the comparison because, except Assumption 1A (smoothness), Assumptions 1 and 3 are also made in [47] and [8].

Note that [47] and [8] also require a (nearly) feasible solution and ensure a (nearly) feasible solution in each outer iteration. Our method does not require a (nearly) feasible solution but cannot ensure a (nearly) feasible solution in all intermediate iterations.

Definition of approximate stationary point

In [8], the authors define an (ε, δ)-**KKT point** for (1.1) with only inequality constraints as follows.

Definition 7 Suppose $\mathbf{c}(\mathbf{x}) \equiv \mathbf{0}$ in (1.1). Given $\varepsilon > 0$ and $\delta > 0$, a point $\widehat{\mathbf{x}}$ is an (ε, δ)-KKT point of (1.1) if there are $\bar{\mathbf{x}} \in \mathrm{dom}(g)$, $\bar{\boldsymbol{\xi}} \in \partial g(\bar{\mathbf{x}})$ and $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^m_+$ such that $f_i(\bar{\mathbf{x}}) \leq 0$ for $i = 1, \ldots, m$, and

$$\left\| \nabla f_0(\bar{\mathbf{x}}) + J_{\mathbf{f}}(\bar{\mathbf{x}})^\top \bar{\boldsymbol{\lambda}} + \bar{\boldsymbol{\xi}} \right\|^2 \leq \varepsilon, \tag{C.1a}$$

$$\sum_{i=1}^{m} |\bar{\lambda}_i f_i(\bar{\mathbf{x}})| \leq \varepsilon, \tag{C.1b}$$

$$\|\widehat{\mathbf{x}} - \bar{\mathbf{x}}\|^2 \leq \delta. \tag{C.1c}$$

A similar definition is considered by [47]. Note that (C.1a) and (C.1b) are identical to (1.3a) and (1.3c) except that the left-hand side of (C.1a) is squared. For the purpose of comparison, in the rest of this subsection, we assume $\widehat{\mathbf{x}}$ is an $(\varepsilon^2, \varepsilon^2)$-KKT point so that the point $\bar{\mathbf{x}}$ associated to $\widehat{\mathbf{x}}$ in Definition 7 satisfies (1.3a) just like an ε-stationary point in Definition 1. Now we discuss about

the connections of $\widehat{\mathbf{x}}$ and its associated point $\bar{\mathbf{x}}$ with an ε-stationary point in Definition 1.

- When the inequalities in (C.1) hold with ε and δ replaced by ε^2 and ε^2, $\bar{\mathbf{x}}$ satisfies (1.3a), $\left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\| = 0$ and $\sum_{i=1}^m |\bar{\lambda}_i f_i(\bar{\mathbf{x}})| \le \varepsilon^2$ where the last two inequalities are stronger than (1.3b) and (1.3c), respectively. However, Definition 7 does not require $\bar{\lambda}_i = 0$ if $f_i(\bar{\mathbf{x}}) < 0$ for $i = 1, \ldots, m$ as Definition 1.
- The algorithms in [47] and [8] can only find the near-KKT point $\widehat{\mathbf{x}}$ but not the associated point $\bar{\mathbf{x}}$. Unfortunately, $\widehat{\mathbf{x}}$ is not necessarily an ε-stationary point even if $\widehat{\mathbf{x}}$ is an $(\varepsilon^2, \varepsilon^2)$-KKT point. Consider the one-dimensional example $\min_{x\in[1,2]} x^2$ which is an instance of (1.1) with $f_0(x) = x^2$, $g(x) = \mathbf{1}_{[1,2]}(x)$, $\mathbf{c}(x) \equiv \mathbf{0}$ and $\mathbf{f}(x) \equiv \mathbf{0}$. For any $\varepsilon \in (0, 1)$, $\widehat{x} = 1 + \varepsilon^2$ is an $(\varepsilon^2, \varepsilon^2)$-KKT point associated to $\bar{x} = 1$. However, $\|\nabla f_0(\widehat{x})\| = 2 + 2\varepsilon^2 > \varepsilon$, which violates (1.3a).
- When $g(\mathbf{x}) \equiv 0$ and (3.4) holds, $\widehat{\mathbf{x}}$ will be an $O(\varepsilon)$-stationary point. In fact, since $\|\widehat{\mathbf{x}} - \bar{\mathbf{x}}\| \le \varepsilon$, by Assumption 1A and (3.4), $\widehat{\mathbf{x}}$ satisfies $\left\|\nabla f_0(\widehat{\mathbf{x}}) + J_{\mathbf{f}}(\widehat{\mathbf{x}})^\top \bar{\lambda}\right\| \le O(\varepsilon)$, $\left\|[\mathbf{f}(\widehat{\mathbf{x}})]_+\right\| \le O(\varepsilon)$ and $\sum_{i=1}^m |\bar{\lambda}_i f_i(\widehat{\mathbf{x}})| = O(\varepsilon)$. Unlike $\bar{\mathbf{x}}$, the point $\widehat{\mathbf{x}}$ is not stronger than an ε-stationary point in terms of the last two inequalities. The algorithms in [47] and [8] can find the $(\varepsilon^2, \varepsilon^2)$-KKT point $\widehat{\mathbf{x}}$ with complexity of $O(\frac{1}{\varepsilon^3})$ which is the same as our complexity under Assumption 4.

Key assumptions on non-convex constraints

The key assumption, called the uniform Slater's condition, made by [47] (see Assumption 1 in [47]) is that

$$\min_{\mathbf{x}'\in\mathcal{X}} \left\{ \max_{i=1,\ldots,m} f_i(\mathbf{x}') + \frac{\rho + \rho_\varepsilon}{2}\|\mathbf{x}' - \mathbf{x}\|^2 \right\} < -\sigma_\varepsilon \tag{C.2}$$

for any ε^2-feasible $\mathbf{x}$ (i.e., $\mathbf{x} \in \mathcal{X}$ and $f_i(\mathbf{x}) \le \varepsilon^2$ for $i = 1, \ldots, m$), where ε is the error of the approximate stationary point found by [47], ρ_ε and σ_ε are positive constants depending on ε, and $\rho = \max_{i=0,\ldots,m} \rho_i$ with ρ_i defined as in Assumption 3. As follows, we use examples to show that the uniform Slater's condition and Assumption 4 do not imply each other.

We consider the following one-dimensional problem

$$\min_{x\in[-0.5,0.5]} f_0(x) \ \text{ s.t. } \ \mathbf{f}(x) := [(x-1)^3 + 1; -(x+1)^3 + 1] \le \mathbf{0},$$

where f_0 is any smooth objective function. It is clear that $\max\{(x-1)^3 + 1, -(x+1)^3 + 1\} \ge 0$ for any x and attains zero at $x = 0$, meaning that $x = 0$ is the only feasible solution to this problem. Considering the left-hand side of (C.2), we can show that, for any $x \in \mathbb{R}$, $\rho > 0$ and $\rho_\varepsilon > 0$,

$$\min_{x'\in[-0.5,0.5]} \left\{ \max\{(x'-1)^3 + 1, -(x'+1)^3 + 1\} + \frac{\rho + \rho_\varepsilon}{2}\|x' - x\|^2 \right\} \ge 0,$$

which indicates the uniform Slater's condition by [47] does not hold. On the contrary, this problem satisfies Assumption 4. In fact, for $x \in (0, 0.5]$, we have $\|[\mathbf{f}(x)]_+\| = (x-1)^3 + 1$ and

$$\begin{aligned}&\operatorname{dist}\left(J_{\mathbf{f}}(x)^\top [\mathbf{f}(x)]_+, -\mathcal{N}_{[-0.5,0.5]}(x)\right) = 3(x-1)^2[(x-1)^3+1] \\ &= 3(x-1)^2\|[\mathbf{f}(x)]_+\| \geq \frac{3}{4}\|[\mathbf{f}(x)]_+\|,\end{aligned}$$

where the last inequality is because $3(x-1)^2 \geq \frac{3}{4}$ on $(0, 0.5]$. This means (6.3) holds on $(0, 0.5]$ with $\nu = \frac{3}{4}$. By symmetricity, (6.3) also holds on $[-0.5, 0)$ with $\nu = \frac{3}{4}$ and it holds trivially at $x = 0$.

Next, we modify the problem above slightly and obtain

$$\min_{x\in[-1,1]} f_0(x) \ \text{ s.t. } \ \mathbf{f}(x) := [(x-1)^3 + 27/64; -(x+1)^3 + 27/64] \leq \mathbf{0}, \tag{C.3}$$

where f_0 is any smooth weakly convex objective function with $\rho_0 = 12$. It can be easily verified that the two constaint functions above are also smooth weakly convex with $\rho_1 = \rho_2 = 12$ so that $\rho = 12$ in (C.2). This problem does not satisfy Assumption 4 because, when $x = 1$, $\|[\mathbf{f}(x)]_+\| = (x-1)^3 + 27/64 = 27/64$ while

$$\operatorname{dist}\left(J_{\mathbf{f}}(x)^\top [\mathbf{f}(x)]_+, -\mathcal{N}_{[-1,1]}(x)\right) = 3(x-1)^2[(x-1)^3 + 27/64] = 0,$$

so that (6.3) does not hold for any $\nu > 0$. On the contrary, this problem satisfies (C.2) with a small enough ε. In fact, since $[-0.25, 0.25]$ is the feasible set of (C.3), by continuity of the constraint functions, there exists $\delta > 0$ such that $[-0.25-\delta, 0.25+\delta]$ is the set of all ε^2-feasible solutions for any small enough ε. Given $x \in [-0.25-\delta, 0.25+\delta]$ and $\rho_\varepsilon = 1$, we have

$$\begin{aligned}&\min_{x'\in[-1,1]} \left\{ \max\left\{(x'-1)^3 + \frac{27}{64}, -(x'+1)^3 + \frac{27}{64}\right\} + \frac{\rho+\rho_\varepsilon}{2}\|x'-x\|^2 \right\} \leq -\frac{37}{64} + \frac{13}{2}x^2 \\ &\leq -\frac{37}{64} + \frac{13}{2}(0.25+\delta)^2,\end{aligned}$$

where the first inequality is obtained by taking $x' = 0$. Note that the limit of $-\frac{37}{64} + \frac{13}{2}(0.25+\delta)^2$ as δ approaches zero is $-\frac{11}{64}$. Hence, for a small enough ε, the corresponding δ will also be small enough so that the right-hand side of the last inequality above will be less than $-\sigma_\varepsilon$ for some $\sigma_\varepsilon \approx \frac{11}{64}$ and for any ε^2-feasible x, indicating that (C.2) holds.

When there are only inequality constraints in (1.1), we say the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at a feasible solution $\mathbf{x}$ to (1.1) if there exists a direction $\mathbf{d} \in \mathbb{R}^d$ such that

$$\nabla f_i(\mathbf{x})^\top \mathbf{d} < 0 \text{ for all } i \text{ such that } f_i(\mathbf{x}) = 0. \tag{C.4}$$

Assuming the initial solution is feasible, [8] proved that any limiting point of the iterates generated by its algorithm is feasible (see Theorem 3.3 in [8]). Then they assume that *MFCQ holds at any limiting point of the iterates* (see Assumption 3.6 and Lemma 3.7 in [8]). However, this assumption is algorithm-dependent while

Assumption 4 only depends on the problem. Hence, the comparison below will depend on where the limiting point is.

We consider the following one-dimensional problem

$$\min_{x\in[-1,1]} -x^3 \ \text{ s.t. } \ x \le 0,\ x^3 \le 0.$$

It is clear that the assumption by [8] does not hold if the limiting point is $x = 0$ (the optimal solution). In fact, when $x = 0$, both constraints are active but the gradient of the second constraint is zero so that no direction $\mathbf{d}$ can satisfy (C.4). On the contrary, this problem satisfies Assumption 4. In fact, (6.3) holds trivially for any $\nu > 0$ when $x \in [-1, 0]$. For $x \in (0, 1]$, we have $\|[\mathbf{f}(x)]_+\| = \sqrt{x^2 + x^6}$ and

$$\operatorname{dist}\left(J_{\mathbf{f}}(x)^\top [\mathbf{f}(x)]_+,\ -\mathcal{N}_{[-1,1]}(x)\right) = x + 3x^5 = \|[\mathbf{f}(x)]_+\| \frac{1+3x^4}{\sqrt{1+x^4}} \ge \|[\mathbf{f}(x)]_+\|,$$

which means (6.3) holds with $\nu = 1$. Then, we consider problem (C.3) again. According to the previous subsection, we know that (C.3) does not satisfy Assumption 4. However, MFCQ holds at any feasible solution to (C.3), namely, any $x \in [-0.25, 0.25]$. Hence, we conclude that the MFCQ assumption of [8] and Assumption 4 do not imply each other.

Appendix 4: Complexity with convex constraints and unbounded domain

In this section, we assume Assumption 1 holds except Assumption 1B. In other words, the domain $\mathcal{X}$ of g can be unbounded, e.g., when $g(\mathbf{x}) = \|\mathbf{x}\|_1$ or $g(\mathbf{x}) \equiv 0$. Throughout this section, we consider the special case of (1.1) with convex constraints, which is formulated as (5.1) and satisfies Assumption 2. In addition, we make the following assumption.

Assumption 6 Function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\mathbf{x})]_+\right\|^2$ has compact level sets on $\mathcal{X}$, that is, for all $\alpha \in \mathbb{R}_+$, the set

$$\mathcal{S}_\alpha := \left\{ \mathbf{x} \in \mathcal{X} \,\middle|\, \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\mathbf{x})]_+\right\|^2 \le \alpha \right\} \tag{D.1}$$

is compact. Moreover, there exist constants $\{B_{f_i}\}_{i=0}^m$ satisfying (3.4a), and $f_0(\mathbf{x}) + g(\mathbf{x})$ is bounded below on $\mathcal{X}$, that is, there exists $\underline{F} \in \mathbb{R}$ such that $f_0(\mathbf{x}) + g(\mathbf{x}) \ge \underline{F}$ for any $\mathbf{x} \in \mathcal{X}$.

This assumption is used to prove the iterates $\{\bar{\mathbf{x}}^{(k)}\}_{k\ge 0}$ of Algorithm 1 will stay in a bounded region (see Lemma 7 below), which is a key property to replace Assumption 1B in the proof of convergence. Set $\mathcal{S}_\alpha$ is compact, for example, when $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\mathbf{x})]_+\right\|^2$ is coercive or strongly convex. For many applications in

machine learning, $\underline{F}$ exists and equals zero. Condition (3.4a) is needed to ensure $\phi_k(\mathbf{x})$ in (4.5) is L_{ϕ_k}-smooth with L_{ϕ_k} in (5.2) so that the AdapAPG method can be applied to subproblem (4.1).

Since the diameter D of $\mathcal{X}$ is not necessarily finite, Lemma 1 needs to be modified as follows. The only changes are that $\hat{\varepsilon}_k D$ on the left-hand sides of (4.7) and (4.8) are replaced by $\hat{\varepsilon}_k^2/\mu_{\phi_k}$ as shown in (D.2) and (D.3). Moreover, the constant term $2B_{f_0} + 2G$ in (4.8) is replaced by $f_0(\mathbf{x}^{(0)}) + g(\mathbf{x}^{(0)}) - \underline{F}$ as shown in (D.3).

Lemma 6 *Suppose ϕ_k in (4.5) is μ_{ϕ_k}-strongly convex. Let $\{\bar{\mathbf{x}}^{(k)}\}$ be generated from Algorithm 1. Then for any $\mathbf{x} \in \mathcal{X}$, it holds that*

$$\phi_k(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \phi_k(\mathbf{x}) - g(\mathbf{x}) \le \frac{\hat{\varepsilon}_k^2}{2\mu_{\phi_k}}, \ \forall k \ge 0 \tag{D.2}$$

and

$$\begin{aligned} & s \sum_{k=0}^{K-1} \frac{\gamma_k}{2} \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{\beta_{K-1}}{2} \left(\|\mathbf{c}(\bar{\mathbf{x}}^{(K)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(K)})]_+\right\|^2 \right) \\ \le & f_0(\mathbf{x}^{(0)}) + g(\mathbf{x}^{(0)}) - \underline{F} + \frac{\beta_0}{2} \left(\|\mathbf{c}(\mathbf{x}^{(0)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 \right) \\ & + \frac{1}{2} \sum_{k=1}^{K-1} (\beta_k - \beta_{k-1}) \left(\|\mathbf{c}(\bar{\mathbf{x}}^{(k)})\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2 \right) + \sum_{k=0}^{K-1} \frac{\hat{\varepsilon}_k^2}{2\mu_{\phi_k}}, \ \forall K \ge 1. \end{aligned} \tag{D.3}$$

Proof According to Line 4 of Algorithm 1, there exists $\bar{\boldsymbol{\xi}}^{(k+1)} \in \partial g(\bar{\mathbf{x}}^{(k+1)})$ such that $\|\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)}\| \le \hat{\varepsilon}_k$. Since ϕ_k is μ_{ϕ_k}-strongly convex, so is $\phi_k + g$. Hence, we obtain (D.2) by noting

$$\begin{aligned} & \phi_k(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \phi_k(\mathbf{x}) - g(\mathbf{x}) \\ & \le \left(\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)} \right)^\top (\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}) - \frac{\mu_{\phi_k}}{2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}\|^2 \\ & \le \frac{\|\nabla \phi_k(\bar{\mathbf{x}}^{(k+1)}) + \bar{\boldsymbol{\xi}}^{(k+1)}\|^2}{2\mu_{\phi_k}} \le \frac{\hat{\varepsilon}_k^2}{2\mu_{\phi_k}}. \end{aligned}$$

Now let $\mathbf{x} = \bar{\mathbf{x}}^{(k)}$ in (D.2) and sum it over $k = 0$ through $K - 1$ to obtain (D.3) by the lower boundedness of $f_0(\mathbf{x}) + g(\mathbf{x})$ in Assumption 6. □

Lemma 7 *Suppose that Assumptions 2 and 6 hold and the parameters $\{\gamma_k\}$, $\{\beta_k\}$ and $\{\hat{\varepsilon}_k\}$ in Algorithm 1 are chosen as in (5.14). Each iterate $\bar{\mathbf{x}}^{(k)}$ will stay in $\mathcal{S}_{\bar{\alpha}}$ during Algorithm 1, where $\mathcal{S}_{\bar{\alpha}}$ is defined in (D.1) and*

$$\bar{\alpha} = \frac{2}{\beta} \left(f_0(\bar{\mathbf{x}}^{(0)}) + g(\bar{\mathbf{x}}^{(0)}) - \underline{F} \right) + \|\mathbf{A}\bar{\mathbf{x}}^{(0)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 + \frac{14}{5\beta^3(\gamma - \rho_0)} \tag{D.4}$$

with $\underline{F}$ from Assumption 6 and β from (5.14).

Proof Under (5.14) and Assumption 2, the function ϕ_k in (4.5) is μ_{ϕ_k}-strongly convex with $\mu_{\phi_k} = \gamma - \rho_0$. According to Lemma 6, we obtain

$$\phi_k(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \phi_k(\bar{\mathbf{x}}^{(k)}) - g(\bar{\mathbf{x}}^{(k)}) \leq \frac{\hat{\varepsilon}_k^2}{\gamma - \rho_0}, \ \forall k \geq 0 \tag{D.5}$$

by choosing $\mathbf{x} = \bar{\mathbf{x}}^{(k)}$ in (D.2). By the definition of ϕ_k in (4.5), subtracting $\underline{F}$ from both sides of the inequality in (D.5) and then dividing both sides by β_k give

$$\begin{aligned}
&\frac{f_0(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \underline{F}}{\beta_k} \\
&\quad + \frac{\gamma_k}{2\beta_k}\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{1}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2\right) \\
&\quad \leq \frac{f_0(\bar{\mathbf{x}}^{(k)}) + g(\bar{\mathbf{x}}^{(k)}) - \underline{F}}{\beta_k} + \frac{1}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2\right) \\
&\quad + \frac{\hat{\varepsilon}_k^2}{(\gamma - \rho_0)\beta_k}, \ \forall k \geq 0.
\end{aligned} \tag{D.6}$$

Since $\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|^2 \geq 0$, $f_0(\bar{\mathbf{x}}^{(k)}) + g(\bar{\mathbf{x}}^{(k)}) - \underline{F} \geq 0$ and $\beta_{k+1} \geq \beta_k$ for any k, (D.6) implies

$$\begin{aligned}
&\frac{f_0(\bar{\mathbf{x}}^{(k+1)}) + g(\bar{\mathbf{x}}^{(k+1)}) - \underline{F}}{\beta_{k+1}} + \frac{1}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2\right) \\
\leq &\frac{f_0(\bar{\mathbf{x}}^{(k)}) + g(\bar{\mathbf{x}}^{(k)}) - \underline{F}}{\beta_k} + \frac{1}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(k)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k)})]_+\right\|^2\right) + \frac{\hat{\varepsilon}_k^2}{(\gamma - \rho_0)\beta_k}, \ \forall k \geq 0.
\end{aligned}$$

Summing the inequality above over $k = 0$ through $K - 1$ gives

$$\begin{aligned}
&\frac{f_0(\bar{\mathbf{x}}^{(K)}) + g(\bar{\mathbf{x}}^{(K)}) - \underline{F}}{\beta_K} + \frac{1}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(K)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(K)})]_+\right\|^2\right) \\
&\quad \leq \frac{f_0(\bar{\mathbf{x}}^{(0)}) + g(\bar{\mathbf{x}}^{(0)}) - \underline{F}}{\beta_0} + \frac{1}{2}\left(\|\mathbf{A}\bar{\mathbf{x}}^{(0)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2\right) + \sum_{k=0}^{K-1} \frac{\hat{\varepsilon}_k^2}{(\gamma - \rho_0)\beta_k}.
\end{aligned} \tag{D.7}$$

Using the facts that $f_0(\bar{\mathbf{x}}^{(K)}) + g(\bar{\mathbf{x}}^{(K)}) - \underline{F} \geq 0$ and that

$$\sum_{k=0}^{K-1} \frac{\hat{\varepsilon}_k^2}{\beta_k} = \sum_{k=0}^{K-1} \frac{1}{\beta^3(k+1)^{3.5}} \leq \frac{1}{\beta^3}\left(1 + \int_1^K x^{-3.5} dx\right) \leq \frac{7}{5\beta^3},$$

we derive from (D.7)

$$\begin{aligned}
\|\mathbf{A}\bar{\mathbf{x}}^{(K)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(K)})]_+\right\|^2 &\leq \frac{2}{\beta}\left(f_0(\bar{\mathbf{x}}^{(0)}) + g(\bar{\mathbf{x}}^{(0)}) - \underline{F}\right) \\
&\quad + \|\mathbf{A}\bar{\mathbf{x}}^{(0)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 + \frac{14}{5\beta^3(\gamma - \rho_0)} = \bar{\alpha}.
\end{aligned}$$

Since K above can be any integer greater or equal to one, we have proved the conclusion for $k \geq 1$. The conclusion when $k = 0$ is trivially true. □

According to Lemma 7, if Assumptions 2 and 6 hold and the parameters $\{\gamma_k\}$, $\{\beta_k\}$ and $\{\hat{\epsilon}_k\}$ are chosen as in (5.14), we must have

$$D_{\bar{\alpha}} = \max_{\mathbf{x},\mathbf{x}' \in \mathcal{S}_{\bar{\alpha}}} \|\mathbf{x} - \mathbf{x}'\| < +\infty, \tag{D.8}$$

where $\mathcal{S}_{\bar{\alpha}}$ is defined in (D.1) with $\bar{\alpha}$ in (D.4). Using Lemma 7 and $D_{\bar{\alpha}}$, we obtain the following bounds for $(\widehat{\mathbf{y}}^{(k)}, \widehat{\lambda}^{(k)})$ similar to the ones in Lemma 3.

Lemma 8 *Suppose Assumptions 1 (except B), 2 and 6 hold and the parameters $\{\gamma_k\}$, $\{\beta_k\}$ and $\{\hat{\epsilon}_k\}$ in Algorithm 1 are chosen as in (5.14). Let $(\widehat{\mathbf{x}}^{(k)}, \widehat{\mathbf{y}}^{(k)}, \widehat{\lambda}^{(k)})$ be the solution satisfying the conditions in (5.4) and $\widehat{\mathbf{y}}^{(k)} \in \text{Range}(\mathbf{A}\mathbf{A}^\top)$ for $k \geq 0$. Then*

$$\|\widehat{\lambda}^{(k)}\| \leq M_\lambda^{\bar{\alpha}}(\gamma_k) := \frac{Q_k^{\bar{\alpha}}}{\min_i |f_i(\mathbf{x}_{\text{feas}})|} \tag{D.9}$$

$$\|\widehat{\mathbf{y}}^{(k)}\| \leq M_y^{\bar{\alpha}}(\gamma_k) := Q_k^{\bar{\alpha}} \|(\mathbf{A}\mathbf{A}^\top)^\dagger \mathbf{A}\| \left(\frac{1}{D_{\bar{\alpha}}} + \frac{1}{\text{dist}(\mathbf{x}_{\text{feas}}, \partial\mathcal{X})} + \frac{\max_i B_{f_i}}{\min_i |f_i(\mathbf{x}_{\text{feas}})|} \right), \tag{D.10}$$

where $Q_k^{\bar{\alpha}} = D_{\bar{\alpha}}(B_{f_0} + \gamma_k D_{\bar{\alpha}} + M)$, $D_{\bar{\alpha}}$ is defined in (D.8), B_{f_i} is defined in (3.4a), and $(\mathbf{A}\mathbf{A}^\top)^\dagger$ denotes the pseudoinverse of $\mathbf{A}\mathbf{A}^\top$.

Proof Since $\widehat{\mathbf{x}}^{(k)}$ defined in (5.3) and $\mathbf{x}_{\text{feas}}$ in Assumption 2 are feasible to (5.1), both of them must be in $\mathcal{S}_0 \subset \mathcal{S}_{\bar{\alpha}}$ for any $k \geq 0$. Then the proof of (D.9) and (D.10) will be the same as Lemma 3 except that D is replaced by $D_{\bar{\alpha}}$. □

Similar to Lemma 4, the next lemma bounds the feasibility violation of iterate $\bar{\mathbf{x}}^{(k+1)}$.

Lemma 9 *Suppose Assumptions 1 (except B) and 2 hold. Given $\gamma_k > \rho_0$ and $\beta_k > 0$ for $k \geq 0$, let ϕ_k be defined in (4.5) with $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $\widehat{\mathbf{x}}^{(k)}$ be defined in (5.3), and $\bar{\mathbf{x}}^{(k+1)}$ be generated as in Algorithm 1. Then for any $k \geq 0$,*

$$\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \leq \frac{2\hat{\epsilon}_k^2}{(\gamma_k - \rho_0)\beta_k} + \frac{4\|\widehat{\mathbf{y}}^{(k)}\|^2}{\beta_k^2} + \frac{4\|\widehat{\lambda}^{(k)}\|^2}{\beta_k^2}. \tag{D.11}$$

Proof The proof is the same as Lemma 4 except that (D.2) with $\mu_{\phi_k} = \gamma_k - \rho_0$ is used in place of (4.7) throughout the proof. □

With the lemmas introduced above, we next analyze the complexity of Algorithm 1 in order to find an ϵ-stationary point of (5.1) when Assumption 1B is replaced by Assumption 6.

Theorem 6 *Suppose that Assumptions 1 (except B), 2 and 6 hold and the parameters* $\{\gamma_k\}, \{\beta_k\}$ *and* $\{\hat{\epsilon}_k\}$ *in Algorithm* 1 *are chosen as in* (5.14). *If* R_k *is defined as in* (4.3), *it holds for any* $K \geq 1$ *that*

$$\begin{aligned}\max\left\{\mathbf{S}_{R_K}, \mathbf{F}_{R_K}, \mathbf{C}_{R_K}\right\} \leq & \frac{3}{2\beta^2(\gamma-\rho_0)K} + \sqrt{\frac{2\gamma\mathcal{C}_1^{\bar{\alpha}}}{K}} + \frac{4\sqrt{1/(2\beta(\gamma-\rho_0))+(M_y^{\bar{\alpha}})^2+(M_\lambda^{\bar{\alpha}})^2}}{\beta\sqrt{K}} \\ & + \frac{8(1/(2\beta(\gamma-\rho_0))+(M_y^{\bar{\alpha}})^2+(M_\lambda^{\bar{\alpha}})^2)}{\beta\sqrt{K}},\end{aligned} \tag{D.12}$$

where $\{(\mathbf{S}_k, \mathbf{F}_k, \mathbf{C}_k)\}_{k\geq 1}$ *is defined in* (4.2), $M_y^{\bar{\alpha}} = M_y^{\bar{\alpha}}(\gamma)$, $M_\lambda^{\bar{\alpha}} = M_\lambda^{\bar{\alpha}}(\gamma)$ *defined in* (D.9) *and* (D.10), *and*

$$\begin{aligned}\mathcal{C}_1^{\bar{\alpha}} = & f_0(\bar{\mathbf{x}}^{(0)}) + g(\bar{\mathbf{x}}^{(0)}) - \underline{F} + \frac{\beta}{2}\|\mathbf{A}\bar{\mathbf{x}}^{(0)} - \mathbf{b}\|^2 \\ & + \frac{\beta}{2}\left\|[\mathbf{f}(\bar{\mathbf{x}}^{(0)})]_+\right\|^2 + \frac{3}{\beta}\Big(1/(\beta(\gamma-\rho_0)) + (M_y^{\bar{\alpha}})^2 + (M_\lambda^{\bar{\alpha}})^2\Big).\end{aligned} \tag{D.13}$$

Proof Notice that ϕ_k is strongly convex when $\gamma_k > \rho_0$. Hence, (D.3) holds.

Since $\gamma_k = \gamma$ for all k, we have from Lemma 8 that $\|\widehat{\lambda}^{(k)}\| \leq M_\lambda^{\bar{\alpha}}$ and $\|\widehat{\mathbf{y}}^{(k)}\| \leq M_y^{\bar{\alpha}}$ for all k. Hence, it follows from (D.11) and (5.14) that

$$\begin{aligned}\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} - \mathbf{b}\|^2 + \left\|[\mathbf{f}(\bar{\mathbf{x}}^{(k+1)})]_+\right\|^2 \leq & \frac{2\hat{\epsilon}_k^2}{(\gamma-\rho_0)\beta_k} + \frac{4((M_y^{\bar{\alpha}})^2+(M_\lambda^{\bar{\alpha}})^2)}{\beta_k^2} \\ \leq & \frac{4\big(1/(2\beta(\gamma-\rho_0)) + (M_y^{\bar{\alpha}})^2 + (M_\lambda^{\bar{\alpha}})^2\big)}{\beta_k^2},\end{aligned} \tag{D.14}$$

for any $k \geq 0$. This inequality is the same as (5.17) except that D in (5.17) is replaced by $1/(2\beta(\gamma-\rho_0))$. The rest of the proof is the same as Theorem 2 except that (5.17) is replaced by (D.14), (4.8) is replaced by (D.3), and constant $\mathcal{C}_1$ is repalced by $\mathcal{C}_1^{\bar{\alpha}}$ throughout the proof. □

According to Theorem 6, the convergence rate of Algorithm 1 is still $O(\frac{1}{\sqrt{K}})$ in terms of the number of outer iterations K, which is the same as Theorem 2. Suppose (4.6) is guaranteed by applying the AdapAPG method in Algorithm 4 in Appendix 1 to (4.1). We can analyze the total complexity of Algorithm 1 based on the complexity of Algorithm 4 in Theorem 1.

Corollary 4 (complexity result) *Under the assumptions of Theorem 6, let*

$$K = \left\lceil \max\left\{ \frac{3}{\beta(\gamma-\rho_0)\varepsilon}, \frac{4}{\varepsilon^2}\left[\sqrt{2\gamma \mathcal{C}_1^{\bar{\alpha}}} + \frac{4\sqrt{\frac{1}{2\beta(\gamma-\rho_0)} + (M_y^{\bar{\alpha}})^2 + (M_\lambda^{\bar{\alpha}})^2}}{\beta} + \frac{8(\frac{1}{2\beta(\gamma-\rho_0)} + (M_y^{\bar{\alpha}})^2 + (M_\lambda^{\bar{\alpha}})^2)}{\beta} \right]^2 \right\} \right\rceil$$
$$= O(1/\varepsilon^2)$$

where $\mathcal{C}_1^{\bar{\alpha}}$ *is defined as in* (D.13). *Then*

$\bar{\mathbf{x}}^{(R_K)}$ *is an* ε*-stationary point of* (5.1). *In addition, if* $\bar{\mathbf{x}}^{(k+1)}$ *satisfying* (4.6) *is found by the AdapAPG method, the total complexity for Algorithm* 1 *to produce* $\bar{\mathbf{x}}^{(R_K)}$ *is* $\tilde{O}\left(1/\varepsilon^{\frac{5}{2}}\right)$.

Proof

With the given K, the right hand side of (D.12) is upper bounded by ε. Hence, $\bar{\mathbf{x}}^{(R_K)}$ is an ε-stationary point of (5.1). The total complexity is obtained by the same procedure in the proof of Corollary 1.

□

Acknowledgements The authors would like to thank the action editor and two anonymous reviewers for their constructive comments and suggestions that greatly help improve the paper. The work of Xu Y. is partly supported by the NSF grant DMS-2053493.

Data availability The datasets generated during and/or analyzed during the current study are available in the LIBSVM Data: Classification (Multi-class) repository, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html.

References

1. Allen-Zhu, Z.: Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 89–97 (2017)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
3. Basu, K., Nandy, P.: Optimal convergence for stochastic optimization with multiple expectation constraints. arXiv preprint arXiv:1906.03401 (2019)
4. Bayandina, A., Dvurechensky, P., Gasnikov, A., Stonyakin, F., Titov, A.: Mirror descent and convex optimization problems with non-smooth inequality constraints. In: Large-Scale and Distributed Optimization, pp. 181–213. Springer (2018)
5. Birgin, E., Martínez, J.: Complexity and performance of an augmented Lagrangian algorithm. arXiv preprint arXiv:1907.02401 (2019)
6. Birgin, E.G., Floudas, C.A., Martínez, J.M.: Global minimization using an augmented Lagrangian method with variable lower-level constraints. Math. Program. **125**(1), 139–162 (2010)
7. Birgin, E.G., Haeser, G., Ramos, A.: Augmented Lagrangians with constrained subproblems and convergence to second-order stationary points. Comput. Optim. Appl. **69**(1), 51–75 (2018)

8. Boob, D., Deng, Q., Lan, G.: Proximal point methods for optimization with nonconvex functional constraints. arXiv preprint arXiv:1908.02734 (2019)
9. Burke, J.V.: An exact penalization viewpoint of constrained optimization. SIAM J. Control Optim. **29**(4), 968–998 (1991)
10. Byrd, R.H., Gould, N.I., Nocedal, J., Waltz, R.A.: On the convergence of successive linear-quadratic programming algorithms. SIAM J. Optim. **16**(2), 471–489 (2005)
11. Cartis, C., Gould, N.I., Toint, P.L.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. **21**(4), 1721–1739 (2011)
12. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of finding first-order critical points in constrained nonlinear optimization. Math. Program. **144**(1–2), 93–106 (2014)
13. Cartis, C., Gould, N.I., Toint, P.L.: Corrigendum: on the complexity of finding first-order critical points in constrained nonlinear optimization. Math. Program. **161**(1–2), 611–626 (2017)
14. Crammer, K., Singer, Y.: On the learnability and design of output codes for multiclass problems. Mach. Learn. **47**(2–3), 201–233 (2002)
15. Curtis, F.E., Gould, N.I., Jiang, H., Robinson, D.P.: Adaptive augmented Lagrangian methods: algorithms and practical numerical experience. Optim. Methods Softw. **31**(1), 157–186 (2016)
16. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. arXiv preprint arXiv:1803.06523 (2018)
17. Davis, D., Drusvyatskiy, D.: Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. arXiv preprint arXiv:1802.02988 (2018)
18. Davis, D., Grimmer, B.: Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. arXiv preprint arXiv:1707.03505 (2017)
19. Di Pillo, G., Grippo, L.: A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints. SIAM J. Control Optim. **23**(1), 72–84 (1985)
20. Di Pillo, G., Grippo, L.: An exact penalty function method with global convergence properties for nonlinear programming problems. Math. Program. **36**(1), 1–18 (1986)
21. Drusvyatskiy, D., Paquette, C.: Efficiency of minimizing compositions of convex functions and smooth maps. Math. Program. (2018)
22. Fernández, D., Solodov, M.V.: Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. SIAM J. Optim. **22**(2), 384–407 (2012)
23. Fletcher, R.: Penalty functions. In: Mathematical Programming The State of the Art, pp. 87–114. Springer (1983)
24. Friedlander, M.P., Saunders, M.A.: A globally convergent linearly constrained Lagrangian method for nonlinear optimization. SIAM J. Optim. **15**(3), 863–897 (2005)
25. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**(4), 2341–2368 (2013). https://doi.org/10.1137/120880811
26. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**(1–2), 59–99 (2016). https://doi.org/10.1007/s10107-015-0871-8
27. Gonçalves, M.L., Melo, J.G., Monteiro, R.D.: Convergence rate bounds for a proximal admm with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. arXiv preprint arXiv:1702.01850 (2017)
28. Gould, N.I.M.: On the convergence of a sequential penalty function method for constrained minimization. SIAM J. Numer. Anal. **26**(1), 107–128 (1989)
29. Grapiglia, G.N., Yuan, Y.x.: On the complexity of an augmented Lagrangian method for nonconvex optimization. arXiv preprint arXiv:1906.05622 (2019)
30. Haeser, G., Liu, H., Ye, Y.: Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. Math. Program. **178**(1–2), 263–299 (2019)
31. Hajinezhad, D., Hong, M.: Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. Math. Program. **176**(1–2), 207–245 (2019)
32. Hong, M.: Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: algorithms, convergence, and applications. arXiv preprint arXiv:1604.00543 (2016)
33. Hong, M., Lee, J.D., Razaviyayn, M.: Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization. arXiv preprint arXiv:1802.08941 (2018)
34. Jiang, B., Lin, T., Ma, S., Zhang, S.: Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. Comput. Optim. Appl. **72**(1), 115–157 (2019)

35. Jiang, B., Meng, X., Wen, Z., Chen, X.: An exact penalty approach for optimization with nonnegative orthogonality constraints. arXiv preprint arXiv:1907.12424 (2019)
36. Kong, W., Melo, J.G., Monteiro, R.D.: Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. SIAM J. Optim. **29**(4), 2566–2593 (2019)
37. Kong, W., Melo, J.G., Monteiro, R.D.: Iteration-complexity of a proximal augmented Lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. arXiv preprint arXiv:2008.07080 (2020)
38. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order penalty methods for convex programming. Math. Program. **138**(1–2), 115–139 (2013)
39. Lan, G., Yang, Y.: Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. arXiv:1805.05411 (2018)
40. Lan, G., Zhou, Z.: Algorithms for stochastic optimization with expectation constraints. arXiv preprint arXiv:1604.03887 (2016)
41. Li, Z., Chen, P.Y., Liu, S., Lu, S., Xu, Y.: Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. arXiv preprint arXiv:2007.01284 (2020)
42. Li, Z., Xu, Y.: Augmented Lagrangian based first-order methods for convex and nonconvex programs: nonergodic convergence and iteration complexity. arXiv preprint arXiv:2003.08880 (2020)
43. Lin, Q., Ma, R., Xu, Y.: Inexact proximal-point penalty methods for non-convex optimization with non-convex constraints. arXiv preprint arXiv:190811518v1 (2019)
44. Lin, Q., Ma, R., Yang, T.: Level-set methods for finite-sum constrained convex optimization. In: International Conference on Machine Learning, pp. 3118–3127 (2018)
45. Lin, Q., Xiao, L.: An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. Comput. Optim. Appl. **60**(3) (2015)
46. Lu, S., Razaviyayn, M., Yang, B., Huang, K., Hong, M.: Snap: Finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems. arXiv preprint arXiv:1907.04450 (2019)
47. Ma, R., Lin, Q., Yang, T.: Proximally constrained methods for weakly convex optimization with weakly convex constraints. arXiv preprint arXiv:1908.01871 (2019)
48. Melo, J.G., Monteiro, R.D.: Iteration-complexity of a Jacobi-type non-euclidean ADMM for multi-block linearly constrained nonconvex programs. arXiv preprint arXiv:1705.07229 (2017)
49. Melo, J.G., Monteiro, R.D.: Iteration-complexity of an inner accelerated inexact proximal augmented Lagrangian method based on the classical Lagrangian function and a full Lagrange multiplier update. arXiv preprint arXiv:2008.00562 (2020)
50. Melo, J.G., Monteiro, R.D., Wang, H.: Iteration-complexity of an inexact proximal accelerated augmented Lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. arXiv preprint arXiv:2006.08048 (2020)
51. Necoara, I., Patrascu, A., Glineur, F.: Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. Optim. Methods Softw. **34**(2), 305–335 (2019)
52. Nesterov, Y.: Introductory lectures on convex optimization : a basic course. Kluwer Academic Publ, Dordrecht (2004)
53. Nesterov, Y.: Barrier subgradient method. Math. Program. **127**(1), 31–56 (2011)
54. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
55. Nouiehed, M., Lee, J.D., Razaviyayn, M.: Convergence to second-order stationarity for constrained non-convex optimization. arXiv preprint arXiv:1810.02024 (2018)
56. O'Neill, M., Wright, S.J.: A log-barrier newton-cg method for bound constrained optimization with complexity guarantees. arXiv preprint arXiv:1904.03563 (2019)
57. Powell, M.J., Yuan, Y.: A recursive quadratic programming algorithm that uses differentiable exact penalty functions. Math. Program. **35**(3), 265–278 (1986)
58. Reddi, S.J., Hefny, A., Sra, S., Póczós, B., Smola, A.: Stochastic variance reduction for nonconvex optimization. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML), pp. 314–323. JMLR.org (2016)
59. Rockafellar, R.: Convex Analysis. Princeton University Press, Princeton Mathematical Series, Princeton (1970)
60. Sahin, M.F., eftekhari, A., Alacaoglu, A., Latorre, F., Cevher, V.: An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. In: Advances in Neural Information Processing Systems (2019)

61. Tran-Dinh, Q.: Proximal alternating penalty algorithms for nonsmooth constrained convex optimization. Comput. Optim. Appl. **72**(1), 1–43 (2019)
62. Tran-Dinh, Q., Cevher, V.: A primal-dual algorithmic framework for constrained convex minimization. arXiv preprint arXiv:1406.5403 (2014)
63. Tran-Dinh, Q., Kyrillidis, A., Cevher, V.: Composite self-concordant minimization. arXiv preprint arXiv 1308 (2013)
64. Tran-Dinh, Q., Kyrillidis, A., Cevher, V.: An inexact proximal path-following algorithm for constrained convex minimization. SIAM J. Optim. **24**(4), 1718–1745 (2014)
65. Tran-Dinh, Q., Kyrillidis, A., Cevher, V.: A single-phase, proximal path-following framework. Math. Oper. Res. **43**(4), 1326–1347 (2018)
66. Wang, F., Xu, Z., Xu, H.K.: Convergence of bregman alternating direction method with multipliers for nonconvex composite problems. arXiv preprint arXiv:1410.8625 (2014)
67. Wang, S., Chang, T.H., Cui, Y., Pang, J.S.: Clustering by orthogonal nmf model and non-convex penalty optimization. arXiv preprint arXiv:1906.00570 (2019)
68. Wang, X., Ma, S., Yuan, Y.X.: Penalty methods with stochastic approximation for stochastic nonlinear programming. Math. Comput. **86**(306), 1793–1820 (2017)
69. Wang, Y., Yin, W., Zeng, J.: Global convergence of Admm in nonconvex nonsmooth optimization. J. Sci. Comput. **78**(1), 29–63 (2019)
70. Wei, X., Neely, M.J.: Primal-dual frank-wolfe for constrained stochastic programs with convex and non-convex objectives. arXiv preprint arXiv:1806.00709 (2018)
71. Wei, X., Yu, H., Ling, Q., Neely, M.: Solving non-smooth constrained programs with lower complexity than $O(1/\varepsilon)$: a primal-dual homotopy smoothing approach. In: Advances in Neural Information Processing Systems, pp. 3995–4005 (2018)
72. Weston, J., Watkins, C.: Multi-class support vector machines. Tech. rep, Citeseer (1998)
73. Xie, Y., Wright, S.J.: Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. arXiv preprint arXiv:1908.00131 (2019)
74. Xu, Y.: First-order methods for constrained convex programming based on linearized augmented Lagrangian function. arXiv preprint arXiv:1711.08020 (2017)
75. Xu, Y.: Primal-dual stochastic gradient method for convex programs with many functional constraints. arXiv preprint arXiv:1802.02724 (2018)
76. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. Math. Program. Ser. A 1–46 (2019)
77. Yang, T., Lin, Q., Zhang, L.: A richer theory of convex constrained optimization with reduced projections and improved rates. In: Proceedings of the 34th International Conference on Machine Learning-vol. 70, pp. 3901–3910. JMLR. org (2017)
78. Yu, H., Neely, M., Wei, X.: Online convex optimization with stochastic constraints. In: Advances in Neural Information Processing Systems, pp. 1428–1438 (2017)
79. Yu, H., Neely, M.J.: A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs. SIAM J. Optim. **27**(2), 759–783 (2017)
80. Zhang, J., Luo, Z.: A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. arXiv preprint arXiv:2006.16440 (2020)
81. Zhang, J., Luo, Z.Q.: A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. SIAM J. Optim. **30**(3), 2272–2302 (2020)
82. Zhang, S., He, N.: On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. arXiv preprint arXiv:1806.04781 (2018)