First-order methods can have almost the same convergence rate as for unconstrained problems when there are $O(1)$ functional constraints

Yangyang Xu

February 18, 2020

Abstract First-order methods (FOMs) have recently been applied and analyzed for solving problems with complicated functional constraints. Existing works show that FOMs for functional constrained problems have lower-order convergence rates than those for unconstrained problems. In particular, an FOM for a smooth strongly-convex problem can have linear convergence, while it can only converge sublinearly for a constrained problem if the projection onto the constraint set is prohibited. In this paper, we point out that the slower convergence is caused by the large number of functional constraints but not the constraints themselves. When there are only $m = O(1)$ functional constraints, we show that an FOM can have almost the same convergence rate as that for solving an unconstrained problem, even without the projection onto the feasible set. In addition, given an $\varepsilon > 0$, we show that a complexity result that is better than a lower bound can be obtained, if there are only $m = o(\varepsilon^{-\frac{1}{2}})$ functional constraints. Our result is surprising but does not contradict to the existing lower complexity bound, because we focus on a specific subclass of problems.

Keywords: first-order method, cutting-plane method, nonlinearly constrained problem, iteration complexity

Mathematics Subject Classification: 65K05, 68Q25, 90C30, 90C60

1 Introduction

In this paper, we consider the constrained convex programming

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}),\ \text{s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})] \le \mathbf{0}, \tag{1}$$

where f is a differentiable convex function with a Lipschitz continuous gradient, h is a simple closed convex function, and each g_i is convex differentiable and has a Lipschitz continuous gradient.

For a smooth strongly-convex linearly-constrained problem $\min_{\mathbf{x}}\{f(\mathbf{x}),\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}\}$, [23] gives a lower complexity bound $O(\frac{1}{\sqrt{\varepsilon}})$ of first-order methods (FOMs) to produce an ε-optimal solution, if $\mathbf{A}$ can be inquired only by the matrix-vector multiplication $\mathbf{A}(\cdot)$ and $\mathbf{A}^\top(\cdot)$. Notice $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\} = \{\mathbf{x} : \mathbf{A}\mathbf{x} \le$

Y. Xu
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180
E-mail: xuy21@rpi.edu

$\mathbf{b}, -\mathbf{A}\mathbf{x} \le -\mathbf{b}\}$. In addition, if $\nabla f(\mathbf{x}) + \mathbf{A}^\top \mathbf{y} = \mathbf{0}$, then $\nabla f(\mathbf{x}) + \mathbf{A}^\top \mathbf{y}^+ - \mathbf{A}^\top \mathbf{y}^- = \mathbf{0}$, where $\mathbf{y}^+ \ge \mathbf{0}$ and $\mathbf{y}^- \ge \mathbf{0}$ denote the positive and negative parts of $\mathbf{y}$. Hence, if the linear-equality constrained problem has a KKT point, then so does the equivalent linear-inequality constrained problem. Therefore, the lower bound in [23] also applies to the inequality constrained problem (1), if $\mathbf{g}$ can be accessed only through its function value and derivative. However, for the special case of $\mathbf{g} \equiv \mathbf{0}$ or $m = 0$, an accelerated proximal gradient method [22] can achieve a complexity result $O(\sqrt{\kappa}|\log\varepsilon|)$ to produce an ε-optimal solution of (1), when f or h is strongly convex. Here, κ denotes the condition number.

The worst-case instance constructed in [23] relies on the condition that m is in the same or higher order of $\frac{1}{\sqrt{\varepsilon}}$. For the case with $m = o(\frac{1}{\sqrt{\varepsilon}})$, the lower bound $O(\frac{1}{\sqrt{\varepsilon}})$ may not hold any more. Examples of (1) with small m include the Neyman-Pearson classification problem [24], fairness-constrained classification [30], and the risk-constrained portfolio optimization [7]. Therefore, we pose the following question while solving a strongly-convex problem in the form of (1):

> Given $\varepsilon > 0$, can an FOM achieve a better complexity result than $O(\frac{1}{\sqrt{\varepsilon}})$ to produce an ε-optimal solution of (1) when $m = o(\frac{1}{\sqrt{\varepsilon}})$, or even achieve $\tilde{O}(\sqrt{\kappa})$ when $m = O(1)$?

Here, an FOM for (1) only uses the function value and derivative information of f and $\mathbf{g}$ and also the proximal mapping of h and its multiples, and $\tilde{O}$ suppresses a polynomial of $|\log\varepsilon|$. We will give an affirmative answer to the above question.

1.1 Algorithmic framework

The FOM that we will design and analyze is based the inexact augmented Lagrangian method (iALM). The classic AL function of (1) is:

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{z}) = F(\mathbf{x}) + \frac{\beta}{2}\left\|[\mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}]_+\right\|^2 - \frac{\|\mathbf{z}\|^2}{2\beta}, \tag{2}$$

where $\mathbf{z}$ is the multiplier vector, and $[\mathbf{a}]_+$ takes the compoment-wise positive part of a vector $\mathbf{a}$. The pseudocode of a first-order iALM is shown in Algorithm 1. Notice that $\mathcal{L}_\beta$ is convex about $\mathbf{x}$ and concave about $\mathbf{z}$. Hence, we can directly apply the accelerated proximal gradient in [22] to solve each $\mathbf{x}$-subproblem. However, that way can only give a complexity result of $O(\frac{1}{\sqrt{\varepsilon}})$ as shown in [26], regardless of the value of m. To have a better overall complexity, we will design a new FOM to solve each $\mathbf{x}$-subproblem by utilizing the condition $m = O(1)$ or $m = o(\frac{1}{\sqrt{\varepsilon}})$.

Algorithm 1: First-order inexact augmented Lagrangian method for (1)

Initialization: choose $\mathbf{x}^0, \mathbf{z}^0$, and $\beta_0 > 0$
for $k = 0, 1, \ldots$ **do**
  Apply a first-order method to find $\mathbf{x}^{k+1}$ as an approximate solution of $\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$.
  Update $\mathbf{z}$ by $\mathbf{z}^{k+1} = [\mathbf{z}^k + \beta_k \mathbf{g}(\mathbf{x}^{k+1})]_+$.
  Choose $\beta_{k+1} \ge \beta_k$.
  if *a stopping condition is satisfied* **then**
    Output $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ and stop

1.2 Related works

We briefly mention some existing works that also study the complexity of FOMs for solving functional constrained problems.

By using the ordinary Lagrangian function, [19,20] analyze a dual subgradient method for general convex problems. The method needs $O(\varepsilon^{-2})$ subgradient evaluations to produce an ε-optimal solution (see the definition in Eq. (4) below). For a smooth problem, [18] studies the complexity of an inexact dual gradient (IDG) method. Suppose that an optimal FOM is applied to each outer-subproblem of IDG. Then to produce an ε-optimal solution, IDG needs $O(\varepsilon^{-\frac{3}{2}})$ gradient evaluations when the problem is convex, and the result can be improved to $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ when the problem is strongly convex. The primal-dual FOM proposed in [29] achieves an $O(\varepsilon^{-1})$ complexity result to produce an ε-optimal solution, and the same-order complexity result has also been established in [27]. Based on a previous work [10] for affinely constrained problems, [16] gives a modified first-order iALM for solving convex cone programs. The overall complexity of the modified method is $O(\varepsilon^{-1}|\log\varepsilon|)$ to produce an ε-KKT point (see Definition 1 below). A similar result has also been shown in [3] for convex conic programs. A proximal iALM is analyzed in [12]. By a linearly-convergent first-order subroutine for primal subproblems, [12] shows that $O(\varepsilon^{-1})$ calls to the subroutine are needed for convex problems and $O(\varepsilon^{-\frac{1}{2}})$ for strongly convex problems, to achieve either an ε-optimal or an ε-KKT point. In terms of function value and derivative evaluations, the complexity result is $O(\varepsilon^{-1}|\log\varepsilon|)$ for the convex case and $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ for the strongly-convex case.

On solving general nonlinear constrained problems, FOMs have also been proposed under the framework of the level-set method [1,14,15]. For convex problems, the level-set based FOMs can also achieve an $O(\varepsilon^{-1})$ complexity result to produce an ε-optimal solution. However, to obtain $\tilde{O}(\varepsilon^{-\frac{1}{2}})$, they require strong convexity of both the objective and the constraint functions.

Under the condition of strong duality, (1) can be equivalently formulated as a non-bilinear saddle-point (SP) problem. In this case, one can apply any FOM that is designed for solving non-bilinear SP problems. The work [8] generalizes the primal-dual method proposed in [5] from the bilinear SP case to the non-bilinear case. If the underlying SP problem is convex-concave, [8] establishes an $O(\varepsilon^{-1})$ complexity result to guarantee ε-duality gap. When the problem is strongly-convex-linear, the result can be improved to $O(\varepsilon^{-\frac{1}{2}})$. Notice that both results apply to the equivalent ordinary-Lagrangian-based SP problem of (1). By the smoothing technique, [9] gives an FOM (with both deterministic and stochastic versions) for solving non-bilinear SP problems. To ensure an ε-duality gap of a strongly-convex-concave problem, the method requires $\tilde{O}(\varepsilon^{-\frac{1}{2}})$ primal first-order oracles and $\tilde{O}(\varepsilon^{-1})$ dual first-order oracles. While applied to the functional constrained problem (1), the method in [9] can obtain an ε-optimal solution by $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$. FOMs for solving the more general variational inequality (VI) problem can also be applied to (1), such as the mirror-prox method in [21], the hybrid extragradient method in [17], and the accelerated method in [6]. All of the three methods can have an $O(\varepsilon^{-1})$ complexity result by assuming smoothness and/or monotonicity of the involved operator.

1.3 Contributions

On solving a functional constrained problem, none of the existing works about FOMs (such as those we mentioned previously) could obtain a complexity result better than $\tilde{O}(\varepsilon^{-\frac{1}{2}})$. Without specifying the regime of m, the task is impossible. We show that when $m = O(1)$ in (1), an FOM can achieve almost the same-order complexity result (with a difference at most a polynomial of $|\log\varepsilon|$) as for solving an unconstrained problem.

When $m = o(\varepsilon^{-\frac{1}{2}})$, we show that a complexity result better than $\tilde{O}(\varepsilon^{-\frac{1}{2}})$ can be obtained. The key step in the design of our algorithm is to formulate each primal subproblem into an equivalent SP problem. The SP formulation is strongly concave about the dual variable, and the strong concavity enables the generation of a cutting plane while searching for an approximate dual solution of the SP problem. Since there are m dual variables, we can apply a cutting-plane method to efficiently find an approximate dual solution when $m = O(1)$ or $m = o(\varepsilon^{-\frac{1}{2}})$.

1.4 Assumptions and notation

Throughout our analysis, we make the following assumptions.

Assumption 1 (smoothness) *f is L_f-smooth, i.e., ∇f is L_f-Lipschitz continuous. In addition, each g_i is smooth, and the Jacobian matrix $J_{\mathbf{g}} = [\nabla g_1^\top; \ldots; \nabla g_m^\top]$ is L_g-Lipschitz continuous.*

Assumption 2 (bounded domain and strong convexity) *The domain of h is bounded with diameter $D_h = \max_{\mathbf{x},\mathbf{y}\in\mathrm{dom}(h)} \|\mathbf{x}-\mathbf{y}\| < \infty$, and h is μ-strongly convex on $\mathrm{dom}(h)$.*

The above two assumptions imply the boundedness of $\mathbf{g}$ and $J_{\mathbf{g}}$ on $\mathrm{dom}(h)$. We use G and B_g respectively for their bounds, namely,

$$G = \max_{\mathbf{x}\in\mathrm{dom}(h)} \|\mathbf{g}(\mathbf{x})\|, \quad B_g = \max_{\mathbf{x}\in\mathrm{dom}(h)} \|J_{\mathbf{g}}(\mathbf{x})\|.$$

Assumption 3 (strong duality) *There is a primal-dual solution $(\mathbf{x}^*, \mathbf{z}^*)$ satisfying the KKT conditions of (1), i.e.,*

$$\mathbf{0} \in \partial F(\mathbf{x}^*) + J_{\mathbf{g}}(\mathbf{x}^*)^\top \mathbf{z}^*,\ \mathbf{z}^* \geq \mathbf{0}, \quad g(\mathbf{x}^*) \leq \mathbf{0}, \quad \mathbf{g}(\mathbf{x}^*)^\top \mathbf{z}^* = 0.$$

Notation. For a real number a, we use $\lceil a \rceil$ to denote the smallest integer that is no less than a and $\lceil a \rceil_+$ the smallest nonnegative integer that is no less than a. $\mathcal{B}_\delta(\mathbf{x})$ denotes a ball with radius δ and center $\mathbf{x}$. If $\mathbf{x} = \mathbf{0}$, we simply use $\mathcal{B}_\delta$. We define $\mathcal{B}_\delta^+$ as the intersection of $\mathcal{B}_\delta$ with the nonnegative orthant, so in the n-dimensional space, $\mathcal{B}_\delta^+ = \mathcal{B}_\delta \cap \mathbb{R}^n_+$. We use $V_m(\delta)$ for the volume of $\mathcal{B}_\delta$ in the m-dimensional space. $[n]$ denotes the set $\{1, \ldots, n\}$. Given a closed convex set $X \subset \mathbb{R}^n$ and a point $\mathbf{x} \in \mathbb{R}^n$, we define $\mathrm{dist}(\mathbf{x}, X) = \min_{\mathbf{y}\in X} \|\mathbf{y}-\mathbf{x}\|$. We use O, Θ, and o with standard meanings, while in the complexity result statement, $\tilde{O}$ has a similar meaning as O but suppresses a polynomial of $|\log\varepsilon|$ for a given error tolerance $\varepsilon > 0$.

Definition 1 (ε-KKT point) Given $\varepsilon > 0$, a point $\bar{\mathbf{x}} \in \mathrm{dom}(h)$ is called an ε-KKT point of (1) if there is $\bar{\mathbf{z}} \geq \mathbf{0}$ such that

$$\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_0(\bar{\mathbf{x}}, \bar{\mathbf{z}})\big) \leq \varepsilon, \quad \|[\mathbf{g}(\bar{\mathbf{x}})]_+\| \leq \varepsilon, \quad \sum_{i=1}^m |\bar{z}_i g_i(\bar{\mathbf{x}})| \leq \varepsilon, \tag{3}$$

where $\mathcal{L}_0(\mathbf{x}, \mathbf{z}) = F(\mathbf{x}) + \mathbf{z}^\top \mathbf{g}(\mathbf{x})$ is the ordinary Lagrangian function of (1).

By the convexity of F and each g_i, and also Assumption 3, one can easily show that an ε-KKT point of (1) must be an $O(\varepsilon)$-optimal solution, where we call a point $\bar{\mathbf{x}} \in \mathrm{dom}(h)$ as an ε-optimal solution of (1) if

$$\big|F(\bar{\mathbf{x}}) - F(\mathbf{x}^*)\big| \leq \varepsilon, \quad \|[\mathbf{g}(\bar{\mathbf{x}})]_+\| \leq \varepsilon. \tag{4}$$

1.5 Outline

The rest of the paper is organized as follows. In section 2, we review Nesterov's optimal FOM and give the convergence rate of the iALM. In section 3, we design new FOMs (that are better than directly applying Nesterov's method) for solving primal subproblems in the iALM. Overall complexity results are shown in section 4, and more discussions are given in section 5.

2 Nesterov's optimal FOM and convergence rate of iALM

In this section, we give Nesterov's optimal FOM that will be used as a subroutine in our algorithm. Also, we establish the convergence rate of the iALM to produce an approximate KKT point.

2.1 Nesterov's optimal FOM for strongly-convex composite problems

Consider the problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, P(\mathbf{x}) := \psi(\mathbf{x}) + r(\mathbf{x}), \tag{5}$$

where ψ is a differentiable convex function with L_ψ-Lipschitz continuous gradient, and r is a closed and μ_r-strongly convex function. Nesterov [22] gives an optimal FOM on solving (5). We rewrite, in a different way but equivalently, the method in Algorithm 2.

Algorithm 2: Nesterov's optimal first-order method for (5): $\widehat{\mathbf{x}} = \mathrm{APG}(\psi, r, \mu_r, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$

Input: minimum Lipschitz $L_{\min} > 0$, increase rate $\gamma_1 > 1$, decrease rate $\gamma_2 \geq 1$, and error tolerance $\bar{\varepsilon} > 0$.
Initialization: $L_0 \geq L_{\min}$, $\widehat{\mathbf{x}}^0 = \mathbf{x}^0 = \mathbf{w}^0 = \mathbf{v}^0 \in \mathrm{dom}(r)$, and $A_0 = 0$.
for $k = 0, 1, \ldots$ **do**
  let $a > 0$ be a solution of the equation $\frac{a^2}{A_k + a} = \frac{2(1+\mu_r A_k)}{L_k}$.
  set $A_{k+1} = A_k + a$, $\widetilde{\mathbf{y}} = \frac{A_k \mathbf{x}^k + a\mathbf{v}^k}{A_{k+1}}$, and $\widetilde{L} = L_k$.
  let $\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{y}}), \mathbf{x}\rangle + \frac{\widetilde{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{y}}\|^2 + r(\mathbf{x})$ and $\widetilde{\mathbf{u}} = \nabla\psi(\widetilde{\mathbf{x}}) - \nabla\psi(\widetilde{\mathbf{y}}) + \widetilde{L}(\widetilde{\mathbf{y}} - \widetilde{\mathbf{x}})$.
  while $\widetilde{L}\langle \widetilde{\mathbf{u}}, \widetilde{\mathbf{y}} - \widetilde{\mathbf{x}}\rangle < \|\widetilde{\mathbf{u}}\|^2$ **do**
    increase $\widetilde{L} \leftarrow \gamma_1 \widetilde{L}$;
    let $\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{y}}), \mathbf{x}\rangle + \frac{\widetilde{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{y}}\|^2 + r(\mathbf{x})$ and $\widetilde{\mathbf{u}} = \nabla\psi(\widetilde{\mathbf{x}}) - \nabla\psi(\widetilde{\mathbf{y}}) + \widetilde{L}(\widetilde{\mathbf{y}} - \widetilde{\mathbf{x}})$.
  let $\widehat{L} = \widetilde{L}$ and $\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{x}}), \mathbf{x}\rangle + \frac{\widehat{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 + r(\mathbf{x})$ ▷ modified step to guarantee near-stationarity at $\widehat{\mathbf{x}}$
  while $\psi(\widehat{\mathbf{x}}) > \psi(\widetilde{\mathbf{x}}) + \langle \nabla\psi(\widetilde{\mathbf{x}}), \widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\rangle + \frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2$ **do**
    increase $\widehat{L} \leftarrow \gamma_1 \widehat{L}$;
    let $\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{x}}), \mathbf{x}\rangle + \frac{\widehat{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 + r(\mathbf{x})$
  set $\mathbf{x}^{k+1} = \widetilde{\mathbf{x}}$, $\widehat{\mathbf{x}}^{k+1} = \widehat{\mathbf{x}}$, $L_{k+1} = \max\{L_{\min}, \widetilde{L}/\gamma_2\}$, and $\mathbf{w}^{k+1} = \mathbf{w}^k - a\nabla\psi(\widetilde{\mathbf{x}})$;
  let $\mathbf{v}^{k+1} = \arg\min_{\mathbf{x}} \frac{1}{2A_{k+1}}\|\mathbf{x} - \mathbf{w}^{k+1}\|^2 + r(\mathbf{x})$.
  if $\mathrm{dist}\big(\mathbf{0}, \partial P(\widehat{\mathbf{x}})\big) \leq \bar{\varepsilon}$ **then**
    return $\widehat{\mathbf{x}}$ and stop.

The results in the next theorem are from Lemma 5 and Theorem 6 of [22].

Theorem 1 *If $\widetilde{L} \geq L_\psi$, the condition in Line 7 of Algorithm 2 will not hold, i.e., the algorithm will exit the first while-loop. The generated sequence $\{\mathbf{x}^k\}_{k\geq 0}$ satisfies*

$$P(\mathbf{x}^{k+1}) - P(\mathbf{x}^*) \leq \frac{\gamma_1 L_\psi}{4}\|\mathbf{x}^0 - \mathbf{x}^*\|^2 \left(1 + \sqrt{\frac{\mu_r}{2\gamma_1 L_\psi}}\right)^{-2k}, \forall\, k \geq 0, \tag{6}$$

where $\mathbf{x}^*$ is the optimal solution of (5).

By the above theorem, we can easily bound the distance of $\widehat{\mathbf{x}}^k$ to stationarity for each k.

Theorem 2 *The generated sequence $\{\widehat{\mathbf{x}}^k\}_{k\geq 0}$ satisfies*

$$\mathrm{dist}\big(\mathbf{0}, \partial P(\widehat{\mathbf{x}}^{k+1})\big) \leq \left(\sqrt{\gamma_1 L_\psi} + \tfrac{L_\psi}{\sqrt{L_{\min}}}\right) \frac{\sqrt{\gamma_1 L_\psi}}{\sqrt{2}} \|\mathbf{x}^0 - \mathbf{x}^*\| \left(1 + \sqrt{\frac{\mu_r}{2\gamma_1 L_\psi}}\right)^{-k}, \forall\, k \geq 0.$$

Proof. First notice that if $\widehat{L} \geq L_\psi$, it must hold $\psi(\widehat{\mathbf{x}}) \leq \psi(\widetilde{\mathbf{x}}) + \langle \nabla\psi(\widetilde{\mathbf{x}}), \widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\rangle + \frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2$, and when this inequality holds, we have (cf. [28, Lemma 2.1]) $P(\widetilde{\mathbf{x}}) - P(\widehat{\mathbf{x}}) \geq \frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2$. Since $P(\widetilde{\mathbf{x}}) - P(\widehat{\mathbf{x}}) \leq P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)$, we have $\frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2 \leq P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)$. By the optimality condition of $\widehat{\mathbf{x}}$, it holds $\mathbf{0} \in \nabla\psi(\widetilde{\mathbf{x}}) + \widehat{L}(\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}) + \partial r(\widehat{\mathbf{x}})$, and thus $\mathrm{dist}(\mathbf{0}, \partial P(\widehat{\mathbf{x}})) \leq \|\nabla\psi(\widehat{\mathbf{x}}) - \nabla\psi(\widetilde{\mathbf{x}})\| + \widehat{L}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\| \leq (L_\psi + \widehat{L})\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|$. Using the inequalities $\frac{\widehat{L}^2}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2 \leq \widehat{L}\big(P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)\big)$ and $\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2 \leq \frac{2}{L_{\min}}\big(P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)\big)$, we have

$$\mathrm{dist}(\mathbf{0}, \partial P(\widehat{\mathbf{x}})) \leq (L_\psi + \widehat{L})\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\| \leq \sqrt{2(P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*))}\left(\sqrt{\widehat{L}} + \tfrac{L_\psi}{\sqrt{L_{\min}}}\right).$$

Therefore, the desired result follows from (6), the fact $\widehat{L} \leq \gamma_1 L_\psi$, and the above inequality with $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^{k+1}$ and $\widetilde{\mathbf{x}} = \mathbf{x}^{k+1}$. □

Corollary 1 *Assume $0 < \mu_r \leq L_\psi$. Given $\bar{\varepsilon} > 0$, $\gamma_1 > 1$, $\gamma_2 \geq 1$ and $L_{\min} > 0$, Algorithm 2 needs at most T evaluations on the objective value of ψ and the gradient $\nabla\psi$ to produce $\widehat{\mathbf{x}}$ such that $\mathrm{dist}(\mathbf{0}, \partial P(\widehat{\mathbf{x}})) \leq \bar{\varepsilon}$, where*

$$T = 2\left(1 + \lceil \log_{\gamma_1} \tfrac{L_\psi}{L_{\min}} \rceil_+\right) \cdot \left[(1+\sqrt{2})\sqrt{\frac{\gamma_1 L_\psi}{\mu_r}} \log \frac{\left(\sqrt{\gamma_1 L_\psi} + \frac{L_\psi}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_\psi}}{\sqrt{2}}\|\mathbf{x}^0 - \mathbf{x}^*\|}{\bar{\varepsilon}}\right]_+.$$

Proof. From Theorem 2, it follows that after at most K iterations, the algorithm will produce a point $\widehat{\mathbf{x}}$ satisfying $\mathrm{dist}(\mathbf{0}, \partial P(\widehat{\mathbf{x}})) \leq \bar{\varepsilon}$, where

$$K = \left[\frac{\log\left[\left(\sqrt{\gamma_1 L_\psi} + \frac{L_\psi}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_\psi}}{\sqrt{2}\bar{\varepsilon}}\|\mathbf{x}^0 - \mathbf{x}^*\|\right]}{\log(1 + \sqrt{\frac{\mu_r}{2\gamma_1 L_\psi}})}\right]_+.$$

Since the conditions in Line 7 and Line 11 of Algorithm 2 will be violated if $\widetilde{L} \geq L_\psi$ and $\widehat{L} \geq L_\psi$, every iteration will evaluate the objective value of ψ and the gradient $\nabla\psi$ at most $2(1 + \lceil \log_{\gamma_1} \frac{L_\psi}{L_{\min}} \rceil_+)$ times. Now using the fact $\log(1+a) \geq \frac{\sqrt{2}a}{1+\sqrt{2}}$, $\forall\, 0 < a \leq \frac{\sqrt{2}}{2}$, we obtain the desired result. □

2.2 Convergence rate of iALM

The next lemma is from Eq. (3.20) and the proof of Lemma 7 of [26].

Lemma 1 *Let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be generated from Algorithm 1 with $\mathbf{z}^0 = \mathbf{0}$. Suppose*

$$\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) \le \min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k) + e_k,\ \forall\, k = 0, 1, \ldots, \tag{7}$$

for an error sequence $\{e_k\}$. Then

$$\|\mathbf{z}^k\|^2 \le 4\|\mathbf{z}^*\|^2 + 4\sum_{t=0}^{k-1} \beta_t e_t,\ \ \textit{and}\ \ \|\mathbf{z}^k\| \le 2\|\mathbf{z}^*\| + \sqrt{2\sum_{t=0}^{k-1} \beta_t e_t},\ \forall\, k \ge 1. \tag{8}$$

By this lemma and also the strong convexity of F, we can show the following result.

Lemma 2 *Let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be generated from Algorithm 1 with $\mathbf{z}^0 = \mathbf{0}$. If $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \le \varepsilon_k$, $\forall\, k \ge 0$ for a sequence $\{\varepsilon_k\}$, then*

$$\|\mathbf{z}^k\|^2 \le 4\|\mathbf{z}^*\|^2 + 4\sum_{t=0}^{k-1} \beta_t \frac{\varepsilon_t^2}{\mu},\ \ \textit{and}\ \ \|\mathbf{z}^k\| \le 2\|\mathbf{z}^*\| + \sqrt{2\sum_{t=0}^{k-1} \beta_t \frac{\varepsilon_t^2}{\mu}},\ \forall\, k \ge 1. \tag{9}$$

Proof. Since F is μ-strongly convex, $\mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$ is also μ-strongly convex about $\mathbf{x}$, and thus we have that (7) holds with $e_t = \frac{\varepsilon_t^2}{\mu}$ from $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \le \varepsilon_k$. Therefore, (9) follows from (8). □

Theorem 3 (convergence rate of iALM) *Let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be generated from Algorithm 1 with $\mathbf{z}^0 = \mathbf{0}$. Suppose $\beta_k = \beta_0 \sigma^k$, $\forall\, k \ge 0$ for some $\sigma > 1$ and $\beta_0 > 0$, and $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \le \bar{\varepsilon}$, $\forall\, k \ge 0$ for a positive number $\bar{\varepsilon}$. Then*

$$\big\|[\mathbf{g}(\mathbf{x}^{k+1})]_+\big\| \le \frac{4\|\mathbf{z}^*\|}{\beta_0 \sigma^k} + \frac{\bar{\varepsilon}(\sqrt{\sigma}+1)\sqrt{\frac{2}{\mu(\sigma-1)}}}{\sqrt{\beta_0 \sigma^k}}, \tag{10}$$

$$\sum_{i=1}^{m} \big|z_i^{k+1} g_i(\mathbf{x}^{k+1})\big| \le \frac{9\|\mathbf{z}^*\|^2}{2\beta_0 \sigma^k} + \frac{\bar{\varepsilon}^2(8\sigma+1)}{2\mu(\sigma-1)}. \tag{11}$$

Proof. From the update of $\mathbf{z}$, it follows that $g_i(\mathbf{x}^{k+1}) \le \frac{z_i^{k+1} - z_i^k}{\beta_k}$ for each $i \in [m]$, and thus by (9), we have

$$\big\|[\mathbf{g}(\mathbf{x}^{k+1})]_+\big\| \le \frac{\|\mathbf{z}^{k+1} - \mathbf{z}^k\|}{\beta_k} \le \frac{\|\mathbf{z}^{k+1}\| + \|\mathbf{z}^k\|}{\beta_k} \le \frac{4\|\mathbf{z}^*\| + \sqrt{2\sum_{t=0}^{k-1} \beta_t \frac{\varepsilon_t^2}{\mu}} + \sqrt{2\sum_{t=0}^{k} \beta_t \frac{\varepsilon_t^2}{\mu}}}{\beta_k}.$$

Plugging into the above inequality $\varepsilon_t = \bar{\varepsilon}$, $\forall\, t \ge 0$ and $\beta_k = \sigma^k \beta_0$, we obtain the inequality in (10).

Furthermore, for each $i \in [m]$, we have

$$\big|z_i^{k+1} g_i(\mathbf{x}^{k+1})\big| = \frac{1}{\beta_k}\big|z_i^{k+1}(z_i^{k+1} - z_i^k)\big| \le \frac{1}{\beta_k}\left((z_i^{k+1})^2 + \frac{(z_i^k)^2}{8}\right),$$

and thus $\sum_{i=1}^m \left|z_i^{k+1} g_i(\mathbf{x}^{k+1})\right| \le \frac{1}{\beta_k}\left(\|\mathbf{z}^{k+1}\|^2 + \frac{\|\mathbf{z}^k\|^2}{8}\right)$. Now we obtain the result in (11) by plugging the first inequality in (9). □

We make a few remarks here. Given $\varepsilon > 0$, choose $\bar{\varepsilon} > 0$ such that $\frac{\bar{\varepsilon}^2(8\sigma+1)}{2\mu(\sigma-1)} < \varepsilon$ in Theorem 3. Notice that $\partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) = \partial_{\mathbf{x}}\mathcal{L}_0(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$. Hence, from (10) and (11), it follows that to ensure $\mathbf{x}^{k+1}$ to be an ε-KKT point, we need $\beta_0\sigma^k = \Theta(\frac{1}{\varepsilon})$ and solve $k = \Theta\big(\log_\sigma \frac{1}{\beta_0\varepsilon}\big)$ $\mathbf{x}$-subproblems. Since the smooth part of $\mathcal{L}_{\beta_k}(\cdot, \mathbf{z}^k)$ has $\Theta(\beta_k)$-Lipschitz continuous gradient, it needs $O(\sqrt{\frac{\beta_k}{\mu}})$ proximal gradient steps if we directly apply Algorithm 2. In this way, we can guarantee an ε-KKT point with total complexity $O(\sqrt{\frac{\kappa}{\varepsilon}}|\log\varepsilon|)$, where κ denotes the condition number in some sense. This complexity result has been established in a few existing works, e.g., [12,16]. It is worse by an order of $\sqrt{\frac{1}{\varepsilon}}$ than the complexity result in Corollary 1 for the unconstrained case. Generally, we cannot improve it any more because the result matches with the lower bound given in [23].

In the rest of the paper, we show that in some cases, a better complexity can be obtained. When $m = O(1)$, we show that we can achieve a complexity result $O(\sqrt{\kappa}|\log\varepsilon|)$, which is the same as the optimal result for the unconstrained case. For a general m, we can achieve $O(m\sqrt{\kappa}|\log\varepsilon|)$, which is better than $O(\sqrt{\frac{\kappa}{\varepsilon}}|\log\varepsilon|)$ in the regime of $m = o(\sqrt{\frac{1}{\varepsilon}})$.

3 Better first-order methods for x-subproblems

When m is small in (1), we do not directly apply Algorithm 2 to solve the $\mathbf{x}$-subproblem $\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$ in Algorithm 1. Instead, we design new and better FOMs that use Algorithm 2 as a subroutine under the framework of a cutting-plane method.

Given $\mathbf{z} \ge \mathbf{0}$, let $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}$. Then the problem $\min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z})$ can be written in the form of

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, \phi(\mathbf{x}) := F(\mathbf{x}) + \frac{\beta}{2}\|[\boldsymbol{\theta}(\mathbf{x})]_+\|^2. \tag{12}$$

Notice that $\frac{1}{2}\|[\boldsymbol{\theta}(\mathbf{x})]_+\|^2 = \max_{\mathbf{y}\ge\mathbf{0}} \left\{\mathbf{y}^\top\boldsymbol{\theta}(\mathbf{x}) - \frac{1}{2}\|\mathbf{y}\|^2\right\}$ and $\mathbf{y} = [\boldsymbol{\theta}(\mathbf{x})]_+$ reaches the maximum. We re-write (12) into

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\ge\mathbf{0}} \Phi(\mathbf{x}, \mathbf{y}) := F(\mathbf{x}) + \beta\left(\mathbf{y}^\top\boldsymbol{\theta}(\mathbf{x}) - \frac{1}{2}\|\mathbf{y}\|^2\right). \tag{13}$$

Define

$$d(\mathbf{y}) = \min_{\mathbf{x}\in\mathbb{R}^n} \Phi(\mathbf{x}, \mathbf{y}), \quad \text{and} \quad \bar{\mathbf{y}} = \arg\max_{\mathbf{y}\ge\mathbf{0}} d(\mathbf{y}). \tag{14}$$

Also, for a given $\mathbf{y} \ge \mathbf{0}$, define $\mathbf{x}(\mathbf{y})$ as the unique minimizer of $\Phi(\cdot, \mathbf{y})$, i.e.,

$$\mathbf{x}(\mathbf{y}) = \arg\min_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y}).$$

The idea of our algorithm design is to first find an approximate solution $\widehat{\mathbf{y}}$ of $\max_{\mathbf{y}\ge\mathbf{0}} d(\mathbf{y})$ and then to find an approximate solution $\widehat{\mathbf{x}}$ of $\min_{\mathbf{x}} \Phi(\mathbf{x}, \widehat{\mathbf{y}})$. By controlling the approximation errors, we can guarantee $\widehat{\mathbf{x}}$ to be a near-stationary point of ϕ. On finding $\widehat{\mathbf{y}}$, we use a cutting-plane method. Since d is strongly concave, a cutting plane can be generated at a query point $\mathbf{y} \ge \mathbf{0}$, though we can only have an estimate of $\nabla d(\mathbf{y})$. Notice that the same idea may not work if we directly play with the augmented (or ordinary) Lagrangian dual function because it is not strongly concave.

3.1 Preparatory lemmas

We first establish a few lemmas.

Lemma 3 *Suppose* $\bar{\mathbf{x}}$ *is the minimizer of* ϕ *in* (12). *Then* $\bar{\mathbf{y}} = [\boldsymbol{\theta}(\bar{\mathbf{x}})]_+$ *is the solution of* $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$, *and* $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ *is the saddle point of* Φ. *In addition, let* $(\mathbf{x}^*, \mathbf{z}^*)$ *be the point in Assumption 3. Then*

$$\|\bar{\mathbf{y}}\| = \|[\boldsymbol{\theta}(\bar{\mathbf{x}})]_+\| \leq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta}. \tag{15}$$

Proof. It is easy to see that $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a saddle point of Φ. We only need to show (15). Since $\bar{\mathbf{x}}$ is the minimizer of ϕ, it holds

$$F(\bar{\mathbf{x}}) + \frac{\beta}{2}\|[\boldsymbol{\theta}(\bar{\mathbf{x}})]_+\|^2 \leq F(\mathbf{x}^*) + \frac{\beta}{2}\|[\boldsymbol{\theta}(\mathbf{x}^*)]_+\|^2 = F(\mathbf{x}^*) + \frac{\beta}{2}\left\|\left[\mathbf{g}(\mathbf{x}^*) + \tfrac{\mathbf{z}}{\beta}\right]_+\right\|^2 \leq F(\mathbf{x}^*) + \frac{\|\mathbf{z}\|^2}{2\beta},$$

where the last inequality holds because $\mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$. By the above inequality and the fact $F(\bar{\mathbf{x}}) - F(\mathbf{x}^*) + \langle \mathbf{z}^*, \mathbf{g}(\bar{\mathbf{x}})\rangle \geq 0$, we have

$$\frac{\beta}{2}\|[\boldsymbol{\theta}(\bar{\mathbf{x}})]_+\|^2 \leq \frac{\|\mathbf{z}\|^2}{2\beta} + \langle \mathbf{z}^*, \mathbf{g}(\bar{\mathbf{x}})\rangle \leq \frac{\|\mathbf{z}\|^2}{2\beta} + \langle \mathbf{z}^*, \boldsymbol{\theta}(\bar{\mathbf{x}})\rangle \leq \frac{\|\mathbf{z}\|^2}{2\beta} + \|\mathbf{z}^*\| \cdot \|[\boldsymbol{\theta}(\bar{\mathbf{x}})]_+\|,$$

which implies the inequality in (15). □

Lemma 4 *For any* $\mathbf{y} \geq \mathbf{0}$, *it holds*

$$\nabla d(\mathbf{y}) = \beta\big(\boldsymbol{\theta}(\mathbf{x}(\mathbf{y})) - \mathbf{y}\big).$$

In addition,

$$\beta\big\langle \mathbf{y}_1 - \mathbf{y}_2, \boldsymbol{\theta}(\mathbf{x}(\mathbf{y}_1)) - \boldsymbol{\theta}(\mathbf{x}(\mathbf{y}_2))\big\rangle \leq -\mu\|\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)\|^2, \ \forall\, \mathbf{y}_1, \mathbf{y}_2 \geq \mathbf{0}, \tag{16}$$

and

$$\|\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)\| \leq \frac{\beta B_g}{\mu}\|\mathbf{y}_1 - \mathbf{y}_2\|, \ \forall\, \mathbf{y}_1, \mathbf{y}_2 \geq \mathbf{0}. \tag{17}$$

Proof. For $i = 1, 2$, denote $\mathbf{x}_i = \mathbf{x}(\mathbf{y}_i)$. From the definition of $\mathbf{x}(\mathbf{y})$ and the μ-strong convexity of F, it holds

$$\begin{aligned} F(\mathbf{x}_1) + \beta\mathbf{y}_1^\top\boldsymbol{\theta}(\mathbf{x}_1) &\leq F(\mathbf{x}_2) + \beta\mathbf{y}_1^\top\boldsymbol{\theta}(\mathbf{x}_2) - \frac{\mu}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2, \\ F(\mathbf{x}_2) + \beta\mathbf{y}_2^\top\boldsymbol{\theta}(\mathbf{x}_2) &\leq F(\mathbf{x}_1) + \beta\mathbf{y}_2^\top\boldsymbol{\theta}(\mathbf{x}_1) - \frac{\mu}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2. \end{aligned}$$

Adding the above two inequalities gives the result in (16). Now using the B_g-Lipschitz continuity of $\boldsymbol{\theta}$, we have (17) from (16) and complete the proof. □

Lemma 5 (approximate dual gradient) *Given* $\widehat{\mathbf{y}} \geq \mathbf{0}$ *and* $\delta \geq 0$, *let* $\widehat{\mathbf{x}}$ *be an approximate minimizer of* $\Phi(\cdot, \widehat{\mathbf{y}})$ *such that* $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \delta$. *Then*

$$\|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\| \leq B_g\frac{\delta}{\mu}, \ \left\|\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big) - \nabla d(\widehat{\mathbf{y}})\right\| \leq \beta B_g\frac{\delta}{\mu}.$$

Proof. From the μ-strong convexity of F, it follows that for each $\mathbf{y} \geq \mathbf{0}$, $\Phi(\cdot, \mathbf{y})$ is μ-strongly convex, and thus $\mu\|\widehat{\mathbf{x}} - \mathbf{x}(\widehat{\mathbf{y}})\| \leq \operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \delta$, which gives $\|\widehat{\mathbf{x}} - \mathbf{x}(\widehat{\mathbf{y}})\| \leq \frac{\delta}{\mu}$. Hence, by the B_g-Lipschitz continuity of $\boldsymbol{\theta}$, we have $\|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\| \leq B_g\frac{\delta}{\mu}$, and thus

$$\big\|\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big) - \nabla d(\widehat{\mathbf{y}})\big\| = \beta\|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\| \leq \beta B_g\frac{\delta}{\mu}.$$

This completes the proof. □

Lemma 6 *Given $\widehat{\mathbf{y}} \geq \mathbf{0}$, it holds*

$$\operatorname{dist}\big(\mathbf{0}, \partial\phi(\widehat{\mathbf{x}})\big) \leq \operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) + \beta\|J_{\boldsymbol{\theta}}(\widehat{\mathbf{x}})\| \cdot \|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\|,\ \forall\, \widehat{\mathbf{x}} \in \operatorname{dom}(h).$$

Proof. It is easy to have $\partial\phi(\widehat{\mathbf{x}}) = \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) + \beta J_{\boldsymbol{\theta}}^{\top}(\widehat{\mathbf{x}})([\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}})$. The desired result now follows from the triangle inequality. □

Lemma 7 *Given $\bar{\varepsilon} > 0$, if $\widehat{\mathbf{y}} \geq \mathbf{0}$ is an approximate solution of $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ such that $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \frac{\bar{\varepsilon}}{3\beta B_g}$, and $\widehat{\mathbf{x}}$ is an approximate minimizer of $\Phi(\cdot, \widehat{\mathbf{y}})$ such that $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\bar{\varepsilon}}{3}\min\{1, \frac{\mu}{\beta B_g^2}\}$, then $\operatorname{dist}\big(\mathbf{0}, \partial\phi(\widehat{\mathbf{x}})\big) \leq \bar{\varepsilon}$.*

Proof. By Lemma 5 and the condition $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\bar{\varepsilon}\mu}{3\beta B_g^2}$, it holds $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - [\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+\| \leq \|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\| \leq \frac{\bar{\varepsilon}}{3\beta B_g}$, and thus by the triangle inequality, $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| \leq \frac{2\bar{\varepsilon}}{3\beta B_g}$. The desired result now follows from Lemma 6 and the assumption $\|J_{\mathbf{g}}(\mathbf{x})\| \leq B_g$, $\forall\, \mathbf{x} \in \operatorname{dom}(h)$. □

3.2 the case with a single constraint

For simplicity, we start with the case of $m = 1$. We show the complexity to produce a point $\widehat{\mathbf{x}}$ satisfying $\operatorname{dist}\big(\mathbf{0}, \partial\phi(\widehat{\mathbf{x}})\big) \leq \bar{\varepsilon}$ for a specified error tolerance $\bar{\varepsilon} > 0$. Although we still use the bold letters $\mathbf{y}, \boldsymbol{\theta}$, they are actually scalars in this subsection.

Assume that we know B_g. Notice that if $\bar{\mathbf{y}}$ is the solution of $\max_{\mathbf{y}\geq 0} d(\mathbf{y})$, then it satisfies the optimality condition $[\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+ = \bar{\mathbf{y}}$. We aim to find a $\widehat{\mathbf{y}} \geq \mathbf{0}$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$ for a given $\delta > 0$.

Lemma 8 *Given $\delta > 0$ and $\widehat{\mathbf{y}} \geq \mathbf{0}$, let $\widehat{\mathbf{x}} \in \operatorname{dom}(h)$ be a point satisfying $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\mu\delta}{4B_g}$. If $\big|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\big| \leq \frac{3\delta}{4}$, then $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$. Otherwise, $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| > \frac{\delta}{2}$, and $\nabla d(\widehat{\mathbf{y}})(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}) > 0$.*

Proof. From Lemma 5 and the condition on $\widehat{\mathbf{x}}$, it follows that

$$\big|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\big| \leq \frac{\delta}{4},\ \text{and}\ \big|\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big) - \nabla d(\widehat{\mathbf{y}})\big| \leq \frac{\beta\delta}{4}. \tag{18}$$

Thus by the nonexpansiveness of $[\cdot]_+$, it holds $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - [\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+| \leq \frac{\delta}{4}$. Furthermore, by the triangle inequality, we have $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$ if $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}| \leq \frac{3\delta}{4}$ and $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| > \frac{\delta}{2}$ otherwise. When $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}| > \frac{3\delta}{4}$, it holds $|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}| > \frac{3\delta}{4}$, and thus from the second inequality in (18), we conclude that $\nabla d(\widehat{\mathbf{y}})$ must have the same sign as $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}$. This completes the proof. □

By this lemma, we design an algorithm that can either return a point $\widehat{\mathbf{y}} \geq 0$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$ or return an interval $Y = [a, b] \subset [0, \infty)$ that contains the solution $\bar{\mathbf{y}}$. The pseudocode is shown in Algorithm 3.

Algorithm 3: Interval search: $Y = \mathrm{IntV}(\beta, \mathbf{z}, \delta, L_{\min}, \gamma_1, \gamma_2)$

1 **Input:** multiplier vector $\mathbf{z} \geq \mathbf{0}$, penalty $\beta > 0$, target accuracy $\delta > 0$, $L_{\min} > 0$, and $\gamma_1 > 1, \gamma_2 \geq 1$
2 **Overhead:** define $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}$, $\Phi(\mathbf{x}, \mathbf{y})$ as in (13), and $\bar{\varepsilon} = \frac{\mu\delta}{4B_g}$.
3 **Initial step:** call Alg. 2: $\widehat{\mathbf{x}} = \mathrm{APG}(\psi, h, \mu, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$ with $\psi = \Phi(\cdot, 0) - h$. ▷ so $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, 0)\big) \leq \frac{\mu\delta}{4B_g}$
4 **if** $[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ \leq \frac{3\delta}{4}$ **then**
5 Return $Y = \{0\}$ and stop. ▷ otherwise, $\nabla d(0)$ is positive
6 Let $a = 0$, $b = \frac{1}{\beta}$ and call Alg. 2: $\widehat{\mathbf{x}} = \mathrm{APG}(\psi, h, \mu, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$ with $\psi = \Phi(\cdot, b) - h$. ▷ set $b = O(\frac{1}{\beta})$
7 **while** $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - b\| > \frac{3\delta}{4}$ *and* $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b > 0$ **do**
8 let $a \leftarrow b$, and increase $b \leftarrow 2b$. ▷ fine to multiply b by a constant $\sigma > 1$
9 call Alg. 2: $\widehat{\mathbf{x}} = \mathrm{APG}(\psi, h, \mu, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$ with $\psi = \Phi(\cdot, b) - h$.
10 **if** $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - b\| \leq \frac{3\delta}{4}$ **then**
11 Return $Y = \{b\}$ and stop. ▷ found $\widehat{\mathbf{y}} = b$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$
12 **else**
13 Return $Y = [a, b]$ and stop. ▷ found an interval containing $\bar{\mathbf{y}}$

Lemma 9 *Given $\delta > 0$, let Y be the return from Algorithm 3. If Y contains a single point $\widehat{\mathbf{y}}$, then $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$. Otherwise, Y is an interval $[a, b]$, and it holds that $\nabla d(a) > 0, \nabla d(b) < 0$, and $\bar{\mathbf{y}} \in [a, b]$.*

Proof. If Y contains a single point $\widehat{\mathbf{y}}$, then we immediately have $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$ from Lemma 8. If Y is an interval $[a, b]$, then from Lemma 8 and the setting in Line 8 of Algorithm 3, we have $\nabla d(a) > 0$ and $\nabla d(b) < 0$. Therefore, the unique solution $\bar{\mathbf{y}}$ must lie in (a, b) by the Mean-Value Theorem and the strong concavity of d. □

The next lemma shows that Algorithm 3 must stop within a finitely many iterations.

Lemma 10 *Given $\delta > 0$, if $b \geq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta}$ and $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, b)\big) \leq \frac{\mu\delta}{4B_g}$, then either $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - b\| \leq \frac{3\delta}{4}$ or $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b < 0$.*

Proof. From Lemma 3, it follows that $\bar{\mathbf{y}} = [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+ \leq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta}$. The result in (16) indicates the decreasing monotonicity of $\boldsymbol{\theta}(\mathbf{x}(\mathbf{y}))$ with respect to $\mathbf{y}$. Hence, if $b \geq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta}$, it must hold $\boldsymbol{\theta}(\mathbf{x}(b)) - b \leq 0$. Now if $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - b| > \frac{3\delta}{4}$, we know from Lemma 8 that $\nabla d(b) = \beta(\boldsymbol{\theta}(\mathbf{x}(b)) - b)$ has the same sign as $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b$. Hence, $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b < 0$, and we complete the proof. □

Remark 1 Suppose Algorithm 3 returns an interval $[a, b]$. Then Lemma 10 indicates that $b \leq \frac{1}{\beta}\max\{1, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|\}$, and in addition, at most $T + 2$ calls are made to Alg. 2, where T is the smallest non-negative integer such that $2^T \geq 2\|\mathbf{z}^*\| + \|\mathbf{z}\|$.

Suppose Algorithm 3 returns an interval $[a, b]$. We can then use the bisection method to obtain a desired point $\widehat{\mathbf{y}}$. The pseudocode is given in Algorithm 4.

By the lemma below, it holds that the return $\widehat{\mathbf{y}}$ from Algorithm 4 must satisfy $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$.

Lemma 11 *Let $Y = [a, b] \subset (0, \infty)$. If $\nabla d(a) > 0$, $\nabla d(b) < 0$, and $b - a \leq \frac{\mu\delta}{\mu + \beta B_g^2}$ for a positive δ, then $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$ for any $\widehat{\mathbf{y}} \in [a, b]$.*

Algorithm 4: Bisection method for $\max_{\mathbf{y}\geq 0} d(\mathbf{y})$: $(\widehat{\mathbf{x}},\widehat{\mathbf{y}}) = \mathrm{BiSec}(\beta,\mathbf{z},\delta,L_{\min},\gamma_1,\gamma_2)$

Input: multiplier vector $\mathbf{z}\geq\mathbf{0}$, penalty $\beta>0$, target accuracy $\delta>0$, $L_{\min}>0$, and $\gamma_1>1,\gamma_2\geq 1$
Overhead: define $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}$, $\Phi(\mathbf{x},\mathbf{y})$ as in (13), and $\bar{\varepsilon} = \frac{\mu\delta}{4B_g}$.
Call Alg. 3: $Y = \mathrm{IntV}(\beta,\mathbf{z},\delta,L_{\min},\gamma_1,\gamma_2)$ and denote it as $[a,b]$. ▷ If Y is a singleton, then $a=b$
while $b-a > \frac{\mu\delta}{\mu+\beta B_g^2}$ **do**
  let $c=\frac{a+b}{2}$ and call Alg. 2: $\widehat{\mathbf{x}} = \mathrm{APG}(\psi,h,\mu,L_{\min},\bar{\varepsilon},\gamma_1,\gamma_2)$ with $\psi = \Phi(\cdot,c)-h$
  if $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - c| \leq \frac{3\delta}{4}$ **then**
    Let $\widehat{\mathbf{y}}=c$, return $(\widehat{\mathbf{x}},\widehat{\mathbf{y}})$, and stop
  else if $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - c > 0$ **then**
    let $a\leftarrow c$
  else
    let $b\leftarrow c$.
Let $\widehat{\mathbf{y}} = \frac{a+b}{2}$ and $\widehat{\mathbf{x}} = \mathrm{APG}(\psi,h,\mu,L_{\min},\bar{\varepsilon},\gamma_1,\gamma_2)$ with $\psi=\Phi(\cdot,\widehat{\mathbf{y}})-h$, return $(\widehat{\mathbf{x}},\widehat{\mathbf{y}})$, and stop.

Proof. For any $\widehat{\mathbf{y}}\in[a,b]$, we have

$$\begin{aligned}\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| &= \|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}} - [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+ + \bar{\mathbf{y}}\| \\ &\leq \|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+\| + \|\widehat{\mathbf{y}}-\bar{\mathbf{y}}\| \\ &\leq \tfrac{\beta B_g^2}{\mu}\|\widehat{\mathbf{y}}-\bar{\mathbf{y}}\| + \|\widehat{\mathbf{y}}-\bar{\mathbf{y}}\|, \end{aligned} \tag{19}$$

where in the second inequality, we have used the Lipschitz continuity of $\boldsymbol{\theta}$, the non-expansiveness of $[\cdot]_+$, and the inequality in (17). Since $\|\widehat{\mathbf{y}}-\bar{\mathbf{y}}\|\leq b-a\leq\frac{\mu\delta}{\mu+\beta B_g^2}$, the desired result follows. □

Remark 2 Since the bisection method halves the interval every time, it takes at most $\lceil\log_2\frac{(b-a)(\mu+\beta B_g^2)}{\mu\delta}\rceil_+$ halves to reduce an initial interval $[a,b]$ to one with length no larger than $\frac{\mu\delta}{\mu+\beta B_g^2}$. Notice $a\geq 0$ and $b\leq\frac{1}{\beta}\max\{1,4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\}$ from Remark 1. Hence, after Y is obtained, Algorithm 4 will call Algorithm 2 at most $\left\lceil\log_2\frac{\max\left\{1,\ 4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\right\}(\mu+\beta B_g^2)}{\beta\mu\delta}\right\rceil_+ + 1$ times.

Theorem 4 (Iteration complexity of BiSec) *Assume $\mu\leq L_f$. Algorithm 4 with input $(\beta,\mathbf{z},\delta,L_{\min},\gamma_1,\gamma_2)$ needs at most T evaluations on f, $\boldsymbol{\theta}$, ∇f, and $J_{\boldsymbol{\theta}}$ to output $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}\geq 0$ satisfying $\mathrm{dist}\big(\mathbf{0},\partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}},\widehat{\mathbf{y}})\big)\leq\bar{\varepsilon}$ and $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+-\widehat{\mathbf{y}}|\leq\delta$, where $\bar{\varepsilon}=\frac{\mu\delta}{4B_g}$, and*

$$T = 2K\left(1+\lceil\log_{\gamma_1}\tfrac{L_{\mathbf{z}}}{L_{\min}}\rceil_+\right)\cdot\left\lceil(1+\sqrt{2})\sqrt{\frac{\gamma_1 L_{\mathbf{z}}}{\mu}}\log\frac{\left(\sqrt{\gamma_1 L_{\mathbf{z}}}+\frac{L_{\mathbf{z}}}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_{\mathbf{z}}}}{\sqrt{2}}D_h}{\bar{\varepsilon}}\right\rceil,$$

with

$$K = 3+\lceil\log_2(2\|\mathbf{z}^*\|+\|\mathbf{z}\|)\rceil_+ + \left\lceil\log_2\frac{\max\left\{1,\ 4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\right\}(\mu+\beta B_g^2)}{\beta\mu\delta}\right\rceil_+, \tag{20}$$

and

$$L_{\mathbf{z}} = L_f + L_g\max\{1,4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\}. \tag{21}$$

Proof. By Remarks 1 and 2, Algorithm 4 calls Algorithm 2 at most K times, where K is given in (20). Notice that the gradient of $\psi = \Phi(\cdot, b) - h$ is Lipschitz continuous with constant $L_f + \beta b L_g$. Since $b \leq \frac{1}{\beta}\max\{1, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|\}$ from Remark 1, we apply Corollary 1 to obtain the desired result. □

3.3 the case with multiple constraints

In this subsection, we consider the case of $m > 1$. Similar to the case of $m = 1$, we use a cutting-plane method to approximately solve $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$. The next lemma is the key. It provides the foundation to generate a cutting plane if a query point is not sufficiently close to the solution $\bar{\mathbf{y}}$.

Lemma 12 *Let $b > 0$, and suppose $\|\bar{\mathbf{y}}\| \leq b$. Given $\delta > 0$ and $\widehat{\mathbf{y}} \geq \mathbf{0}$, let $\widehat{\mathbf{x}} \in \mathrm{dom}(h)$ be a point satisfying $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$. If $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| \leq \frac{3\delta}{4}$, then $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \delta$. Otherwise, $\big\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\big\| > \frac{\delta}{2}$, and also $\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}}\rangle \geq 0$ for any $\mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+$, where $\eta = \min\{b, \eta_+\}$ with η_+ being the positive root of the equation*

$$\frac{\mu + \beta B_g^2}{\mu}\left(\eta + \sqrt{\frac{2\eta B_d}{\beta}}\right) = \frac{\delta}{4}, \tag{22}$$

and with $B_d = \max_{\mathbf{y}\in\mathcal{B}_b^+} \nabla d(\mathbf{y})$.

Proof. By the same arguments in the proof of Lemma 8, we can show that $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \delta$ if $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| \leq \frac{3\delta}{4}$ and $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| > \frac{\delta}{2}$ otherwise. Hence, we only need to show $\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}}\rangle \geq 0$ for any $\mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+$ in the latter case, and we prove this by contradiction.

Suppose $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| > \frac{3\delta}{4}$ and the following condition holds

$$\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}}\rangle < 0, \text{ for some } \mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+. \tag{23}$$

By the β-strong concavity of d, it holds

$$d(\mathbf{y}) \leq d(\widehat{\mathbf{y}}) + \langle \nabla d(\widehat{\mathbf{y}}), \mathbf{y} - \widehat{\mathbf{y}}\rangle - \frac{\beta}{2}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2. \tag{24}$$

From the Mean-Value Theorem, it follows that there is $\widetilde{\mathbf{y}}$ between $\mathbf{y}$ and $\bar{\mathbf{y}}$ such that $d(\mathbf{y}) - d(\bar{\mathbf{y}}) = \langle \nabla d(\widetilde{\mathbf{y}}), \mathbf{y} - \bar{\mathbf{y}}\rangle \geq -\eta B_d$. Since $d(\bar{\mathbf{y}}) \geq d(\widehat{\mathbf{y}})$, we have $d(\widehat{\mathbf{y}}) - d(\mathbf{y}) \leq \eta B_d$. Hence, (23) and (24) imply

$$\frac{\beta}{2}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 \leq \eta B_d + \langle \beta(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}) - \nabla d(\widehat{\mathbf{y}}), \widehat{\mathbf{y}} - \mathbf{y}\rangle.$$

From Lemma 5 and the condition $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}$, it follows $\|\beta(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}) - \nabla d(\widehat{\mathbf{y}})\| \leq \frac{\beta\mu\delta}{8(\mu+\beta B_g^2)}$, which together with the above inequality and the Cauchy-Schwartz inequality gives

$$\frac{\beta}{2}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 \leq \eta B_d + \frac{\beta\mu\delta}{8(\mu + \beta B_g^2)}\|\widehat{\mathbf{y}} - \mathbf{y}\|.$$

Solving the above inequality, we have $\|\mathbf{y}-\widehat{\mathbf{y}}\| \le \sqrt{\frac{2\eta B_d}{\beta}} + \frac{\mu\delta}{4(\mu+\beta B_g^2)}$, and since $\|\mathbf{y}-\bar{\mathbf{y}}\| \le \eta$, it holds $\|\bar{\mathbf{y}}-\widehat{\mathbf{y}}\| \le \eta + \sqrt{\frac{2\eta B_d}{\beta}} + \frac{\mu\delta}{4(\mu+\beta B_g^2)}$. Now using (19), we have

$$\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \le \frac{\mu+\beta B_g^2}{\mu}\left(\eta + \sqrt{\frac{2\eta B_d}{\beta}} + \frac{\mu\delta}{4(\mu+\beta B_g^2)}\right) = \frac{\mu+\beta B_g^2}{\mu}\left(\eta + \sqrt{\frac{2\eta B_d}{\beta}}\right) + \frac{\delta}{4} \le \frac{\delta}{2}, \quad (25)$$

where the last inequality follows from the choice of η.

However, we know that when $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| > \frac{3\delta}{4}$, it holds $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| > \frac{\delta}{2}$, and the inequality in (25) contradicts to this fact. Therefore, the assumption in (23) cannot hold. This completes the proof. □

Suppose $\|\bar{\mathbf{y}}\| \le b$ for some $b > 0$. For a given $\widehat{\mathbf{y}} \ge \mathbf{0}$, let $\widehat{\mathbf{x}}$ satisfy the condition required in Lemma 12. Then if $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| > \frac{3\delta}{4}$, we find a half-space containing the set $\mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+$, whose volume is at least $4^{-m}V_m(\eta)$ if $\eta \le b$. Therefore, we can apply a cutting-plane method to find a near-optimal $\widehat{\mathbf{y}}$. For simplicity, we use the ellipsoid method. The pseudocode is shown in Algorithm 5.

Algorithm 5: Ellipsoid Method for $\max_{\mathbf{y}\ge\mathbf{0}} d(\mathbf{y})$: $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \text{FLAG}) = \text{Ellipsoid}(\beta, \mathbf{z}, \delta, b, L_{\min}, \gamma_1, \gamma_2)$

Input: multiplier vector $\mathbf{z} \ge \mathbf{0}$, penalty $\beta > 0$, target accuracy $\delta > 0$, $b > 0$, $L_{\min} > 0$, and $\gamma_1 > 1, \gamma_2 \ge 1$

Overhead: define $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}$, $\Phi(\mathbf{x},\mathbf{y})$ as in (13), $\bar{\varepsilon} = \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$, and FLAG $= 0$.

Let η_+ be the positive root of (22) and $\eta \leftarrow \min\{b, \eta_+\}$, and set $k = 0$.

Set $\mathcal{E}_0 = \{\mathbf{y} \in \mathbb{R}^m : (\mathbf{y}-\widehat{\mathbf{y}})^\top \mathbf{B}^{-1}(\mathbf{y}-\widehat{\mathbf{y}}) \le 1\}$ with $\mathbf{B} = b^2\mathbf{I}$ and $\widehat{\mathbf{y}} = \mathbf{0}$ ▷ initial ellipsoid

while *the volume of* $\mathcal{E}_k > 4^{-m}V_m(\eta)$ **do**

  if $\widehat{\mathbf{y}} \not\ge \mathbf{0}$ **then**

    Let $\mathbf{a} = -\mathbf{e}_{i_0}$ where $i_0 = \arg\min_{i\in[m]} \widehat{y}_i$ ▷ add a cutting plane $y_{i_0} \ge \widehat{y}_{i_0}$

    Set $\mathcal{E}_{k+1} = \{\mathbf{y} \in \mathbb{R}^m : (\mathbf{y}-\widehat{\mathbf{y}})^\top \mathbf{B}^{-1}(\mathbf{y}-\widehat{\mathbf{y}}) \le 1\}$ with updated $\mathbf{B}$ and $\widehat{\mathbf{y}}$

$$\mathbf{B} \leftarrow \frac{m^2}{m^2-1}\left(\mathbf{B} - \frac{2}{(m+1)\mathbf{a}^\top\mathbf{B}\mathbf{a}}\mathbf{B}\mathbf{a}(\mathbf{B}\mathbf{a})^\top\right), \quad \widehat{\mathbf{y}} \leftarrow \widehat{\mathbf{y}} - \frac{1}{m+1}\frac{\mathbf{B}\mathbf{a}}{\sqrt{\mathbf{a}^\top\mathbf{B}\mathbf{a}}} \quad (26)$$

  else if $\|\widehat{\mathbf{y}}\| > b$ **then**

    Let $\mathbf{a} = \widehat{\mathbf{y}}$ ▷ add a cutting plane $\langle \widehat{\mathbf{y}}, \mathbf{y}-\widehat{\mathbf{y}}\rangle \le 0$

    Set $\mathcal{E}_{k+1} = \{\mathbf{y} \in \mathbb{R}^m : (\mathbf{y}-\widehat{\mathbf{y}})^\top \mathbf{B}^{-1}(\mathbf{y}-\widehat{\mathbf{y}}) \le 1\}$ with $\mathbf{B}$ and $\widehat{\mathbf{y}}$ updated by (26)

  else

    Call Alg. 2: $\widehat{\mathbf{x}} = \text{APG}(\psi, h, \mu, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$ with $\psi = \Phi(\cdot, \widehat{\mathbf{y}}) - h$

    if $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| \le \frac{3\delta}{4}$ **then**

      FLAG $= 1$, return $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \text{FLAG})$, and stop ▷ found $\widehat{\mathbf{y}}$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \le \delta$

    else

      Let $\mathbf{a} = \widehat{\mathbf{y}} - \boldsymbol{\theta}(\widehat{\mathbf{x}})$ ▷ add a cutting plane $\langle \widehat{\mathbf{y}} - \boldsymbol{\theta}(\widehat{\mathbf{x}}), \mathbf{y}-\widehat{\mathbf{y}}\rangle \le 0$

      Set $\mathcal{E}_{k+1} = \{\mathbf{y} \in \mathbb{R}^m : (\mathbf{y}-\widehat{\mathbf{y}})^\top \mathbf{B}^{-1}(\mathbf{y}-\widehat{\mathbf{y}}) \le 1\}$ with $\mathbf{B}$ and $\widehat{\mathbf{y}}$ updated by (26)

  Increase $k \leftarrow k+1$.

From Lemma 12 and the property of the ellipsoid method (cf. [4]), we can easily show the finite convergence of Algorithm 5, and furthermore, we can estimate its total complexity by using Corollary 1.

Theorem 5 *Let $(\beta, \mathbf{z}, \delta, b, L_{\min}, \gamma_1, \gamma_2)$ be the input of Algorithm 5. Then Algorithm 5 will stop within at most $\left\lceil 2m(m+1)\log\frac{4b}{\eta}\right\rceil$ iterations, where η is defined in Line 3 of the algorithm. In addition, assume $0 < \mu \le L_\psi := L_f + \beta b L_g$. If $\|\bar{\mathbf{y}}\| \le b$, it must return* FLAG $= 1$ *and a vector $\widehat{\mathbf{y}} \ge \mathbf{0}$ satisfying $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \le \delta$ with at most T evaluations of f, ∇f, $\boldsymbol{\theta}$, and $J_{\boldsymbol{\theta}}$, where*

$$T = 2\left\lceil 2m(m+1)\log\tfrac{4b}{\eta}\right\rceil \cdot \left(1 + \lceil \log_{\gamma_1} \tfrac{L_\psi}{L_{\min}}\rceil_+\right) \cdot \left\lceil (1+\sqrt{2})\sqrt{\tfrac{\gamma_1 L_\psi}{\mu}}\log\frac{\left(\sqrt{\gamma_1 L_\psi} + \frac{L_\psi}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_\psi}}{\sqrt{2}} D_h}{\bar{\varepsilon}}\right\rceil, \quad (27)$$

with $\bar{\varepsilon} = \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8(\mu B_g + \beta B_g^3)}\}$.

Proof. By the property of the ellipsoid method, we have (cf. [4, Eq. 2.11])

$$\mathrm{vol}(\mathcal{E}_k) \le e^{-\frac{1}{2(m+1)}}\mathrm{vol}(\mathcal{E}_{k-1}) \le e^{-\frac{k}{2(m+1)}}\mathrm{vol}(\mathcal{E}_0),\ \forall\, k \ge 1.$$

Hence, to satisfy the stopping condition $\mathrm{vol}(\mathcal{E}_k) \le 4^{-m}V_m(\eta)$, it suffices to have $e^{-\frac{k}{2(m+1)}}\mathrm{vol}(\mathcal{E}_0) \le 4^{-m}V_m(\eta)$. Since $\mathcal{E}_0$ is a ball of radius b, this requirement is equivalent to $e^{-\frac{k}{2(m+1)}} \le \left(\frac{\eta}{4b}\right)^m$, which holds if $k \ge \left\lceil 2m(m+1)\log\frac{4b}{\eta}\right\rceil$. We below estimate the number of evaluations of the function value and gradient.

Notice that when Algorithm 2 is called, $\|\widehat{\mathbf{y}}\| \le b$, and thus the smooth function ψ has $(L_f + \beta L_g b)$-Lipschitz continuous gradient. Since Algorithm 2 is called at most $\left\lceil 2m(m+1)\log\frac{4b}{\eta}\right\rceil$ times, we have from Corollary 1 that the total number of function and gradient evaluations is T given in (27). □

By Theorem 5, we can guarantee to find a desired approximate solution $\widehat{\mathbf{y}}$ by gradually increasing the search radius b. The algorithm is shown below.

Algorithm 6: **S**earch by **t**he **E**llipsoid **M**ethod for $\max_{\mathbf{y}\ge\mathbf{0}} d(\mathbf{y})$: $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = \mathrm{StEM}(\beta, \mathbf{z}, \delta, L_{\min}, \gamma_1, \gamma_2)$

1 **Input:** multiplier vector $\mathbf{z} \ge \mathbf{0}$, penalty $\beta > 0$, target accuracy $\delta > 0$, $L_{\min} > 0$, and $\gamma_1 > 1, \gamma_2 \ge 1$
2 **Overhead:** define $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}}{\beta}$, $\Phi(\mathbf{x}, \mathbf{y})$ as in (13), and set $k = 0$, $b_0 = \frac{1}{\beta}$ and FLAG $= 0$.
3 **while** FLAG $= 0$ **do**
4 Call Alg. 5: $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \mathrm{FLAG}) = \mathrm{Ellipsoid}(\beta, \mathbf{z}, \delta, b_k, L_{\min}, \gamma_1, \gamma_2)$.
5 Let $b_{k+1} \leftarrow 2b_k$ and increase $k \leftarrow k+1$.
6 **Output** $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$.

Theorem 6 *Let $(\beta, \mathbf{z}, \delta, L_{\min}, \gamma_1, \gamma_2)$ be the input of Algorithm 6. Assume $\delta \le \frac{8(\mu+\beta B_g^2)}{\beta\mu}$. Then the output $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ of Algorithm 6 must satisfy $\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \le \bar{\varepsilon}$, $\widehat{\mathbf{y}} \ge \mathbf{0}$ and $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \le \delta$. In addition, it needs at most T evaluations of f, ∇f, $\boldsymbol{\theta}$, and $J_{\boldsymbol{\theta}}$ to give the output, where*

$$T \le CK + C(1+\sqrt{2})\sqrt{\gamma_1}\log\frac{\left(\sqrt{\gamma_1 L_{\max}} + \frac{L_{\max}}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_{\max}}}{\sqrt{2}} D_h}{\bar{\varepsilon}}\left(K\sqrt{\tfrac{L_f}{\mu}} + \frac{\sqrt{L_g}\max\left\{1, \frac{2\sqrt{2\|\mathbf{z}^*\|+\|\mathbf{z}\|}}{\sqrt{2}-1}\right\}}{\sqrt{\mu}}\right), \quad (28)$$

with the constants defined as $\bar{\varepsilon} = \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$, *and*

$$L_{\max} = L_f + L_g(4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|),\ C = 2\lceil 2m(m+1)\log R\rceil \cdot \left(1 + \lceil \log_{\gamma_1} \tfrac{L_{\max}}{L_{\min}}\rceil_+\right),$$
$$K = \lceil \log_2(2\|\mathbf{z}^*\| + \|\mathbf{z}\|)\rceil_+ + 1,\ R = \tfrac{64(2\|\mathbf{z}^*\|+\|\mathbf{z}\|)}{\beta}\left(\tfrac{4(\beta G+4\|\mathbf{z}^*\|+3\|\mathbf{z}\|)(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \tfrac{\mu+\beta B_g^2}{\mu\delta}\right).$$

Proof. By the quadratic formula, we can easily have the positive root of (22) to be

$$\eta_+ = \frac{\left(\frac{\mu\delta}{\mu+\beta B_g^2}\right)^2}{4\left(\sqrt{\frac{2B_d}{\beta}} + \sqrt{\frac{2B_d}{\beta} + \frac{\mu\delta}{\mu+\beta B_g^2}}\right)^2} \geq \frac{\left(\frac{\mu\delta}{\mu+\beta B_g^2}\right)^2}{8\left(\frac{4B_d}{\beta} + \frac{\mu\delta}{\mu+\beta B_g^2}\right)}.$$

Hence, it holds that

$$\frac{b}{\eta_+} \leq \frac{8b\left(\frac{4B_d}{\beta} + \frac{\mu\delta}{\mu+\beta B_g^2}\right)}{\left(\frac{\mu\delta}{\mu+\beta B_g^2}\right)^2} = 8b\left(\frac{4B_d(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \frac{\mu+\beta B_g^2}{\mu\delta}\right).$$

When $b \geq \frac{1}{\beta}$, the right hand side of the above inequality is greater than one by the assumption $\delta \leq \frac{8(\mu+\beta B_g^2)}{\beta\mu}$, and since $\eta = \min\{\eta_+, b\}$ in Algorithm 5, we have

$$\tfrac{b}{\eta} \leq 8b\left(\tfrac{4B_d(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \tfrac{\mu+\beta B_g^2}{\mu\delta}\right) \leq 8b\left(\tfrac{4(\beta G+\|\mathbf{z}\|+\beta b)(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \tfrac{\mu+\beta B_g^2}{\mu\delta}\right), \tag{29}$$

where we have used $\nabla d(\mathbf{y}) = \beta(\mathbf{g}(\mathbf{x}(\mathbf{y})) + \frac{\mathbf{z}}{\beta} - \mathbf{y})$ and thus the bound of $\nabla d(\mathbf{y})$ over $\mathcal{B}_b^+$ satisfies $B_d \leq \beta G + \|\mathbf{z}\| + \beta b$. Furthermore, by Lemma 3 and Theorem 5, Algorithm 5 must return FLAG $= 1$ and a vector $\widehat{\mathbf{y}}$ satisfying $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \delta$ when $b \geq \frac{2\|\mathbf{z}^*\|+\|\mathbf{z}\|}{\beta}$. Therefore, Algorithm 6 must stop after making at most K calls to Algorithm 5, where K is the smallest positive integer such that $2^{K-1} \geq 2\|\mathbf{z}^*\| + \|\mathbf{z}\|$, i.e., $K = \lceil \log_2(2\|\mathbf{z}^*\| + \|\mathbf{z}\|)\rceil_+ + 1$. In addition, $b_k < \frac{4\|\mathbf{z}^*\|+2\|\mathbf{z}\|}{\beta}$ for each $0 \leq k \leq K-1$.

For the k-th call to Algorithm 5, let η_k denote the used η, $L_{\psi_k} = L_f + \beta L_g b_k$ the gradient Lipschitz constant of the smooth function ψ, and T_k the total number of gradient and function evaluations. Then $L_{\psi_k} \leq L_{\max}$, and from (29), $\frac{4b_k}{\eta_k} \leq R$ for each $0 \leq k \leq K-1$. In addition, we have from (27) that

$$T_k \leq 2\lceil 2m(m+1)\log R\rceil \cdot \left(1 + \lceil \log_{\gamma_1} \tfrac{L_{\max}}{L_{\min}}\rceil_+\right) \cdot \left\lceil (1+\sqrt{2})\sqrt{\tfrac{\gamma_1 L_{\psi_k}}{\mu}} \log \frac{\left(\sqrt{\gamma_1 L_{\max}} + \frac{L_{\max}}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_{\max}}}{\sqrt{2}} D_h}{\widetilde{\delta}} \right\rceil$$
$$\leq C + C(1+\sqrt{2})\sqrt{\tfrac{\gamma_1 L_{\psi_k}}{\mu}} \log \frac{\left(\sqrt{\gamma_1 L_{\max}} + \frac{L_{\max}}{\sqrt{L_{\min}}}\right)\frac{\sqrt{\gamma_1 L_{\max}}}{\sqrt{2}} D_h}{\widetilde{\delta}}.$$

Notice that $\sqrt{L_{\psi_k}} \leq \sqrt{L_f} + \sqrt{\beta L_g b_k}$ and, thus

$$\textstyle\sum_{k=0}^{K-1} \sqrt{L_{\psi_k}} \leq K\sqrt{L_f} + \sum_{k=0}^{K-1} \sqrt{\beta L_g b_k} = K\sqrt{L_f} + \sqrt{L_g}\frac{\sqrt{2^K}-1}{\sqrt{2}-1}$$
$$\leq K\sqrt{L_f} + \sqrt{L_g}\max\left\{1, \tfrac{2\sqrt{2\|\mathbf{z}^*\|+\|\mathbf{z}\|}}{\sqrt{2}-1}\right\}.$$

Therefore, T must satisfy the condition in (28) since $T \leq \sum_{k=0}^{K-1} T_k$. □

Remark 3 In terms of the dependence on m, the number T in (28) is proportional to m^2. We can improve it to the order of m if a more advanced cutting-plane method is used, such as the one using volumetric center in [25], and the one using analytic center in [2], and the faster cutting plane method in [11].

4 Overall iteration complexity of the first-order augmented Lagrangian method

In this section, we specify the implementation details in Algorithm 1. We use the method derived in section 3 as the subroutine to find each $\mathbf{x}^{k+1}$. In addition, we choose a geometrically increasing sequence $\{\beta_k\}$ and stop the algorithm once an ε-KKT point is obtained. The pseudocode is given in Algorithm 7.

Algorithm 7: Optimal first-order inexact augmented Lagrangian method for (1) with $m = O(1)$

Input: $\beta_0 > 0$, $\sigma > 1$, tolerance $\varepsilon > 0$, $L_{\min} > 0$, $\gamma_1 > 1$, and $\gamma_2 \geq 1$
Initialization: choose $\mathbf{x}^0 \in \mathrm{dom}(h)$, and set $\mathbf{z}^0 = \mathbf{0}$
for $k = 0, 1, \ldots$ **do**
  Choose $\varepsilon_k \leq \min\left\{\varepsilon, \frac{24B_g(\mu+\beta_k B_g^2)}{\mu}\right\}$ and set $\delta_k = \frac{\varepsilon_k}{3\beta_k B_g}$.
  if $m = 1$ **then**
    Call Alg. 4: $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) = \mathrm{BiSec}(\beta_k, \mathbf{z}^k, \delta_k, L_{\min}, \gamma_1, \gamma_2)$
  else
    Call Alg. 6: $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) = \mathrm{StEM}(\beta_k, \mathbf{z}^k, \delta_k, L_{\min}, \gamma_1, \gamma_2)$
  if $m = 1$ *and* $\frac{\mu}{4\beta_k B_g^2} > 1$, *or* $m > 1$ *and* $\min\left\{\frac{\mu}{4\beta_k B_g^2}, \frac{\mu^2}{8\beta_k B_g^2(\mu+\beta_k B_g^2)}\right\} > 1$ **then**
    Call Alg. 2: $\mathbf{x}^{k+1} = \mathrm{APG}(\psi, h, \mu, L_{\min}, \varepsilon_k/3, \gamma_1, \gamma_2)$ with $\psi(\mathbf{x}) = f(\mathbf{x}) + \beta_k\langle \mathbf{y}^{k+1}, \mathbf{g}(\mathbf{x})\rangle$.
  Update $\mathbf{z}$ by $\mathbf{z}^{k+1} = [\mathbf{z}^k + \beta_k \mathbf{g}(\mathbf{x}^{k+1})]_+$.
  Let $\beta_{k+1} \leftarrow \sigma\beta_k$.
  if $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ *is an* ε*-KKT point of* (1) **then**
    Output $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ and stop

The next theorem gives a bound on the number of calls to the subroutine.

Theorem 7 *Suppose that Assumptions 1 through 3 hold. Let* $(\beta_0, \sigma, \varepsilon, \gamma_1, \gamma_2)$ *be the input of Algorithm 7 and* $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\geq 0}$ *be the generated sequence. Then* $\mathrm{dist}\big(\mathbf{0}, \partial\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \leq \varepsilon_k$ *for each* $k \geq 0$. *If* $\varepsilon_k = \bar{\varepsilon} = \min\left\{\varepsilon, \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\right\}$ *for all* $k \geq 0$, *then after at most* $K-1$ *iterations, Algorithm 7 will produce an* ε*-KKT point of* (1), *where*

$$K = \max\left\{\left\lceil \log_\sigma \frac{9\|\mathbf{z}^*\|^2}{\beta_0\varepsilon}\right\rceil_+, \left\lceil \log_\sigma \frac{1+2\sqrt{\|\mathbf{z}^*\|}}{\sqrt{\beta_0\varepsilon}}\right\rceil_+\right\} + 1. \tag{30}$$

Proof. For each $k \geq 0$, define

$$\boldsymbol{\theta}_k(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}^k}{\beta_k},\ \phi_k(\mathbf{x}) = F(\mathbf{x}) + \frac{\beta_k}{2}\left\|[\boldsymbol{\theta}_k(\mathbf{x})]_+\right\|,\quad \Phi_k(\mathbf{x}, \mathbf{y}) = F(\mathbf{x}) + \beta_k\left(\mathbf{y}^\top\boldsymbol{\theta}_k(\mathbf{x}) - \frac{1}{2}\|\mathbf{y}\|^2\right).$$

When $m=1$, if $(\mathbf{x}^{k+1},\mathbf{y}^{k+1})$ is obtained in Line 6 of Alg. 7, then we have from Theorem 4 that

$$\mathrm{dist}\big(\mathbf{0},\partial_{\mathbf{x}}\Phi_k(\mathbf{x}^{k+1},\mathbf{y}^{k+1})\big)\le\frac{\mu\delta_k}{4B_g},\ \text{ and }\ \big|[\boldsymbol{\theta}_k(\mathbf{x}(\mathbf{y}^{k+1}))]_+-\mathbf{y}^{k+1}\big|\le\delta_k.$$

Furthermore, notice that if $\frac{\mu}{4\beta_k B_g^2}>1$, we will do Line 10 in Alg. 7 to obtain a new $\mathbf{x}^{k+1}$ that satisfies $\mathrm{dist}\big(\mathbf{0},\partial_{\mathbf{x}}\Phi_k(\mathbf{x}^{k+1},\mathbf{y}^{k+1})\big)\le\frac{\varepsilon_k}{3}$. Now by Lemma 7 and the choice of δ_k, we have $\mathrm{dist}\big(\mathbf{0},\partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1},\mathbf{z}^k)\big)=\mathrm{dist}\big(\mathbf{0},\partial\phi_k(\mathbf{x}^{k+1})\big)\le\varepsilon_k$.

When $m>1$, by the choice of ε_k and δ_k, it holds $\delta_k\le\frac{8(\mu+\beta_k B_g^2)}{\beta_k\mu}$ for each k. Hence, we can use Theorem 6 and Lemma 7 to show $\mathrm{dist}\big(\mathbf{0},\partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1},\mathbf{z}^k)\big)\le\varepsilon_k$ by the same arguments as in the case of $m=1$.

Therefore, for $m\ge1$, if $\varepsilon_k=\bar\varepsilon$ for all k, we have from Theorem 3 that the inequalities in (10) and (11) hold. By the choice of $\bar\varepsilon$, it holds $\frac{\bar\varepsilon^2}{\mu(\sigma-1)}\le\frac{\varepsilon}{8\sigma+1}$. Since $K-1\ge\log_\sigma\frac{9\|\mathbf{z}^*\|^2}{\beta_0\varepsilon}$, then $\frac{9\|\mathbf{z}^*\|^2}{2\beta_0\sigma^{K-1}}\le\frac{\varepsilon}{2}$, and thus $\sum_{i=1}^m|z_i^K g_i(\mathbf{x}^K)|\le\varepsilon$. In addition, since $K-1\ge\log_\sigma\frac{1+2\sqrt{\|\mathbf{z}^*\|}}{\sqrt{\beta_0\varepsilon}}$, we have $\sigma^{K-1}\ge\frac{1+2\sqrt{\|\mathbf{z}^*\|}}{\sqrt{\beta_0\varepsilon}}\ge\frac{\sqrt{\frac{\varepsilon}{\beta_0}}+\sqrt{\frac{\varepsilon}{\beta_0}+\frac{16\varepsilon\|\mathbf{z}^*\|}{\beta_0}}}{2\varepsilon}$, which implies $\frac{4\|\mathbf{z}^*\|}{\beta_0\sigma^{K-1}}+\frac{\sqrt{\varepsilon}}{\sqrt{\beta_0\sigma^{K-1}}}\le\varepsilon$. Now noticing $\frac{\sqrt{2}(\sqrt{\sigma}+1))}{\sqrt{8\sigma+1}}\le1$, we obtain $\|[\mathbf{g}(\mathbf{x}^K)]_+\|\le\varepsilon$ and complete the proof. □

By Theorem 7, we establish the overall iteration complexity of Algorithm 7 to produce an ε-KKT point of (1). We first give the result for the case of $m=1$.

Theorem 8 (Iteration complexity when $m=1$) *Suppose that Assumptions 1 through 3 hold, and $m=1$ in (1). Let $(\beta_0,\sigma,\varepsilon,\gamma_1,\gamma_2)$ be the input of Algorithm 7 and $\{(\mathbf{x}^k,\mathbf{y}^k,\mathbf{z}^k)\}_{k\ge0}$ be the generated sequence. If $\varepsilon_k=\bar\varepsilon=\min\left\{\varepsilon,\sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\right\}$ for all $k\ge0$, then Algorithm 7 needs at most $T_{\rm total}=O\big(\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^3\big)$ evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$ to produce an ε-KKT point of (1).*

Proof. Let K be the integer given in (30) and $L_{\mathbf{z}^k}=L_f+L_g\max\{1,4\|\mathbf{z}^*\|+2\|\mathbf{z}^k\|\}$ for $0\le k\le K-1$. Also, let T_k be the number of evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$ during the k-th iteration of Algorithm 7. From Theorem 4 and the setting $\delta_k=\frac{\varepsilon_k}{3\beta_k B_g}$, we have that the complexity incurred by Line 6 of Algorithm 7 is $O(\sqrt{\frac{L_{\mathbf{z}^k}}{\mu}}|\log\varepsilon|^2)$. In addition, the complexity incurred by Line 10 is $O\big(\sqrt{\frac{L_f+L_g(\|\mathbf{z}^*\|+\|\mathbf{z}^k\|)}{\mu}}|\log\varepsilon|\big)$. From (9) with $\varepsilon_k=\bar\varepsilon$, it follows $\|\mathbf{z}^k\|=O(\|\mathbf{z}^*\|)$, and thus $L_{\mathbf{z}^k}=O(L_f+L_g(1+\|\mathbf{z}^*\|))$ for $0\le k\le K-1$. Therefore, $T_k=O\big(\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2\big)$. Since $K=O(|\log\varepsilon|)$ in (30), the total complexity $T_{\rm total}=\sum_{k=0}^{K-1}T_k=O\big(\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^3\big)$, which completes the proof. □

Remark 4 If β_0 is taken in the order of $\frac{1}{\varepsilon}$, then $K=O(1)$ in (30). In this case, the total complexity of Algorithm 7 is $O\big(\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2\big)$ to produce an ε-KKT point.

Similarly, we can show the complexity result for the case of $m>1$ by using Theorem 6.

Theorem 9 (Iteration complexity when $m>1$) *Suppose that Assumptions 1 through 3 hold, and $m>1$ in (1). Let $(\beta_0,\sigma,\varepsilon,\gamma_1,\gamma_2)$ be the input of Algorithm 7 and $\{(\mathbf{x}^k,\mathbf{y}^k,\mathbf{z}^k)\}_{k\ge0}$ be the generated sequence. If $\varepsilon_k=\bar\varepsilon=\min\left\{\varepsilon,\sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\right\}$ for all $k\ge0$, then Algorithm 7 needs at most $T_{\rm total}=O\big(m^2\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^3\big)$ evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$ to produce an ε-KKT point of (1).*

Remark 5 Similar to Remark 4, the total complexity can be improved to $O\big(m^2\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2\big)$ if $\beta_0=\Theta(\frac{1}{\varepsilon})$. Ignoring the term $|\log\varepsilon|$, our result is better than the best known result $O\big(\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu\varepsilon}}|\log\varepsilon|\big)$ if $m=o(\varepsilon^{-\frac{1}{4}})$. As we discussed in Remark 3, the dependence on m^2 can be improved to m if a more advanced cutting plane method is used. In this case, we can obtain a result $O\big(m\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2\big)$ that is better than $O\big(\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu\varepsilon}}|\log\varepsilon|\big)$ if $m=o(\varepsilon^{-\frac{1}{2}})$ by ignoring the logarithmic term $|\log\varepsilon|$.

5 Concluding remarks

We have proposed a first-order method (FOM) for solving strongly-convex problems with m functional constraints. If $m=O(1)$, our method can achieve a complexity result of $\tilde{O}(\sqrt{\kappa})$, where κ denotes the condition number in some sense. In general, a complexity result of $\tilde{O}(m^2\sqrt{\kappa})$ has been established. To give an ε-KKT point, our result is better than an existing lower bound if $m=o(\varepsilon^{-\frac{1}{4}})$. Out result can be further improved to $\tilde{O}(m\sqrt{\kappa})$ by using a more advanced cutting-plane method as the key ingredient in our algorithm.

Although we only studied the strongly-convex case, our result can be easily extended to the convex case and the weakly-convex case. Suppose $m=O(1)$. For the convex case, we can perturb the objective by adding a small error-dependent quadratic term to have a strongly-convex perturbed problem as in [10]. This way, we can obtain an FOM that will produce an ε-KKT point with overall complexity $\tilde{O}(\varepsilon^{-\frac{1}{2}})$. For the weakly-convex case, we can follow [13] to have an FOM based on the proximal-point method framework, which can achieve an ε-KKT point with complexity $\tilde{O}(\varepsilon^{-2})$. Therefore, we can have almost the same complexity results (with a difference at most a polynomial of $|\log\varepsilon|$) as for solving an unconstrained problem, in the strongly-convex case, or the convex case, or the weakly-convex case.

References

1. A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy. Level-set methods for convex optimization. *Mathematical Programming*, 174(1-2):359–390, 2019. 3
2. D. S. Atkinson and P. M. Vaidya. A cutting plane algorithm for convex programming that uses analytic centers. *Mathematical Programming*, 69(1-3):1–43, 1995. 17
3. N. S. Aybat and G. Iyengar. An augmented lagrangian method for conic convex programming. *arXiv preprint arXiv:1302.6322*, 2013. 3
4. R. G. Bland, D. Goldfarb, and M. J. Todd. The ellipsoid method: A survey. *Operations research*, 29(6):1039–1091, 1981. 14, 15
5. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. 3
6. Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017. 3
7. R. Gandy. *Portfolio optimization with risk constraints*. PhD thesis, Universität Ulm, 2005. 2
8. E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401v4*, 2018. 3
9. L. T. K. Hien, R. Zhao, and W. B. Haskell. An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. *arXiv preprint arXiv:1711.03669v3*, 2017. 3
10. G. Lan and R. D. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016. 3, 19
11. Y. T. Lee, A. Sidford, and S. C.-w. Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1049–1065. IEEE, 2015. 17

12. F. Li and Z. Qu. An inexact proximal augmented lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization. *arXiv preprint arXiv:1909.09582*, 2019. 3, 8
13. Q. Lin, R. Ma, and Y. Xu. Inexact proximal-point penalty methods for non-convex optimization with non-convex constraints. *arXiv preprint arXiv:1908.11518*, 2019. 19
14. Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *International Conference on Machine Learning*, pages 3112–3121, 2018. 3
15. Q. Lin, S. Nadarajah, and N. Soheili. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization*, 28(4):3290–3311, 2018. 3
16. Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented lagrangian methods for convex conic programming. *arXiv preprint arXiv:1803.09941*, 2018. 3, 8
17. R. D. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010. 3
18. I. Necoara and V. Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5):1232–1243, 2014. 3
19. A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009. 3
20. A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009. 3
21. A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. 3
22. Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. 2, 5
23. Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming, Series A (online first)*, pages 1–35, 2019. 1, 2, 8
24. P. Rigollet and X. Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12(Oct):2831–2855, 2011. 2
25. P. M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical programming*, 73(3):291–341, 1996. 17
26. Y. Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *Mathematical Programming, Series A (online first)*, pages 1–46, 2019. 2, 7
27. Y. Xu. First-order methods for constrained convex programming based on linearized augmented lagrangian function. *INFORMS Journal on Optimization (to appear)*, 2020. 3
28. Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013. 6
29. H. Yu and M. J. Neely. A primal-dual type algorithm with the $O(1/t)$ convergence rate for large scale constrained convex programs. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 1900–1905. IEEE, 2016. 3
30. M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015. 2