SIAM J. OPTIM.
Vol. 32, No. 3, pp. 1759–1790

FIRST-ORDER METHODS FOR PROBLEMS WITH $O(1)$ FUNCTIONAL CONSTRAINTS CAN HAVE ALMOST THE SAME CONVERGENCE RATE AS FOR UNCONSTRAINED PROBLEMS*

YANGYANG XU†

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 the existing lower complexity bound because we focus on a specific subclass of problems. Experimental results on quadratically constrained quadratic programs demonstrate our theory.

Key words. first-order method, cutting-plane method, nonlinearly constrained problem, iteration complexity

MSC codes. 65K05, 68Q25, 90C30, 90C60

DOI. 10.1137/20M1371579

1. Introduction.

In this paper, we consider the constrained convex programming

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}),\ \text{s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \dots, g_m(\mathbf{x})] \le \mathbf{0}, \tag{1.1}$$

where f is a differentiable strongly 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 $\min_{\mathbf{x}}\{f(\mathbf{x}),\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}\}$, the authors of [32] give 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 $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\} = \{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}, -\mathbf{A}\mathbf{x} \le -\mathbf{b}\}$. 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 [32] also applies to the inequality constrained problem (1.1) 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, 31] can achieve a complexity result $O(\sqrt{\kappa}|\log\varepsilon|)$ to produce an ε-optimal solution of (1.1) when f is strongly convex. Here, κ denotes the condition number.

*Received by the editors October 6, 2020; accepted for publication (in revised form) March 31, 2022; published electronically August 1, 2022.

Funding: This work was partly supported by NSF grant DMS-2053493.

https://doi.org/10.1137/20M1371579

†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (xuy21@rpi.edu).

The worst-case instance constructed in [32] 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 no longer hold. Examples of (1.1) with small m include the Neyman–Pearson classification problem [33], the fairness-constrained classification [43], and the risk-constrained portfolio optimization [10]. Therefore, we pose the following question while solving a strongly convex problem in the form of (1.1):

> Given $\varepsilon > 0$, can an FOM achieve a better complexity result than $O(\frac{1}{\sqrt{\varepsilon}})$ to produce an ε-optimal solution of (1.1) when $m = o(\frac{1}{\sqrt{\varepsilon}})$, or even achieve $\tilde{O}(\sqrt{\kappa})$ when $m = O(1)$?

Here, an FOM for (1.1) 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 on the inexact augmented Lagrangian method (iALM). The classic AL function of (1.1) is

$$\mathcal{L}_\beta(\mathbf{x},\mathbf{z}) = F(\mathbf{x}) + \tfrac{\beta}{2}\left\|[\mathbf{g}(\mathbf{x}) + \tfrac{\mathbf{z}}{\beta}]_+\right\|^2 - \tfrac{\|\mathbf{z}\|^2}{2\beta}, \tag{1.2}$$

where $\mathbf{z}$ is the multiplier vector, and $[\mathbf{a}]_+$ takes the componentwise 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 strongly convex about $\mathbf{x}$ and concave about $\mathbf{z}$. Hence, we can directly apply the accelerated proximal gradients in [22, 31] to solve each $\mathbf{x}$-subproblem. However, that way can only give a complexity result of $O(\frac{1}{\sqrt{\varepsilon}})$ as shown in [40], regardless of the value of m. To have a better overall complexity, we will design a new cutting-plane based 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 (1.1).

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 $\beta_{k+1} \geq \beta_k$.
  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, the authors of [27, 28] 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 (1.6) below). For a smooth problem, the authors of [26] study 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. For convex problems, the primal-dual FOM proposed in [42] achieves an $O(\varepsilon^{-1})$ complexity result to produce an ε-optimal solution, and the same-order complexity result has also been established in [39]. Based on a previous work [15] for affinely constrained problems, the authors of [23] give 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 1.1 below). A similar result has also been shown in [3] for convex conic programs. A proximal iALM is analyzed in [16]. By a linearly convergent first-order subroutine for primal subproblems, the authors of [16] show 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. Complexity results of FOMs for nonconvex problems with functional constraints have also been established; see, e.g., [6, 7, 14, 17, 18, 19, 24, 35]. To produce an ε-KKT point, the best-known result is $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ when the constraints are convex [17, 19] and $\tilde{O}(\varepsilon^{-3})$ when the constraints are nonconvex and satisfy a certain regularity condition [19].

On solving general nonlinear constrained problems, FOMs have also been proposed under the framework of the level-set method [2, 20, 21]. 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. Nesterov gives a level-set-type FOM in [30] for functional constrained problems. For strongly convex problems, the method can produce an ε-optimal solution by $O(\sqrt{\kappa}|\log\varepsilon|\log\kappa)$ first-order oracles [30, eq. 2.3.26], where κ is the condition number. This oracle complexity result differs from a lower-bound result for unconstrained problems only by a factor of $\log\kappa$. However, the book [30] requires strong convexity for the objective function and all the constraint functions. In contrast, we will only need strong convexity for the objective, while the constraint functions can be merely convex. In addition, the method in [30] assumes exact solutions to a sequence of quadratically constrained quadratic programs.

Under the condition of strong duality, (1.1) can be equivalently formulated as a nonbilinear saddle-point (SP) problem. In this case, one can apply any FOM that is designed for solving nonbilinear SP problems. The work [12] generalizes the primal-dual method proposed in [8] from the bilinear SP case to the nonbilinear case. If the underlying SP problem is convex-concave, the work [12] establishes an $O(\varepsilon^{-1})$ complexity result to guarantee an 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 (1.1). By the smoothing technique, the authors of [13] give an FOM (with both deterministic and stochastic versions) for solving nonbilinear 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 (1.1), the method in [13] 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 (1.1), such as the mirror-prox method in [29], the hybrid extragradient method in [25], and the accelerated method in [9]. 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 with a strongly convex objective and convex constraint functions, 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 (1.1), an FOM can achieve almost the same-order complexity result (with a difference of 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 $m = o(\varepsilon^{-\frac{1}{2}})$. In addition, we extend the idea of a cutting-plane based FOM to the convex and nonconvex cases. For these two cases, we show that an FOM for problems with $O(1)$ functional constraints can also achieve almost the same-order complexity result as for solving unconstrained problems.

1.4. Assumptions and notation. Throughout our analysis for strongly convex problems, 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 convexity). *The domain of h is bounded with a diameter $D_h = \max_{\mathbf{x},\mathbf{y}\in\mathrm{dom}(h)} \|\mathbf{x}-\mathbf{y}\| < \infty$. The functions h and $\{g_i\}$ are all convex.*

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

$$G = \max_{\mathbf{x}\in\mathrm{dom}(h)} \|\mathbf{g}(\mathbf{x})\|, \quad B_g = \max_{\mathbf{x}\in\mathrm{dom}(h)} \|J_{\mathbf{g}}(\mathbf{x})\|. \tag{1.3}$$

ASSUMPTION 3 (strong convexity). *The smooth function f is μ-strongly convex with $\mu > 0$.*

ASSUMPTION 4 (strong duality). *There is a primal-dual solution $(\mathbf{x}^*, \mathbf{z}^*)$ satisfying the KKT conditions of (1.1), i.e., $\mathbf{0} \in \partial F(\mathbf{x}^*) + J_{\mathbf{g}}(\mathbf{x}^*)^\top \mathbf{z}^*$, $\mathbf{z}^* \geq \mathbf{0}$, $g(\mathbf{x}^*) \leq \mathbf{0}$, $\mathbf{g}(\mathbf{x}^*)^\top \mathbf{z}^* = 0$.*

When Assumption 4 holds, it is easy to have (cf. [38, eq. 2.4])

$$F(\mathbf{x}) - F(\mathbf{x}^*) + \langle \mathbf{z}^*, \mathbf{g}(\mathbf{x})\rangle \geq 0 \ \forall\, \mathbf{x} \in \mathrm{dom}(h). \tag{1.4}$$

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 $X \subseteq \mathbb{R}^n$ and a point $\mathbf{x} \in \mathbb{R}^n$, we define $\mathrm{dist}(\mathbf{x}, X) = \min_{\mathbf{y}\in X} \|\mathbf{y} - \mathbf{x}\|$. For any vector $\mathbf{x}$, $\mathrm{Diag}(\mathbf{x})$ denotes a diagonal matrix with $\mathbf{x}$ on the diagonal, and for any square matrix $\mathbf{A}$, $\mathrm{diag}(\mathbf{A})$ is a vector that takes the diagonal of $\mathbf{A}$. 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 1.1 (ε-KKT point). *Given $\varepsilon > 0$, a point $\bar{\mathbf{x}} \in \mathrm{dom}(h)$ is called an ε-KKT point of* (1.1) *if there is $\bar{\mathbf{z}} \geq \mathbf{0}$ such that*

$$\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_0(\bar{\mathbf{x}}, \bar{\mathbf{z}})\big) \leq \varepsilon, \quad \|[\mathbf{g}(\bar{\mathbf{x}})]_+\| \leq \varepsilon, \quad \sum_{i=1}^{m} |\bar{z}_i g_i(\bar{\mathbf{x}})| \leq \varepsilon, \tag{1.5}$$

where $\mathcal{L}_0(\mathbf{x}, \mathbf{z}) = F(\mathbf{x}) + \mathbf{z}^\top \mathbf{g}(\mathbf{x})$ is the ordinary Lagrangian function of (1.1).

By the convexity of F and each g_i, and also Assumption 4, one can easily show that an ε-KKT point of (1.1) must be an $O(\varepsilon)$-optimal solution, where we call a point $\bar{\mathbf{x}} \in \mathrm{dom}(h)$ an ε-optimal solution of (1.1) if

$$\big|F(\bar{\mathbf{x}}) - F(\mathbf{x}^*)\big| \leq \varepsilon, \quad \|[\mathbf{g}(\bar{\mathbf{x}})]_+\| \leq \varepsilon. \tag{1.6}$$

1.5. Outline. The rest of the paper is organized as follows. In section 2, we review an adaptive accelerated proximal gradient (APG) method and give the convergence rate of the iALM. In section 3, we design new FOMs (that are better than directly applying the APG method) for solving primal subproblems in the iALM. Overall complexity results are shown in section 4. Extensions to convex and nonconvex cases are given in section 5. Numerical experiments are conducted in section 6 to demonstrate our theory, and section 7 concludes the paper.

2. An optimal FOM and convergence rate of iALM.

In this section, we give an optimal FOM with line search 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. An optimal FOM for strongly convex composite problems. Consider the problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, P(\mathbf{x}) := \psi(\mathbf{x}) + r(\mathbf{x}), \tag{2.1}$$

where ψ is a differentiable μ_ψ-strongly convex function with an L_ψ-Lipschitz continuous gradient, and r is a closed convex function. Several optimal FOMs have been given in the literature for solving (2.1), e.g., in [22, 31]. In this paper, we choose the APG method with line search in [22], and we rewrite it in Algorithm 2 with a few modified steps for our purpose to produce near-stationary points. One can also use the APG method in [31].

The results in the next theorem are from Theorem 1 of [22].

THEOREM 2.1. *The generated sequence $\{\mathbf{x}^k\}_{k\geq 0}$ by Algorithm 2 satisfies*

$$P(\mathbf{x}^{k+1}) - P(\mathbf{x}^*) \leq \left(1 - \sqrt{\frac{\mu_\psi}{\gamma_1 L_\psi}}\right)^{k+1} \left(P(\mathbf{x}^0) - P(\mathbf{x}^*) + \frac{\mu_\psi}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right) \ \forall\, k \geq 0, \tag{2.2}$$

where $\mathbf{x}^*$ is the optimal solution of (2.1).

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

THEOREM 2.2. *The generated sequence $\{\widehat{\mathbf{x}}^k\}_{k\geq 0}$ satisfies*

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

Algorithm 2: An optimal FOM with line search for (2.1): $\widehat{\mathbf{x}} = \mathrm{APG}(\psi, r, \mu_\psi, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$.

Input: minimum Lipschitz $L_{\min} > 0$, increase rate $\gamma_1 > 1$, decrease rate $\gamma_2 \geq 1$, and error tolerance $\bar{\varepsilon} > 0$.
Prestep: choose any $\widetilde{\mathbf{y}} = \mathbf{y}^0 \in \mathrm{dom}(r)$ and let $\widetilde{L} = L_{\min}/\gamma_1$
repeat
  $\widetilde{L} \leftarrow \gamma_1 \widetilde{L}$ and let $\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{y}}), \mathbf{x}\rangle + \frac{\widetilde{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{y}}\|^2 + r(\mathbf{x})$
until $\psi(\widetilde{\mathbf{x}}) \leq \psi(\widetilde{\mathbf{y}}) + \langle \nabla\psi(\widetilde{\mathbf{y}}), \widetilde{\mathbf{x}} - \widetilde{\mathbf{y}}\rangle + \frac{\widetilde{L}}{2}\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{y}}\|^2$
Initialization: let $\mathbf{x}^{-1} = \mathbf{x}^0 = \widetilde{\mathbf{x}}$, $L_0 = \max\{L_{\min}, \widetilde{L}/\gamma_2\}$, and $\alpha_{-1} = 1$
for $k = 0, 1, \ldots$ **do**
  $\widetilde{L} \leftarrow L_k/\gamma_1$
  repeat
    $\widetilde{L} \leftarrow \gamma_1 \widetilde{L}$, $\alpha_k \leftarrow \sqrt{\mu_\psi/\widetilde{L}}$, and $\widetilde{\mathbf{y}} \leftarrow \mathbf{x}^k + \frac{\alpha_k(1-\alpha_{k-1})}{\alpha_{k-1}(1+\alpha_k)}(\mathbf{x}^k - \mathbf{x}^{k-1})$
    let $\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{y}}), \mathbf{x}\rangle + \frac{\widetilde{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{y}}\|^2 + r(\mathbf{x})$
  until $\psi(\widetilde{\mathbf{x}}) \leq \psi(\widetilde{\mathbf{y}}) + \langle \nabla\psi(\widetilde{\mathbf{y}}), \widetilde{\mathbf{x}} - \widetilde{\mathbf{y}}\rangle + \frac{\widetilde{L}}{2}\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{y}}\|^2$
  $\widehat{L} \leftarrow \widetilde{L}/\gamma_1$;
  repeat
    increase $\widehat{L} \leftarrow \gamma_1 \widehat{L}$;
    let $\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \langle \nabla\psi(\widetilde{\mathbf{x}}), \mathbf{x}\rangle + \frac{\widehat{L}}{2}\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 + r(\mathbf{x})$; ▷ modified step to guarantee near-stationarity at $\widehat{\mathbf{x}}$
  until $\psi(\widehat{\mathbf{x}}) \leq \psi(\widetilde{\mathbf{x}}) + \langle \nabla\psi(\widetilde{\mathbf{x}}), \widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\rangle + \frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2$
  set $\mathbf{x}^{k+1} = \widetilde{\mathbf{x}}$, $\widehat{\mathbf{x}}^{k+1} = \widehat{\mathbf{x}}$, and $L_{k+1} = \max\{L_{\min}, \widetilde{L}/\gamma_2\}$;
  if $\mathrm{dist}\big(\mathbf{0}, \partial P(\widehat{\mathbf{x}})\big) \leq \bar{\varepsilon}$ **then**
    return $\widehat{\mathbf{x}}$ and stop.

Proof. First notice that if $\widehat{L} \geq L_\psi$, it must hold that $\psi(\widehat{\mathbf{x}}) \leq \psi(\widetilde{\mathbf{x}}) + \langle \nabla\psi(\widetilde{\mathbf{x}}), \widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\rangle + \frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2$, and when this inequality holds, we have (cf. [41, Lem. 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 $\frac{\widehat{L}}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2 \leq P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)$, which together with the fact $\widehat{L} \geq L_{\min}$ implies

$$\frac{\widehat{L}^2}{2}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2 \leq \widehat{L}\big(P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)\big), \quad \|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|^2 \leq \tfrac{2}{L_{\min}}\big(P(\widetilde{\mathbf{x}}) - P(\mathbf{x}^*)\big). \tag{2.3}$$

In addition, from the optimality condition of $\widehat{\mathbf{x}}$, it follows that $\mathbf{0} \in \nabla\psi(\widetilde{\mathbf{x}}) + \widehat{L}(\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}) + \partial r(\widehat{\mathbf{x}})$, and thus

$$\mathrm{dist}(\mathbf{0}, \partial P(\widehat{\mathbf{x}})) \leq \|\nabla\psi(\widehat{\mathbf{x}}) - \nabla\psi(\widetilde{\mathbf{x}})\| + \widehat{L}\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\| \leq (L_\psi + \widehat{L})\|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|. \tag{2.4}$$

By (2.3) and (2.4), 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 (2.2), the fact that $\widehat{L} \leq \gamma_1 L_\psi$, and the above inequality with $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^{k+1}$ and $\widetilde{\mathbf{x}} = \mathbf{x}^{k+1}$. ∎

From [4, Thm. 3.1], we have

$$P(\mathbf{x}^0) - P(\mathbf{x}^*) \leq \tfrac{\gamma_1 L_\psi \|\mathbf{y}^0 - \mathbf{x}^*\|^2}{2}. \tag{2.5}$$

Hence, we can obtain the following complexity result by Theorem 2.2 together with (2.5).

COROLLARY 2.3. *Assume that* $\mathrm{dom}(r)$ *is bounded with a diameter*

$$D_r = \max_{\mathbf{x}_1,\mathbf{x}_2\in\mathrm{dom}(r)} \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

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*

$$T = \left(1 + \lceil \log_{\gamma_1} \tfrac{L_\psi}{L_{\min}} \rceil_+\right)\left(1 + 2\left\lceil 2\sqrt{\tfrac{\gamma_1 L_\psi}{\mu_\psi}} \log\left(\tfrac{D_r}{\bar{\varepsilon}}\left(\sqrt{\gamma_1 L_\psi} + \tfrac{L_\psi}{\sqrt{L_{\min}}}\right)\sqrt{2\gamma_1 L_\psi + \mu_\psi}\right)\right\rceil_+\right).$$

Proof. Since $\mathrm{dom}(r)$ has a diameter D_r, we have from Theorem 2.2 and (2.5) that

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

Hence, if $k + 1 \geq K$, then $\mathrm{dist}\big(\mathbf{0}, \partial P(\widehat{\mathbf{x}}^{k+1})\big) \leq \bar{\varepsilon}$, where

$$K = \left\lceil \frac{2\log\left(\frac{D_r}{\bar{\varepsilon}}\left(\sqrt{\gamma_1 L_\psi} + \frac{L_\psi}{\sqrt{L_{\min}}}\right)\sqrt{2\gamma_1 L_\psi + \mu_\psi}\right)}{\log(1 - \sqrt{\frac{\mu_\psi}{\gamma_1 L_\psi}})^{-1}} \right\rceil_+;$$

namely, after at most K iterations, the algorithm will produce a point $\widehat{\mathbf{x}}$ satisfying $\mathrm{dist}(\mathbf{0}, \partial P(\widehat{\mathbf{x}})) \leq \bar{\varepsilon}$.

Notice that the conditions in lines 5, 12, and 17 of Algorithm 2 will hold if $\widetilde{L} \geq L_\psi$ and $\widehat{L} \geq L_\psi$. Hence, 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 that $\log(1-a)^{-1} \geq a$ for all $0 < a < 1$, we obtain the desired result by also counting the objective and gradient evaluations to obtain $\mathbf{x}^0$. ∎

2.2. Convergence rate of iALM. The next lemma is from [40, eq. 3.20] and the proof of [40, Lem. 7].

LEMMA 2.4. *Let* $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ *be generated from Algorithm* 1 *with* $\mathbf{z}^0 = \mathbf{0}$. *Suppose*

$$\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) \leq \min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k) + e_k \;\forall\, k = 0, 1, \ldots \tag{2.6}$$

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

$$\|\mathbf{z}^k\|^2 \leq 4\|\mathbf{z}^*\|^2 + 4\sum_{t=0}^{k-1}\beta_t e_t, \;\; \textit{and} \;\; \|\mathbf{z}^k\| \leq 2\|\mathbf{z}^*\| + \sqrt{2\sum_{t=0}^{k-1}\beta_t e_t} \;\forall\, k \geq 1. \tag{2.7}$$

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

LEMMA 2.5. *Let* $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ *be generated from Algorithm* 1 *with* $\mathbf{z}^0 = \mathbf{0}$. *If*

$$\mathrm{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \leq \varepsilon_k \;\forall\, k \geq 0$$

for a sequence $\{\varepsilon_k\}$, *then*

$$\|\mathbf{z}^k\|^2 \leq 4\|\mathbf{z}^*\|^2 + 4\sum_{t=0}^{k-1}\beta_t\frac{\varepsilon_t^2}{\mu}, \;\; \textit{and} \;\; \|\mathbf{z}^k\| \leq 2\|\mathbf{z}^*\| + \sqrt{2\sum_{t=0}^{k-1}\beta_t\frac{\varepsilon_t^2}{\mu}} \;\forall\, k \geq 1. \tag{2.8}$$

Proof. If $\mathbf{x}_*^{k+1}$ is the minimizer of $\mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$ about $\mathbf{x}$, then $\mathbf{0} \in \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}_*^{k+1}, \mathbf{z}^k)$. Also, it follows from $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \le \varepsilon_k$ that there is $\mathbf{v} \in \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)$ and $\|\mathbf{v}\| \le \varepsilon_k$. Since F is μ-strongly convex, $\mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$ is also μ-strongly convex about $\mathbf{x}$. Then we have $\langle \mathbf{v}, \mathbf{x}^{k+1} - \mathbf{x}_*^{k+1} \rangle \ge \mu \|\mathbf{x}^{k+1} - \mathbf{x}_*^{k+1}\|^2$, which together with the Cauchy–Schwarz inequality gives $\|\mathbf{x}^{k+1} - \mathbf{x}_*^{k+1}\| \le \frac{\|\mathbf{v}\|}{\mu} \le \frac{\varepsilon_k}{\mu}$. Now, by the convexity of $\mathcal{L}_{\beta_k}(\cdot, \mathbf{z}^k)$, it holds that $\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) - \mathcal{L}_{\beta_k}(\mathbf{x}_*^{k+1}, \mathbf{z}^k) \le \langle \mathbf{v}, \mathbf{x}^{k+1} - \mathbf{x}_*^{k+1} \rangle \le \frac{\varepsilon_k^2}{\mu}$, and thus we have that (2.6) holds with $e_t = \frac{\varepsilon_t^2}{\mu}$. Hence, (2.8) follows from (2.7). ∎

THEOREM 2.6 (convergence rate of iALM). *Let $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ be generated from Algorithm 1 with $\mathbf{z}^0 = \mathbf{0}$. Suppose $\beta_k = \beta_0 \sigma^k$ for all $k \ge 0$ for some $\sigma > 1$ and $\beta_0 > 0$, and $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \le \bar{\varepsilon}$ for all $k \ge 0$ for a positive number $\bar{\varepsilon}$. Then*

$$\big\|[\mathbf{g}(\mathbf{x}^{k+1})]_+\big\| \le \tfrac{4\|\mathbf{z}^*\|}{\beta_0 \sigma^k} + \frac{\bar{\varepsilon}(\sqrt{\sigma}+1)\sqrt{\frac{2}{\mu(\sigma-1)}}}{\sqrt{\beta_0 \sigma^k}}, \tag{2.9}$$

$$\sum_{i=1}^m \big|z_i^{k+1} g_i(\mathbf{x}^{k+1})\big| \le \frac{9\|\mathbf{z}^*\|^2}{2\beta_0 \sigma^k} + \frac{\bar{\varepsilon}^2(8\sigma+1)}{2\mu(\sigma-1)}. \tag{2.10}$$

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 (2.8) we have

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

Plugging into the above inequality $\varepsilon_t = \bar{\varepsilon}$ for all $t \ge 0$ and $\beta_k = \beta_0 \sigma^k$, we obtain the inequality in (2.9).

Furthermore, for each $i \in [m]$, we have $|z_i^{k+1} g_i(\mathbf{x}^{k+1})| \le \frac{1}{\beta_k} |z_i^{k+1}(z_i^{k+1} - z_i^k)|$. Notice that z_i^k and z_i^{k+1} are both nonnegative. If $z_i^{k+1} \ge z_i^k$, then it is obvious to have $|z_i^{k+1}(z_i^{k+1} - z_i^k)| \le (z_i^{k+1})^2$, and if $z_i^{k+1} < z_i^k$, it holds that $|z_i^{k+1}(z_i^{k+1} - z_i^k)| = -(z_i^{k+1})^2 + z_i^k z_i^{k+1} \le (z_i^{k+1})^2 + \frac{(z_i^k)^2}{8}$ by Young's inequality. Hence, $|z_i^{k+1} g_i(\mathbf{x}^{k+1})| \le \frac{1}{\beta_k}((z_i^{k+1})^2 + \frac{(z_i^k)^2}{8})$, and thus

$$\sum_{i=1}^m \big|z_i^{k+1} g_i(\mathbf{x}^{k+1})\big| \le \frac{1}{\beta_k} \left(\|\mathbf{z}^{k+1}\|^2 + \frac{\|\mathbf{z}^k\|^2}{8} \right).$$

Now we obtain the result in (2.10) by plugging the first inequality in (2.8). ∎

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 2.6. 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 (2.9) and (2.10), it follows that to ensure that $\mathbf{x}^{k+1}$ is an ε-KKT point, we need $\beta_0 \sigma^k = \Theta(\frac{1}{\varepsilon})$ and to 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 a $\Theta(\beta_k)$-Lipschitz continuous gradient, it needs $O(\sqrt{\frac{\beta_k}{\mu}})$ proximal gradient steps if we directly apply Algorithm 2. This way, we can guarantee an ε-KKT point with a 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., [16, 23]. It is worse by an order of $\sqrt{\frac{1}{\varepsilon}}$ than the complexity result in Corollary 2.3 for the unconstrained case. Generally, we cannot improve it any more because the result matches with the lower bound given in [32].

In the rest of the paper, we show that in some special cases a better complexity can be obtained. When $m = O(1)$, we show that we can achieve a complexity result $O(\sqrt{\kappa}|\log\varepsilon|^3)$, which is in almost the same order as the optimal result for the unconstrained case. For a general m, we can achieve $O\big(m\sqrt{\kappa}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$, which is better than $O(\sqrt{\frac{\kappa}{\varepsilon}}|\log\varepsilon|)$ in the regime of $m = o(\sqrt{\frac{1}{\varepsilon}})$ by ignoring the logarithmic terms.

3. Better first-order methods for x-subproblems. When m is small in (1.1), 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 in the framework of a cutting-plane method. Our key idea is to reformulate the $\mathbf{x}$-subproblem into a strongly convex–strongly concave SP problem, which has a unique primal-dual solution. For the SP formulation, we first find a sufficiently accurate dual solution by a cutting-plane based FOM. Then we find a sufficiently accurate primal solution based on the obtained approximate dual solution.

Below, we give more precise description of how to design better FOMs. Given $\mathbf{z} \geq \mathbf{0}$, let

$$\boldsymbol{\theta}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \tfrac{\mathbf{z}}{\beta}.$$

From (1.3) and the mean-value theorem, it follows that $\boldsymbol{\theta}$ is B_g-Lipschitz continuous, namely,

$$\|\boldsymbol{\theta}(\mathbf{x}_1) - \boldsymbol{\theta}(\mathbf{x}_2)\| \leq B_g\|\mathbf{x}_1 - \mathbf{x}_2\| \ \forall\, \mathbf{x}_1, \mathbf{x}_2. \tag{3.1}$$

With $\boldsymbol{\theta}$, we can rewrite the problem $\min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z})$ into

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, \phi(\mathbf{x}) := F(\mathbf{x}) + \tfrac{\beta}{2}\|[\boldsymbol{\theta}(\mathbf{x})]_+\|^2. \tag{3.2}$$

Notice that $\frac{1}{2}\|[\boldsymbol{\theta}(\mathbf{x})]_+\|^2 = \max_{\mathbf{y}\geq\mathbf{0}} \left\{\mathbf{y}^\top\boldsymbol{\theta}(\mathbf{x}) - \frac{1}{2}\|\mathbf{y}\|^2\right\}$ and $\mathbf{y} = [\boldsymbol{\theta}(\mathbf{x})]_+$ reaches the maximum. We rewrite (3.2) into

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\geq\mathbf{0}}\ \Phi(\mathbf{x}, \mathbf{y}) := F(\mathbf{x}) + \beta\left(\mathbf{y}^\top\boldsymbol{\theta}(\mathbf{x}) - \tfrac{1}{2}\|\mathbf{y}\|^2\right). \tag{3.3}$$

Define

$$d(\mathbf{y}) = \min_{\mathbf{x}\in\mathbb{R}^n} \Phi(\mathbf{x}, \mathbf{y}) \ \text{ and } \ \bar{\mathbf{y}} = \arg\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y}). \tag{3.4}$$

Notice that d is β-strongly concave, so $\bar{\mathbf{y}}$ is the unique maximizer of d. Also, for a given $\mathbf{y} \geq \mathbf{0}$, define $\mathbf{x}(\mathbf{y})$ as the unique minimizer of $\Phi(\cdot, \mathbf{y})$, i.e.,

$$\mathbf{x}(\mathbf{y}) = \arg\min_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y}). \tag{3.5}$$

In our algorithm design, we first find an approximate solution $\widehat{\mathbf{y}}$ of $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ and then 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 $\mathbf{y} \geq \mathbf{0}$, though we can only have an estimate of $\nabla d(\mathbf{y})$ by approximately solving $\min_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y})$. It is unclear whether the same idea works 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. The next lemma indicates that the complexity of solving $\min_{\mathbf{x}} \Phi(\mathbf{x},\mathbf{y})$ by the APG method can be independent of β if $\|\mathbf{y}\|$ is in the same order of $\|\bar{\mathbf{y}}\|$. This fact is the key for us to design a better FOM for solving ALM subproblems.

LEMMA 3.1. *Suppose* $\bar{\mathbf{x}}$ *is the minimizer of* ϕ *in* (3.2). *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* 4. *Then*

$$\|\bar{\mathbf{y}}\| = \|[\boldsymbol{\theta}(\bar{\mathbf{x}})]_+\| \leq \tfrac{2\|\mathbf{z}^*\|+\|\mathbf{z}\|}{\beta}. \tag{3.6}$$

Proof. It is easy to see that $\bar{\mathbf{y}} = [\boldsymbol{\theta}(\bar{\mathbf{x}})]_+$ is the solution of $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ and $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ is a saddle point of Φ; cf. [34, Cor. 37.3.2]. We only need to show (3.6). Since $\bar{\mathbf{x}}$ is the minimizer of ϕ, it holds that

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

where the last inequality holds because $\mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$ and $\mathbf{z} \geq \mathbf{0}$. By the above inequality and (1.4), we have

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

which implies the inequality in (3.6). ∎

Our cutting-plane based FOM for solving $\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ needs a sufficiently accurate approximation of $\nabla d(\mathbf{y})$ at any query point $\mathbf{y}$. We first give the formula of $\nabla d(\mathbf{y})$ in Lemma 3.2 and then provide a way to approximate it with a desired accuracy in Lemma 3.3.

LEMMA 3.2. *For any* $\mathbf{y} \geq \mathbf{0}$, *it holds that*

$$\nabla d(\mathbf{y}) = \beta\big(\boldsymbol{\theta}(\mathbf{x}(\mathbf{y})) - \mathbf{y}\big), \tag{3.7}$$

where $\mathbf{x}(\mathbf{y})$ *is defined in* (3.5). *In addition, the following two inequalities hold:*

$$\beta\big\langle \mathbf{y}_1 - \mathbf{y}_2, \boldsymbol{\theta}(\mathbf{x}(\mathbf{y}_1)) - \boldsymbol{\theta}(\mathbf{x}(\mathbf{y}_2))\big\rangle \leq -\mu\|\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)\|^2 \ \forall\, \mathbf{y}_1,\mathbf{y}_2 \geq \mathbf{0}, \tag{3.8}$$

$$\|\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)\| \leq \tfrac{\beta B_g}{\mu}\|\mathbf{y}_1 - \mathbf{y}_2\| \ \forall\, \mathbf{y}_1,\mathbf{y}_2 \geq \mathbf{0}. \tag{3.9}$$

Proof. The result in (3.7) follows from the Danskin theorem (cf. [5]). We only need to show (3.8) and (3.9).

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 that

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

Adding the above two inequalities gives the result in (3.8). Now, using the B_g-Lipschitz continuity of $\boldsymbol{\theta}$, we have (3.9) from (3.8) and complete the proof. ∎

LEMMA 3.3 (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*

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

Hence, $\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big)$ *is a good approximation of* $\nabla d(\widehat{\mathbf{y}})$ *when* δ *is small.*

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 from (3.7), $\big\|\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big) - \nabla d(\widehat{\mathbf{y}})\big\| = \beta\|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\| \leq \beta B_g\frac{\delta}{\mu}$. This completes the proof. ∎

In order to have a verifiable stopping condition, we will compute the violation of first-order optimality conditions. The following two lemmas quantify the accuracy levels of solving $\widehat{\mathbf{y}} \approx \arg\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$ and $\widehat{\mathbf{x}} \approx \arg\min_{\mathbf{x}} \Phi(\mathbf{x}, \widehat{\mathbf{y}})$ in order to find a desired-accurate stationary point of (3.2). These results will be used to estimate the worst-case complexity result.

LEMMA 3.4. *Given $\widehat{\mathbf{y}} \geq \mathbf{0}$, it holds that*

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

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 and the Cauchy–Schwarz inequality. ∎

LEMMA 3.5. *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. Since $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\bar{\varepsilon}\mu}{3\beta B_g^2}$, we use Lemma 3.3 with $\delta = \frac{\bar{\varepsilon}\mu}{3\beta B_g^2}$ to have $\|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\| \leq \frac{\bar{\varepsilon}}{3\beta B_g}$. In addition, from the nonexpansiveness of $[\cdot]_+$, it follows that $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - [\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+\| \leq \frac{\bar{\varepsilon}}{3\beta B_g}$. Because $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \frac{\bar{\varepsilon}}{3\beta B_g}$, we have from the triangle inequality that $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - \widehat{\mathbf{y}}\| \leq \frac{2\bar{\varepsilon}}{3\beta B_g}$. The desired result now follows from Lemma 3.4 and $\|J_{\mathbf{g}}(\mathbf{x})\| \leq B_g$ for all $\mathbf{x} \in \operatorname{dom}(h)$. ∎

3.2. The case with a single constraint. For simplicity and ease of understanding, we start with the case of $m = 1$, so the bold letters $\mathbf{y}, \boldsymbol{\theta}$ are actually scalars in this subsection. 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 $\bar{\varepsilon} > 0$. By Lemma 3.5, we can first find a $\widehat{\mathbf{y}} \geq \mathbf{0}$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \frac{\bar{\varepsilon}}{3\beta B_g}$ and then approximately solve $\min_{\mathbf{x}} \Phi(\mathbf{x}, \widehat{\mathbf{y}})$ to obtain $\widehat{\mathbf{x}}$.

Our idea of finding a desired approximate solution $\widehat{\mathbf{y}}$ is to first obtain an interval that contains the solution $\bar{\mathbf{y}} = \arg\max_{\mathbf{y}\geq 0} d(\mathbf{y})$ and then to apply a bisection method. The following lemma shows that for a given $\widehat{\mathbf{y}} \geq 0$, we can either check whether it is a desired approximate solution or obtain the sign of $\nabla d(\widehat{\mathbf{y}})$ so that we know the search direction has a desired solution.

LEMMA 3.6. *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 3.3 and the condition on $\widehat{\mathbf{x}}$, it follows that

$$\big|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))\big| \leq \tfrac{\delta}{4} \ \text{ and } \ \big|\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big) - \nabla d(\widehat{\mathbf{y}})\big| \leq \tfrac{\beta\delta}{4}. \tag{3.10}$$

Hence, by the nonexpansiveness of $[\cdot]_+$, it holds that $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - [\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+| \leq \frac{\delta}{4}$. Then, 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 must hold that $|\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}| > \frac{3\delta}{4}$ because $\widehat{\mathbf{y}} \geq 0$, and thus $|\beta(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}})| > \frac{3\beta\delta}{4}$. Therefore, from the second inequality in (3.10), we conclude that $\nabla d(\widehat{\mathbf{y}})$ must have the same sign as $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}$ because otherwise $\big|\beta\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}\big) - \nabla d(\widehat{\mathbf{y}})\big| \geq |\beta(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}})| > \frac{3\beta\delta}{4}$. This completes the proof. ∎

By this lemma, we design an interval search 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 $Y = [a, b] \subseteq [0, \infty)$ 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 (3.3), 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 $\bar{\mathbf{y}}$

Once the stopping condition in line 4 or 10 is satisfied, then by Lemma 3.6 we immediately obtain a desired $\widehat{\mathbf{y}}$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$. The next lemma shows that the algorithm must exit the while loop within finitely many iterations.

LEMMA 3.7. *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 3.1, it follows that $\bar{\mathbf{y}} = [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+ \leq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta}$. The result in (3.8) indicates the decreasing monotonicity of $\boldsymbol{\theta}(\mathbf{x}(\mathbf{y}))$ with respect to $\mathbf{y}$. Hence, if $b \geq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta}$, then $\boldsymbol{\theta}(\mathbf{x}(b)) \leq \boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}})) \leq \frac{2\|\mathbf{z}^*\| + \|\mathbf{z}\|}{\beta} \leq b$, and thus $\boldsymbol{\theta}(\mathbf{x}(b)) - b \leq 0$. Now if $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - b| > \frac{3\delta}{4}$, we know from Lemma 3.6 that $\nabla d(b)\big(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b\big) > 0$, and thus $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b < 0$ since $\nabla d(b) = \beta(\boldsymbol{\theta}(\mathbf{x}(b)) - b) \leq 0$. This completes the proof. ∎

When Algorithm 3 exits the while loop, it can output a single point or an interval. The lemma below shows that if an interval is returned, then it will contain the solution $\bar{\mathbf{y}}$.

LEMMA 3.8. *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 the condition in either line 4 or 10 of Algorithm 3 is satisfied, and we immediately have $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$ from Lemma 3.6.

Now suppose that Y is an interval $[a, b]$. From Lemma 3.6 and the setting in line 8 of Algorithm 3, we always have $\nabla d(a) > 0$. When the algorithm exits the while

loop and returns an interval, we have $\|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - b\| > \frac{3\delta}{4}$ but $\boldsymbol{\theta}(\widehat{\mathbf{x}}) - b \le 0$. Then it follows from Lemma 3.6 that $\nabla d(b) < 0$. Therefore, the unique solution $\bar{\mathbf{y}}$ must lie in (a, b) by the mean-value theorem and the strong concavity of d. ∎

Remark 3.1. Suppose Algorithm 3 returns an interval $[a, b]$. Then Lemma 3.7 indicates that $b \le \frac{1}{\beta}\max\{1, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|\}$, and in addition, at most $T + 2$ calls are made to Algorithm 2, where T is the smallest nonnegative integer such that $2^T \ge 2\|\mathbf{z}^*\| + \|\mathbf{z}\|$.

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.

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

Input: multiplier vector $\mathbf{z} \ge \mathbf{0}$, penalty $\beta > 0$, target accuracy $\delta > 0$, $L_{\min} > 0$, 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 (3.3), 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 $|[\boldsymbol{\theta}(\widehat{\mathbf{x}})]_+ - c| \le \frac{3\delta}{4}$ **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.

By Lemma 3.6 and the lemma below, it holds that the returned point $\widehat{\mathbf{y}}$ from Algorithm 4 must satisfy $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \le \delta$.

LEMMA 3.9. *Let $Y = [a, b] \subseteq (0, \infty)$. If $\nabla d(a) > 0$, $\nabla d(b) < 0$, and $b - a \le \frac{\mu\delta}{\mu + \beta B_g^2}$ for a positive δ, then $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \le \delta$ for any $\widehat{\mathbf{y}} \in [a, b]$.*

Proof. Recall from Lemma 3.1 that $\bar{\mathbf{y}} = [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+$. Hence, for any $\widehat{\mathbf{y}} \in [a, b]$, we have

$$\begin{aligned}
\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| &= \|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}} - [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+ + \bar{\mathbf{y}}\| \\
&\le \|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - [\boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))]_+\| + \|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\| \\
&\le \|\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}})) - \boldsymbol{\theta}(\mathbf{x}(\bar{\mathbf{y}}))\| + \|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\| \\
&\le B_g\|\mathbf{x}(\widehat{\mathbf{y}}) - \mathbf{x}(\bar{\mathbf{y}})\| + \|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\| \\
&\le \tfrac{\beta B_g^2}{\mu}\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\| + \|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\|, \qquad (3.11)
\end{aligned}$$

where we have used the nonexpansiveness of $[\cdot]_+$ in the second inequality, the third inequality follows from (3.1), and the last inequality holds because of (3.9). Now, since $\bar{\mathbf{y}} \in [a, b]$, we have $\|\widehat{\mathbf{y}} - \bar{\mathbf{y}}\| \le b - a \le \frac{\mu\delta}{\mu + \beta B_g^2}$, and hence the desired result follows. ∎

Remark 3.2. 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 3.1. Hence, after Y is obtained, Algorithm 4 will call Algorithm 2 at most $\left\lceil \log_2 \frac{\max\{1,\ 4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\}(\mu+\beta B_g^2)}{\beta\mu\delta} \right\rceil_+ + 1$ times.

Below we establish the complexity result of Algorithm 4 to return $\widehat{\mathbf{y}}$.

THEOREM 3.10 (iteration complexity of BiSec). *Under Assumptions* 1–4*, Algorithm* 4 *needs at most* T *evaluations on* f*,* $\boldsymbol{\theta}$*,* ∇f*, and* $J_{\boldsymbol{\theta}}$ *to output* $\widehat{\mathbf{x}}$ *and* $\widehat{\mathbf{y}} \geq 0$ *that satisfy* $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \bar{\varepsilon}$ *and* $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$*, where* $\bar{\varepsilon} = \frac{\mu\delta}{4B_g}$*, and*

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

with $L_{\mathbf{z}} = L_f + L_g \max\{1, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|\}$ *and*

$$K = 3 + \lceil \log_2(2\|\mathbf{z}^*\| + \|\mathbf{z}\|) \rceil_+ + \left\lceil \log_2 \frac{\max\left\{1,\ 4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\right\}(\mu+\beta B_g^2)}{\beta\mu\delta} \right\rceil_+ . \tag{3.12}$$

Proof. By Remarks 3.1 and 3.2, Algorithm 4 calls Algorithm 2 at most K times, where K is given in (3.12). 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 3.1, we apply Corollary 2.3 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 $\bar{\mathbf{y}} = \arg\max_{\mathbf{y}\geq\mathbf{0}} d(\mathbf{y})$.

LEMMA 3.11. *Let* $b > 0$*, and suppose* $\|\bar{\mathbf{y}}\| \leq b$*. Given* $\delta > 0$ *and* $\widehat{\mathbf{y}} \geq \mathbf{0}$*, let* $\widehat{\mathbf{x}} \in \operatorname{dom}(h)$ *be a point satisfying* $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$*. 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* $\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}}\rangle > 0$ *for any* $\mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+$*, where* $\eta = \min\{b, \eta_+\}$*, and* η_+ *is the positive root of the equation*

$$\frac{\mu+\beta B_g^2}{\mu}\left(\eta + \sqrt{\frac{2\eta B_d}{\beta}}\right) = \frac{\delta}{4}, \quad \text{with} \quad B_d = \max_{\mathbf{y}\in\mathcal{B}_b^+} \nabla d(\mathbf{y}). \tag{3.13}$$

Proof. By the same arguments in the proof of Lemma 3.6, 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 that $\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}}\rangle > 0$ 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:

$$\langle \boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}, \mathbf{y} - \widehat{\mathbf{y}}\rangle \leq 0 \text{ for some } \mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+. \tag{3.14}$$

By the β-strong concavity of d, it holds that

$$d(\mathbf{y}) \leq d(\widehat{\mathbf{y}}) + \langle \nabla d(\widehat{\mathbf{y}}), \mathbf{y} - \widehat{\mathbf{y}}\rangle - \tfrac{\beta}{2}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2. \tag{3.15}$$

From the mean-value theorem, it follows that there is $\widetilde{\mathbf{y}}$ between $\mathbf{y}$ and $\bar{\mathbf{y}}$ such that $d(\mathbf{y}) - d(\bar{\mathbf{y}}) = \langle \nabla d(\widetilde{\mathbf{y}}), \mathbf{y} - \bar{\mathbf{y}}\rangle \geq -\eta B_d$, where the inequality holds because $\mathbf{y} \in \mathcal{B}_\eta(\bar{\mathbf{y}})$

and $\widetilde{\mathbf{y}}$ must fall in $\mathcal{B}_b^+$. Since $d(\bar{\mathbf{y}}) \geq d(\widehat{\mathbf{y}})$, we have $d(\widehat{\mathbf{y}}) - d(\mathbf{y}) \leq d(\bar{\mathbf{y}}) - d(\mathbf{y}) \leq \eta B_d$. Hence, (3.14) and (3.15) imply

$$\tfrac{\beta}{2}\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 \leq \eta B_d + \langle \beta(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}) - \nabla d(\widehat{\mathbf{y}}), \widehat{\mathbf{y}} - \mathbf{y}\rangle. \tag{3.16}$$

From Lemma 3.3 and the condition $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \leq \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}$, it follows that $\|\beta(\boldsymbol{\theta}(\widehat{\mathbf{x}}) - \widehat{\mathbf{y}}) - \nabla d(\widehat{\mathbf{y}})\| \leq \frac{\beta\mu\delta}{8(\mu+\beta B_g^2)}$, which together with (3.16) and the Cauchy–Schwarz inequality gives

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

Solving the above inequality, we have $\|\mathbf{y}-\widehat{\mathbf{y}}\| \leq \sqrt{\frac{2\eta B_d}{\beta}} + \frac{\mu\delta}{4(\mu+\beta B_g^2)}$, and since $\|\mathbf{y}-\bar{\mathbf{y}}\| \leq \eta$, it holds that $\|\bar{\mathbf{y}} - \widehat{\mathbf{y}}\| \leq \eta + \sqrt{\frac{2\eta B_d}{\beta}} + \frac{\mu\delta}{4(\mu+\beta B_g^2)}$. Now, noting that (3.11) also holds for the case of $m > 1$ as its proof does not rely on $m = 1$, we have

$$\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \leq \tfrac{\mu+\beta B_g^2}{\mu}\left(\eta + \sqrt{\tfrac{2\eta B_d}{\beta}} + \tfrac{\mu\delta}{4(\mu+\beta B_g^2)}\right) = \tfrac{\mu+\beta B_g^2}{\mu}\left(\eta + \sqrt{\tfrac{2\eta B_d}{\beta}}\right) + \tfrac{\delta}{4} \leq \tfrac{\delta}{2}, \tag{3.17}$$

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 that $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| > \frac{\delta}{2}$, and (3.17) contradicts this fact. Therefore, the assumption in (3.14) cannot hold. This completes the proof. ∎

Suppose $\|\bar{\mathbf{y}}\| \leq b$ for some $b > 0$. For a given $\widehat{\mathbf{y}} \geq \mathbf{0}$, let $\widehat{\mathbf{x}}$ satisfy the condition required in Lemma 3.11. 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 $\eta \leq b$. Therefore, we can apply a cutting-plane method to find a near-optimal $\widehat{\mathbf{y}}$. In order to have a good scalability to m, we choose the volumetric-center cutting-plane (VCCP) method [1, 37]. Below we first give the more efficient version of VCCP in [1] and then adapt it to solve our problem.

Volumetric-center cutting-plane (VCCP) method. Let $\mathcal{C}$ be a convex set in $\mathbb{R}^m$. Suppose that there is a *separation oracle*. Given a point $\tilde{\mathbf{y}} \in \mathbb{R}^m$, the separation oracle can either tell $\tilde{\mathbf{y}} \in \mathcal{C}$ or return one vector $\mathbf{a}$ such that $\mathbf{a}^\top\mathbf{y} > \mathbf{a}^\top\tilde{\mathbf{y}}$ for all $\mathbf{y} \in \mathcal{C}$. By using the oracle, VCCP aims to solve the feasibility problem: find a point $\mathbf{y} \in \mathcal{C}$ or show that the volume of $\mathcal{C}$ is less than a given positive number ρ.

Let $\mathcal{P} = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{A}\mathbf{y} \geq \mathbf{b}\}$ be a polytope with nonempty interior. For each interior point $\mathbf{y}$ in $\mathcal{P}$, i.e., $\mathbf{A}\mathbf{y} - \mathbf{b} > \mathbf{0}$, the volumetric barrier function is defined as

$$V(\mathbf{y}) = \tfrac{1}{2}\log\left(\det\left(\mathbf{A}^\top\mathbf{S}(\mathbf{y})^{-2}\mathbf{A}\right)\right), \text{ with } \mathbf{S}(\mathbf{y}) = \operatorname{Diag}(\mathbf{A}\mathbf{y} - \mathbf{b}), \tag{3.18}$$

where $\det(\cdot)$ denotes the determinant. The minimizer of $V(\cdot)$ is called the volumetric center (VC) of $\mathcal{P}$. Let

$$\mathbf{Q}(\mathbf{y}) = \mathbf{A}^\top\mathbf{S}(\mathbf{y})^{-2}\operatorname{Diag}(\mathbf{p}(\mathbf{y}))\mathbf{A}, \text{ with } \mathbf{p}(\mathbf{y}) = \operatorname{diag}\left(\mathbf{S}(\mathbf{y})^{-1}\mathbf{A}(\mathbf{A}^\top\mathbf{S}(\mathbf{y})^{-2}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{S}(\mathbf{y})^{-1}\right). \tag{3.19}$$

With these notations, the pseudocode of the VCCP is given in Algorithm 5, where we define $\mathbf{S}^k = \mathbf{S}(\mathbf{y}^k)$, $\mathbf{Q}^k = \mathbf{Q}(\mathbf{y}^k)$, $V^k(\mathbf{y}^k) = V(\mathbf{y}^k)$, $\mathbf{p}^k = \mathbf{p}(\mathbf{y}^k)$, and $p^k_{\min} = \min_{1\leq i\leq m} p_i^k$ by using (3.18) and (3.19) for $\mathcal{P} = \mathcal{P}^k$.

The lemma below is obtained from Lemma 3.1 in [1] and its proof.

Algorithm 5: Volumetric-center cutting-plane (VCCP) method.

Initialization: choose a polytope $\mathcal{P}^0 = \{\mathbf{y} : \mathbf{A}^0\mathbf{y} \geq \mathbf{b}^0\}$ that has a VC $\mathbf{y}^0$ in the interior of $\mathcal{P}^0$, choose $p_{\min} \in (0,1)$, $\tau > 0$, and $0 < c_1 \leq c_2$; set $k = 0$.
while $V^k(\mathbf{y}^k) < V^k_{\max}$ **do**
  if $p^k_{\min} \geq p_{\min}$ **then**
    Call the separation oracle to check whether $\mathbf{y}^k \in \mathcal{C}$. If so, return $\mathbf{y}^k$ and stop. Otherwise, obtain $\tilde{\mathbf{a}}$ from the oracle such that $\tilde{\mathbf{a}}^\top\mathbf{y} > \tilde{\mathbf{a}}^\top\mathbf{y}^k \ \forall \mathbf{y} \in \mathcal{C}$. Let $\mathbf{A}^{k+1} = [\mathbf{A}^k; \tilde{\mathbf{a}}^\top]$ and $\mathbf{b}^{k+1} = [\mathbf{b}^k; \tilde{b}]$ with

$$\tilde{b} = \tilde{\mathbf{a}}^\top\mathbf{y}^k - \tfrac{1}{\sqrt{\tau}}\sqrt{\tilde{\mathbf{a}}^\top\left((\mathbf{A}^k)^\top(\mathbf{S}^k)^{-2}\mathbf{A}^k\right)^{-1}\tilde{\mathbf{a}}}; \tag{3.20}$$

  else
    Suppose $p^k_j = p^k_{\min}$. Let $[\mathbf{A}^{k+1}, \mathbf{b}^{k+1}]$ be obtained by removing the jth row from $[\mathbf{A}^k, \mathbf{b}^k]$;
  Let $\mathcal{P}^{k+1} = \{\mathbf{y} : \mathbf{A}^{k+1}\mathbf{y} \geq \mathbf{b}^{k+1}\}$; start from $\mathbf{y}^k$ and apply a sequence of pure Newton steps to find $\mathbf{y}^{k+1}$ as an approximate VC of $\mathcal{P}^{k+1}$ such that

$$\|(\mathbf{Q}^{k+1})^{-1}\nabla V^{k+1}(\mathbf{y}^{k+1})\|_{\mathbf{Q}^{k+1}} \leq \min\left\{c_1, (2\sqrt{p^{k+1}_{\min}} - p^{k+1}_{\min})^{\frac{1}{2}}c_2\right\}; \tag{3.21}$$

  set $k \leftarrow k+1$.

LEMMA 3.12. *Suppose $\mathcal{C} \subseteq \mathcal{P}^0$ and $c_1 \leq c_2 \leq 0.03$. Let $V^k_{\max} = \log\frac{V_m(1)}{\rho} + m\log(n_k) + 0.00135$, where n_k is the number of rows of $\mathbf{A}^k$ and $V_m(1)$ is the volume of a unit ball in $\mathbb{R}^m$. If Algorithm 5 terminates because $V^k(\mathbf{y}^)) \geq V^k_{\max}$ for some k, then the volume of $\mathcal{C}$ is smaller than ρ.*

Also we have the following theorem from [1].

THEOREM 3.13. *Suppose that $\mathbf{A}^0$ has $2m$ rows. Let $p_{\min} = 0.005, \tau = 0.007, c_1 = 0.0001, c_2 = 0.00027$, and $V^k_{\max} = \log\frac{V_m(1)}{\rho} + m\log(n_k) + 0.00135$ in Algorithm 5 with $\rho \in (0, V_m(1))$. Then at most five Newton steps are needed to ensure the condition in (3.21). In addition, Algorithm 5 must terminate in*

$$\left\lceil \Gamma\left(m\log m + \log\frac{V_m(1)}{\rho} + 6m - V^0(\mathbf{y}^0)\right) + 16m + 1\right\rceil$$

calls to the separation oracle, where $\Gamma \leq 5406$ is a universal constant.

Proof. From (3.8) to (3.9) in the proof of [1, Theorem 3.2], we have that $V^k(\mathbf{y}^k) \geq V^k_{\max}$ occurs if

$$V^0(\mathbf{y}^0) + \tfrac{k}{2}\Delta V - \tfrac{m}{2}(\Delta V^+ + \Delta V^-) \geq \log\tfrac{V_m(1)}{\rho} + m\log(1 + \tfrac{1}{p_{\min}}) + m\log m + 0.00135, \tag{3.22}$$

where $\Delta V^+ = 0.00301$, $\Delta V^- = 0.00264$, and $\Delta V = \Delta V^+ - \Delta V^- = 0.00037$ by Theorems 6.4 and 6.5 and Corollary 6.6 in [1]. We complete the proof by solving (3.22) for k and noting that $\log(1 + \frac{1}{p_{\min}}) \leq 6$. ∎

Remark 3.3. From the proof of Theorem 6.4 in [1], if each $\mathbf{y}^k$ is the VC of $\mathcal{P}^k$, then $\Delta V^+ = \frac{1}{2}\log(1+\tau)$ and $\Delta V^- = \frac{1}{2}\log(1-p_{\min})$. In this case, the constant Γ can be significantly reduced by increasing τ. For example, let $\tau = 2$ and $p_{\min} = 0.005$. Then $\Delta V^+ < 0.5494$, $\Delta V^- < 0.0027$, and $\Delta V > 0.5466$. To have (3.22), it suffices to let $k \geq 3.66\big(m\log m + \log\frac{V_m(1)}{\rho} + 6m - V^0(\mathbf{y}^0)\big) + 2m + 1$. Notice that if $\tau = \infty$ in (3.20), the generated cut $(\tilde{\mathbf{a}}, \tilde{b})$ will pass through $\mathbf{y}^k$. Roughly speaking, a larger

τ gives a deeper cut and reduces the constant Γ in Theorem 3.13, but more Newton iterations will be needed to find a sufficiently accurate VC.

From Lemma 3.12 and Theorem 3.13, we conclude that if $\mathcal{C} \subseteq \mathcal{P}^0$ and the volume of $\mathcal{C}$ is no smaller than ρ, then Algorithm 5 must be able to find a point $\tilde{\mathbf{y}} \in \mathcal{C}$. The proof of the above theorem is essentially by the logic that $V^k(\mathbf{y}^k) \geq V^k_{\max}$ will eventually occur if a point in $\mathcal{C}$ is never found. Below we exploit this idea and adapt the VCCP method to solve our problem in Algorithm 6, where n_k denotes the number of rows of $\mathbf{A}^k$ for each $k \geq 0$. Notice that from Lemma 3.11, if $\|\bar{\mathbf{y}}\| \leq b$ and $\mathcal{C} := \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+ \subseteq \mathcal{P}^0$, then the cut $(\tilde{\mathbf{a}}, \tilde{b})$ generated from line 18 satisfies $\tilde{\mathbf{a}}^\top \mathbf{y} > \tilde{b}$ for all $\mathbf{y} \in \mathcal{C}$ and thus $\mathcal{C} \subseteq \mathcal{P}^k$ for all $k \geq 0$. The checking in lines 7 and 9 ensures that the subproblem solved in line 12 will be strongly convex and have a bounded smoothness constant. Also notice that different from what we do in Algorithm 5, we fix $p_{\min} = 0.005$, $c_1 = 0.0001$, $c_2 = 0.00027$ but only leave τ to be tuned in Algorithm 6.

Algorithm 6: VCCP method for $\max_{\mathbf{y} \geq \mathbf{0}} d(\mathbf{y})$:
$(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \text{FLAG}) = \text{VCCP}(\beta, \mathbf{z}, \delta, b, L_{\min}, \gamma_1, \gamma_2)$.

1 **Input:** multiplier vector $\mathbf{z} \geq \mathbf{0}$, penalty $\beta > 0$, target accuracy $\delta > 0$, $b > 0$, $L_{\min} > 0$, 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 (3.3), $\bar{\varepsilon} = \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$, and FLAG $= 0$.
3 Let η_+ be the positive root of (3.13), $\eta \leftarrow \min\{b, \eta_+\}$, and $\rho = 4^{-m} V_m(\eta)$; set $k = 0$.
4 Set $\mathbf{A}^0 = [\mathbf{I}; -\mathbf{I}], \mathbf{b}^0 = [\mathbf{0}_m; -b\mathbf{1}_n]$; let $\mathcal{P}^0 = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{A}^0\mathbf{y} \geq \mathbf{b}^0\}$, $\mathbf{y}^0 = \frac{b}{2}\mathbf{1}$; choose $\tau \geq 0.007$
5 **while** $V^k(\mathbf{y}^k) < V^k_{\max} := \log \frac{V_m(1)}{\rho} + m \log(n_k) + 0.00135$ **do**
6 **if** $p^k_{\min} \geq 0.005$ **then**
7 **if** $\mathbf{y}^k \not\geq \mathbf{0}$ **then**
8 Let $\tilde{\mathbf{a}} = \mathbf{e}_{i_0}$, where $i_0 = \arg\min_{i \in [m]} y^k_i$ ▷ to ensure a check point in $\mathbb{R}^m_+$
9 **else if** $\|\mathbf{y}^k\| > b$ **then**
10 Let $\tilde{\mathbf{a}} = -\mathbf{y}^k$ ▷ to ensure a check point in $\mathcal{B}_b$
11 **else**
12 Call Alg. 2: $\mathbf{x}^k = \text{APG}(\psi, h, \mu, L_{\min}, \bar{\varepsilon}, \gamma_1, \gamma_2)$ with $\psi = \Phi(\cdot, \mathbf{y}^k) - h$
13 **if** $\|[\boldsymbol{\theta}(\mathbf{x}^k)]_+ - \mathbf{y}^k\| \leq \frac{3\delta}{4}$ **then**
14 Let $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = (\mathbf{x}^k, \mathbf{y}^k)$ and FLAG $= 1$;
15 Return $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \text{FLAG})$, and stop ▷ found $\widehat{\mathbf{y}}$ such that $|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}| \leq \delta$
16 **else**
17 Let $\tilde{\mathbf{a}} = \boldsymbol{\theta}(\mathbf{x}^k) - \mathbf{y}^k$
18 Let $\mathbf{A}^{k+1} = [\mathbf{A}^k; \tilde{\mathbf{a}}^\top]$ and $\mathbf{b}^{k+1} = [\mathbf{b}^k; \tilde{b}]$ with $\tilde{b}$ given by (3.20)
19 **else**
20 Suppose $p^k_j = p^k_{\min}$. Let $[\mathbf{A}^{k+1}, \mathbf{b}^{k+1}]$ be obtained by removing the jth row from $[\mathbf{A}^k, \mathbf{b}^k]$
21 Let $\mathcal{P}^{k+1} = \{\mathbf{y} : \mathbf{A}^{k+1}\mathbf{y} \geq \mathbf{b}^{k+1}\}$; start from $\mathbf{y}^k$ and apply a sequence of pure Newton steps to find $\mathbf{y}^{k+1}$ as an approximate VC of $\mathcal{P}^{k+1}$ such that (3.21) holds with $c_1 = 0.0001$ and $c_2 = 0.00027$.
22 Increase $k \leftarrow k + 1$.
23 Let $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = (\mathbf{x}^k, \mathbf{y}^k)$ and return $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \text{FLAG})$

Similar to Theorem 3.13, we are able to show the finite convergence of Algorithm 6.

THEOREM 3.14. *Under Assumptions* 1–4, *Algorithm* 6 *with* $\tau = 0.007$ *will stop within* N *iterations, where* $N = \lceil \Gamma\big(m\log m + m\log\frac{\sqrt{2}b}{\eta} + 6m\big) + 16m + 1\rceil$, η *is defined in line* 3 *of the algorithm, and* $\Gamma \le 5406$ *is a universal constant. In addition, if* $\|\bar{\mathbf{y}}\| \le b$, *Algorithm* 6 *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*

(3.23)
$$T = N\left(1 + \lceil \log_{\gamma_1} \tfrac{L_\psi}{L_{\min}} \rceil_+\right)\left(1 + 2\left\lceil 2\sqrt{\tfrac{\gamma_1 L_\psi}{\mu}} \log\left(\tfrac{D_h}{\bar{\varepsilon}}\left(\sqrt{\gamma_1 L_\psi} + \tfrac{L_\psi}{\sqrt{L_{\min}}}\right)\sqrt{2\gamma_1 L_\psi + \mu}\right)\right\rceil_+\right),$$

with $L_\psi := L_f + \beta b L_g$, *and* $\bar{\varepsilon} = \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8(\mu B_g + \beta B_g^3)}\}$.

Proof. First, notice that $\mathbf{y}^0$ is the VC of $\mathcal{P}^0$. Second, it is straightforward to compute $V^0(\mathbf{y}^0) = m\log\frac{2\sqrt{2}}{b}$ and $\log\frac{V_m(1)}{\rho} = m\log\frac{4}{\eta}$. Hence, from the proof of Theorem 3.13, $V^k(\mathbf{y}^k) \ge V^k_{\max}$ must occur if $k \ge \lceil \Gamma\big(m\log m + m\log\frac{\sqrt{2}b}{\eta} + 6m\big) + 16m + 1\rceil$, where $\Gamma \le 5406$ is a universal constant.

It is obvious that $\mathcal{C} := \mathcal{B}_\eta(\bar{\mathbf{y}}) \cap \mathcal{B}_b^+ \subseteq \mathcal{P}^0$ by the choice of $\mathcal{P}^0$. Below we argue that $\mathcal{C} \subseteq \mathcal{P}^k$ for all $k \ge 0$ before the algorithm stops. First, if $\mathcal{P}^{k+1}$ is obtained by deleting one row from the system of $\mathcal{P}^k$, then $\mathcal{P}^k \subseteq \mathcal{P}^{k+1}$; second, if $\tilde{\mathbf{a}}$ is obtained from line 8 or line 10 of Algorithm 6, the generated cut $(\tilde{\mathbf{a}}, \tilde{b})$ will not cut any point from $\mathcal{C}$; third, if $\tilde{\mathbf{a}}$ is obtained from line 17, by Lemma 3.11, the generated cut $(\tilde{\mathbf{a}}, \tilde{b})$ will not cut any point from $\mathcal{C}$ either. Therefore, if $\mathcal{C} \subseteq \mathcal{P}^k$, then $\mathcal{C} \subseteq \mathcal{P}^{k+1}$, and thus by induction $\mathcal{C} \subseteq \mathcal{P}^k$ for all $k \ge 0$. Now, since the volume of $\mathcal{C}$ is no smaller than ρ, we conclude that there must be a point $\mathbf{x}^k$ from line 12 of Algorithm 6 such that the condition in line 13 is satisfied. Hence, the algorithm will return FLAG $= 1$ and a vector $\widehat{\mathbf{y}} \ge \mathbf{0}$ satisfying $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \le \delta$ by Lemma 3.11.

Finally, notice that when Algorithm 2 is called in line 12, $\|\mathbf{y}^k\| \le b$, and thus the smooth function ψ has an $(L_f + \beta L_g b)$-Lipschitz continuous gradient. Since Algorithm 2 is called at most N times, we have from Corollary 2.3 that the total number of function and gradient evaluations is T given in (3.23). ∎

As discussed in Remark 3.3, the constant Γ can be reduced to 3.66 if $\tau = 2$ is used, like in our numerical experiments. By Theorem 3.14, we can guarantee finding a desired approximate solution $\widehat{\mathbf{y}}$ by gradually increasing the search radius b. The algorithm is shown below.

Algorithm 7: Search by the VCCP method for $\max_{\mathbf{y}\ge\mathbf{0}} d(\mathbf{y})$:
$(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = \text{SVCCP}(\beta, \mathbf{z}, \delta, L_{\min}, \gamma_1, \gamma_2)$.

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 (3.3), and set $k = 0$, $b_0 = \frac{1}{\beta}$ and FLAG $= 0$.
3 **while** FLAG $= 0$ **do**
4 Call Alg. 6: $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \text{FLAG}) = \text{VCCP}(\beta, \mathbf{z}, \delta, b_k, L_{\min}, \gamma_1, \gamma_2)$.
5 Let $b_{k+1} \leftarrow 2b_k$ and increase $k \leftarrow k + 1$.
6 **Output** $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$.

THEOREM 3.15. *Under Assumptions* 1–4, *if* $\delta \le \frac{8(\mu+\beta B_g^2)}{\beta\mu}$, *then the output* $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ *of Algorithm* 7 *must satisfy* $\text{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})\big) \le \bar{\varepsilon}$, $\widehat{\mathbf{y}} \ge \mathbf{0}$, *and* $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \le \delta$, *where* $\bar{\varepsilon} = \min\{\frac{\mu\delta}{4B_g}, \frac{\mu^2\delta}{8B_g(\mu+\beta B_g^2)}\}$. *In addition, it needs at most* T *evaluations of* f,

∇f, $\boldsymbol{\theta}$, *and* $J_{\boldsymbol{\theta}}$ *to give the output, where*

(3.24)
$$T \le 3CK + 4C\sqrt{\gamma_1}\log\left(\frac{D_h}{\bar{\varepsilon}}\left(\sqrt{\gamma_1 L_{\max}} + \frac{L_{\max}}{\sqrt{L_{\min}}}\right)\sqrt{2\gamma_1 L_{\max}+\mu}\right)\left(K\sqrt{\frac{L_f}{\mu}} + \frac{\sqrt{L_g}\max\left\{1, \frac{2\sqrt{2}\|\mathbf{z}^*\|+\|\mathbf{z}\|}{\sqrt{2}-1}\right\}}{\sqrt{\mu}}\right),$$

with the constants defined as

$$\begin{aligned}
L_{\max} &= L_f + L_g(4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|),\\
C &= \lceil \Gamma(m\log m + m\log R + 6m) + 16m + 1\rceil \cdot \left(1 + \lceil \log_{\gamma_1}\tfrac{L_{\max}}{L_{\min}}\rceil_+\right),\\
R &= \frac{8\sqrt{2}(\max\{1,\, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|\})}{\beta}\left(\frac{4(\beta G + \|\mathbf{z}\| + \max\{1,\, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}\|\})(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \frac{\mu+\beta B_g^2}{\mu\delta}\right),\\
K &= \lceil \log_2(2\|\mathbf{z}^*\| + \|\mathbf{z}\|)\rceil_+ + 1,
\end{aligned}$$

and $\Gamma \le 5406$ *is a universal constant.*

Proof. By the quadratic formula, we can easily have the positive root of (3.13) to be

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

Hence, it holds that

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

When $b \ge \frac{1}{\beta}$, the right-hand side of the above inequality is greater than one by the assumption $\delta \le \frac{8(\mu+\beta B_g^2)}{\beta\mu}$, and since $\eta = \min\{\eta_+, b\}$ in Algorithm 6, we have

(3.25)
$$\tfrac{b}{\eta} = \max\{\tfrac{b}{\eta_+}, 1\} \le 8b\left(\frac{4B_d(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \frac{\mu+\beta B_g^2}{\mu\delta}\right) \le 8b\left(\frac{4(\beta G+\|\mathbf{z}\|+\beta b)(\mu+\beta B_g^2)^2}{\beta(\mu\delta)^2} + \frac{\mu+\beta B_g^2}{\mu\delta}\right),$$

where we have used $\nabla d(\mathbf{y}) = \beta(\mathbf{g}(\mathbf{x}(\mathbf{y})) + \frac{\mathbf{z}}{\beta} - \mathbf{y})$ in (3.7), and thus the bound of $\nabla d(\mathbf{y})$ over $\mathcal{B}_b^+$ satisfies $B_d \le \beta G + \|\mathbf{z}\| + \beta b$ with G defined in (1.3).

Furthermore, by Lemma 3.1 and Theorem 3.14, Algorithm 6 must return FLAG = 1 and a vector $\widehat{\mathbf{y}}$ satisfying $\|[\boldsymbol{\theta}(\mathbf{x}(\widehat{\mathbf{y}}))]_+ - \widehat{\mathbf{y}}\| \le \delta$ when $b \ge \frac{2\|\mathbf{z}^*\|+\|\mathbf{z}\|}{\beta}$. Since $b_0 = \frac{1}{\beta}$ and $b_{k+1} = 2b_k$, Algorithm 7 must stop after making at most K calls to Algorithm 6, where K is the smallest positive integer such that $2^{K-1} \ge 2\|\mathbf{z}^*\| + \|\mathbf{z}\|$, i.e., $K = \lceil \log_2(2\|\mathbf{z}^*\| + \|\mathbf{z}\|)\rceil_+ + 1$. In addition, from $b_{k+1} = 2b_k$, it holds that

$$b_k = \tfrac{2^k}{\beta} \le \tfrac{\max\{1, 4\|\mathbf{z}^*\|+2\|\mathbf{z}\|\}}{\beta} \text{ for each } 0 \le k \le K-1. \tag{3.26}$$

In the kth call to Algorithm 6, let η_k denote the η used in line 3 of Algorithm 6, $L_{\psi_k} = L_f + \beta L_g b_k$ the gradient Lipschitz constant of the smooth function ψ, and T_k the total number of gradient and function evaluations. Then, by (3.26) and the definition of $L_{\max}$, we have $L_{\psi_k} \le L_{\max}$. Also, from (3.25), (3.26), and the definition of R, it follows that $\frac{\sqrt{2}b_k}{\eta_k} \le R$ for each $0 \le k \le K-1$. Moreover, we have from (3.23) that

$$\begin{aligned}
T_k &\le C\left(1 + 2\left\lceil 2\sqrt{\tfrac{\gamma_1 L_{\psi_k}}{\mu}}\log\left(\frac{D_h}{\bar{\varepsilon}}\left(\sqrt{\gamma_1 L_{\psi_k}} + \frac{L_{\psi_k}}{\sqrt{L_{\min}}}\right)\sqrt{2\gamma_1 L_{\psi_k}+\mu}\right)\right\rceil_+\right)\\
&\le 3C + 4C\sqrt{\tfrac{\gamma_1 L_{\psi_k}}{\mu}}\log\left(\frac{D_h}{\bar{\varepsilon}}\left(\sqrt{\gamma_1 L_{\max}} + \frac{L_{\max}}{\sqrt{L_{\min}}}\right)\sqrt{2\gamma_1 L_{\max}+\mu}\right).
\end{aligned}$$

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

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

Therefore, T must satisfy the condition in (3.24) since $T \le \sum_{k=0}^{K-1} T_k$. ∎

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 8. Notice that for each k we aim to find $\mathbf{x}^{k+1}$ such that $\operatorname{dist}\big(\mathbf{0}, \partial\phi_k(\mathbf{x}^{k+1})\big) \le \varepsilon_k$, where ϕ_k is defined in (4.3) as the objective of the kth ALM subproblem. In line 10, in case μ is big or β_k is small, we call the APG in order to ensure this by Lemma 3.5.

Algorithm 8: Cutting-plane first-order iALM for problems in the form of (1.1) with $m = O(1)$.

1 **Input:** $\beta_0 > 0$, $\sigma > 1$, tolerance $\varepsilon > 0$, $L_{\min} > 0$, $\gamma_1 > 1$, and $\gamma_2 \ge 1$
2 **Initialization:** choose $\mathbf{x}^0 \in \operatorname{dom}(h)$, and set $\mathbf{z}^0 = \mathbf{0}$
3 **for** $k = 0, 1, \ldots$ **do**
4 Choose $\varepsilon_k \le \min\big\{\varepsilon, \frac{24B_g(\mu+\beta_k B_g^2)}{\mu}\big\}$ and set $\delta_k = \frac{\varepsilon_k}{3\beta_k B_g}$.
5 **if** $m = 1$ **then**
6 Call Alg. 4: $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) = \mathrm{BiSec}(\beta_k, \mathbf{z}^k, \delta_k, L_{\min}, \gamma_1, \gamma_2)$
7 **else**
8 Call Alg. 7: $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) = \mathrm{SVCCP}(\beta_k, \mathbf{z}^k, \delta_k, L_{\min}, \gamma_1, \gamma_2)$
9 **if** $m = 1$ *and* $\frac{\mu}{4\beta_k B_g^2} > 1$, *or* $m > 1$ *and* $\min\Big\{\frac{\mu}{4\beta_k B_g^2}, \frac{\mu^2}{8\beta_k B_g^2(\mu+\beta_k B_g^2)}\Big\} > 1$ **then**
10 Call Alg. 2: $\mathbf{x}^{k+1} = \mathrm{APG}(\psi, h, \mu, L_{\min}, \varepsilon_k/3, \gamma_1, \gamma_2)$ with $\psi(\mathbf{x}) = f(\mathbf{x}) + \beta_k\langle \mathbf{y}^{k+1}, \mathbf{g}(\mathbf{x})\rangle$.
11 Update $\mathbf{z}$ by $\mathbf{z}^{k+1} = [\mathbf{z}^k + \beta_k \mathbf{g}(\mathbf{x}^{k+1})]_+$.
12 Let $\beta_{k+1} \leftarrow \sigma\beta_k$.
13 **if** $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ *is an* ε*-KKT point of* (1.1) **then**
14 Output $(\bar{\mathbf{x}}, \bar{\mathbf{z}}) = (\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ and stop

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

THEOREM 4.1. *Let Assumptions* 1–4 *hold,* $(\beta_0, \sigma, \varepsilon, \gamma_1, \gamma_2)$ *be the input of Algorithm* 8*, and* $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\ge 0}$ *be the generated sequence. Then* $\operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)) \le \varepsilon_k$ *for each* $k \ge 0$*. Suppose* $\bar{\varepsilon} = \min\{\varepsilon, \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\} \le \big\{\varepsilon, \frac{24B_g(\mu+\beta_k B_g^2)}{\mu}\big\}$ *for all* $k \ge 0$*. Let* $\varepsilon_k = \bar{\varepsilon}$ *for all* $k \ge 0$*. Then, after at most* $K-1$ *iterations, Algorithm* 8 *will produce an* ε*-KKT point of* (1.1)*, where*

$$K = \max\left\{\left\lceil \log_\sigma \tfrac{9\|\mathbf{z}^*\|^2}{\beta_0\varepsilon}\right\rceil_+, \left\lceil \log_\sigma \tfrac{8\|\mathbf{z}^*\|}{\beta_0\varepsilon}\right\rceil_+, \left\lceil \log_\sigma \tfrac{4}{\beta_0\varepsilon}\right\rceil_+\right\} + 1. \tag{4.1}$$

In addition, the output multiplier vector $\bar{\mathbf{z}}$ *satisfies*

$$\|\bar{\mathbf{z}}\| \le 2\|\mathbf{z}^*\| + \sqrt{\tfrac{2\sigma^2}{8\sigma+1}}\max\big\{3\|\mathbf{z}^*\|, 2\sqrt{2\|\mathbf{z}^*\|}, 2\big\}. \tag{4.2}$$

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

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

$$\Phi_k(\mathbf{x},\mathbf{y}) = F(\mathbf{x}) + \beta_k\left(\mathbf{y}^\top\boldsymbol{\theta}_k(\mathbf{x}) - \frac{1}{2}\|\mathbf{y}\|^2\right). \tag{4.3}$$

When $m = 1$, if $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$ is obtained in line 6 of Algorithm 8, then we have from Theorem 3.10 that

$$\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi_k(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})\big) \leq \tfrac{\mu\delta_k}{4B_g} \text{ and } \left|[\boldsymbol{\theta}_k(\mathbf{x}(\mathbf{y}^{k+1}))]_+ - \mathbf{y}^{k+1}\right| \leq \delta_k,$$

where $\mathbf{x}(\mathbf{y}^{k+1}) = \arg\min_{\mathbf{x}} \Phi_k(\mathbf{x}, \mathbf{y}^{k+1})$. Furthermore, note that if $\frac{\mu}{4\beta_k B_g^2} > 1$, we will do line 10 in Algorithm 8 to get a new $\mathbf{x}^{k+1}$ satisfying $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\Phi_k(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})\big) \leq \frac{\varepsilon_k}{3}$. Now, by Lemma 3.5 and the choice of $\delta_k = \frac{\varepsilon_k}{3\beta_k B_g}$, we get $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) = \operatorname{dist}\big(\mathbf{0}, \partial\phi_k(\mathbf{x}^{k+1})\big) \leq \varepsilon_k$.

When $m > 1$, by the choice of ε_k and δ_k, it holds that $\delta_k \leq \frac{8(\mu+\beta_k B_g^2)}{\beta_k\mu}$ for each k. Hence, we can use Theorem 3.15 and Lemma 3.5 in order to show that $\operatorname{dist}\big(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k)\big) \leq \varepsilon_k$ by the same arguments as in the case of $m = 1$.

Therefore, for $m \geq 1$, if $\varepsilon_k = \bar{\varepsilon}$ for all k, we have from Theorem 2.6 that the inequalities in (2.9) and (2.10) hold. By the choice of $\bar{\varepsilon}$, it holds that $\frac{\bar{\varepsilon}^2(8\sigma+1)}{2\mu(\sigma-1)} \leq \frac{\varepsilon}{2}$. Since $K - 1 \geq \log_\sigma \frac{9\|\mathbf{z}^*\|^2}{\beta_0\varepsilon}$, then $\frac{9\|\mathbf{z}^*\|^2}{2\beta_0\sigma^{K-1}} \leq \frac{\varepsilon}{2}$, and thus we have from (2.10) that $\sum_{i=1}^m |z_i^K g_i(\mathbf{x}^K)| \leq \varepsilon$. In addition, noticing $\frac{\sqrt{2}(\sqrt{\sigma}+1))}{\sqrt{8\sigma+1}} \leq 1$ and $\bar{\varepsilon} \leq \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}$, we have $\bar{\varepsilon}(\sqrt{\sigma}+1)\sqrt{\frac{2}{\mu(\sigma-1)}} \leq \sqrt{\varepsilon}$, and thus (2.9) implies

$$\left\|[\mathbf{g}(\mathbf{x}^K)]_+\right\| \leq \tfrac{4\|\mathbf{z}^*\|}{\beta_0\sigma^{K-1}} + \tfrac{\sqrt{\varepsilon}}{\sqrt{\beta_0\sigma^{K-1}}}.$$

Now, by the setting of K in (4.1), we have that both terms on the right-hand side of the above inequality are no greater than $\varepsilon/2$. Hence, $\|[\mathbf{g}(\mathbf{x}^K)]_+\| \leq \varepsilon$, and thus $\mathbf{x}^K$ must be an ε-KKT point of (1.1).

To show (4.2), we have from the second inequality in (2.8) and the fact that $\varepsilon_k = \bar{\varepsilon} \leq \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}$ for all k that

$$\|\mathbf{z}^k\| \leq 2\|\mathbf{z}^*\| + \sqrt{\tfrac{2\beta_0\bar{\varepsilon}^2}{\mu}\tfrac{\sigma^k-1}{\sigma-1}} \leq 2\|\mathbf{z}^*\| + \sqrt{\tfrac{2\beta_0\varepsilon\sigma^k}{8\sigma+1}}\ \forall\, k \geq 1.$$

Hence, for each $1 \leq k \leq K$ with the K given in (4.1), it holds that

$$\|\mathbf{z}^k\| \leq 2\|\mathbf{z}^*\| + \sqrt{\tfrac{2\beta_0\varepsilon\sigma^K}{8\sigma+1}} \leq 2\|\mathbf{z}^*\| + \sqrt{\tfrac{2\sigma^2}{8\sigma+1}}\max\left\{3\|\mathbf{z}^*\|,\ 2\sqrt{2\|\mathbf{z}^*\|},\ 2\right\}.$$

Since the output $\bar{\mathbf{z}}$ must be one of $\{\mathbf{z}^k\}_{k=1}^K$, we complete the proof. ∎

By Theorem 4.1, we establish the overall iteration complexity of Algorithm 8 to produce an ε-KKT point of (1.1). Notice that if $m = 1$, the complexity result in Theorem 3.15 is in the same order as that in Theorem 3.10. Hence, we state the complexity result of Algorithm 8 for $m = 1$ and $m > 1$ together.

THEOREM 4.2 (oracle complexity). *Suppose that Assumptions* 1–4 *hold. Let* $(\beta_0, \sigma, \varepsilon, \gamma_1, \gamma_2)$ *be the input of Algorithm* 8 *and* $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k\geq 0}$ *be the generated sequence. Suppose* $\bar{\varepsilon} = \min\{\varepsilon, \sqrt{\frac{\varepsilon\mu(\sigma-1)}{8\sigma+1}}\} \leq \{\varepsilon, \frac{24B_g(\mu+\beta_k B_g^2)}{\mu}\}$ *for all* $k \geq 0$. *Let* $\varepsilon_k = \bar{\varepsilon}$ *for all* $k \geq 0$. *Then, to produce an* ε*-KKT point of* (1.1), *Algorithm* 8 *needs at most* $T_{\text{total}} = O\big(m\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$ *evaluations on* f, ∇f, $\mathbf{g}$, *and* $J_{\mathbf{g}}$.

Proof. Let K be the integer given in (4.1), and let $L_{\mathbf{z}^k} = L_f + L_g\max\{1, 4\|\mathbf{z}^*\| + 2\|\mathbf{z}^k\|\}$ for $0 \leq k \leq K-1$. Also, let T_k be the number of evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$ during the kth iteration of Algorithm 8. From Theorem 3.10 and the setting $\delta_k = \frac{\varepsilon_k}{3\beta_k B_g}$, we have that the complexity incurred by line 6 of Algorithm 8 is $O(\sqrt{\frac{L_{\mathbf{z}^k}}{\mu}}|\log\varepsilon|^2)$. Also, from Theorem 3.15, the complexity incurred by line 8 is $O\big(m\sqrt{\frac{L_{\mathbf{z}^k}}{\mu}}|\log\varepsilon|(\log m + |\log\varepsilon|)\big)$ by noting that $\log R = O(|\log\varepsilon|)$. In addition, the complexity incurred by line 10 is $O\big(\sqrt{\frac{L_{\mathbf{z}^k}}{\mu}}|\log\varepsilon|\big)$. From (2.8) with $\varepsilon_t = \bar{\varepsilon}$ for all t, it follows that $\|\mathbf{z}^k\| = O(\|\mathbf{z}^*\|)$, and thus $L_{\mathbf{z}^k} = O(L_f + L_g(1+\|\mathbf{z}^*\|))$ for $0 \leq k \leq K-1$. Therefore, $T_k = O\big(m\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|(\log m + |\log\varepsilon|)\big)$. Since $K = O(|\log\varepsilon|)$ in (4.1), the total complexity is $\sum_{k=0}^{K-1} T_k = O\big(m\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$, which completes the proof. ∎

Remark 4.1. If β_0 is taken in the order of $\frac{1}{\varepsilon}$, then $K = O(1)$ in (4.1). In this case, the total oracle complexity of Algorithm 8 is $O\big(m\sqrt{\frac{L_f+L_g(1+\|\mathbf{z}^*\|)}{\mu}}|\log\varepsilon|(\log m + |\log\varepsilon|)\big)$ to produce an ε-KKT point. The complexity result is in a lower order than the best one $O(\varepsilon^{-\frac{1}{2}})$ in the literature if $m = O(\varepsilon^{-q})$ with $q < \frac{1}{2}$. This affirmatively answers the question we posed in the beginning. Notice that finding an approximate VC of a polytope in Algorithm 6 takes $\Theta(m^3)$ operations by Newton's method, as the number of constraints defining each polytope is $\Theta(m)$, as shown in [1]. This cost can be negligible for a high-dimensional problem, i.e., when n is very big, for which case the cost of querying an oracle can be much higher. Take the quadratically constrained quadratic program in (6.1) as an example. Computing the gradients of the objective and constraint functions needs $\Theta(mn^2)$ operations, far more than $\Theta(m^3)$ if $n \gg m$.

5. Extensions to convex or nonconvex problems.

In this section, we extend the idea of the cutting-plane based FOM to constrained problems with a convex or nonconvex objective. Similar to the strongly convex case, we show that FOMs for solving problems with $O(1)$ nonlinear functional constraints can achieve a complexity result of almost the same order as for solving unconstrained problems.

5.1. Extension to the convex case.

We still consider the problem in (1.1). Suppose that the conditions in Assumptions 1 and 2 hold. Instead of the strong convexity in Assumption 3, we assume the convexity of f in this subsection.

Given a target accuracy $\varepsilon > 0$, to find an ε-KKT point of (1.1), we follow [15] and solve a perturbed strongly convex problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} F_\varepsilon(\mathbf{x}) := f_\varepsilon(\mathbf{x}) + h(\mathbf{x}),\ \text{s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})] \leq \mathbf{0}, \tag{5.1}$$

where

$$f_\varepsilon(\mathbf{x}) = f(\mathbf{x}) + \frac{\varepsilon}{4D_h}\|\mathbf{x} - \mathbf{x}^0\|^2 \text{ with } \mathbf{x}^0 \in \text{dom}(h). \tag{5.2}$$

Let $\bar{\mathbf{x}} \in \mathrm{dom}(h)$ be an $\frac{\varepsilon}{2}$-KKT point of (5.1); i.e., there is $\bar{\mathbf{z}} \geq \mathbf{0}$ such that

$$\mathrm{dist}\left(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_0(\bar{\mathbf{x}}, \bar{\mathbf{z}}) + \tfrac{\varepsilon}{2D_h}(\bar{\mathbf{x}} - \mathbf{x}^0)\right) \leq \tfrac{\varepsilon}{2}, \quad \|[\mathbf{g}(\bar{\mathbf{x}})]_+\| \leq \tfrac{\varepsilon}{2}, \quad \sum_{i=1}^m |\bar{z}_i g_i(\bar{\mathbf{x}})| \leq \frac{\varepsilon}{2},$$

where $\mathcal{L}_0$ is the Lagrange function of (1.1). Since $\|\frac{\varepsilon}{2D_h}(\bar{\mathbf{x}} - \mathbf{x}^0)\| \leq \frac{\varepsilon}{2}$, $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$ must satisfy the conditions in (1.5), and thus $\bar{\mathbf{x}}$ is an ε-KKT point of (1.1). Based on this observation, we can apply Algorithm 8 to the perturbed problem (5.1). By Theorem 4.2 and noticing that f_ε in (5.2) is $\frac{\varepsilon}{2D_h}$-strongly convex, we obtain the following complexity result.

THEOREM 5.1 (complexity result for convex cases). *Assume that the conditions in Assumptions* 1 *and* 2 *hold and that f is convex. Given $\varepsilon > 0$, suppose that* (5.1) *has a KKT point $\mathbf{x}^*_\varepsilon$ with a corresponding multiplier $\mathbf{z}^*_\varepsilon$. Apply Algorithm* 8 *to find an $\frac{\varepsilon}{2}$-KKT point $\bar{\mathbf{x}}$ of* (5.1). *Then $\bar{\mathbf{x}}$ is an ε-KKT point of* (1.1), *and the total number of evaluations on f, ∇f, $\mathbf{g}$, and $J_{\mathbf{g}}$ is $O\big(m\sqrt{\frac{D_h\left(L_f + L_g(1+\|\mathbf{z}^*_\varepsilon\|)\right)}{\varepsilon}}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$.*

5.2. Extension to the nonconvex case. In this subsection, we assume Assumptions 1 and 2 but do not assume the convexity of f. For the nonconvex case, we follow [19] and design an FOM within the framework of the proximal-point method; namely, we solve a sequence of problems in the form of

$$\bar{\mathbf{x}}^{k+1} \approx \arg\min_{\mathbf{x}\in\mathbb{R}^n} \big\{F_k(\mathbf{x}) := f(\mathbf{x}) + L_f\|\mathbf{x} - \bar{\mathbf{x}}^k\|^2 + h(\mathbf{x}), \text{ s.t. } \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})] \leq \mathbf{0}\big\}. \tag{5.3}$$

Under Assumptions 1 and 2, the above problem is convex, and its objective is L_f-strongly convex. Hence, we can apply Algorithm 8 to find $\bar{\mathbf{x}}^{k+1}$. Let $\mathbf{x}^{k+1}_*$ be the unique optimal solution to (5.3). To ensure the existence of a corresponding multiplier for each k and also a uniform bound, we assume Slater's condition on the original problem (1.1).

ASSUMPTION 5 (Slater's condition). *There is $\mathbf{x}_{\mathrm{feas}} \in \mathrm{relint}(h)$ such that $g_i(\mathbf{x}_{\mathrm{feas}}) < 0$ for all $i = 1, \ldots, m$.*

With Slater's condition, the solution $\mathbf{x}^{k+1}_*$ to (5.3) must be a KKT point (cf. [34]). Let $\mathbf{z}^{k+1}_* \geq \mathbf{0}$ be a corresponding multiplier. We give a uniform bound of $\mathbf{z}^{k+1}_*$ below.

LEMMA 5.2 (uniform bound of multipliers). *Assume Assumptions* 1, 2, *and* 5. *Let $\mathbf{x}^*$ be a minimizer of* (1.1), *and let $\mathbf{x}^{k+1}_*$ be the KKT point of* (5.3) *with a corresponding Lagrangian multiplier $\mathbf{z}^{k+1}_*$. Then*

$$\|\mathbf{z}^{k+1}_*\| \leq B_{\mathbf{z}} := \frac{F(\mathbf{x}_{\mathrm{feas}}) - F(\mathbf{x}^*) + L_f D_h^2}{\min_i\left(-g_i(\mathbf{x}_{\mathrm{feas}})\right)} \ \forall\, k \geq 0. \tag{5.4}$$

Proof. From the KKT system, we have that

$$-\sum_{i=1}^m (z^{k+1}_*)_i \nabla g_i(\mathbf{x}^{k+1}_*) \in \partial F_k(\mathbf{x}^{k+1}_*), \quad (z^{k+1}_*)_i g_i(\mathbf{x}^{k+1}_*) = 0 \ \forall\, i = 1, \ldots, m. \tag{5.5}$$

Then we have

$$\begin{aligned}
\sum_{i=1}^m (z^{k+1}_*)_i g_i(\mathbf{x}_{\mathrm{feas}}) &\geq \sum_{i=1}^m (z^{k+1}_*)_i \Big(g_i(\mathbf{x}^{k+1}_*) + \big\langle \mathbf{x}_{\mathrm{feas}} - \mathbf{x}^{k+1}_*, \nabla g_i(\mathbf{x}^{k+1}_*)\big\rangle\Big) \\
&= \left\langle \mathbf{x}_{\mathrm{feas}} - \mathbf{x}^{k+1}_*, \sum_{i=1}^m (z^{k+1}_*)_i \nabla g_i(\mathbf{x}^{k+1}_*)\right\rangle \\
&\geq F_k(\mathbf{x}^{k+1}_*) - F_k(\mathbf{x}_{\mathrm{feas}}),
\end{aligned} \tag{5.6}$$

where the first inequality is from the convexity of each g_i and the nonnegativity of $\mathbf{z}_*^{k+1}$, the equality holds because of the second equation in (5.5), and the last inequality follows from the convexity of F_k and the first equation in (5.5).

Since the diameter of $\mathrm{dom}(h)$ is D_h, it holds that

$$\begin{aligned} -F_k(\mathbf{x}_*^{k+1}) + F_k(\mathbf{x}_{\rm feas}) &= F(\mathbf{x}_{\rm feas}) + L_f\|\mathbf{x}_{\rm feas} - \bar{\mathbf{x}}^k\|^2 - F(\mathbf{x}_*^{k+1}) - L_f\|\mathbf{x}_*^{k+1} - \bar{\mathbf{x}}^k\|^2 \\ &\le F(\mathbf{x}_{\rm feas}) - F(\mathbf{x}_*^{k+1}) + L_f D_h^2. \end{aligned} \tag{5.7}$$

Notice that $F(\mathbf{x}_*^{k+1}) \ge F(\mathbf{x}^*)$. Hence, $F(\mathbf{x}_{\rm feas}) - F(\mathbf{x}_*^{k+1}) \le F(\mathbf{x}_{\rm feas}) - F(\mathbf{x}^*)$, and from (5.7), it follows that $-F_k(\mathbf{x}_*^{k+1}) + F_k(\mathbf{x}_{\rm feas}) \le F(\mathbf{x}_{\rm feas}) - F(\mathbf{x}^*) + L_f D_h^2$. Now we have from (5.6) that

$$\|\mathbf{z}_*^{k+1}\|_1 \le \tfrac{-F_k(\mathbf{x}_*^{k+1}) + F_k(\mathbf{x}_{\rm feas})}{\min_i\left(-g_i(\mathbf{x}_{\rm feas})\right)} \le \tfrac{F(\mathbf{x}_{\rm feas}) - F(\mathbf{x}^*) + L_f D_h^2}{\min_i\left(-g_i(\mathbf{x}_{\rm feas})\right)},$$

and we complete the proof by $\|\mathbf{z}_*^{k+1}\|_2 \le \|\mathbf{z}_*^{k+1}\|_1$. ∎

Similar to our discussion in section 5.1, we notice that if $\bar{\mathbf{x}}^{k+1}$ is an $\frac{\varepsilon}{2}$-KKT point of (5.3) and also $2L_f\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\| \le \frac{\varepsilon}{2}$, then $\bar{\mathbf{x}}^{k+1}$ is an ε-KKT point of (1.1). Below we show that the sum of $\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2$ can be controlled if each $\bar{\mathbf{x}}^{k+1}$ is obtained with sufficient accuracy, and thus a near-KKT point of (1.1) can be produced.

THEOREM 5.3 (complexity result for nonconvex cases). *Assume Assumptions* 1, 2, *and* 5. *Let* $\mathbf{x}^*$ *be a minimizer of* (1.1). *Let* $\varepsilon > 0$ *be given, and let* $\bar{\mathbf{x}}^0 \in \mathrm{dom}(h)$. *Generate the sequence* $\{(\bar{\mathbf{x}}^k, \bar{\mathbf{z}}^k)\}_{k\ge1}$ *by applying Algorithm* 8 *to* (5.3) *with the target accuracy* $\tilde{\varepsilon} = \min\left\{\frac{\varepsilon}{2},\ \frac{3\varepsilon^2}{128L_f(D_h+2\bar{B}_{\mathbf{z}})}\right\}$, *where*

$$\bar{B}_{\mathbf{z}} := 2B_{\mathbf{z}} + \sqrt{\tfrac{2\sigma^2}{8\sigma+1}}\max\left\{3B_{\mathbf{z}},\ 2\sqrt{2B_{\mathbf{z}}},\ 2\right\}, \tag{5.8}$$

with $B_{\mathbf{z}}$ *defined in* (5.4). *Then, after solving at most* K *proximal point subproblems as that in* (5.3), *we can find an* ε*-KKT point of* (1.1), *where*

$$K = \left\lceil \tfrac{128L_f(F(\bar{\mathbf{x}}^0) - F(\mathbf{x}^*) + L_f D_h^2 + \bar{B}_{\mathbf{z}}\|[\mathbf{g}(\bar{\mathbf{x}}^0)]_+\|)}{3\varepsilon^2} \right\rceil. \tag{5.9}$$

In addition, the total number of evaluations on f, ∇f, $\mathbf{g}$, *and* $J_{\mathbf{g}}$ *is* $O\big(\frac{m}{\varepsilon^2}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$.

Proof. Since each $(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{z}}^{k+1})$ is an output from Algorithm 8 applied to (5.3) and with a target accuracy $\tilde{\varepsilon}$, then $\bar{\mathbf{x}}^{k+1}$ is an $\tilde{\varepsilon}$-KKT point of the problem in (5.3), and thus there is a subgradient $\tilde{\nabla}F_k(\mathbf{x}^{k+1}) \in \partial F_k(\bar{\mathbf{x}}^{k+1})$ such that

$$\|\tilde{\nabla}F_k(\bar{\mathbf{x}}^{k+1}) + J_{\mathbf{g}}^\top(\bar{\mathbf{x}}^{k+1})\bar{\mathbf{z}}^{k+1}\| \le \tilde{\varepsilon}, \quad \|\mathbf{g}(\bar{\mathbf{x}}^{k+1})\| \le \tilde{\varepsilon}\ \forall\, k \ge 0. \tag{5.10}$$

From the first inequality in (5.10) and recalling that the diameter of $\mathrm{dom}(h)$ is D_h, we have

$$\left\langle \bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k, \tilde{\nabla}F_k(\bar{\mathbf{x}}^{k+1}) + J_{\mathbf{g}}^\top(\bar{\mathbf{x}}^{k+1})\bar{\mathbf{z}}^{k+1}\right\rangle \le D_h\tilde{\varepsilon}.$$

Hence, by the L_f-strong convexity of F_k and convexity of each g_i, we have

$$\begin{aligned} D_h\tilde{\varepsilon} &\ge \left\langle \bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k, \tilde{\nabla}F_k(\bar{\mathbf{x}}^{k+1}) + J_{\mathbf{g}}^\top(\bar{\mathbf{x}}^{k+1})\bar{\mathbf{z}}^{k+1}\right\rangle \\ &\ge F_k(\bar{\mathbf{x}}^{k+1}) - F_k(\bar{\mathbf{x}}^k) + \tfrac{L_f}{2}\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \langle \bar{\mathbf{z}}^{k+1}, \mathbf{g}(\bar{\mathbf{x}}^{k+1}) - \mathbf{g}(\bar{\mathbf{x}}^k)\rangle \\ &= F(\bar{\mathbf{x}}^{k+1}) - F(\bar{\mathbf{x}}^k) + \tfrac{3L_f}{2}\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \langle \bar{\mathbf{z}}^{k+1}, \mathbf{g}(\bar{\mathbf{x}}^{k+1}) - \mathbf{g}(\bar{\mathbf{x}}^k)\rangle. \end{aligned} \tag{5.11}$$

By (4.2) and (5.4), we have $\|\bar{\mathbf{z}}^{k+1}\| \le \bar{B}_{\mathbf{z}}$ for all $k \ge 0$, where $\bar{B}_{\mathbf{z}}$ is given in (5.8). Hence, it follows from the second inequality in (5.10) that $\langle \bar{\mathbf{z}}^{k+1}, \mathbf{g}(\bar{\mathbf{x}}^{k+1}) - \mathbf{g}(\bar{\mathbf{x}}^k)\rangle \ge -2\tilde{\varepsilon}\bar{B}_{\mathbf{z}}$ for all $k \ge 1$. Now summing up (5.11) gives

$$\frac{3L_f}{2}\sum_{k=0}^{K-1}\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2 \le KD_h\tilde{\varepsilon}+F(\bar{\mathbf{x}}^0)-F(\bar{\mathbf{x}}^K)+(2K-1)\tilde{\varepsilon}\bar{B}_{\mathbf{z}}+\bar{B}_{\mathbf{z}}\|[\mathbf{g}(\bar{\mathbf{x}}^0)]_+\|, \tag{5.12}$$

where we have used $\langle \bar{\mathbf{z}}^1, \mathbf{g}(\bar{\mathbf{x}}^0)\rangle \le \|\bar{\mathbf{z}}^1\| \cdot \|[\mathbf{g}(\bar{\mathbf{x}}^0)]_+\| \le \bar{B}_{\mathbf{z}}\|[\mathbf{g}(\bar{\mathbf{x}}^0)]_+\|$.

Because $\mathbf{x}_*^K$ is a KKT point of (5.3) with a corresponding multiplier $\mathbf{z}_*^K$, we have from (1.4) that

$$F_{K-1}(\bar{\mathbf{x}}^K) - F_{K-1}(\mathbf{x}_*^K) + \left\langle \mathbf{z}_*^K, \mathbf{g}(\bar{\mathbf{x}}^K)\right\rangle \ge 0.$$

Plugging $F_{K-1}(\cdot) = F(\cdot) + L_f\|\cdot - \bar{\mathbf{x}}^{K-1}\|^2$ into the above equation gives

$$F(\bar{\mathbf{x}}^K) + L_f\|\bar{\mathbf{x}}^K - \bar{\mathbf{x}}^{K-1}\|^2 - F(\mathbf{x}_*^K) - L_f\|\mathbf{x}_*^K - \bar{\mathbf{x}}^{K-1}\|^2 + \left\langle \mathbf{z}_*^K, \mathbf{g}(\bar{\mathbf{x}}^K)\right\rangle \ge 0.$$

Now, using (5.4), $\|\mathbf{g}(\bar{\mathbf{x}}^K)\| \le \tilde{\varepsilon}$, $\|\bar{\mathbf{x}}^K - \bar{\mathbf{x}}^{K-1}\|^2 \le D_h^2$, and the fact that $F(\mathbf{x}_*^K) \ge F(\mathbf{x}^*)$, we have from the above inequality that $-F(\bar{\mathbf{x}}^K) \le -F(\mathbf{x}^*) + L_f D_h^2 + \tilde{\varepsilon}B_{\mathbf{z}}$. This inequality together with (5.12) gives

$$\frac{3L_f}{2}\sum_{k=0}^{K-1}\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 \le KD_h\tilde{\varepsilon} + F(\bar{\mathbf{x}}^0) - F(\mathbf{x}^*) + L_f D_h^2 + 2K\tilde{\varepsilon}\bar{B}_{\mathbf{z}} + \bar{B}_{\mathbf{z}}\|[\mathbf{g}(\bar{\mathbf{x}}^0)]_+\|. \tag{5.13}$$

Multiplying L_f to both sides of the above inequality and taking the square root, we have

$$\min_{0\le k<K} L_f\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\| \le \sqrt{\tfrac{2}{3}L_f(D_h\tilde{\varepsilon} + 2\bar{B}_{\mathbf{z}}\tilde{\varepsilon})}+\sqrt{\tfrac{2}{3}\frac{L_f\left(F(\bar{\mathbf{x}}^0)-F(\mathbf{x}^*)+L_fD_h^2+\bar{B}_{\mathbf{z}}\|[\mathbf{g}(\bar{\mathbf{x}}^0)]_+\|\right)}{K}}. \tag{5.14}$$

Therefore, by the setting of $\tilde{\varepsilon}$ and K, we have $\min_{0\le k<K} L_f\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\| \le \frac{\varepsilon}{4}$. Suppose $L_f\|\bar{\mathbf{x}}^{k_0+1} - \bar{\mathbf{x}}^{k_0}\| \le \frac{\varepsilon}{4}$. Then, by our discussion above Theorem 5.3, $\bar{\mathbf{x}}^{k_0+1}$ is an ε-KKT point of (1.1). From Theorem 4.2, the complexity of solving one problem as that in (5.3) is $O\big(m|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$, and thus the total complexity is $O(Km|\log\varepsilon|^2(\log m + |\log\varepsilon|)) = O\big(\frac{m}{\varepsilon^2}|\log\varepsilon|^2(\log m + |\log\varepsilon|)\big)$. This completes the proof. ∎

6. Experimental results. In this section, we demonstrate the established theory by performing numerical experiments on solving the following quadratically constrained quadratic program (QCQP):

$$\min_{\mathbf{x}\in\mathbb{R}^n} \tfrac{1}{2}\mathbf{x}^\top\mathbf{Q}_0\mathbf{x}+\mathbf{x}^\top\mathbf{c}_0,\ \text{s.t. } \tfrac{1}{2}\mathbf{x}^\top\mathbf{Q}_j\mathbf{x}+\mathbf{x}^\top\mathbf{c}_j+d_j \le 0,\ j=1,\dots,m;\ x_i\in[l_i,u_i],\ i=1,\dots,n. \tag{6.1}$$

In the experiment, $\mathbf{Q}_0$ is generated to be positive definite, $\mathbf{Q}_j$ is positive semidefinite but rank-deficient for each $j = 1,\dots,m$, and $l_i = -10$ and $u_i = 10$ for each i. All d_j are negative, so Slater's condition holds. In addition, we conduct tests on solving the elastic-net regularized Neyman–Pearson classification problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{N_+}\sum_{\mathbf{a}\in\mathcal{N}_+}\log(1+\exp(-\mathbf{a}^\top\mathbf{x})) + \lambda_1\|\mathbf{x}\|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|^2,\ \text{s.t. } \frac{1}{N_-}\sum_{\mathbf{a}\in\mathcal{N}_-}\log(1+\exp(\mathbf{a}^\top\mathbf{x})) \le \alpha, \tag{6.2}$$

where $\mathcal{N}_+$ and $\mathcal{N}_-$ respectively denote the sets of positive and negative samples, and N_+ and N_- are their cardinality. The tests in sections 6.1 and 6.2 are conducted on a quad-core iMAC with 8GB memory, and those in section 6.3 are conducted on a Windows PC with 10 CPU cores and 128GB memory.

6.1. Comparison of different first-order iALMs. We first compare two implementations of the iALM in Algorithm 1 to solve (6.1). One directly applies the APG method in Algorithm 2 to solve each ALM subproblem, and we call it the "APG-based iALM." The other uses the proposed cutting-plane based FOM to solve subproblems; namely, we implement Algorithm 8 , and we call it the "cutting-plane iALM." For both implementations, we set $\beta_k = 10^{k-1}$ for each outer iteration $k \geq 1$ and run the iALM to five outer iterations. The target accuracy for a near-KKT point is set to $\varepsilon = 10^{-4}$. In the implementation of the APG-based iALM, due to the quadratic penalty term, we apply Algorithm 2 with line search for a local smoothness constant and set the parameters to $\gamma_1 = 1.5, \gamma_2 = 2, L_{\min} = 1$. In the implementation of the cutting-plane iALM, we use Algorithm 2 to solve problems in the form of (3.5), for which we can explicitly compute the global smoothness constant, and thus we simply set $L_{\min}$ to the global smoothness constant. In addition, we set $\tau = 2$ in Algorithm 6 when it is called. Notice that Algorithm 6 works for any $\tau \geq 0.007$. However, empirically we find that a small τ will result in more calls to the separation oracle, while a too-big τ will cause trouble for finding a sufficiently accurate VC. $\tau = 2$ gives a good tradeoff.

We test three groups of QCQP instances, each of which has $n = 1000$. The first group has $m = 1$ constraint, the second has $m = 2$, and the third has $m = 5$. For each group, we conduct three independent trials. For each instance, we report the number of gradient and function evaluations, the primal residual, dual residual, and complementarity violation, which are denoted as #grad, #func, pres, dres, and compl, for solving each ALM subproblem. In order to demonstrate the worst-case theoretical result, we use a randomly generated initial point while solving each ALM subproblem. The performance of the iALM can be much better if the warm-start technique is adopted. The results are shown in Tables 1–3. For the cutting-plane iALM, its #func is zero and not shown in the tables because we feed the APG an explicitly computed smoothness constant and no line search is performed.

From the results, we see that as the penalty parameter increases, the APG-based iALM needs significantly more iterations to solve the subproblems, while the cutting-plane iALM does not suffer from the big penalty parameter. However, the cutting-plane iALM has worse scalability to m, and this matches with our theory.

6.2. Comparison to a primal-dual method with line search. In this subsection, we compare the proposed cutting-plane based iALM with the primal-dual method with line search in [12] on solving (6.1) and on solving (6.2). The latter is called APDB. It is a single-loop first-order method and can achieve the optimal complexity result $O(\varepsilon^{-\frac{1}{2}})$ for solving strongly convex problems with nonlinear functional constraints.

In the experiment for solving (6.1), we generate three groups of QCQP instances in the same way as that in the previous test, and in each group we conduct 10 independent trials. The setting of the proposed iALM is the same as in the previous test. For APDB, we set $\gamma_0 = 1$, $\eta = 0.7$ and select the best τ_0 from $\{0.1, 0.01, 0.001\}$; see Algorithm 2.3 in [12] for the specific meaning of these parameters. In order to have a fair comparison, we terminate APDB once it produces a 10^{-8}-KKT point. The results are plotted in Figure 1. From the figure, we see that when $m = 1$ or $m = 2$,

TABLE 1

Results by the APG-based first-order iALM and the proposed cutting-plane based first-order iALM for solving QCQP (6.1) *with* $m = 1$ *and* $n = 1000$.

		APG-based iALM					Proposed cutting-plane iALM			
Out. iter.	β	#grad	#func	pres	dres	compl	#grad	pres	dres	compl
Trial 1		Total running time = 774.2 sec.					Total running time = 12.4 sec.			
1	1	5056	9420	5.13e-02	9.65e-05	2.63e-03	2136	5.13e-02	6.40e-11	2.64e-03
2	10	16802	31298	1.65e-06	9.46e-05	8.46e-08	1434	4.23e-07	9.20e-11	2.17e-08
3	10^2	55359	103112	5.40e-08	9.77e-05	2.77e-09	1068	4.22e-10	2.63e-10	2.17e-11
4	10^3	179877	335030	6.51e-09	9.96e-05	3.34e-10	1080	0.00e+00	1.74e-08	4.84e-11
5	10^4	584145	1087988	0.00e+00	9.95e-05	4.57e-11	1104	2.29e-11	9.23e-09	1.17e-12
Trial 2		Total running time = 760.0 sec.					Total running time = 12.1 sec.			
1	1	4969	9258	5.78e-02	9.78e-05	3.34e-03	1926	5.78e-02	4.94e-09	3.34e-03
2	10	16466	30672	2.10e-06	9.99e-05	1.21e-07	1440	5.85e-07	3.41e-10	3.38e-08
3	10^2	54617	101730	4.57e-08	9.85e-05	2.64e-09	1050	0.00e+00	7.90e-09	4.03e-10
4	10^3	177171	329990	6.44e-09	9.93e-05	3.72e-10	1074	0.00e+00	2.18e-07	1.42e-10
5	10^4	580377	1080970	0.00e+00	1.00e-04	4.06e-11	1104	2.75e-10	1.84e-09	1.59e-11
Trial 3		Total running time = 780.9 sec.					Total running time = 12.4 sec.			
1	1	5100	9502	4.37e-02	9.66e-05	1.91e-03	2088	4.37e-02	2.53e-09	1.91e-03
2	10	17035	31732	0.00e+00	9.33e-05	8.08e-08	1428	4.34e-07	7.52e-09	1.90e-08
3	10^2	56348	104954	1.43e-07	9.79e-05	6.25e-09	1092	0.00e+00	2.75e-13	2.36e-10
4	10^3	182583	340070	0.00e+00	9.63e-05	5.12e-10	1122	4.33e-09	4.76e-07	1.89e-10
5	10^4	595012	1108228	1.81e-10	9.99e-05	7.92e-12	1164	0.00e+00	1.88e-09	2.01e-11

TABLE 2

Results by the APG-based first-order iALM and the proposed cutting-plane based first-order iALM for solving QCQP (6.1) *with* $m = 2$ *and* $n = 1000$.

		APG-based iALM					Proposed cutting-plane iALM			
Out. iter.	β	#grad	#func	pres	dres	compl	#grad	pres	dres	compl
trial 1		total running time = 1348.0 sec.					total running time = 51.0 sec.			
1	1	5551	10342	4.45e-02	8.71e-05	1.40e-03	3342	4.45e-02	1.06e-09	1.40e-03
2	10	18330	34144	0.00e+00	9.62e-05	6.47e-08	3384	3.19e-07	9.17e-09	9.98e-09
3	10^2	59680	111160	8.81e-08	9.77e-05	2.71e-09	3522	6.01e-09	9.15e-10	2.44e-10
4	10^3	194236	361774	0.00e+00	9.94e-05	9.15e-11	3582	1.36e-10	3.84e-09	6.17e-12
5	10^4	629359	1172200	0.00e+00	9.99e-05	7.65e-12	3678	2.66e-11	1.60e-09	8.13e-13
Trial 2		Total running time = 1299.4 sec.					Total running time = 49.5 sec.			
1	1	5362	9990	6.60e-02	9.05e-05	3.10e-03	3180	6.60e-02	8.27e-09	3.10e-03
2	10	17646	32870	2.74e-06	9.26e-05	1.34e-07	3282	6.17e-07	2.67e-10	2.91e-08
3	10^2	57832	107718	1.41e-08	9.82e-05	2.79e-09	3372	5.91e-10	9.05e-11	2.61e-11
4	10^3	187544	349310	0.00e+00	9.88e-05	2.70e-10	3450	4.97e-10	6.76e-09	2.34e-11
5	10^4	606432	1129498	9.88e-11	9.97e-05	7.38e-12	3528	1.82e-11	5.23e-09	1.95e-12
Trial 3		Total running time = 1337.1 sec.					Total running time = 49.2 sec.			
1	1	5464	10180	5.50e-02	9.51e-05	2.25e-03	3156	5.50e-02	6.27e-09	2.25e-03
2	10	18039	33602	1.78e-06	9.90e-05	8.15e-08	3324	5.16e-07	1.76e-10	2.07e-08
3	10^2	59505	110834	2.88e-08	9.95e-05	1.86e-09	3384	5.93e-09	8.30e-09	2.49e-10
4	10^3	192301	358170	3.78e-09	9.99e-05	1.45e-10	3504	0.00e+00	1.02e-09	3.00e-11
5	10^4	627235	1168244	6.81e-11	1.00e-04	9.17e-12	3528	5.23e-11	2.78e-09	1.70e-12

the proposed iALM needs fewer gradient evaluations than APDB to give a solution of similar or higher accuracy, and when $m = 5$, APDB needs fewer gradient evaluations. In addition, different from the proposed iALM, APDB needs fewer gradient evaluations as m increases. Hence, APDB may be even more efficient than the proposed iALM as m further increases.

In the experiment for solving (6.2), we use `arcene` and `spambase` datasets, both of which are from the UCI repository,[1] and we set $\alpha = 0.5$. Each sample is normalized. In order to achieve at least 90% prediction accuracy for the positive dataset, we tune the regularization parameters to $\lambda_1 = \lambda_2 = 10^{-3}$ for the `arcene` dataset and to $\lambda_1 = \lambda_2 = 10^{-4}$ for the `spambase` dataset. The APDB is applied to an SP problem formulated by using the ordinary Lagrangian function of (6.2). As the logistic loss function has bounded gradient and Hessian, we explicitly compute the global Lipschitz constants and adopt constant stepsize for both APDB and the proposed iALM. We set $\beta_k = 10^{k-1}$ for iALM and run it to five outer iterations. The target accuracy for a near-KKT point is $\varepsilon = 10^{-5}$. For APDB, we set the maximum number of iterations to 10^5 and terminate it if an ε-KKT solution is produced. In the SP formulation solved by APDB, we set an upper bound of its dual variable to twice of the value of the dual variable returned by the iALM. This known upper bound benefits APDB.

[1] The data can be downloaded from https://archive.ics.uci.edu/ml/datasets.php.

TABLE 3

Results by the APG-based first-order iALM and the proposed cutting-plane based first-order iALM for solving QCQP (6.1) *with* $m = 5$ *and* $n = 1000$.

		APG-based iALM					Proposed cutting-plane iALM			
Out. iter.	β	#grad	#func	pres	dres	compl	#grad	pres	dres	compl
Trial 1		Total running time = 2833.1 sec.					Total running time = 156.8 sec.			
1	1	5537	10316	7.93e-02	9.91e-05	2.90e-03	6714	7.93e-02	2.91e-09	2.90e-03
2	10	18417	34306	1.12e-06	9.83e-05	4.28e-08	6984	8.93e-07	4.32e-09	3.27e-08
3	10^2	60058	111864	5.83e-08	9.62e-05	2.25e-09	7158	4.64e-09	1.50e-09	2.02e-10
4	10^3	195894	364862	3.14e-09	9.88e-05	1.64e-10	7314	4.37e-10	4.28e-09	1.64e-11
5	10^4	640357	1192684	9.40e-10	9.97e-05	3.51e-11	7614	2.79e-11	8.77e-09	1.74e-12
Trial 2		Total running time = 2786.0 sec.					Total running time = 160.7 sec.			
1	1	5537	10316	6.77e-02	8.21e-05	2.42e-03	6900	6.77e-02	6.16e-09	2.42e-03
2	10	18170	33846	6.24e-07	9.21e-05	2.43e-08	7110	7.39e-07	2.64e-09	2.75e-08
3	10^2	59607	111024	2.66e-08	9.73e-05	1.71e-09	7224	2.81e-09	9.46e-09	1.90e-10
4	10^3	194483	362234	1.21e-08	9.99e-05	3.19e-10	7512	6.61e-10	4.34e-09	2.53e-11
5	10^4	636109	1184772	7.58e-11	9.94e-05	1.76e-11	7698	3.94e-11	7.84e-09	1.73e-12
Trial 3		Total running time = 2820.0 sec.					Total running time = 155.3 sec.			
1	1	5595	10424	8.47e-02	8.51e-05	3.26e-03	6594	8.47e-02	9.82e-09	3.26e-03
2	10	18461	34388	7.78e-07	9.55e-05	3.07e-08	6882	8.64e-07	5.52e-09	3.33e-08
3	10^2	60422	112542	3.78e-09	9.93e-05	4.10e-09	7116	3.42e-09	1.52e-10	1.83e-10
4	10^3	196869	366678	7.70e-09	9.87e-05	3.05e-10	7260	7.35e-11	5.28e-09	1.91e-11
5	10^4	640997	1193876	3.63e-10	9.95e-05	1.37e-11	7488	6.86e-11	6.05e-09	2.72e-12

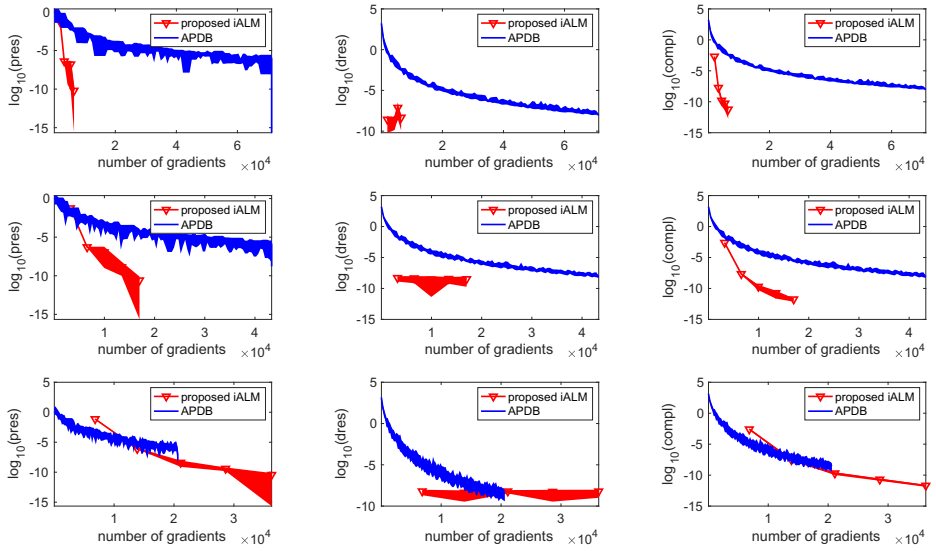

FIG. 1. *Results by the proposed cutting-plane based iALM and the APDB method in* [12] *on solving QCQP instances of size* $n = 1000$ *and* $m \in \{1, 2, 5\}$. *The solid curve in each figure plots the mean of* 10 *independent trials. First row:* $m = 1$*; second row:* $m = 2$*; third row:* $m = 5$. *First column: primal residual; second column: dual residual; third column: complementarity violation.*

The results are reported in Figure 2, from which we see that the proposed iALM takes significantly fewer gradient evaluations than APDB to produce a similarly accurate KKT solution. Moreover, we achieve 91.67% accuracy for the positive samples and 83.33% for the negative samples in the `arcene` dataset, and the final obtained solution has only 427 nonzeros out of 10,000. For the `spambase` dataset, we achieve 90.89% accuracy for the positive samples and 74.44% for the negative samples, and the final solution has 51 nonzeros out of 57 because a small regularization parameter is used.

6.3. Comparison to the interior-point method. In this subsection, we compare the proposed cutting-plane iALM to SDPT3 [36] on solving (6.1). SDPT3 is a primal-dual infeasible interior-point method. Interior-point methods can give high-accurate solutions to convex problems but do not often have a good scalability to

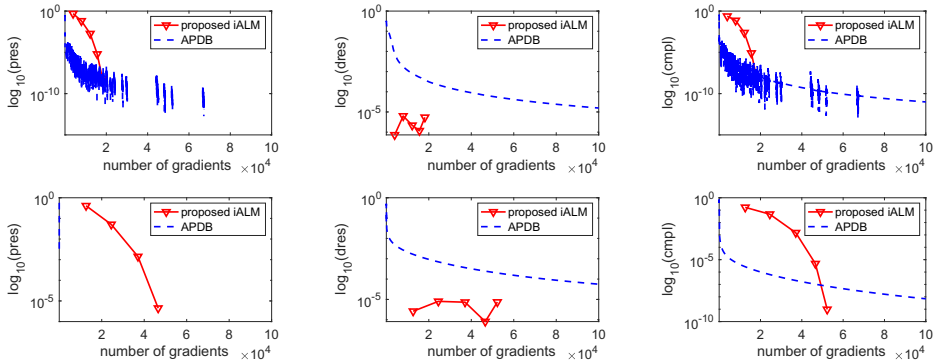

FIG. 2. *Results by the proposed cutting-plane based iALM and the APDB method in* [12] *on solving instances of Neyman–Pearson problem* (6.2) *with arcene dataset (first row) and spambase dataset (second row). First column: primal residual; second column: dual residual; third column: complementarity violation.* †*Missing parts on the curves by APDB correspond to zero residuals, and for spambase, the primal residual by the proposed iALM at the last outer iteration is zero.*

the problem dimension. In this test, we generate instances of (6.1) with $m = 2$ and $n \in \{1000, 5000, 10000\}$. For each (m, n), we generate five QCQP instances independently in the same way as that in previous tests. The parameters of the proposed iALM are set the same as previously, except how we choose the global smoothness constant of (3.5). Notice that for the QCQP (6.1) the corresponding subproblem (3.5) has the Hessian matrix $\mathbf{H} = \mathbf{Q}_0 + \sum_{i=1}^m y_i \mathbf{Q}_i$. A tight smoothness constant is $\|\mathbf{H}\|$. For a small n, computing the spectral norm is not so expensive. However, it can be very expensive when n is big. Hence, we set the global smoothness constant to $\|\mathbf{H}\|$ for $n = 1000$ and to $\|\mathbf{Q}_0\| + \sum_{i=1}^m y_i \|\mathbf{Q}_i\|$ for $n \in \{5000, 10000\}$. Since $\mathbf{y}$ changes during the proposed iALM, the former setting needs to compute the spectral norm of a sequence of $n \times n$ matrices, while the latter one only needs to compute $\{\|\mathbf{Q}_i\|\}_{0 \le i \le m}$ once at the beginning of the algorithm. This way, we can save the time of computing the spectral norm but will obtain larger smoothness constants that lead to smaller stepsize and eventually result in more gradient evaluations. We call SDPT3 by using CVX [11] and set the `precision` to "high."

To compare the performance of the cutting-plane iALM and SDPT3, we report their running time and violation to the KKT system at the output solution. For the former method, we also report its number of gradient evaluations. The results are shown in Table 4. From the table, we see that the cutting-plane iALM can yield similar or more accurate solutions than SDPT3. When $n = 1000$, SDPT3 is significantly faster, but for $n \in \{5000, 10000\}$ the cutting-plane first-order iALM takes much shorter time than SDPT3.

7. Concluding remarks. We have proposed a cutting-plane based 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 of the underlying problem in some sense. In general, a complexity result of $\tilde{O}(m\sqrt{\kappa})$ has been established. To give an ε-KKT point, our result is better than an existing lower bound if $m = o(\varepsilon^{-\frac{1}{2}})$. We have also extended the idea of the cutting-plane based FOM to convex and nonconvex cases. Similarly, when $m = O(1)$, we obtained almost the same-order complexity results (with a difference of a polynomial of $|\log \varepsilon|$) as for solving an unconstrained problem.

TABLE 4

Results by the proposed cutting-plane based first-order iALM and the interior-point method SDPT3 on solving instances of (6.1). *"NaN" means that SDPT3 could not solve that instance successfully.*

	Proposed cutting-plane iALM					SDPT3			
Trial	Time(h:m:s)	#grad	pres	dres	compl	Time(h:m:s)	pres	dres	compl
Problem size: $m = 2, n = 1000$									
1	0:0:35	16776	0.00e+00	1.13e-10	3.42e-12	0:0:11	3.30e-10	1.03e-09	4.12e-11
2	0:0:36	16812	0.00e+00	1.89e-09	8.75e-13	0:0:16	2.14e-10	4.40e-10	9.25e-12
3	0:0:35	17004	4.09e-11	1.19e-09	1.91e-12	0:0:11	0.00e+00	2.04e-09	8.31e-11
4	0:0:36	16698	3.53e-11	2.69e-09	2.27e-12	0:0:11	0.00e+00	8.00e-09	1.61e-08
5	0:0:35	16578	2.32e-11	3.19e-09	2.77e-12	0:0:17	1.58e-09	8.16e-10	9.10e-11
Problem size: $m = 2, n = 5000$									
1	0:11:9	21630	2.58e-11	5.85e-10	1.24e-12	0:40:44	0.00e+00	8.26e-09	5.71e-10
2	0:11:11	21642	3.58e-11	9.17e-10	1.63e-12	0:52:23	6.55e-08	1.18e-09	2.84e-09
3	0:11:6	21504	1.95e-11	6.10e-10	7.19e-13	0:50:39	5.45e-08	NaN	NaN
4	0:11:12	21678	3.13e-11	4.67e-09	1.04e-12	0:40:38	0.00e+00	1.12e-08	1.59e-09
5	0:11:7	21516	1.99e-11	8.59e-09	9.04e-13	0:36:17	2.71e-08	1.10e-08	1.28e-09
Problem size: $m = 2, n = 10000$									
1	1:16:1	22332	0.00e+00	6.32e-10	2.37e-13	5:55:22	2.41e-07	3.10e-08	1.33e-08
2	1:8:33	22296	4.99e-12	6.36e-09	1.30e-12	6:20:3	0.00e+00	4.60e-10	3.19e-09
3	0:58:16	22296	1.73e-11	2.54e-09	7.43e-13	6:13:5	0.00e+00	3.44e-08	8.17e-09
4	0:58:9	22368	2.05e-11	1.14e-08	9.17e-13	7:3:39	0.00e+00	2.16e-08	7.70e-09
5	1:15:19	22182	7.95e-12	1.04e-08	1.30e-12	8:31:16	0.00e+00	3.70e-08	1.48e-09

Acknowledgments. The author would like to thank the two anonymous referees for their careful reviews and helpful comments/suggestions. He also would like to thank Dr. Necdet Serhat Aybat and Dr. Erfan Yazdandoost Hamedani for sharing their code of the APDB method.

REFERENCES

[1] K. M. ANSTREICHER, *On Vaidya's volumetric cutting plane method for convex programming*, Math. Oper. Res., 22 (1997), pp. 63–89.

[2] A. Y. ARAVKIN, J. V. BURKE, D. DRUSVYATSKIY, M. P. FRIEDLANDER, AND S. ROY, *Level-set methods for convex optimization*, Math. Program., 174 (2019), pp. 359–390.

[3] N. S. AYBAT AND G. IYENGAR, *An Augmented Lagrangian Method for Conic Convex Programming*, preprint, https://arxiv.org/abs/1302.6322, 2013.

[4] A. BECK AND M. TEBOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[5] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[6] D. BOOB, Q. DENG, AND G. LAN, *Stochastic first-order methods for convex and nonconvex functional constrained optimization*, Math. Program., (2022), https://doi.org/10.1007/s10107-021-01742-y.

[7] C. CARTIS, N. I. M GOULD, AND P. L. TOINT, *On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming*, SIAM J. Optim., 21 (2011), pp. 1721–1739, https://doi.org/10.1137/11082381X.

[8] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.

[9] Y. CHEN, G. LAN, AND Y. OUYANG, *Accelerated schemes for a class of variational inequalities*, Math. Program., 165 (2017), pp. 113–149.

[10] R. GANDY, *Portfolio Optimization with Risk Constraints*, Ph.D. thesis, Universität Ulm, Ulm, Germany, 2005.

[11] M. GRANT, S. BOYD, AND Y. YE, *CVX: MATLAB Software for Disciplined Convex Programming*, 2008.

[12] E. Y. HAMEDANI AND N. S. AYBAT, *A primal-dual algorithm with line search for general convex-concave saddle point problems*, SIAM J. Optim., 31 (2021), pp. 1299–1329, https://doi.org/10.1137/18M1213488.

[13] L. T. K. HIEN, R. ZHAO, AND W. B. HASKELL, *An Inexact Primal-Dual Smoothing Framework for Large-Scale Non-bilinear Saddle Point Problems*, preprint, https://arxiv.org/abs/1711.03669v3, 2017.

[14] W. KONG, J. G. MELO, AND R. D. C. MONTEIRO, *Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite pro-*

grams, SIAM J. Optim., 29 (2019), pp. 2566–2593, https://doi.org/10.1137/18M1171011.

[15] G. Lan and R. D. Monteiro, *Iteration-complexity of first-order augmented Lagrangian methods for convex programming*, Math. Program., 155 (2016), pp. 511–547.

[16] F. Li and Z. Qu, *An inexact proximal augmented Lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization*, Math. Program. Comput., 13 (2021), pp. 583–644.

[17] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu, *Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2170–2178.

[18] Z. Li and Y. Xu, *Augmented Lagrangian–based first-order methods for convex-constrained programs with weakly convex objective*, INFORMS J. Optim., 3 (2021), pp. 373–397.

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

[20] Q. Lin, R. Ma, and T. Yang, *Level-set methods for finite-sum constrained convex optimization*, in International Conference on Machine Learning, 2018, pp. 3112–3121.

[21] Q. Lin, S. Nadarajah, and N. Soheili, *A level-set method for convex optimization with a feasible solution path*, SIAM J. Optim., 28 (2018), pp. 3290–3311, https://doi.org/10.1137/17M1152334.

[22] Q. Lin and L. Xiao, *An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization*, Comput. Optim. Appl., 60 (2015), pp. 633–674.

[23] Z. Lu and Z. Zhou, *Iteration-Complexity of First-Order Augmented Lagrangian Methods for Convex Conic Programming*, preprint, https://arxiv.org/abs/1803.09941, 2018.

[24] J. G. Melo, R. D. Monteiro, and H. Wang, *Iteration-Complexity of an Inexact Proximal Accelerated Augmented Lagrangian Method for Solving Linearly Constrained Smooth Nonconvex Composite Optimization Problems*, preprint, https://arxiv.org/abs/2006.08048, 2020.

[25] R. D. C. Monteiro and B. F. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 2755–2787, https://doi.org/10.1137/090753127.

[26] I. Necoara and V. Nedelcu, *Rate analysis of inexact dual first-order methods application to dual decomposition*, IEEE Trans. Automat. Control, 59 (2014), pp. 1232–1243.

[27] A. Nedić and A. Ozdaglar, *Approximate primal solutions and rate analysis for dual subgradient methods*, SIAM J. Optim., 19 (2009), pp. 1757–1780, https://doi.org/10.1137/070708111.

[28] A. Nedić and A. Ozdaglar, *Subgradient methods for saddle-point problems*, J. Optim. Theory Appl., 142 (2009), pp. 205–228.

[29] 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 J. Optim., 15 (2004), pp. 229–251, https://doi.org/10.1137/S1052623403425629.

[30] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[31] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.

[32] Y. Ouyang and Y. Xu, *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*, Math. Program., 185 (2021), pp. 1–35.

[33] P. Rigollet and X. Tong, *Neyman-Pearson classification, convexity and stochastic constraints*, J. Mach. Learn. Res., 12 (2011), pp. 2831–2855.

[34] R. T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.

[35] M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher, *An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2019, pp. 13965–13977.

[36] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, *On the implementation and usage of SDPT3–A MATLAB software package for semidefinite-quadratic-linear programming, version* 4.0, in Handbook on Semidefinite, Conic and Polynomial Optimization, Springer, New York, 2012, pp. 715–754.

[37] P. M. Vaidya, *A new algorithm for minimizing convex functions over convex sets*, Math. Programming, 73 (1996), pp. 291–341.

[38] Y. Xu, *Primal-dual stochastic gradient method for convex programs with many functional constraints*, SIAM J. Optim., 30 (2020), pp. 1664–1692, https://doi.org/10.1137/18M1229869.

[39] Y. Xu, *First-order methods for constrained convex programming based on linearized augmented Lagrangian function*, INFORMS J. Optim., 3 (2021), pp. 89–117.
[40] Y. Xu, *Iteration complexity of inexact augmented lagrangian methods for constrained convex programming*, Math. Program., 185 (2021), pp. 199–244.
[41] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789, https://doi.org/10.1137/120887795.
[42] 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 Proceedings of the 55th Conference on Decision and Control (CDC), 2016, pp. 1900–1905.
[43] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, *Fairness Constraints: Mechanisms for Fair Classification*, preprint, https://arxiv.org/abs/1507.05259, 2015.