DAMPED PROXIMAL AUGMENTED LAGRANGIAN METHOD FOR WEAKLY-CONVEX PROBLEMS WITH CONVEX CONSTRAINTS

HARI DAHAL, WEI LIU, YANGYANG XU*

Abstract. We give a damped proximal augmented Lagrangian method (DPALM) for solving problems with a weakly-convex objective and convex linear/nonlinear constraints. Instead of taking a full stepsize, DPALM adopts a damped dual stepsize to ensure the boundedness of dual iterates. We show that DPALM can produce a (near) ε-KKT point within $O(\varepsilon^{-2})$ outer iterations if each DPALM subproblem is solved to a proper accuracy. In addition, we establish overall iteration complexity of DPALM when the objective is either a regularized smooth function or in a regularized compositional form. For the former case, DPALM achieves the complexity of $\widetilde{\mathcal{O}}\left(\varepsilon^{-2.5}\right)$ to produce an ε-KKT point by applying an accelerated proximal gradient (APG) method to each DPALM subproblem. For the latter case, the complexity of DPALM is $\widetilde{\mathcal{O}}\left(\varepsilon^{-3}\right)$ to produce a near ε-KKT point by using an APG to solve a Moreau-envelope smoothed version of each subproblem. Our outer iteration complexity and the overall complexity either generalize existing best ones from unconstrained or linear-constrained problems to convex-constrained ones, or improve over the best-known results on solving the same-structured problems. Furthermore, numerical experiments on linearly/quadratically constrained non-convex quadratic programs and linear-constrained robust nonlinear least squares are conducted to demonstrate the empirical efficiency of the proposed DPALM over several state-of-the art methods.

Keywords: weakly-convex optimization, first-order methods, proximal augmented Lagrangian method, functional constrained problems

Mathematics Subject Classification: 49M05, 49M37, 90C06, 90C30, 90C26, 90C60

1. Introduction. Given the rapid increase of data volume in modern applications, there has been a substantial surge in interest of designing first-order methods (FOMs). Traditionally, a significant portion of research in optimization has been concentrated in the realm of convex problems. However, there has been a noticeable and accelerating shift towards the investigation of non-convex optimization during the past decade. This trend is primarily attributable to the applications and the recognition that most contemporary optimization challenges indeed fall within the category of non-convex problems.

In this paper, we consider to design new FOMs for non-convex constrained optimization in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}), \text{ s.t. } \; \mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{g}(\mathbf{x}) := [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})] \le \mathbf{0}, \tag{P}$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{b} \in \mathbb{R}^n$, F is continuous on its domain $\mathcal{X} := \mathrm{dom}(F)$, h is closed convex, and $g_i : \mathbb{R}^d \to \mathbb{R}$ is closed convex for each $i = 1, 2, \ldots, m$. We will assume that h is simple and admits an easy proximal mapping, each g_i is smooth on an open set containing $\mathcal{X}$, and f is ρ-weakly convex and may be nondifferentiable; see Definition 1.1 below.

The problem (P) is rather general and has many interesting applications, such as non-convex quadratic programs with linear and/or nonlinear constraints, reformulation of the nonnegative matrix completion by variable splitting [47], hyperspectral image denoising [4], classification problem with ROC-based constraints [17], and the Neyman-Pearson classification [38].

In the realm of non-convex optimization, locating a global optimizer is usually a computationally intractable task [7]. As a practical alternative, the goal is often directed towards identifying a stationary point. On solving (P), we aim at finding a (near) ε-KKT point for a given $\varepsilon > 0$; see Definition 1.3.

*{dahalh, liuw16, xuy21}@rpi.edu, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180

1.1. Algorithmic Framework. The FOM that we will design is based on the framework of a damped proximal augmented Lagrangian method (DPALM). The augmented Lagrangian (AL) function of (P) is

$$\mathcal{L}_\beta(\mathbf{x};\mathbf{y},\mathbf{z}) = F(\mathbf{x}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + \tfrac{\beta}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \tfrac{\beta}{2}\left\|\left[\mathbf{g}(\mathbf{x}) + \tfrac{\mathbf{z}}{\beta}\right]_+\right\|^2 - \tfrac{\|\mathbf{z}\|^2}{2\beta}, \tag{1.1}$$

where $\beta > 0$ is a penalty parameter, $[\mathbf{a}]_+$ takes the component-wise positive part of a vector $\mathbf{a}$, and $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m$ are Lagrangian multipliers. Define a proximal AL function as

$$\widetilde{\mathcal{L}}_\beta(\mathbf{x};\mathbf{y},\mathbf{z}) = \mathcal{L}_\beta(\mathbf{x};\mathbf{y},\mathbf{z}) + \rho\|\mathbf{x}-\mathbf{x}^k\|^2, \tag{1.2}$$

which is ρ-strongly convex due to the ρ-weak convexity of F. Notice that we use ρ in (1.2) for convenience. It can be any number that is strictly larger than $\frac{\rho}{2}$. With $\widetilde{\mathcal{L}}_\beta$, we present the DPALM framework in Algorithm 1, where we adopt the convention of $\frac{v}{0} = +\infty$ for any $v > 0$.

Algorithm 1: A Damped Proximal Augmented Lagrangian Method (DPALM) for (P)

Initialize $\mathbf{x}^0 \in \mathcal{X} = \mathrm{dom}(F)$, $\mathbf{y}^0$, and $\mathbf{z}^0 \geq \mathbf{0}$.
Choose a positive sequence $\{\beta_k\}_{k\geq 0}$ and positive summable sequences $\{v_k\}_{k\geq 0}, \{w_k\}_{k\geq 0}$.
while *a stopping criterion is not met,* **do**
  Step 1: Obtain an approximate solution $\mathbf{x}^{k+1}$ of problem $\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$.
  Step 2: Set $\mathbf{y}^{k+1} = \mathbf{y}^k + \alpha_k(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b})$ with $\alpha_k := \min\{\beta_k, v_k/\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|\}$.
  Step 3: Set $\mathbf{z}^{k+1} = \mathbf{z}^k + \gamma_k \max\{-\mathbf{z}^k/\beta_k, \mathbf{g}\left(\mathbf{x}^{k+1}\right)\}$ with $\gamma_k := \min\{\beta_k, w_k/\|[\mathbf{g}(\mathbf{x}^{k+1})]_+\|\}$.
Output: $\mathbf{x}^{k+1}$.

In Algorithm 1, we use damped stepsizes α_k and γ_k instead of a full stepsize β_k for the $\mathbf{y}$- and $\mathbf{z}$-updates in Steps 2 and 3, in order to ensure the boundedness of the $\mathbf{y}$- and $\mathbf{z}$-iterates. The main cost of Algorithm 1 is in computing $\mathbf{x}^{k+1}$ in Step 1. By "an approximate solution", we mean that $\mathbf{x}^{k+1}$ is either a near-stationary point of problem $\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$ or close to a near-stationary point. Our analysis given later indicates that our DPALM framework would work if f is weakly convex and the subgradient of f or its smoothed version is uniformly bounded on $\mathcal{X}$. However, we will assume a certain structure on f such that some FOM can be applied to efficiently compute $\mathbf{x}^{k+1}$. More precisely, we consider three cases of f: (i) f is smooth; (ii) f is a composition of a convex function l with a smooth mapping $\mathbf{c}$, i.e., $f = l \circ \mathbf{c}$; (iii) f is a general weakly-convex function. Details will be specified in Sect. 3.

Throughout this paper, we use the following setting for β_k. Choose some $\beta_0 > 0$ and let

$$\beta_k = \beta_0\sqrt{k+1}, \forall\, k \geq 0. \tag{1.3}$$

This setting is inspired from [27]. However, Algorithm 1 will work for any other increasing sequence $\{\beta_k\}$.

1.2. Related Work. Significant efforts on FOMs for non-convex optimization have been dedicated to unconstrained or simple-constrained settings, as evidenced by a notable body of research [1, 8–13, 23, 37, 52]. For problem (P), these methods are inapplicable or inefficient as projecting onto the constraint set of (P) can be prohibitively expensive. Also, many existing FOMs for affine and/or nonlinear functional constrained optimization deal with the convex case; see, for example, [2, 3, 21, 22, 28–30, 33, 36, 40, 42, 44, 45] for a deterministic case and [24, 41, 43, 48] for a stochastic case. Below we review existing FOMs for solving non-convex optimization in the form of (P).

FOMs for solving affinely constrained non-convex optimization, i.e., problem (P) with $\mathbf{g} \equiv \mathbf{0}$, have been studied extensively such as in [14–16,18,31,32,50,51]. Hajinezhad and Hong [15] introduce a perturbed proximal primal-dual algorithm (PProx-PDA) with a complexity result of $\widetilde{\mathcal{O}}(\varepsilon^{-4})$[1]. The FOM in [32] is based on the inexact proximal accelerated augmented Lagrangian (IPAAL) method. It can produce an ε-KKT point within $\widetilde{\mathcal{O}}\left(\varepsilon^{-3}\right)$ iterations. Kong et al. [19] give an FOM based on a quadratic penalty accelerated inexact proximal point method and show a complexity result of $\mathcal{O}\left(\varepsilon^{-3}\right)$. On a special class of affine-constrained non-convex optimization, where the regularizer in the objective is the indicator function of a polyhedral set, Zhang and Luo [50, 51] introduce an FOM based on a proximal alternating direction method of multipliers (ADMM) and show that their method can generate an ε-KKT solution within $\mathcal{O}\left(\varepsilon^{-2}\right)$ iterations. The methods in [14, 16, 18, 31] are all variants of ADMM and can produce an ε-KKT solution within $O(\varepsilon^{-2})$ iterations under a certain assumption about the matrices in the affine constraint.

For regularized non-convex smooth optimization with convex nonlinear constraints, Li et al. [26] design an FOM, called HiAPeM, by applying a hybrid of ALM and a quadratic penalty method to a sequence of proximal point subproblems. Under Slater's condition, HiAPeM is able to produce an ε-KKT point with complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-2.5})$. It is demonstrated in [26] that more frequent use of ALM can yield better empirical performance. However, obtaining the $\widetilde{\mathcal{O}}(\varepsilon^{-2.5})$ complexity result requires to use the quadratic penalty method more frequently. This is different from our proposed DPALM, which is solely ALM based and can yield better practical performance. Kong et al. [20] consider a convex cone-constrained regularized non-convex smooth optimization problem. With an appropriate convex cone, the problem considered in [20] can have the same constraints as those in our considered problem (P). However, the objective function in [20] is a special case of what we consider. Similar to our method, the FOM in [20], called NL-IAPIAL, is also based on the proximal ALM framework. Compared to our method, NL-IAPIAL increases the penalty parameter much faster, which is doubled once a condition holds; see Eqn. (35) in [20]. Another key difference from our method is that NL-IAPIAL always uses β_k as the dual stepsize. Due to these differences, NL-IAPIAL has a higher complexity than ours. It requires $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ first-order oracles to produce an ε-KKT point.

For non-convex optimization with non-convex constraints, Sahin et al. [39] design an ALM-based FOM. Under a regularity condition that ensures near primal feasibility at a near stationary point of an ALM subproblem, their FOM can produce an ε-KKT point by $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ calls to the first-order oracle. The two works [27] and [25] assume the same regularity condition as that in [39], but differently their FOMs achieve an $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ complexity result, by adopting the framework of a proximal point penalty method and ALM respectively. The FOM in [25] is also analyzed for problems with convex constraints, in which case its complexity becomes $\widetilde{\mathcal{O}}(\varepsilon^{-\frac{5}{2}})$. Without a regularity condition but instead assuming a feasible initial point, the method in [27] achieves a complexity result of $\widetilde{\mathcal{O}}(\varepsilon^{-4})$. For solving non-smooth weakly convex problems with convex or weakly convex nonlinear constraint, Huang and Lin propose a single-loop switching subgradient method in [17]. They introduce a switching stepsize rule to accompany the switching subgradient and show that their method can find a near ε-KKT point in $\mathcal{O}(\varepsilon^{-4})$ iterations, by assuming a (uniform) Slater's condition. Curtis and Overton [6] give a method based on the sequential quadratic programming (SQP) for solving problems where the objective and constraints are locally Lipschitz and continuously differentiable. Global convergence to stationarity is shown.

The two works [11] and [49] are most closely related to ours, though their considered problems are special cases of (P). In [11], Drusvyatskiy and Paquette consider the regularized compositional optimization in the form of $\min_{\mathbf{x}} F(\mathbf{x}) := l(\mathbf{c}(\mathbf{x})) + r(\mathbf{x})$, where l and r are closed convex functions, and $\mathbf{c}$ is a smooth

[1] Throughout this paper, $\widetilde{\mathcal{O}}$ hides a logarithmic term.

mapping. A Moreau-envelope based smoothing prox-linear method is analyzed in [11], and it reaches a complexity result of $\tilde{\mathcal{O}}(\varepsilon^{-3})$ to produce a near ε-stationary point. We notice that the constraint in (P) can be encoded into the objective by adding an indicator composed function $\iota_{\{\mathbf{0}\}}(\mathbf{A}\mathbf{x}-\mathbf{b})+\iota_{\mathbb{R}^m_-}(\mathbf{g}(\mathbf{x}))$. By Moreau-envelope smoothing, i.e., replacing the indicator functions $\iota_{\{\mathbf{0}\}}(\cdot)$ and $\iota_{\mathbb{R}^m_-}(\cdot)$ by their Moreau envelope (see Definition 1.1), the composed function can be smoothed to $\frac{1}{2\nu}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\frac{1}{2\nu}\|[\mathbf{g}(\mathbf{x})]_+\|^2$, where $\nu>0$ is a smoothing parameter. Hence, the smoothing prox-linear method in [11] can be applied to (P) with a regularized compositional objective as we consider in Sect. 3.2, but it is based on the quadratic penalty of the constraints. In contrast, our method is based on the proximal ALM framework and can perform significantly better; see the experimental results in Sect. 4.3. Zeng et al. [49] introduce a Moreau Envelope Augmented Lagrangian Method (MEAL) for problems with a weakly-convex objective and linear constraints. MEAL can be viewed as a gradient update method on the Moreau envelope of the AL function. Assuming an exact solution of each proximal ALM subproblem, MEAL can produce an ε-KKT point within $O(\varepsilon^{-2})$ outer iterations, when either an implicit Lipschitz subgradient property or an implicit bounded subgradient property on the objective function holds. An inexact version, named iMEAL, is also given in [49], and it only requires an ε_k-stationary solution for the k-th proximal ALM subproblem. The same-order outer iteration complexity results are shown for iMEAL, provided that $\sum_{k=0}^{\infty}\varepsilon_k^2<\infty$. When the objective function is composite, i.e., a smooth term plus a convex regularizer, [49] also presents a linearized variant of MEAL, which has the same-order outer iteration complexity as MEAL and iMEAL. The (inexact) MEAL becomes an (inexact) proximal ALM when its stepsize $\eta=1$; see the updates in Eqn. (5) and Eqn. (7) of [49]. On the special linear-constrained case, iMEAL can achieve the same-order outer iteration complexity result as our proposed DPALM. However, it needs to set the penalty parameter to $\beta_k=\Theta(\varepsilon^{-2}),\forall\, k\geq 0$ when the objective function satisfies an implicit bounded subgradient property. In contrast, our proposed DPALM only needs to set $\beta_k=\Theta(\sqrt{k+1}),\forall\, k\geq 0$, which increases to $\Theta(\varepsilon^{-1})$ eventually to produce a (near) ε-KKT point. Though an overall first-order oracle complexity result is not explicitly shown in [49] for iMEAL, due to the higher-order penalty parameters, it will be higher than our complexity by at least an order of $\varepsilon^{-\frac{1}{2}}$ if iMEAL applies the same first-order subroutine as our method.

1.3. Contributions. Our contributions lie in both algorithm design and complexity analysis. They are summarized as follows.

(i) We propose a damped proximal augmented Lagrangian method (DPALM) to solve problems in the form of (P), which has a weakly-convex objective and linear and/or nonlinear convex constraints. At each iteration of DPALM, a strongly convex subproblem is formed by adding a proximal term with an appropriate proximal parameter to the AL function. The primal variable is updated to a desired-accurate solution of the subproblem, and the dual variables (or Lagrangian multipliers) are then updated by performing a dual gradient ascent step but with damped stepsizes instead of using the penalty parameter as a full stepsize. The damped dual stepsizes are important to ensure the boundedness of the dual iterates and further enable us to have guaranteed convergence, even when the objective is non-smooth non-convex.

(ii) We show that under Slater's condition, for any given $\varepsilon>0$, DPALM can produce a near ε-KKT point of problem (P) (see Definition 1.3) within $O(\varepsilon^{-2})$ outer iterations, when f in (P) is weakly convex and has uniform boundedness on its subgradients, and if for each DPALM subproblem, a near-optimal solution is found with a desired accuracy. This result generalizes that in [49] from an affine-constrained problem to an affine and/or convex functional constrained one. In addition, the value of the penalty parameter that we require is in a lower order than that in [49] to ensure a (near) ε-KKT point, under the non-smooth case. This leads our method to have a lower overall complexity.

(iii) For the case where f in (P) is smooth but may be non-convex, we apply Nesterov's APG method to find a near stationary solution of each DPALM subproblem and establish an $\widetilde{\mathcal{O}}(\varepsilon^{-\frac{5}{2}})$ complexity result to produce an ε-KKT point. This result improves the $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ complexity obtained in [20] for a proximal ALM based FOM. It matches the complexity results in [26] and [27] for either an ALM-penalty-hybrid method or a quadratic penalty based FOM, which often yields worse empirical performance than our proximal ALM based method as demonstrated in Sect. 4.

(iv) For the case where f is in a compositional form, we apply Nesterov's APG method to a Morean-envelope smoothed version of each DPALM subproblem and establish an $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ complexity result to produce a near ε-KKT point. This result generalizes that in [11] for solving an unconstrained compositional problem. Though with an appropriate outer convex function and an appropriate inner vector function, the compositional term can encode the constraints of (P), the smoothing method in [11] will become a quadratic penalty based method and performs significantly worse than our proposed proximal ALM based method, as demonstrated in Sect. 4.3.

1.4. Notations and Definitions. The relative interior and boundary of a set $\mathcal{X}$ are denoted by $\mathrm{relint}(\mathcal{X})$ and $\mathrm{bd}(\mathcal{X})$, respectively. We use $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ to denote the normal cone of $\mathcal{X}$ at $\mathbf{x}$ and define $\mathcal{B}_r = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq r\}$ for some $r > 0$, where $\|\cdot\|$ denotes the Euclidean norm.

DEFINITION 1.1 (Weakly convex function and Moreau envelope [11]). *A function f is called ρ-weakly convex for some $\rho \geq 0$ if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex. For a ρ-weakly convex function f, its Moreau envelope and the proximal mapping for any $\nu \in (0, 1/\rho)$ are defined by $f_\nu(\mathbf{x}) := \min_{\mathbf{z}} \{f(\mathbf{z}) + \frac{1}{2\nu}\|\mathbf{z} - \mathbf{x}\|^2\}$ and $\mathrm{prox}_{\nu f}(\mathbf{x}) := \arg\min_{\mathbf{z}}\{f(\mathbf{z}) + \frac{1}{2\nu}\|\mathbf{z} - \mathbf{x}\|^2\}$, respectively.*

DEFINITION 1.2 (Subdifferential [5]). *For a locally Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, its subdifferential at $\mathbf{x}$ is $\partial f(\mathbf{x}) := \{\lim_{\mathbf{x}' \to \mathbf{x}} \nabla f(\mathbf{x}') : f \text{ is differentiable at } \mathbf{x}'\}$.*

DEFINITION 1.3 ((near) ε-KKT point). *Given $\varepsilon > 0$, a point $\mathbf{x}$ is an ε-KKT point of (P) if there are $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m_+$ such that*

$$\max\left\{\mathrm{dist}(\mathbf{0}, \partial F(\mathbf{x}) + J_{\mathbf{g}}(\mathbf{x})^\top \mathbf{z} + \mathbf{A}^\top \mathbf{y}),\ \sqrt{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \|[\mathbf{g}(\mathbf{x})]_+\|^2},\ \textstyle\sum_{i=1}^m |z_i g_i(\mathbf{x})|\right\} \leq \varepsilon, \tag{1.4}$$

where $J_{\mathbf{g}}(\mathbf{x})$ denotes the Jacobi matrix of $\mathbf{g}$ at $\mathbf{x}$. We say that $\bar{\mathbf{x}}$ is a near ε-KKT point of (P) if it is ε-close to an ε-KKT point $\mathbf{x}$, i.e., $\|\bar{\mathbf{x}} - \mathbf{x}\| \leq \varepsilon$.

1.5. Organization. The rest of this paper is organized as follows. Some preliminary results are shown in Sect. 2. Iteration complexity results are established in Sect. 3 for three cases of f, and Sect. 4 gives experimental results. Finally, the paper is concluded in Sect. 5.

2. Preliminary Analysis. In this section, we show some results that will be used in Sect. 3 to establish iteration complexity results of our FOMs. We first show the boundedness of the generated Lagrangian multipliers; see Lemma 2.1. Then we give a result that will be used to bound complementary slackness error; see Lemma 2.2. Finally, a bound on the cumulative change of the AL function along the iterates is shown; see Lemma 2.4. This result will be used to bound dual infeasibility.

Throughout the paper, we make the following assumptions.

ASSUMPTION 1. *In (P), the domain of F, denoted by $\mathcal{X} := \mathrm{dom}(F)$, is compact, and its diameter is denoted by $D = \max_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}} \|\mathbf{x}_1 - \mathbf{x}_2\| < \infty$. Also, f is ρ-weakly convex on $\mathcal{X}$ for some $\rho > 0$, h is closed convex, and F is bounded on $\mathcal{X}$, i.e., $\max_{\mathbf{x} \in \mathcal{X}} |F(\mathbf{x})| < \infty$.*

ASSUMPTION 2. *For each $i = 1, \ldots, m$, g_i in (P) is L_g-smooth on an open set $\mathcal{U}$ containing $\mathcal{X}$, i.e., $\|\nabla g_i(\mathbf{x}) - \nabla g_i(\mathbf{y})\| \leq L_g \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{U}$.*

ASSUMPTION 3. *There is* $\mathbf{x}_{\text{feas}} \in \text{relint}\,(\mathcal{X})$ *such that* $\mathbf{A}\mathbf{x}_{\text{feas}} = \mathbf{b}$ *and* $g_i(\mathbf{x}_{\text{feas}}) < 0, \forall\, i = 1, \ldots, m$.

Under Assumptions 1 and 2, there must exist a positive constant B_g such that

$$\max\left\{|g_i(\mathbf{x})|, \|\nabla g_i(\mathbf{x})\|\right\} \leq B_g, \forall\, \mathbf{x} \in \mathcal{X}, \forall\, i = 1, \ldots, m, \tag{2.1}$$

$$|g_i(\widehat{\mathbf{x}}) - g_i(\widetilde{\mathbf{x}})| \leq B_g \|\widehat{\mathbf{x}} - \widetilde{\mathbf{x}}\|, \forall\, \widehat{\mathbf{x}}, \widetilde{\mathbf{x}} \in \mathcal{X}, \forall\, i = 1, \ldots, m. \tag{2.2}$$

The next lemma shows the boundedness of $\{\mathbf{y}^k\}$ and $\{\mathbf{z}^k\}$.

LEMMA 2.1. *Under Assumptions 1–3, let* $\{\mathbf{x}^k\}$, $\{\mathbf{y}^k\}$, $\{\mathbf{z}^k\}$, $\{\alpha_k\}$, $\{\gamma_k\}$, $\{w_k\}$, *and* $\{v_k\}$ *be from Algorithm 1. Define* $C_{\mathbf{y}} := \|\mathbf{y}^0\| + \sum_{k=0}^{\infty} v_k$ *and* $C_{\mathbf{z}} := \left\|\mathbf{z}^0\right\| + \sum_{k=0}^{\infty} w_k$. *It then holds that* $\|\mathbf{y}^k\| \leq C_{\mathbf{y}}$ *and* $\|\mathbf{z}^k\| \leq C_{\mathbf{z}}$ *for all* $k \geq 0$. *In addition, the sequence* $\{\mathbf{z}^k\}$ *is nonnegative.*

Proof. From the updating rule of $\mathbf{y}^{k+1}$ and the definition of α_k given in Step 2 of Algorithm 1, we have $\left\|\mathbf{y}^{k+1}\right\| = \left\|\mathbf{y}^k + \alpha_k(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b})\right\| \leq \left\|\mathbf{y}^k\right\| + v_k$ for all $k \geq 0$. Summing up this inequality gives $\left\|\mathbf{y}^k\right\| \leq \|\mathbf{y}^0\| + \sum_{i=0}^{k} v_i \leq C_{\mathbf{y}}$ for all $k \geq 0$.

By the updating rule of $\mathbf{z}^{k+1}$, it holds $\mathbf{z}^{k+1} \geq (1 - \frac{\gamma_k}{\beta_k})\mathbf{z}^k$ for all $k \geq 0$. Using this relation recursively, we have that the sequence $\{\mathbf{z}^k\}$ is nonnegative from $\mathbf{z}^0 \geq \mathbf{0}$ and $\gamma_k \leq \beta_k, \forall\, k \geq 0$. To show the boundedness of $\{\mathbf{z}^k\}$, we denote

$$J_1^k := \left\{i : -z_i^k/\beta_k \geq g_i\left(\mathbf{x}^{k+1}\right)\right\}, J_2^k := \left\{i : -z_i^k/\beta_k < g_i\left(\mathbf{x}^{k+1}\right)\right\}. \tag{2.3}$$

Then it follows that

$$\begin{aligned}\left\|\mathbf{z}^{k+1}\right\|^2 &= \sum_{i \in J_1^k} \left(1 - \frac{\gamma_k}{\beta_k}\right)^2 \left(z_i^k\right)^2 + \sum_{i \in J_2^k} \left(z_i^k + \gamma_k g_i\left(\mathbf{x}^{k+1}\right)\right)^2 \leq \sum_{i \in J_1^k} \left(z_i^k\right)^2 + \sum_{i \in J_2^k} \left(z_i^k + \gamma_k \left[g_i\left(\mathbf{x}^{k+1}\right)\right]_+\right)^2 \\ &\leq \left\|\mathbf{z}^k\right\|^2 + \gamma_k^2 \left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2 + 2\gamma_k \left\|\mathbf{z}^k\right\| \cdot \left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|,\end{aligned}$$

where the first inequality is from $\gamma_k \leq \beta_k$ and $g_i\left(\mathbf{x}^{k+1}\right) \leq \left[g_i\left(\mathbf{x}^{k+1}\right)\right]_+$, and the last inequality holds by expanding the square term and combining the like terms. Hence, $\left\|\mathbf{z}^{k+1}\right\| \leq \left\|\mathbf{z}^k\right\| + \gamma_k \|[\mathbf{g}(\mathbf{x}^{k+1})]_+\| \leq \left\|\mathbf{z}^k\right\| + w_k$ by the definition of γ_k in Algorithm 1. Summing up this inequality gives $\left\|\mathbf{z}^k\right\| \leq \|\mathbf{z}^0\| + \sum_{i=0}^{k} w_i \leq C_{\mathbf{z}}$ for all $k \geq 0$. This completes the proof. □

In the rest of this section, we assume the following condition for some constant C_P:

$$\left\|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\right\|^2 + \left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2 \leq C_P^2/\beta_k^2, \ \forall\, k \geq 0, \tag{2.4}$$

which will be proved for the three cases considered in Sect. 3 with a detailed formula of C_P.

LEMMA 2.2. *Under Assumptions 1–3, let* $\{\mathbf{x}^k\}$ *and* $\{\mathbf{z}^k\}$ *be generated from Algorithm 1 such that* (2.4) *holds. Then for any* $k \geq 0$ *and any* $\mathbf{x} \in \mathcal{X}$, *it holds*

$$\textstyle\sum_{i=1}^m \left|[z_i^k + \beta_k g_i(\mathbf{x})]_+ g_i(\mathbf{x})\right| \leq C_{\mathbf{z}}^2/\beta_k + \frac{5\beta_k}{4} \sum_{i=1}^m [g_i(\mathbf{x})]_+^2, \tag{2.5}$$

where $C_{\mathbf{z}}$ *is defined in Lemma 2.1. In addition, it holds*

$$\textstyle\sum_{i=1}^m \left|[z_i^k + \beta_k g_i\left(\mathbf{x}^{k+1}\right)]_+ g_i\left(\mathbf{x}^{k+1}\right)\right| \leq (C_{\mathbf{z}}^2 + 5C_P^2/4)/\beta_k. \tag{2.6}$$

Proof. For a given $\mathbf{x}$, let $J_+ := \{i : g_i(\mathbf{x}) \geq 0\}$, $J_- := \{i : g_i(\mathbf{x}) < 0\}$, $J_3^k := \{i : -z_i^k/\beta_k \geq g_i(\mathbf{x})\}$, and $J_4^k := \{i : -z_i^k/\beta_k < g_i(\mathbf{x})\}$. We then have

$$\begin{aligned}
&\sum_{i=1}^{m} \left|[z_i^k + \beta_k g_i(\mathbf{x})]_+ g_i(\mathbf{x})\right| = \sum_{i\in J_4^k\cap J_+} \left(z_i^k + \beta_k g_i(\mathbf{x})\right) g_i(\mathbf{x}) - \sum_{i\in J_4^k\cap J_-} \left(z_i^k + \beta_k g_i(\mathbf{x})\right) g_i(\mathbf{x}) \\
\leq &\sum_{i\in J_4^k\cap J_+} \left(z_i^k g_i(\mathbf{x}) + \beta_k g_i^2(\mathbf{x})\right) + \sum_{i\in J_4^k\cap J_-} \left(z_i^k\right)^2/\beta_k \\
\leq &\sum_{i\in J_4^k\cap J_+} \frac{\left(z_i^k\right)^2}{\beta_k} + \frac{\beta_k}{4} \sum_{i\in J_4^k\cap J_+} g_i^2(\mathbf{x}) + \beta_k \sum_{i\in J_4^k\cap J_+} g_i^2(\mathbf{x}) + \sum_{i\in J_4^k\cap J_-} \frac{\left(z_i^k\right)^2}{\beta_k} \\
\leq &\frac{1}{\beta_k} \sum_{i\in J_4^k} \left(z_i^k\right)^2 + \frac{5\beta_k}{4} \sum_{i\in J_4^k\cap J_+} g_i^2(\mathbf{x}) \leq \frac{1}{\beta_k} \sum_{i=1}^{m} \left(z_i^k\right)^2 + \frac{5\beta_k}{4} \sum_{i=1}^{m} [g_i(\mathbf{x})]_+^2,
\end{aligned} \tag{2.7}$$

where the first inequality holds because $-(z_i^k + \beta_k g_i(\mathbf{x}))g_i(\mathbf{x}) \leq -z_i^k g_i(\mathbf{x}) \leq (z_i^k)^2/\beta_k$ for all $i \in J_4^k \cap J_-$, the second inequality is from Young's inequality, the third inequality is obtained by combining J_+ and J_-, and the last inequality holds by Lemma 2.1.

When $\mathbf{x} = \mathbf{x}^{k+1}$, we use (2.4) to further bound $\sum_{i=1}^m [g_i(\mathbf{x})]_+^2$ and complete the proof. □

The next lemma is proved in Appendix A. It will be used to show the cumulative change of the AL functions in Lemma 2.4.

LEMMA 2.3. *Under Assumption 1-3, let $\{\mathbf{x}^k\}$, $\{\mathbf{y}^k\}$, $\{\mathbf{z}^k\}$, $\{w_k\}$, and $\{\gamma_k\}$ be from Algorithm 1 such that (2.4) holds. Then for any integer $K > 0$, it holds that*

$$\sum_{k=0}^{K-1} \left(\frac{\beta_{k+1}}{2} \left\| \left[\mathbf{g}\left(\mathbf{x}^{k+1}\right) + \frac{\mathbf{z}^{k+1}}{\beta_{k+1}} \right]_+ \right\|^2 - \frac{\beta_k}{2} \left\| \left[\mathbf{g}\left(\mathbf{x}^{k+1}\right) + \frac{\mathbf{z}^{k}}{\beta_{k}} \right]_+ \right\|^2 \right) \leq \frac{1}{4\beta_0}(3C_P^2 + 6C_{\mathbf{z}}C_P + 7C_{\mathbf{z}}^2), \tag{2.8}$$

and

$$\sum_{k=0}^{K-1} \left\langle \mathbf{y}^{k+1} - \mathbf{y}^k, \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} \right\rangle \leq \frac{C_{\mathbf{y}}C_P}{\beta_0}, \quad \sum_{k=0}^{K-1} \frac{\beta_{k+1} - \beta_k}{2} \left\|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\right\|^2 \leq \frac{3C_P^2}{4\beta_0}, \tag{2.9}$$

where $C_{\mathbf{z}}$ is given in Lemma 2.1, and C_P is the constant in (2.4).

LEMMA 2.4. *Under the same assumptions of Lemma 2.3, it holds for all integers $K > \widetilde{K} \geq 0$ that*

$$\textstyle\sum_{k=\widetilde{K}}^{K-1} \left(\mathcal{L}_{\beta_k}(\mathbf{x}^k; \mathbf{y}^k, \mathbf{z}^k) - \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k)\right) \leq C_{\mathbf{x}}, \tag{2.10}$$

where

$$\begin{aligned}
C_{\mathbf{x}} := {} & 2\max_{\mathbf{x}\in\mathcal{X}} |F(\mathbf{x})| + \tfrac{1}{4\beta_0}(14C_P^2 + 6C_{\mathbf{z}}C_P + 12C_{\mathbf{y}}C_P + 11C_{\mathbf{z}}^2) \\
& + \tfrac{\beta_0}{2} \left(\|\mathbf{A}\mathbf{x}^0 - \mathbf{b}\|^2 + \left\| \left[g(\mathbf{x}^0) + \tfrac{\mathbf{z}^0}{\beta_0} \right]_+ \right\|^2 \right) + \left| \langle \mathbf{y}^0, \mathbf{A}\mathbf{x}^0 - \mathbf{b} \rangle \right|
\end{aligned}$$

with $C_{\mathbf{z}}$ and $C_{\mathbf{y}}$ given in Lemma 2.1 and C_P given in (2.4).

Proof. From the definition of $\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k)$, it follows that

$$
\begin{aligned}
&\textstyle\sum_{k=\widetilde{K}}^{K-1}\left(\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k)-\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k)\right)\\
&= F(\mathbf{x}^{\widetilde{K}})-F\left(\mathbf{x}^K\right)+\left\langle \mathbf{y}^{\widetilde{K}},\mathbf{A}\mathbf{x}^{\widetilde{K}}-\mathbf{b}\right\rangle-\left\langle \mathbf{y}^{K-1},\mathbf{A}\mathbf{x}^K-\mathbf{b}\right\rangle+\underbrace{\sum_{k=\widetilde{K}}^{K-2}\left\langle \mathbf{y}^{k+1}-\mathbf{y}^k,\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\right\rangle}_{\textbf{term 1}}\\
&\quad+\underbrace{\sum_{k=\widetilde{K}}^{K-2}\frac{\beta_{k+1}-\beta_k}{2}\left\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\right\|^2}_{\textbf{term 2}}+\underbrace{\sum_{k=\widetilde{K}}^{K-2}\left(\frac{\beta_{k+1}}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)+\frac{\mathbf{z}^{k+1}}{\beta_{k+1}}\right]_+\right\|^2-\frac{\beta_k}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)+\frac{\mathbf{z}^{k}}{\beta_{k}}\right]_+\right\|^2\right)}_{\textbf{term 3}}\\
&\quad+\underbrace{\frac{\beta_{\widetilde{K}}}{2}\|\mathbf{A}\mathbf{x}^{\widetilde{K}}-\mathbf{b}\|^2+\frac{\beta_{\widetilde{K}}}{2}\left\|\left[g(\mathbf{x}^{\widetilde{K}})+\frac{\mathbf{z}^{\widetilde{K}}}{\beta_{\widetilde{K}}}\right]_+\right\|^2}_{\textbf{term 4}}-\frac{\beta_{K-1}}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{K}\right)+\frac{\mathbf{z}^{K-1}}{\beta_{K-1}}\right]_+\right\|^2-\frac{\beta_{K-1}}{2}\|\mathbf{A}\mathbf{x}^K-\mathbf{b}\|^2\\
&\le 2\max_{\mathbf{x}\in\mathcal{X}}|F(\mathbf{x})|+\left|\langle \mathbf{y}^0,\mathbf{A}\mathbf{x}^0-\mathbf{b}\rangle\right|+\frac{C_{\mathbf{y}}C_P}{\beta_0}+\frac{C_{\mathbf{y}}C_P}{\beta_0}+\frac{C_{\mathbf{y}}C_P}{\beta_0}+\frac{3C_P^2}{4\beta_0}+\frac{1}{4\beta_0}(3C_P^2+6C_{\mathbf{z}}C_P+7C_{\mathbf{z}}^2)\\
&\qquad+\frac{\beta_0}{2}\|\mathbf{A}\mathbf{x}^0-\mathbf{b}\|^2+\frac{\beta_0}{2}\left\|\left[g(\mathbf{x}^0)+\frac{\mathbf{z}^0}{\beta_0}\right]_+\right\|^2+\frac{2C_P^2}{\beta_0}+\frac{C_{\mathbf{z}}^2}{\beta_0}.
\end{aligned}
$$

Below we explain how the second inequality is obtained. By Lemma 2.1 and (2.4), we have $\langle -\mathbf{y}^{K-1},\mathbf{A}\mathbf{x}^K-\mathbf{b}\rangle\le\|\mathbf{y}^{K-1}\|\|\mathbf{A}\mathbf{x}^K-\mathbf{b}\|\le C_{\mathbf{y}}C_P/\beta_0$. Similarly, $\left\langle \mathbf{y}^{\widetilde{K}},\mathbf{A}\mathbf{x}^{\widetilde{K}}-\mathbf{b}\right\rangle\le\left|\langle \mathbf{y}^0,\mathbf{A}\mathbf{x}^0-\mathbf{b}\rangle\right|+C_{\mathbf{y}}C_P/\beta_0$ by discussing the cases of $\widetilde{K}=0$ or $\widetilde{K}>0$. **term 1** and **term 2** are bounded by using (2.9); **term 3** is bounded by using (2.8); **term 4** is upper bounded by $2C_P^2/\beta_0+C_{\mathbf{z}}^2/(\beta_0)$ for $\widetilde{K}>0$ from the definition of β_k in (1.3), the bound in (2.4), and $\|[\mathbf{a}+\mathbf{b}]_+\|^2\le\|\mathbf{a}+\mathbf{b}\|^2\le 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$. Adding all the obtained upper bounds and simplifying the summation gives the desired result. □

3. Iteration Complexity Results for Three Cases. In this section, we assume a certain structure on f in (P) and specify how to compute $\mathbf{x}^{k+1}$ in Algorithm 1 so that the condition in (2.4) is satisfied. More specifically, with the assumed structure, we are able to apply Nesterov's APG method [35] to either directly solve $\min_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$ or its Moreau-envelope based smoothed version, or we can apply a subgradient method to solve $\min_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$. Following this, we will present a result of dual infeasibility and thus obtain the iteration complexity to produce a (near) ε-KKT point, by also using (2.4) and Lemma 2.2. We will consider three cases of f: smooth, or compositional, or general weakly convex, and h is always assumed to satisfy the following conditions.

ASSUMPTION 4. *There exists $r_h>0$ such that $\partial h(\mathbf{x})\neq\phi$, and $\partial h(\mathbf{x})\subseteq\mathcal{N}_{\mathcal{X}}(\mathbf{x})+\mathcal{B}_{r_h},\forall\,\mathbf{x}\in\mathcal{X}$.*

3.1. Regularized Smooth Objective. We first consider the case where f is smooth.

ASSUMPTION 5. *In (P), f is L_f-smooth in an open set that contains $\mathcal{X}$.*

Under Assumptions 1 and 5, there must exist a positive constant B_f such that $\|\nabla f(\mathbf{x})\|\le B_f,\forall\,\mathbf{x}\in\mathcal{X}$. At the k-th iteration of Algorithm 1, we directly apply Nesterov's APG method, i.e., Algorithm 2 in Appendix B,

to the following subproblem

$$\min_{\mathbf{x}\in\mathcal{X}} \ \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k) = \widetilde{f}^k(\mathbf{x}) + \widetilde{h}^k(\mathbf{x}), \tag{3.1}$$

where $\widetilde{\mathcal{L}}_{\beta_k}$ is defined in (1.2), and we set

$$\widetilde{f}^k(\mathbf{x}) = f(\mathbf{x}) + \frac{\rho}{2}\left\|\mathbf{x}-\mathbf{x}^k\right\|^2 + (\mathbf{y}^k)^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + \frac{\beta_k}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \frac{\beta_k}{2}\left\|\left[\mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}^k}{\beta_k}\right]_+\right\|^2 - \frac{\|\mathbf{z}^k\|^2}{2\beta_k}, \tag{3.2}$$

$$\widetilde{h}^k(\mathbf{x}) = h(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2. \tag{3.3}$$

Given $\varepsilon > 0$, we compute $\mathbf{x}^{k+1}$ by Algorithm 2 as an ε_k-stationary point of subproblem (3.1), i.e.,

$$\operatorname{dist}\left(\mathbf{0}, \partial_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k)\right) \le \varepsilon_k := \min\left\{\frac{\varepsilon}{8}, \sqrt{\frac{\rho}{2\beta_k}}\right\}, \forall\, k \ge 0. \tag{3.4}$$

Then we have the following lemma that shows (2.4) with a specified C_P.

LEMMA 3.1. *Under Assumptions 1–5, let $\mathbf{x}^{k+1}$ satisfy (3.4). Then the condition in (2.4) holds with*

$$C_P := 2\left(\sqrt{C_{\mathbf{y}}^2 + C_{\mathbf{z}}^2} + \sqrt{Q^2/\min_i|g_i^2(\mathbf{x}_{\rm feas})| + Q^2\|(\mathbf{A}\mathbf{A}^\top)^\dagger\mathbf{A}\|^2 C_1^2}\right) + 1, \tag{3.5}$$

$$Q := D(B_f + 2\rho D + r_h), \qquad C_1 := 1/D + 1/\operatorname{dist}(\mathbf{x}_{\rm feas}, \operatorname{bd}(\mathcal{X})) + B_g/\min_i|g_i(\mathbf{x}_{\rm feas})|. \tag{3.6}$$

Here, $C_{\mathbf{y}}, C_{\mathbf{z}}$ are given in Lemma 2.1, and $(\mathbf{A}\mathbf{A}^\top)^\dagger$ denotes the pseudo inverse of $\mathbf{A}\mathbf{A}^\top$.

Proof. Consider the following problem

$$\min_{\mathbf{x}} \ F(\mathbf{x}) + \rho\left\|\mathbf{x}-\mathbf{x}^k\right\|^2, \quad \text{s.t. } \mathbf{A}\mathbf{x}=\mathbf{b},\ \mathbf{g}(\mathbf{x}) \le \mathbf{0}. \tag{3.7}$$

Since F is ρ-weakly convex, the objective function in (3.7) is ρ-strongly convex. In addition, the constraints are convex. Thus (3.7) has a unique solution $\bar{\mathbf{x}}_*^k$, and under Assumption 2, there must exist a multiplier $\bar{\mathbf{p}}_*^k = (\bar{\mathbf{y}}_*^k, \bar{\mathbf{z}}_*^k)$ corresponding to $\bar{\mathbf{x}}_*^k$. Define $\bar{\mathbf{y}}^{k+1} := \mathbf{y}^k + \beta_k\left(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\right)$, $\bar{\mathbf{z}}^{k+1} := \left[\mathbf{z}^k + \beta_k\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+$, $\bar{\mathbf{p}}^{k+1} := \left(\bar{\mathbf{y}}^{k+1}, \bar{\mathbf{z}}^{k+1}\right)$, and $\mathbf{p}^k = \left(\mathbf{y}^k, \mathbf{z}^k\right)$. We have

$$\left\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\right\|^2 + \left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2 \le \frac{1}{\beta_k^2}\left\|\bar{\mathbf{p}}^{k+1}-\mathbf{p}^k\right\|^2 \le \frac{1}{\beta_k^2}\left(\left\|\bar{\mathbf{p}}^{k+1}-\bar{\mathbf{p}}_*^k\right\| + \left\|\bar{\mathbf{p}}_*^k-\mathbf{p}^k\right\|\right)^2. \tag{3.8}$$

By the ρ-strong convexity of $\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$ and (3.4), it holds from [26, Remark 1] that

$$\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) \le \min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k) + \frac{\varepsilon_k^2}{\rho}. \tag{3.9}$$

Hence, from Lemma A.1, it follows that

$$\left\|\bar{\mathbf{p}}^{k+1}-\bar{\mathbf{p}}_*^k\right\|^2 \le \left\|\mathbf{p}^k-\bar{\mathbf{p}}_*^k\right\|^2 + \frac{2\beta_k\varepsilon_k^2}{\rho}. \tag{3.10}$$

Moreover, using [27, Lemma 3] to our case gives $\|\bar{\mathbf{z}}_*^k\| \le \frac{Q}{\min_i|g_i(\mathbf{x}_{\rm feas})|}$, $\|\bar{\mathbf{y}}_*^k\| \le Q\|(\mathbf{A}\mathbf{A}^\top)^\dagger\mathbf{A}\|C_1$, where Q and C_1 are defined in (3.6).

Noticing $\|\mathbf{p}^k\| = \sqrt{\|\mathbf{y}^k\|^2 + \|\mathbf{z}^k\|^2} \le \sqrt{C_{\mathbf{y}}^2 + C_{\mathbf{z}}^2}$, we obtain that

$$\|\bar{\mathbf{p}}_*^k - \mathbf{p}^k\| \le \sqrt{C_{\mathbf{y}}^2 + C_{\mathbf{z}}^2} + \sqrt{\frac{Q^2}{(\min_i|g_i(\mathbf{x}_{\rm feas})|)^2} + Q^2\|(\mathbf{A}\mathbf{A}^\top)^\dagger\mathbf{A}\|^2 C_1^2},$$

which together with (3.8) and (3.10) implies

$$\sqrt{\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|^2 + \|[\mathbf{g}(\mathbf{x}^{k+1})]_+\|^2} \le \frac{1}{\beta_k}\left(2\|\bar{\mathbf{p}}_*^k - \mathbf{p}^k\| + \varepsilon_k\sqrt{\frac{2\beta_k}{\rho}}\right) \overset{(3.4)}{\le} \frac{1}{\beta_k}\left(2\|\bar{\mathbf{p}}_*^k - \mathbf{p}^k\| + 1\right) \overset{(3.5)}{\le} \frac{C_P}{\beta_k}.$$

This completes the proof. □

The following lemma shows how to achieve a desired bound on dual infeasibility.

LEMMA 3.2. *Given $\varepsilon > 0$, under Assumptions 1-5, let $\{\mathbf{x}^k\},\{\mathbf{y}^k\},\{\mathbf{z}^k\}$ be generated by Algorithm 1 such that the condition in (3.4) is satisfied. Then for $K_2 := \lceil 5C_{\mathbf{x}}\rho\varepsilon^{-2}\rceil$ and any integer $\widetilde{K}_1 \ge 0$, where $C_{\mathbf{x}}$ is defined in Lemma 2.4, it must hold that* $\min_{\widetilde{K}_1\le k\le \widetilde{K}_1+K_2-1} \operatorname{dist}(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k)) \le \varepsilon$.

Proof. Denote $\mathbf{x}_*^{k+1} = \arg\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$. By (3.4), there is $\xi \in \partial_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k)$ such that $\|\xi\| \le \varepsilon_k$. Then we have

$$\begin{aligned}
&\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) + \rho\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 = \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) \le \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}_*^{k+1};\mathbf{y}^k,\mathbf{z}^k) + \frac{\varepsilon_k^2}{\rho}\\
\le&\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \frac{\rho}{2}\|\mathbf{x}_*^{k+1}-\mathbf{x}^k\|^2 + \frac{\varepsilon_k^2}{\rho} \qquad (3.11)\\
=&\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \frac{\rho}{2}\|\mathbf{x}_*^{k+1}-\mathbf{x}^k\|^2 + \frac{\varepsilon_k^2}{\rho} \le \mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \frac{\rho}{4}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2 + \frac{\rho}{2}\|\mathbf{x}^{k+1}-\mathbf{x}_*^{k+1}\|^2 + \frac{\varepsilon_k^2}{\rho}\\
\le&\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \frac{\rho}{4}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2 + \frac{\rho}{2}\frac{\|\xi\|^2}{\rho^2} + \frac{\varepsilon_k^2}{\rho} \le \mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \frac{\rho}{4}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2 + \frac{3\varepsilon_k^2}{2\rho},
\end{aligned}$$

where the first, second and fourth inequalities follow from the ρ-strong convexity of $\widetilde{\mathcal{L}}_{\beta_k}$, and the third inequality is by the triangle inequality, and the last inequality holds since $\|\xi\| \le \varepsilon_k$.

Summing up the above inequality over $k = \widetilde{K}_1,\ldots,\widetilde{K}_1+K_2-1$ gives

$$\frac{5\rho}{4}\sum_{k=\widetilde{K}_1}^{\widetilde{K}_1+K_2-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \le \sum_{k=\widetilde{K}_1}^{\widetilde{K}_1+K_2-1}\left(\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) + \frac{3\varepsilon_k^2}{2\rho}\right). \qquad (3.12)$$

Since $\varepsilon_k \le \varepsilon/8, \forall\, k \ge 0$, it holds that $\frac{4}{5\rho K_2}\sum_{k=\widetilde{K}_1}^{\widetilde{K}_1+K_2-1}\frac{3\varepsilon_k^2}{2\rho} \le \varepsilon^2/(32\rho^2)$ for all $\widetilde{K}_1 \ge 0$. Hence, we have from Lemma 2.4 and (3.12) that $\min_{\widetilde{K}_1\le k\le\widetilde{K}_1+K_2-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\| \le \sqrt{4C_{\mathbf{x}}/(5\rho K_2) + \varepsilon^2/(32\rho^2)}$. Now notice $\partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) = \partial_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) - 2\rho(\mathbf{x}^{k+1}-\mathbf{x}^k)$. We have

$$\min_{\widetilde{K}_1\le k\le\widetilde{K}_1+K_2-1}\operatorname{dist}\left(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k)\right)$$

$$= \min_{\widetilde{K}_1 \le k \le \widetilde{K}_1+K_2-1} \operatorname{dist}\left(\mathbf{0}, \partial_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k) - 2\rho\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)\right)$$
$$\overset{(3.4)}{\le} \min_{\widetilde{K}_1 \le k \le \widetilde{K}_1+K_2-1} \left(\varepsilon_k + 2\rho\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|\right) \le \varepsilon/8 + 2\rho\sqrt{4C_{\mathbf{x}}/(5\rho K_2) + \varepsilon^2/(32\rho^2)} \le \varepsilon,$$

where the last inequality holds because $K_2 \ge 5C_{\mathbf{x}}\rho\varepsilon^{-2}$. This completes the proof. □

Now we are ready to show the outer iteration complexity of Algorithm 1 when f satisfies Assumption 5. The result is given in the theorem below.

THEOREM 3.3 (Outer iteration complexity result I). *Given $\varepsilon > 0$, under Assumptions 1-5, let $\{\mathbf{x}^k\}$, $\{\mathbf{y}^k\}$, and $\{\mathbf{z}^k\}$ be generated by Algorithm 1 such that (3.4) holds. Then for some $k < K = K_1 + K_2$, $\mathbf{x}^{k+1}$ is an ε-KKT point of problem (P), where $K_1 := \left\lceil \max\left\{\frac{C_P^2}{\beta_0^2\varepsilon^2}, \frac{(4C_{\mathbf{z}}+5C_P^2)^2}{16\beta_0^2\varepsilon^2}\right\}\right\rceil$, $K_2 := \left\lceil 5C_{\mathbf{x}}\rho\varepsilon^{-2}\right\rceil$, C_P is given in (3.5), $C_{\mathbf{z}}$ is given in Lemma 2.1, and $C_{\mathbf{x}}$ is defined in Lemma 2.4.*

Proof. From Lemma 3.1, we have (2.4) and thus by (1.3), it holds

$$\sqrt{\|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\|^2 + \left\|[\mathbf{g}(\mathbf{x}^{k+1})]_+\right\|^2} \le \frac{C_P}{\beta_{K_1}} = \frac{C_P}{\beta_0\sqrt{K_1+1}} \le \varepsilon, \forall\, k \ge K_1, \tag{3.13}$$

where the second inequality holds because $K_1 \ge \frac{C_P^2}{\beta_0^2\varepsilon^2}$. Denote $\bar{\mathbf{y}}^k := \mathbf{y}^k + \beta_k(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b})$, and $\bar{\mathbf{z}}^k := [\mathbf{z}^k + \beta_k\mathbf{g}(\mathbf{x}^{k+1})]_+$. Then by (1.3) and (2.6), it holds

$$\sum_{i=i}^{m} |\bar{z}_i^k g_i(\mathbf{x}^{k+1})| \le \frac{1}{\beta_{K_1}}\left(C_{\mathbf{z}}^2 + \frac{5C_P^2}{4}\right) = \frac{1}{\beta_0\sqrt{K_1+1}}\left(C_{\mathbf{z}}^2 + \frac{5C_P^2}{4}\right) \le \varepsilon, \forall\, k \ge K_1, \tag{3.14}$$

where we have used $K_1 \ge \frac{(4C_{\mathbf{z}}+5C_P^2)^2}{16\beta_0^2\varepsilon^2}$ to obtain the second inequality.

Moreover, notice $\partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k) = \partial_{\mathbf{x}}\mathcal{L}_0(\mathbf{x}^{k+1}; \bar{\mathbf{y}}^k, \bar{\mathbf{z}}^k)$. Thus letting $\widetilde{K}_1 = K_1$ in Lemma 3.2 yields

$$\min_{K_1 \le k \le K_1+K_2-1} \operatorname{dist}\left(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_0(\mathbf{x}^{k+1}; \bar{\mathbf{y}}^k, \bar{\mathbf{z}}^k)\right) \le \varepsilon. \tag{3.15}$$

Now let $k' = \underset{K_1 \le k \le K_1+K_2-1}{\arg\min} \operatorname{dist}\left(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_0(\mathbf{x}^{k+1}; \bar{\mathbf{y}}^k, \bar{\mathbf{z}}^k)\right)$. We conclude from (3.13), (3.14), and (3.15) that $\mathbf{x}^{k'+1}$ is an ε-KKT point of problem P with multipliers $\bar{\mathbf{y}}^{k'}$ and $\bar{\mathbf{z}}^{k'}$ by Definition 1.3. □

To obtain the total complexity of Algorithm 1 for solving problem (P), we still need to evaluate the number of inner iterations for solving subproblem (3.1), by using Algorithm 2 as the subroutine, such that (3.4) is met. From [26, Eqn. (3.10)], it follows that $\widetilde{f}^k$ in (3.2) is $L_{\widetilde{f}^k}$-smooth with $L_{\widetilde{f}^k} = L_f + \rho + \sqrt{m}L_g C_{\mathbf{z}} + \sqrt{k+1}\beta_0(\|\mathbf{A}^\top\mathbf{A}\| + mB_g(B_g + L_g))$. In addition, because f is ρ-weakly convex, $\widetilde{f}^k$ is convex. Also, $\widetilde{h}^k$ is ρ-strongly convex. Hence, we have the following lemma directly from Theorem B.1.

LEMMA 3.4. *Under Assumptions 1-5 and for a given $\varepsilon_k > 0$, Algorithm 2 with $\gamma_u = 2$ applied to subproblem (3.1) can find a solution $\mathbf{x}^{k+1}$ that satisfies the criteria in (3.4) within*

$$T_1^k = \left\lceil \max\left\{\frac{1}{\log 2}, 2\sqrt{\frac{L_{\widetilde{f}^k}}{\rho}}\right\} \log\frac{9DL_{\widetilde{f}^k}\sqrt{L_{\widetilde{f}^k}/\rho}}{\varepsilon_k}\right\rceil + 1 \tag{3.16}$$

iterations, where $L_{\widetilde{f}^k} := C_3 + \sqrt{k+1}\beta_0 C_2$ *is the Lipschitz constant of* $\nabla \widetilde{f}^k$ *with* $C_2 := \|\mathbf{A}^\top \mathbf{A}\| + mB_g(B_g + L_g)$ *and* $C_3 = L_f + \rho + \sqrt{m}L_g C_{\mathbf{z}}$.

Combining Theorem 3.3 and Lemma 3.4, we are ready to show the total complexity of Algorithm 1.

THEOREM 3.5 (Total complexity result I). *For a given* $\varepsilon > 0$, *under Assumptions 1-5, Algorithm 1, with Algorithm 2 as a subroutine to compute* $\mathbf{x}^{k+1}$, *can produce an* ε*-KKT point of problem* (P) *by* T_1^{total} *proximal gradient steps. Here,* T_1^{total} *satisfies*

$$T_1^{\text{total}} \le 2K + K\left(2\sqrt{C_3/\rho} + 2\sqrt{\sqrt{K}\beta_0 C_2/\rho} + \frac{1}{\log 2}\right)\log\left(9{\varepsilon_K}^{-1} D\sqrt{\frac{1}{\rho}}(C_3 + \sqrt{K}\beta_0 C_2)^{\frac{3}{2}}\right),$$

where K *is given in Theorem 3.3,* ε_k *is defined in* (3.4), *and* C_2 *and* C_3 *are given in Lemma 3.4.*

Proof. By Theorem 3.3, Algorithm 1 can find an ε-KKT point of problem (P) within K outer iterations. Hence, by Lemma 3.4, the total number T_1^{total} of inner iterations satisfies

$$\begin{aligned}
T_1^{\text{total}} \le& \sum_{k=0}^{K-1} T_1^k \le \sum_{k=0}^{K-1}\left[1 + \left(2\sqrt{\frac{L_{\widetilde{f}^k}}{\rho}} + \frac{1}{\log 2}\right)\log\frac{9L_{\widetilde{f}^k} D\sqrt{\frac{L_{\widetilde{f}^k}}{\rho}}}{\varepsilon_k}\right] \\
\le& 2K + K\left(2\sqrt{\frac{L_{\widetilde{f}^{K-1}}}{\rho}} + \frac{1}{\log 2}\right)\log\frac{9L_{\widetilde{f}^{K-1}} D\sqrt{\frac{L_{\widetilde{f}^{K-1}}}{\rho}}}{\varepsilon_K} \\
\le& 2K + K\left(2\sqrt{C_3/\rho} + 2\sqrt{\sqrt{K}\beta_0 C_2/\rho} + \frac{1}{\log 2}\right)\log\left(9{\varepsilon_K}^{-1} D\sqrt{\frac{1}{\rho}}(C_3 + \sqrt{K}\beta_0 C_2)^{\frac{3}{2}}\right),
\end{aligned}$$

where the second inequality comes from $L_{\widetilde{f}^k} \le L_{\widetilde{f}^{K-1}}, \forall\, k \le K-1$ by the definition of $L_{\widetilde{f}^k}$ in Lemma 3.4, and $\varepsilon_K \le \varepsilon_k, \forall\, k \le K$ from (3.4) and (1.3), the third inequality holds by $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}, \forall\, a, b \ge 0$. This completes the proof.

REMARK 3.1. *From Theorem 3.5 and because* $K = O(\varepsilon^{-2})$, *we have* $T_1^{\text{total}} = \widetilde{\mathcal{O}}(K^{5/4}) = \widetilde{\mathcal{O}}(\varepsilon^{-2.5})$. □

3.2. Regularized Compositional Objective. In this subsection, we make the following structural assumption on the function f in (P).

ASSUMPTION 6. *In* (P), f *is in a compositional form of* $f = l \circ \mathbf{c}$, *where* $\mathbf{c} : \mathbb{R}^d \to \mathbb{R}^p$ *is one* L_c*-smooth mapping, i.e.,* $\|J_{\mathbf{c}}(\mathbf{x}_1) - J_{\mathbf{c}}(\mathbf{x}_2)\|_F \le L_c\|\mathbf{x}_1 - \mathbf{x}_2\|, \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, *and* $l : \mathbb{R}^p \to \mathbb{R}$ *is a convex, potentially non-smooth,* M_l*-Lipschitz continuous function.*

Under Assumption 6, we have that the weak convexity constant of f satisfies $\rho \le M_l L_c$ by [11, Lemma 4.2]. Without the smoothness of f, an FOM will not produce a near-stationary point of problem (3.1) as claimed in [11]. Hence, we aim at finding a point that is close to a near-KKT point of (P), and we utilize the smoothing strategy adopted in [11]. Details are described below.

Given a point $\bar{\mathbf{x}} \in \mathbf{R}^d$, we define the prox-linear function $f^0 : \mathbb{R}^d \mapsto \mathbb{R}$ and a smoothing function $f^\nu : \mathbb{R}^d \mapsto \mathbb{R}$ of f by

$$f^0(\mathbf{x}; \bar{\mathbf{x}}) := l(\mathbf{c}(\bar{\mathbf{x}}) + J_{\mathbf{c}}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})), \quad f^\nu(\mathbf{x}; \bar{\mathbf{x}}) := l^\nu(\mathbf{c}(\bar{\mathbf{x}}) + J_{\mathbf{c}}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})), \tag{3.17}$$

where l^ν is the Moreau envelope of l with $0 < \nu \le 1$. Then, f^ν is a smooth function and $f^0(\mathbf{x}; \mathbf{x}) = f(\mathbf{x})$. Since $\mathcal{X} = \text{dom}(F)$ is bounded, we have from Assumptions 1, 3 and 6 that

$$\|J_{\mathbf{c}}(\mathbf{x})\|_F \le \|\nabla \mathbf{c}\| := \|J_{\mathbf{c}}(\mathbf{x}_{\text{feas}})\|_F + L_c D, \forall\, \mathbf{x} \in \mathcal{X}. \tag{3.18}$$

The following properties hold directly from Lemmas 2.1 and 3.2 of [11].

LEMMA 3.6. *Let f^0 and f^ν be defined in (3.17). It holds that*

(a) *f^ν is $M_l\|\nabla \mathbf{c}\|$ Lipschitz continuous;*

(b) *∇f^ν is $\|\nabla \mathbf{c}\|/\nu$ Lipschitz continuous;*

(c) $0 \le f^0(\mathbf{x};\bar{\mathbf{x}}) - f^\nu(\mathbf{x};\bar{\mathbf{x}}) \le \frac{M_l^2\nu}{2}$, $-\frac{\rho}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \le f(\mathbf{x}) - f^0(\mathbf{x};\bar{\mathbf{x}}) \le \frac{\rho}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2$.

With the setting in (3.17) and the properties in Lemma 3.6, we compute $\mathbf{x}^{k+1}$ in Algorithm 1 by applying Algorithm 2 to solve the following subproblem:

$$\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}^{\nu_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k) := \widehat{f}^k(\mathbf{x}) + \widehat{h}^k(\mathbf{x}), \tag{3.19}$$

such that $\mathbf{x}^{k+1}$ is an ε_k-stationary point of problem (3.19), i.e.,

$$\operatorname{dist}(\mathbf{0}, \partial_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}^{\nu_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k)) \le \varepsilon_k, \tag{3.20}$$

where

$$\widehat{f}^k(\mathbf{x}) = f^{\nu_k}(\mathbf{x};\mathbf{x}^k) + (\mathbf{y}^k)^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) + \frac{\beta_k}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \frac{\beta_k}{2}\left\|[\mathbf{g}(\mathbf{x}) + \frac{\mathbf{z}^k}{\beta_k}]_+\right\|^2 - \frac{\|\mathbf{z}^k\|^2}{2\beta_k}, \tag{3.21}$$

$$\widehat{h}^k(\mathbf{x}) = h(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2. \tag{3.22}$$

Notice that $\widehat{h}^k(\mathbf{x})$ is ρ-strongly convex and by [26, equation (3.10)], $\widehat{f}^k(\mathbf{x})$ is convex and $(\|\nabla \mathbf{c}\|/\nu_k + \rho + \sqrt{m}L_g C_{\mathbf{z}} + \sqrt{k+1}\beta_0 C_2)$-smooth with C_2 given in Lemma 3.4. In addition, by Lemma 3.6(c), it holds that

$$\|\nabla f^{\nu_k}(\mathbf{x};\mathbf{x}^k)\| \le \widetilde{B}_f := M_l\|\nabla \mathbf{c}\|, \forall\, \mathbf{x} \in \mathcal{X}.$$

We then have the following result that shows (2.4) with a specified C_P.

LEMMA 3.7. *Under Assumptions 1–4 and 6, suppose that (3.20) is satisfied with $0 \le \varepsilon_k \le \sqrt{\rho/(2\beta_k)}$ and $0 < \nu_k \le 1$ for all $k \ge 0$. Then the condition in (2.4) holds with*

$$C_P := \frac{1}{2}\Big(\sqrt{C_{\mathbf{y}}^2 + C_{\mathbf{z}}^2} + \sqrt{\widetilde{Q}^2/\min_i |g_i^2(\mathbf{x}_{\rm feas})| + \widetilde{Q}^2\|(\mathbf{A}\mathbf{A}^\top)^\dagger \mathbf{A}\|^2 C_1^2}\Big) + 1, \tag{3.23}$$

where $\widetilde{Q} := D(\widetilde{B}_f + 2\rho D + r_h)$, C_1 is defined in (3.6), and $C_{\mathbf{y}}$ and $C_{\mathbf{z}}$ are given in Lemma 2.1.

Proof. Consider the strongly convex problem $\min_{\mathbf{x}}\{f^{\nu_k}(\mathbf{x};\mathbf{x}^k) + h(\mathbf{x}) + \frac{\rho}{2}\left\|\mathbf{x}-\mathbf{x}^k\right\|^2,\ \text{s.t. } \mathbf{A}\mathbf{x}=\mathbf{b},\ \mathbf{g}(\mathbf{x}) \le \mathbf{0}\}$. Then the claim can be obtained by essentially the same arguments as those in the proof of Lemma 3.1. The only difference is that we shall replace B_f by $\widetilde{B}_f$. □

With Lemma 3.7, we are able to show the outer iteration complexity of Algorithm 1. Define

$$C_4 := \max\left\{\frac{3}{2}, 3\sqrt{\|\mathbf{A}\|^2 + mB_g^2}/(2\rho), 4B_g\sqrt{m\beta_0\sqrt{C_5}}/\rho, 64mB_g^2\beta_0\sqrt{\rho C_{\mathbf{x}}}/\rho^2, 8B_g\sqrt{m\beta_0}/\rho, 3/(4\rho)\right\}, \tag{3.24}$$

$$C_5 := \max\left\{4C_P^2\beta_0^{-2}, 4(C_{\mathbf{z}}^2 + 5C_P^2/2)^2\beta_0^{-2}\right\}, \tag{3.25}$$

where C_P is given in (3.23) and $C_{\mathbf{x}}$ in Lemma 2.4. Then we set ε_k as follows:

$$\varepsilon_k := \min\left\{\frac{\varepsilon}{16C_4}, \sqrt{\frac{\rho}{2\beta_k}}\right\}, \tag{3.26}$$

where β_k is given in (1.3). In addition, we define

$$\begin{aligned} &\mathbf{x}_+^k = \arg\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}^0(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k), \quad \widehat{\mathbf{x}}^k = \arg\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}^{\nu_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k) \\ &\mathcal{G}_{1/\rho}\left(\mathbf{x}^k\right) = \rho\left(\mathbf{x}_+^k - \mathbf{x}^k\right), \quad \mathcal{G}_{1/\rho}^{\nu_k}\left(\mathbf{x}^k\right) = \rho(\widehat{\mathbf{x}}^k - \mathbf{x}^k). \end{aligned} \tag{3.27}$$

By using the same technique as that in [11, Eqn. (4.10)], it holds that with $\widetilde{\mathbf{x}}^k = \mathrm{prox}_{\mathcal{L}_{\beta_k}(\cdot;\mathbf{y}^k,\mathbf{z}^k)/(2\rho)}\left(\mathbf{x}^k\right)$,

$$\begin{cases} \|\widetilde{\mathbf{x}}^k - \mathbf{x}^k\| & \le \frac{2}{\rho}\left\|\mathcal{G}_{1/\rho}\left(\mathbf{x}^k\right)\right\|, \\ \mathcal{L}_{\beta_k}(\widetilde{\mathbf{x}}^k;\mathbf{y}^k,\mathbf{z}^k) & \le \mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k), \\ \mathrm{dist}(\mathbf{0},\partial\mathcal{L}_{\beta_k}(\widetilde{\mathbf{x}}^k;\mathbf{y}^k,\mathbf{z}^k)) & \le 4\left\|\mathcal{G}_{1/\rho}\left(\mathbf{x}^k\right)\right\|. \end{cases} \tag{3.28}$$

Thus if $\left\|\mathcal{G}_{1/\rho}\left(\mathbf{x}^k\right)\right\|$ is small, $\widetilde{\mathbf{x}}^k$ is nearly stationary for problem $\min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$, and $\mathbf{x}^k$ is close to $\widetilde{\mathbf{x}}^k$. Though $\widetilde{\mathbf{x}}^k$ may be difficult to obtain, we do not compute it, and the sole purpose of introducing $\widetilde{\mathbf{x}}^k$ is to certify the quality of $\mathbf{x}^k$.

THEOREM 3.8 (Outer iteration complexity result II). *Given $\varepsilon > 0$, under Assumptions 1-4 and 6, let $\{\mathbf{x}^k\},\{\mathbf{y}^k\}$, and $\{\mathbf{z}^k\}$ be generated by Algorithm 1 such that the condition in (3.20) is satisfied, with ε_k given in (3.26) and $\nu_k = \nu := \min\{1, \varepsilon^2/(64C_4^2\rho M_l^2)\}$, where C_4 is defined in (3.24). Then for some $k \le K := \bar{K}_1 + \bar{K}_2$, $\mathbf{x}^k$ is a near ε-KKT point of problem (P), where $\bar{K}_1 := \lceil C_5\varepsilon^{-2}\rceil$, $\bar{K}_2 := \lceil 64\rho C_{\mathbf{x}} C_4^2\varepsilon^{-2}\rceil$ with $C_{\mathbf{x}}$ and C_5 given in Lemma 2.4 and (3.25), respectively.*

Proof. From $f^\nu(\mathbf{x}^k;\mathbf{x}^k) \le f^0(\mathbf{x}^k;\mathbf{x}^k) = f(\mathbf{x}^k)$, it holds $\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) \ge \widetilde{\mathcal{L}}_{\beta_k}^\nu(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k)$. Hence, by the ρ-strong convexity of $\widetilde{\mathcal{L}}_{\beta_k}^\nu(\cdot;\mathbf{y}^k,\mathbf{z}^k)$, the definition of $\widehat{\mathbf{x}}^k$ in (3.27), and the condition in (3.20), we have

$$\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) \ge \widetilde{\mathcal{L}}_{\beta_k}^\nu(\widehat{\mathbf{x}}^k;\mathbf{y}^k,\mathbf{z}^k) + \frac{\rho}{2}\left\|\mathbf{x}^k - \widehat{\mathbf{x}}^k\right\|^2 \ge \widetilde{\mathcal{L}}_{\beta_k}^\nu(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) + \frac{\rho}{2}\left\|\mathbf{x}^k - \widehat{\mathbf{x}}^k\right\|^2 - \frac{\varepsilon_k^2}{\rho}. \tag{3.29}$$

In addition, by Lemma 3.6(c), it holds $f^\nu(\mathbf{x}^{k+1};\mathbf{x}^k) + \frac{\rho}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \ge f^0(\mathbf{x}^{k+1};\mathbf{x}^k) - M_l^2\nu/2 + \frac{\rho}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \ge f(\mathbf{x}^{k+1}) - M_l^2\nu/2$. Hence, (3.29) indicates $\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) \ge \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) - \frac{M_l^2\nu}{2} + \frac{1}{2\rho}\|\mathcal{G}_{1/\rho}^\nu(\mathbf{x}^k)\|^2 - \frac{\varepsilon_k^2}{\rho}$. Summing up this inequality over k from $\bar{K}_1$ to $K-1$ and using Lemma 2.4 yields

$$\begin{aligned} \textstyle\sum_{k=\bar{K}_1}^{K-1}\|\mathcal{G}_{1/\rho}^\nu(\mathbf{x}^k)\|^2 &\le 2\rho\textstyle\sum_{k=\bar{K}_1}^{K-1}\left(\mathcal{L}_{\beta_k}(\mathbf{x}^k;\mathbf{y}^k,\mathbf{z}^k) - \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1};\mathbf{y}^k,\mathbf{z}^k) + M_l^2\nu/2 + \varepsilon_k^2/\rho\right) \\ &\le 2\rho C_{\mathbf{x}} + 2\rho\textstyle\sum_{k=\bar{K}_1}^{K-1}\left(M_l^2\nu/2 + \varepsilon_k^2/\rho\right), \forall\, K > \bar{K}_1. \end{aligned}$$

The above inequality together with $\nu \le \frac{\varepsilon^2}{64C_4^2\rho M_l^2}$ and $\varepsilon_k \le \varepsilon/(16C_4), \forall\, k \ge 0$ implies that

$$\min_{\bar{K}_1 \le k \le K-1}\|\mathcal{G}_{1/\rho}^\nu(\mathbf{x}^k)\| \le \sqrt{2\rho C_{\mathbf{x}}/(K-\bar{K}_1) + \varepsilon^2/(32C_4^2)}. \tag{3.30}$$

Moreover, it follows from [11, Theorem 6.5] and $\nu \le \frac{\varepsilon^2}{64\rho C_4^2 M_l^2}$ that

$$\left\|\mathcal{G}_{1/\rho}\left(\mathbf{x}^k\right)\right\| \le \left\|\mathcal{G}_{1/\rho}^\nu\left(\mathbf{x}^k\right)\right\| + \sqrt{M_l^2\nu\rho/2} \le \left\|\mathcal{G}_{1/\rho}^\nu\left(\mathbf{x}^k\right)\right\| + \varepsilon/(8C_4). \tag{3.31}$$

Combining (3.30) and (3.31), we have $\min_{\bar{K}_1 \le k \le K-1}\|\mathcal{G}_{1/\rho}(\mathbf{x}^k)\| \le \frac{\varepsilon}{8C_4} + \sqrt{2\rho C_{\mathbf{x}}/(K-\bar{K}_1) + \varepsilon^2/(32C_4^2)}$. Noting $K = \bar{K}_1 + \bar{K}_2 \ge \bar{K}_1 + 64\rho C_{\mathbf{x}} C_4^2\varepsilon^{-2}$, we obtain $\min_{\bar{K}_1 \le k \le K-1}\|\mathcal{G}_{1/\rho}(\mathbf{x}^k)\| \le 3\varepsilon/(8C_4)$. Let $k' =$

$\arg\min_{\bar{K}_1\le k\le K-1}\|\mathcal{G}_{1/\rho}(\mathbf{x}^k)\|$. Then from (3.28) and $C_4\ge\frac{3}{2}$, it holds $\mathrm{dist}(\mathbf{0},\partial_{\mathbf{x}}\mathcal{L}_{\beta_{k'}}(\widetilde{\mathbf{x}}^{k'};\mathbf{y}^{k'},\mathbf{z}^{k'}))\le\varepsilon$ and $\|\widetilde{\mathbf{x}}^{k'}-\mathbf{x}^{k'}\|\le\frac{3\varepsilon}{4\rho C_4}$. Denote $\bar{\mathbf{y}}^{k'}=\mathbf{y}^{k'-1}+\beta_{k'}(\mathbf{A}\widetilde{\mathbf{x}}^{k'}-\mathbf{b})$, and $\bar{\mathbf{z}}^{k'}=[\mathbf{z}^{k'-1}+\beta_{k'-1}\mathbf{g}(\widetilde{\mathbf{x}}^{k'})]_+$. Then noticing $\partial_{\mathbf{x}}\mathcal{L}_{\beta_{k'}}(\widetilde{\mathbf{x}}^{k'};\mathbf{y}^{k'},\mathbf{z}^{k'})=\partial_{\mathbf{x}}\mathcal{L}_0(\widetilde{\mathbf{x}}^{k'};\bar{\mathbf{y}}^{k'},\bar{\mathbf{z}}^{k'})$, we have

$$\partial_{\mathbf{x}}\mathcal{L}_0(\widetilde{\mathbf{x}}^{k'};\bar{\mathbf{y}}^{k'},\bar{\mathbf{z}}^{k'})\le\varepsilon. \tag{3.32}$$

Furthermore, since $k'\ge\bar{K}_1$, it follows from Young's inequality, (2.4), and (1.3) that

$$\begin{aligned}
&\left\|\mathbf{A}\widetilde{\mathbf{x}}^{k'}-\mathbf{b}\right\|^2+\left\|[\mathbf{g}(\widetilde{\mathbf{x}}^{k'})]_+\right\|^2\\
\le&2\left\|\mathbf{A}\mathbf{x}^{k'}-\mathbf{b}\right\|^2+2\left\|[\mathbf{g}(\mathbf{x}^{k'})]_+\right\|^2+2\left\|\mathbf{A}(\widetilde{\mathbf{x}}^{k'}-\mathbf{x}^{k'})\right\|^2+2\left\|[\mathbf{g}(\widetilde{\mathbf{x}}^{k'})]_+-[\mathbf{g}(\mathbf{x}^{k'})]_+\right\|^2\\
\le&\frac{2C_P^2}{\beta^2_{\bar{K}_1}}+2(\|\mathbf{A}\|^2+mB_g^2)\frac{9\varepsilon^2}{16\rho^2C_4^2}\le\varepsilon^2,
\end{aligned}\tag{3.33}$$

where the last inequality follows from $\bar{K}_1\ge 4C_P^2\beta_0^{-2}\varepsilon^{-2}$ and $C_4\ge 3\sqrt{\|\mathbf{A}\|^2+mB_g^2}/(2\rho)$. Finally, notice

$$\begin{aligned}
&\sum_{i=1}^m|\bar{z}_i^{k'}g_i(\widetilde{\mathbf{x}}^{k'})|=\sum_{i=1}^m\left|[z_i^{k'-1}+\beta_{k'-1}g_i(\widetilde{\mathbf{x}}^{k'})]_+g_i(\widetilde{\mathbf{x}}^{k'})\right|\le\frac{1}{\beta_{k'-1}}\sum_{i=1}^m(z_i^{k'-1})^2+\frac{5\beta_{k'-1}}{4}\sum_{i=1}^m[g_i(\widetilde{\mathbf{x}}^{k'})]_+^2\\
\le&\frac{1}{\beta_{k'-1}}\sum_{i=1}^m(z_i^{k'-1})^2+\frac{5\beta_{k'-1}}{2}\sum_{i=1}^m[g_i(\mathbf{x}^{k'})]_+^2+\frac{5\beta_{k'-1}}{2}\left\|[\mathbf{g}(\widetilde{\mathbf{x}}^{k'})]_+-[\mathbf{g}(\mathbf{x}^{k'})]_+\right\|^2\\
\le&\frac{1}{\beta_{k'-1}}\left(C_{\mathbf{z}}^2+\frac{5C_P^2}{2}\right)+\frac{45\sqrt{K}mB_g^2\beta_0\varepsilon^2}{2(4\rho C_4)^2}\le\frac{\varepsilon}{2}+\frac{45mB_g^2\beta_0\varepsilon}{64\rho^2C_4^2}\left(\sqrt{C_5}+8\sqrt{\rho C_{\mathbf{x}}}C_4+2\right)\le\varepsilon.
\end{aligned}\tag{3.34}$$

Here, the first inequality holds by the same arguments to show (2.7); the third inequality follows from Lemma 2.1, the inequality in (2.4), $\beta_{k'}\le\sqrt{K+1}\beta_0$, the Lipschitz continuity of $[\mathbf{g}]_+$, and $\|\widetilde{\mathbf{x}}^{k'}-\mathbf{x}^{k'}\|\le\frac{3\varepsilon}{4\rho C_4}$; the fourth inequality results from $k'\ge\bar{K}_1\ge 4(C_{\mathbf{z}}^2+5C_P^2/2)^2\beta_0^{-2}\varepsilon^{-2}$, $\beta_{k'-1}=\beta_0\sqrt{k'}\ge\beta_0\sqrt{\bar{K}_1}$, and the definition of K; the last inequality holds because of the choice of C_4 in (3.24).

We obtain from (3.32)–(3.34) that $\widetilde{\mathbf{x}}^{k'}$ is an ε-KKT point of problem (P). Since $\|\widetilde{\mathbf{x}}^{k'}-\mathbf{x}^{k'}\|\le\frac{3\varepsilon}{4\rho C_4}\le\varepsilon$, then $\mathbf{x}^{k'}$ is a near ε-KKT point of problem (P) by Definition 1.3. This completes the proof. □

Below we give the number of iterations for solving (3.19) by Algorithm 2 such that (3.20) is met. From [26, Eqn. (3.10)], it follows that $\widehat{f}^k$ in (3.21) is $L_{\widehat{f}^k}$-smooth with $L_{\widehat{f}^k}=\|\nabla\mathbf{c}\|/\nu_k+\sqrt{m}L_gC_{\mathbf{z}}+\sqrt{k+1}\beta_0C_2$, where C_2 given in Lemma 3.4. Hence, we have the following lemma directly from Theorem B.1.

LEMMA 3.9. *Given $\varepsilon_k>0$ and $\nu_k>0$, under Assumptions 1–4 and 6, Algorithm 2 with $\gamma_u=2$ applied to (3.19) can find a solution $\mathbf{x}^{k+1}$ that satisfies the criteria in (3.20) within*

$$T_2^k=\left\lceil\max\left\{\frac{1}{\log 2},2\sqrt{\frac{L_{\widehat{f}^k}}{\rho}}\right\}\log\frac{9DL_{\widehat{f}^k}\sqrt{L_{\widehat{f}^k}/\rho}}{\varepsilon_k}\right\rceil+1 \tag{3.35}$$

iterations, where $L_{\widehat{f}^k}=C_6+\sqrt{k+1}\beta_0C_2$, with $C_6:=\|\nabla\mathbf{c}\|/\nu_k+\sqrt{m}L_gC_{\mathbf{z}}$, C_2 given in Lemma 3.4 and $\|\nabla\mathbf{c}\|$ given in (3.18).

Combining Theorem 3.8 and Lemma 3.9, we are ready to show the total complexity of Algorithm 1.

THEOREM 3.10 (Total complexity result II). *For a given $\varepsilon > 0$, under Assumptions 1–4 and 6, Algorithm 1, with Algorithm 2 as a subroutine to compute $\mathbf{x}^{k+1}$ by solving (3.19), can find a near ε-KKT point of problem (P) by T_2^{total} proximal gradient steps. Here, T_2^{total} satisfies*

$$T_2^{\text{total}} \le 2K + K\left(2\sqrt{C_6/\rho} + 2\sqrt{\sqrt{K}\beta_0 C_2/\rho} + \frac{1}{\log 2}\right)\log\left(9\varepsilon_K{}^{-1} D\sqrt{\frac{1}{\rho}}(C_6 + \sqrt{K}\beta_0 C_2)^{\frac{3}{2}}\right),$$

where C_2 and C_6 are defined in Lemmas 3.4 and 3.9, K is given in Theorem 3.8, and ε_k is defined in (3.26).

Proof. Set $\nu_k = \nu := \min\{1, \varepsilon^2/(64C_4^2\rho M_l^2)\}$ as in Theorem 3.8 and notice $T_2^{\text{total}} \le \sum_{k=0}^{K-1} T_2^k$, where T_2^k is given in (3.35). We obtain the desired result by the same arguments in the proof of Theorem 3.5. □

REMARK 3.2. *By $K = O(\varepsilon^{-2})$, $\nu_k = \nu = O(\varepsilon^2)$, we have $T_2^{\text{total}} = \widetilde{\mathcal{O}}(K\sqrt{\|\nabla \mathbf{c}\|/\nu} + K^{5/4}) = \widetilde{\mathcal{O}}(\varepsilon^{-3})$.*

3.3. General Weakly-Convex Objective. In this subsection, we consider a general case with a weakly-convex objective. We make the following assumption.

ASSUMPTION 7. *In (P), f satisfies $\|\xi\| \le B_f, \forall\, \xi \in \partial f(\mathbf{x}), \forall\, \mathbf{x} \in \mathcal{X}$.*

Without a smoothness structure, we do not expect an FOM to produce an ε_k-stationary point of the problem $\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}; \mathbf{y}^k, \mathbf{z}^k)$ as in (3.4). Instead, we compute a near-optimal solution $\mathbf{x}^{k+1}$ satisfying (3.9).

Let $\mathbf{x}_*^{k+1} = \arg\min_{\mathbf{x}} \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}; \mathbf{y}^k, \mathbf{z}^k)$. Then by the ρ-strong convexity of $\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}; \mathbf{y}^k, \mathbf{z}^k)$, it holds

$$\|\mathbf{x}_*^{k+1} - \mathbf{x}^{k+1}\|^2 \le \frac{2}{\rho}\left(\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k) - \widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}_*^{k+1}; \mathbf{y}^k, \mathbf{z}^k)\right) \overset{(3.9)}{\le} \frac{2\varepsilon_k^2}{\rho^2}, \forall\, k \ge 0. \tag{3.36}$$

The next theorem gives the outer iteration complexity of Algorithm 1 for the general weakly-convex case.

THEOREM 3.11 (Outer iteration complexity result III). *Given $\varepsilon > 0$, under Assumptions 1-4 and 7, let $\{\mathbf{x}^k\}$, $\{\mathbf{y}^k\}$, and $\{\mathbf{z}^k\}$ be generated by Algorithm 1 such that (3.9) holds with $\varepsilon_k := \min\{\frac{\varepsilon}{4}, \frac{\rho\varepsilon}{\sqrt{2}}, \sqrt{\frac{\rho}{2\beta_k}}\}$ for all $k \ge 0$. Then for some $k < K := \widetilde{K}_1 + \widetilde{K}_2$, $\mathbf{x}^{k+1}$ is a near ε-KKT point of problem (P), where $\widetilde{K}_1 := \lceil \max\{C_P^2\beta_0^{-2}, (C_{\mathbf{z}}^2 + 5C_P^2/4)^2\beta_0^{-2}\}\varepsilon^{-2}\rceil$ and $\widetilde{K}_2 := \lceil 16C_{\mathbf{x}}\rho\varepsilon^{-2}\rceil$, with $C_{\mathbf{y}}, C_{\mathbf{z}}$, $C_{\mathbf{x}}$, and C_P given in Lemma 2.1, Lemma 2.4, and (3.6).*

Proof. Notice that the claim in Lemma 3.1 follows from the near-stationarity condition of $\mathbf{x}^{k+1}$ in (3.4), Assumptions 1-4, and the boundedness of ∇f. Hence, by the same arguments and using the definition of $\mathbf{x}_*^{k+1}$, we have

$$\left\|\mathbf{A}\mathbf{x}_*^{k+1} - \mathbf{b}\right\|^2 + \left\|\left[\mathbf{g}\left(\mathbf{x}_*^{k+1}\right)\right]_+\right\|^2 \le C_P^2/\beta_k^2,\ \forall\, k \ge 0. \tag{3.37}$$

Then by (2.5) in Lemma 2.2 and (3.37), it holds that

$$\sum_{i=1}^m |[z_i^k + \beta_k g_i(\mathbf{x}_*^{k+1})]_+ g_i(\mathbf{x}_*^{k+1})| \le \left(C_{\mathbf{z}}^2 + 5C_P^2/4\right)/\beta_k,\ \forall\, k \ge 0. \tag{3.38}$$

In addition, by (3.11), we have $\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}_*^{k+1}; \mathbf{y}^k, \mathbf{z}^k) \le \mathcal{L}_{\beta_k}(\mathbf{x}^k; \mathbf{y}^k, \mathbf{z}^k) - \frac{\rho}{2}\left\|\mathbf{x}^k - \mathbf{x}_*^{k+1}\right\|^2$, which together with (3.9), gives

$$\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}; \mathbf{y}^k, \mathbf{z}^k) + \rho\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \frac{\rho}{2}\|\mathbf{x}_*^{k+1} - \mathbf{x}^k\|^2 \le \mathcal{L}_{\beta_k}(\mathbf{x}^k; \mathbf{y}^k, \mathbf{z}^k) + \frac{\varepsilon_k^2}{\rho}. \tag{3.39}$$

On the other hand, starting from (3.9) and by the same arguments in the proof of Lemma 3.1, we have the claim in Lemma 3.1. Thus the results in Lemmas 2.3 and 2.4 still hold. In particular, we have (2.10). Hence, sum up (3.39) over k from $\widetilde{K}_1$ to $K-1$ and use (2.10) to have $\frac{\rho}{2}\sum_{k=\widetilde{K}_1}^{K-1}\|\mathbf{x}_*^{k+1}-\mathbf{x}^k\|^2 \le \sum_{k=\widetilde{K}_1}^{K-1}\frac{\varepsilon_k^2}{\rho}+C_{\mathbf{x}}$. Since $\varepsilon_k \le \frac{\varepsilon}{4}$, it holds $\sum_{k=\widetilde{K}_1}^{K-1}\frac{\varepsilon_k^2}{\rho} \le \frac{\widetilde{K}_2\varepsilon^2}{16\rho}$. Let $k' = \arg\min_{\widetilde{K}_1\le k\le K-1}\|\mathbf{x}_*^{k+1}-\mathbf{x}^k\|^2$. Then $\|\mathbf{x}_*^{k'+1}-\mathbf{x}^{k'}\| \le \sqrt{\frac{\varepsilon^2}{8\rho^2}+\frac{2C_{\mathbf{x}}}{\rho\widetilde{K}_2}} \le \frac{\varepsilon}{2\rho}$ by $\widetilde{K}_2 \ge 16C_{\mathbf{x}}\rho\varepsilon^{-2}$. Now notice $\mathbf{0} \in \partial_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x}_*^{k+1};\mathbf{y}^k,\mathbf{z}^k) = \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}_*^{k+1};\mathbf{y}^k,\mathbf{z}^k) + 2\rho(\mathbf{x}_*^{k+1}-\mathbf{x}^k)$ and $\partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}_*^{k+1};\mathbf{y}^k,\mathbf{z}^k) = \partial_{\mathbf{x}}\mathcal{L}_0(\mathbf{x}_*^{k+1};\bar{\mathbf{y}}^k,\bar{\mathbf{z}}^k)$ with $\bar{\mathbf{y}}^k := \mathbf{y}^k+\beta_k(\mathbf{A}\mathbf{x}_*^{k+1}-\mathbf{b})$, and $\bar{\mathbf{z}}^k := [\mathbf{z}^k+\beta_k\mathbf{g}(\mathbf{x}_*^{k+1})]_+$. We obtain $\mathrm{dist}(\mathbf{0},\partial_{\mathbf{x}}\mathcal{L}_0(\mathbf{x}_*^{k'+1};\bar{\mathbf{y}}^{k'},\bar{\mathbf{z}}^{k'})) \le \varepsilon$. This claim, together with (3.37), (3.38), and the choice of $\widetilde{K}_1$, indicates that $\mathbf{x}_*^{k'+1}$ is an ε-KKT point of problem (P). Moreover, from (3.36) and $\varepsilon_k \le \rho\varepsilon/\sqrt{2}$, it follows that $\|\mathbf{x}_*^{k'+1}-\mathbf{x}^{k'+1}\| \le \varepsilon$. Therefore, $\mathbf{x}^{k'+1}$ is a near ε-KKT point of problem (P) by Definition 1.3, and we complete the proof. □

REMARK 3.3. *We make a few remarks about Theorem 3.11 and its implications. First, in a general case, one can apply a subgradient method [34] to find $\mathbf{x}^{k+1}$ such that (3.9) holds, due to the strong convexity of $\widetilde{\mathcal{L}}_{\beta_k}(\cdot\,;\mathbf{y}^k,\mathbf{z}^k)$. Our $O(\varepsilon^{-2})$ outer iteration complexity result matches with that in [49], but we allow a smaller penalty parameter for nondifferentiable problems and thus can potentially achieve a lower overall complexity result. Second, when there are certain special structures on f such as those in Sect. 3.1 and Sect. 3.2, one can apply a more efficient way to obtain $\mathbf{x}^{k+1}$ and achieve a lower complexity. The best way to compute $\mathbf{x}^{k+1}$ will depend on the structure on f. For example, when $f=\max\{f_1,f_2\}$ where f_1 and f_2 are both ρ-weakly convex and smooth, one can apply the Moreau-envelope based smoothing approach in Sect. 3.2 and achieve an overall complexity of $\widetilde{O}(\varepsilon^{-3})$ to produce a near ε-KKT point. However, a potentially better way is to have a more efficient subroutine to solve each strongly convex subproblem $\min_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k)$ by exploiting the special structure of f. Notice $f(\mathbf{x})+\frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2 = \max_{\lambda\in[0,1]}\lambda(f_1(\mathbf{x})+\frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2)+(1-\lambda)(f_1(\mathbf{x})+\frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2)$. With $H_k(\mathbf{x}) := h(\mathbf{x})+\frac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2+(\mathbf{y}^k)^\top(\mathbf{A}\mathbf{x}-\mathbf{b})+\frac{\beta_k}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\frac{\beta_k}{2}\|[\mathbf{g}(\mathbf{x})+\frac{\mathbf{z}^k}{\beta_k}]_+\|^2-\frac{\|\mathbf{z}^k\|^2}{2\beta_k}$, it holds*

$$\min_{\mathbf{x}}\widetilde{\mathcal{L}}_{\beta_k}(\mathbf{x};\mathbf{y}^k,\mathbf{z}^k) = \max_{\lambda\in[0,1]}\min_{\mathbf{x}}\lambda(f_1(\mathbf{x})+\tfrac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2)+(1-\lambda)(f_1(\mathbf{x})+\tfrac{\rho}{2}\|\mathbf{x}-\mathbf{x}^k\|^2)+H_k(\mathbf{x}).$$

By exploiting the 1-dimension of λ, we can follow [46] and apply a bisection method to search for the optimal λ, by which we can then find the desired $\mathbf{x}^{k+1}$ in $\widetilde{O}(\sqrt{\beta_k})$ iterations. Hence, by Theorem 3.11, the total iteration complexity is $\widetilde{O}(\varepsilon^{-2.5})$ to produce a near ε-KKT point. We leave details to interested readers.

4. Numerical Experiments. In this section, we conduct numerical experiments to demonstrate the effectiveness of our algorithm, named DPALM. We apply it to the non-convex linearly constrained quadratic problem (LCQP), non-convex quadratically constrained quadratic problem (QCQP), and linearly constrained robust nonlinear least square. All of these tests are performed in MATLAB 2022a on an iMAC with 40GB memory. We report primal infeasibility and dual infeasibility for problems with only linear constraints. We also report complementary slackness error for problems with nonlinear constraints.

4.1. Non-convex Linearly-Constrained Quadratic Program (LCQP). In this subsection, we compare DPALM to HiAPeM [26], LiMEAL [49] and NL-IAPIAL [20] on LCQP in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^d}\frac{1}{2}\mathbf{x}^\top\mathbf{Q}_0\mathbf{x}+\mathbf{c}_0^\top\mathbf{x},\ \text{s.t. } \mathbf{A}\mathbf{x}=\mathbf{b},\ x_i\in[l_i,u_i],\ \forall\, i=1,\dots,d, \tag{4.1}$$

where $\mathbf{A}\in\mathbb{R}^{n\times d}, \mathbf{c}_0\in\mathbb{R}^d$, and $\mathbf{Q}_0\in\mathbb{R}^{d\times d}$ is a symmetric and indefinite matrix with the smallest eigenvalue $-\rho<0$. Thus the objective of (4.1) is ρ-weakly convex.

In our experiments, we set $n = 10$, $d = 1000$, and $l_i = -5, u_i = 5, \forall\, i$. We vary $\rho \in \{0.1, 1, 10\}$. Each algorithm is terminated if an ε-KKT solution is found or after 10^4 outer iterations, where we set $\varepsilon = 10^{-3}$. For DPALM and NL-IAPIAL, we tune the initial penalty parameter β_0 by picking the best one from $\{0.01, 0.1, 1, 10\}$ for each value of ρ. This way, we have $\beta_0 = 0.01, 0.1, 10$ and $\beta_0 = 0.01, 0.1, 1$ corresponding to $\rho = 0.1, 1, 10$ respectively for DPALM and NL-IAPIAL. LiMEAL uses a fixed penalty parameter β. We pick the best β from $\{0.1, 1, 10, 100\}$ and set its parameter $\eta = 1.5$, which appears to work the best for LiMEAL. This way, we have $\beta = 0.1, 1, 100$ corresponding to $\rho = 0.1, 1, 10$ for LiMEAL. For HiAPeM, we use its default value $\beta_0 = 0.01$ but set its parameter N_0 to 10^4. Here, N_0 is the number of calls to ALM as subroutine in the initial stage of HiAPeM. Since we set the maximum number of outer iterations to 10^4, using $N_0 = 10^4$ means that we run HiAPeM by solely using ALM to solve its proximal point subproblems. This yields the best performance for HiAPeM, as demonstrated in [26]. All algorithms use Nesterov's APG in Algorithm 2 to solve their core strongly convex subproblems. Notice that our implementation of LiMEAL is different from that in the numerical experiment of [49] but instead we follow its update given in Eqn. (9).

For each value of ρ, we generate 10 independent random LCQP instances. In Table 1, we present the primal and dual infeasibility, running time (in seconds), and the number of gradient evaluations (shortened by pres, dres, time, #Grad, respectively), averaged over 10 instances, for each method. In Figure 1, we plot primal and dual infeasibility versus the number of gradient evaluations in one random instance for all compared methods. From our results in the table and the figure, we see that our method takes fewer gradient evaluations (and less running time) than all other methods to produce the same-accurate KKT point. A larger value of ρ means more non-convexity and thus a harder instance. However, we notice that all compared methods take fewer gradient evaluations for $\rho = 10$ than $\rho = 0.1, 1$. This is possibly because we tune the algorithm parameter for each ρ, or because we aim at producing a near-KKT point instead of a global optimal solution.

TABLE 1

Average results and variance by the proposed algorithm DPALM, HiAPeM in [26] with $N_0 = 10^4$, LiMEAL in [49], and NL-IAPIAL in [20] on solving 10 instances of ρ-weakly convex LCQP in (4.1) of size $n = 10$ and $d = 1000$, where $\rho \in \{0.1, 1.0, 10\}$.

	pres	dres	time	#Grad	pres	dres	time	#Grad	pres	dres	time	#Grad	pres	dres	time	#Grad
	HiAPeM with $N_0 = 1000$				NL-IAPIAL				LiMEAL				DPALM			
weak convexity constant: $\rho = 0.1$																
avg.	4.15e-4	9.87e-4	203.63	124840	6.92e-7	9.99e-4	25.47	90366	9.48e-4	0.06	93.23	144666	5.82e-3	9.82e-4	10.35	40168
var.	4.97e-8	1.08e-9	7.40e3	2.41e10	1.48e-8	4.23e-8	658.73	1.71e9	4.89e-9	0.15	6.46e3	5.72e6	5.71e6	4.42e-8	1.89e-8	223.90	8.54e8
weak convexity constant: $\rho = 1.0$																
avg.	7.40e-5	9.95e-4	749.48	509020	9.44e-7	9.90e-4	151.15	713015	7.94e-6	0.01	140.21	381753	3.90e-6	9.94e-4	42.23	176762
var.	1.16e-7	1.44e-10	1.76e4	1.91e10	8.24e-10	6.72e-9	517.62	1.22e9	1.76e-6	6.17	5.18e3	3.74e7	4.01e-9	8.09e-9	256.90	3.13e8
weak convexity constant: $\rho = 10$																
avg.	1.47e-4	8.41e-4	46.04	62192	9.03e-5	3.42e-4	17.91	83705	3.90e-5	4.10e-4	12.18	47423	1.31e-4	7.28e-4	7.13	31838
var.	5.10e-8	6.48e-11	4.36e4	4.89e10	1.16e-10	1.92e-9	361.01	1.36e8	2.97e-7	16.95	589.99	3.74e7	2.11e-10	4.45e-10	275.27	1.23e8

4.2. Non-convex Quadratically-Constrained Quadratic Program (QCQP). In this subsection, we compare the proposed DPALM method in Algorithm 1 to HiAPeM in [26] and NL-IAPIAL in [20] on solving non-convex instances of QCQP in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x} + \mathbf{c}_0^\top \mathbf{x}, \text{ s.t. } \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_j \mathbf{x} + \mathbf{c}_j^\top \mathbf{x} + \mathbf{d}_j \le 0, \forall\, j \in [m]; x_i \in [l_i, u_i], \forall\, i \in [d]. \tag{4.2}$$

Here, $\mathbf{Q}_j$ is positive semidefinite for each $j \ge 1$, but $\mathbf{Q}_0$ is indefinite and has the smallest eigenvalue $-\rho < 0$. Thus the objective of (4.2) is ρ-weakly convex but the constraints are convex. In the experiment, we set $m = 10, d = 1000, l_i = -5$ and $u_i = 5, \forall\, i$ and generate the data of vectors and matrices randomly.

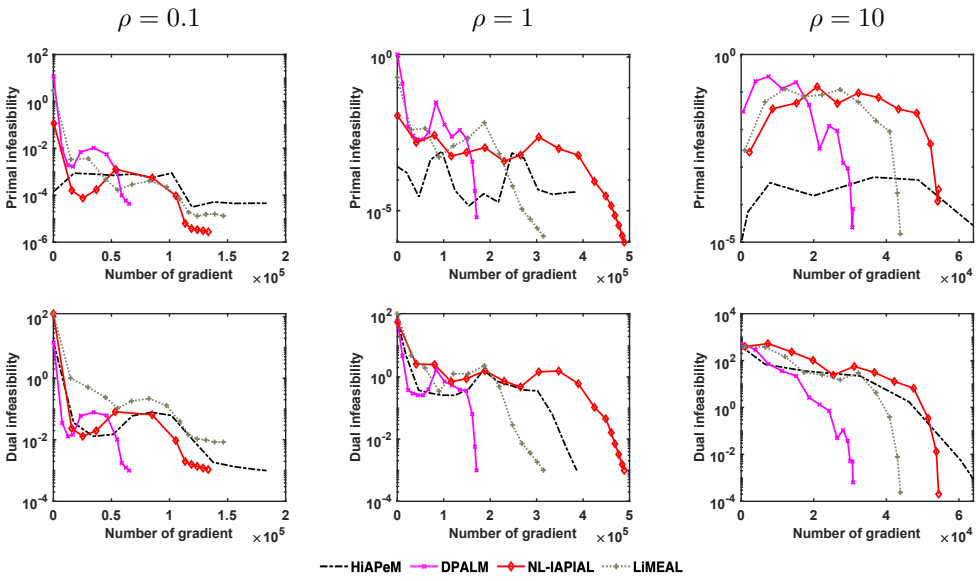

FIG. 1. *Primal and dual infeasibility vs. Number of gradient evaluations by the proposed DPALM, the HiAPeM method in [26], the LiMEAL in [49], and the NL-IAPIAL method in [20] on solving instances of (4.1) with different weak convexity constant* $\rho \in \{0.1, 1, 10\}$.

TABLE 2

Results by the proposed algorithm DPALM, the HiAPeM method in [26] with $N_0 = 10^4$*, and NL-IAPIAL method in [20] on solving instances of* ρ*-weakly convex QCQP (4.2) of size* $m = 10$ *and* $d = 1000$*, where* $\rho \in \{0.1, 1.0, 10\}$.

	pres	dres	compslack	time	#Grad	pres	dres	compslack	time	#Grad	pres	dres	compslack	time	#Grad
	HiAPeM with $N_0 = 1000$					NL-IAPIAL					DPALM				
weak convexity constant: $\rho = 0.1$															
avg.	8.27e-5	5.05e-4	9.04e-5	55.79	9642	1.14e-4	5.78e-5	5.79e-7	17.37	4475	1.97e-4	1.92e-4	2.33e-4	11.42	2947
var.	1.78e-10	9.94e-10	8.95e-12	27.13	1579	4.60e-8	1.63e-10	9.38e-9	9.99	5.65e5	3.62e-8	1.00e-8	6.75e-8	2.67	1.57e5
weak convexity constant: $\rho = 1.0$															
avg.	3.34e-4	8.00e-4	9.04e-5	86.44	16225	2.45e-6	7.42e-4	5.79e-7	14.88	4240	1.22e-4	6.25e-4	3.08e-5	7.02	1931
var.	4.06e-9	1.19e-8	3.92e-10	9.22	1.51e5	6.81e-13	1.51e-08	3.54e-14	2.88	2.24e5	8.40e-9	1.69e-8	4.74e-10	0.14	740
weak convexity constant: $\rho = 10$															
avg.	3.82e-4	9.05e-4	6.61e-4	184.71	30462	3.13e-11	8.98e-4	4.88e-11	638.49	172490	1.14e-4	4.97e-4	2.59e-4	15.69	3874
var.	1.75e-9	2.17e-9	5.71e-9	286.11	5.09e5	9.79e-23	2.14e-9	2.86e-22	2759.60	7.44e10	5.27e-10	2.74e-9	2.94e-9	11.12	1.51e4

We vary $\rho \in \{0.1, 1, 10\}$ and for each value of ρ, we generate 10 instances independently at random. Each algorithm is terminated if an ε-KKT solution is found or after 10^4 outer iterations, where we set $\varepsilon = 10^{-3}$. For DPALM and NL-IAPIAL, we pick the best β_0 from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, resulting in $\beta_0 = 10^{-4}$ for the former and $\beta_0 = 0.1$ for the latter. The setting of HiAPeM is the same as that in the previous test with $\beta_0 = 0.01$ and $N_0 = 10^4$. In addition to pres, dres, time, and #grad, we also report complementarity violation, shortened as compslack. Average results with variance are shown in Table 2 and also, we plot the results in Figure 2 for one instance. From the results, we see that to produce a near-KKT point at the same accuracy, our algorithm takes fewer gradient evaluations and less time than HiAPeM and NL-IAPIAL. This advantage becomes more significant as ρ increases.

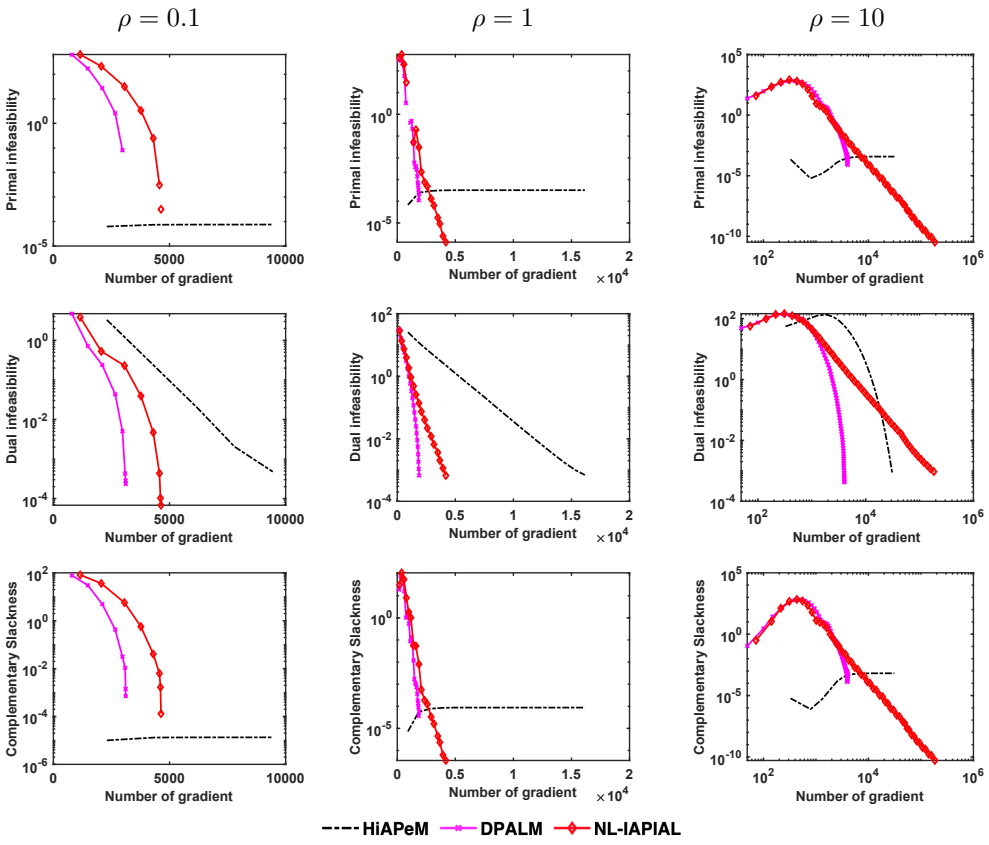

FIG. 2. *Primal infeasibility, dual infeasibility, and complementary slackness error vs. Number of gradient evaluations by the proposed DPALM, the HiAPeM method in [26], and NLIAPIAL method in [20] on solving instances of* (4.2) *with different weak convexity constant* $\rho \in \{0.1, 1, 10\}$.

4.3. Linear Constrained Robust Nonlinear Least Square. In this subsection, we test the proposed DPALM method and compare it to the inexact Prox-Linear method [11, Alg. 2] on solving a linearly constrained robust nonlinear least square:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{f}(\mathbf{x})\|_1, \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}, x_i \in [l_i, u_i], \forall i = 1, 2, \ldots, d \tag{4.3}$$

where $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is a smooth mapping, and $\mathbf{A} \in \mathbb{R}^{n\times d}$. To apply the method in [11], we reformulate the problem in (4.3) to $\min_{\mathbf{x}\in[\mathbf{l},\mathbf{u}]} \widehat{h}(\widehat{\mathbf{f}}(\mathbf{x}))$, where $\mathbf{l} = [l_1, l_2, \ldots, l_d], \mathbf{u} = [u_1, u_2, \ldots, u_d]$, $\widehat{\mathbf{f}}(\mathbf{x}) = \begin{pmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{pmatrix}$, and $\widehat{h} : \mathbb{R}^{m+n} \to \mathbb{R}$ is defined as $\widehat{h}\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \|\mathbf{y}_1\|_1 + \delta_{\{\mathbf{0}\}}(\mathbf{y}_2)$ for any $\mathbf{y}_1 \in \mathbb{R}^m, \mathbf{y}_2 \in \mathbb{R}^n$. Then we smooth $\widehat{h}$ by its Moreau envelope $\widehat{h}_\nu$ for a small $\nu > 0$.

In our experiment, we set $\mathbf{f}(\mathbf{x}) = (f_1, f_2, \ldots, f_m)$ with $f_i = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_i \mathbf{x} + \mathbf{c}_i^\top \mathbf{x}$ in (4.3). Here, each $\mathbf{Q}_i$ is a positive-definite matrix but notice that the composed function $\|\mathbf{f}(\mathbf{x})\|_1$ is nonconvex nonsmooth. The

weak convexity constant $\rho = M_l L_f$, where $M_l = \sqrt{m}$ is the Lipschitz constant of $l(\mathbf{x}) = \|\mathbf{x}\|_1$ and L_f is the smoothness constant of $\mathbf{f}(\mathbf{x})$. We set $l_i = -5, u_i = 5, \forall\, i$, $m = n = 10$ and $d = 1000$ and generate a random instance. From (3.27), (3.28), (3.31), and the ρ-strong convexity of each subproblem, one can show $\mathrm{dist}(\mathbf{0}, \partial \mathcal{L}_{\beta_k}(\widetilde{\mathbf{x}}^k; \mathbf{y}^k, \mathbf{z}^k)) \le 4\rho\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + 4\sqrt{\nu_k \rho/2} + \sqrt{2}\varepsilon_k/\rho$, where ε_k is the error tolerance for solving the k-th subproblem. Hence, when $4\sqrt{\nu_k\rho/2} + \sqrt{2}\varepsilon_k/\rho$ is small, we can use $\rho\|\mathbf{x}^k - \mathbf{x}^{k-1}\|$ as a measure of dual infeasibility. For both methods, we use a small constant smoothing parameter $\nu = 10^{-3}$. For DPALM, we simply set $\beta_0 = 1$. They are terminated once $\max\{\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|, \rho\|\mathbf{x}^k - \mathbf{x}^{k-1}\|\} \le \varepsilon$ for some k, where $\varepsilon = 10^{-2}$ is set. In Figure 3, we plot the primal infeasibility and dual infeasibilty measured by $\rho\|\mathbf{x}^k - \mathbf{x}^{k-1}\|$ versus the number of gradient evaluations. It clearly shows that our method takes far fewer gradient evaluations than the prox-linear method to reach the same ε accuracy, though both methods have the same order of oracle complexity in theory.

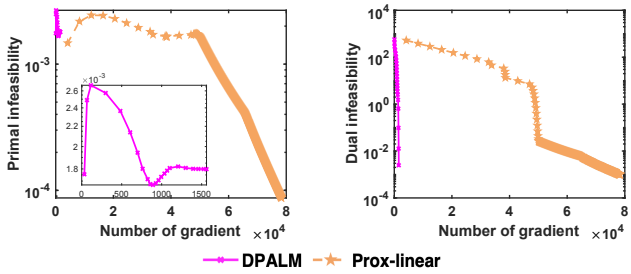

FIG. 3. *Primal and Dual infeasibility vs. Number of gradient evaluations by the proposed DPALM and the Prox-Linear method in [11] on solving an instance of* (4.3).

5. Conclusions. We have presented a damped proximal augmented Lagrangian method (DPALM) for solving problems, which have a weakly-convex objective and convex affine/nonlinear constraints. We show that DPALM can produce a near ε-KKT point under Slater's condition by solving $O(\varepsilon^{-2})$ strongly-convex subproblems, each to a desired accuracy. In addition, we have established the overall iteration complexity of DPALM for two cases where f is smooth or a convex function composed with a smooth mapping. For the smooth case, with an APG method applied to each subproblem, DPALM achieves an $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ complexity result to produce an ε-KKT point, which improves an existing $\tilde{\mathcal{O}}(\varepsilon^{-3})$ result for proximal ALM based method and matches with the best-known result by quadratic penalty based methods. For the compositional case, with an APG applied to a Moreau-envelope smoothed subproblem, DPALM achieves a complexity result of $\tilde{\mathcal{O}}(\varepsilon^{-3})$ to produce a near ε-KKT point, which is new for solving functional constrained compositional problems.

Appendix A. A Key Lemma and Proof of Lemma 2.3. The next lemma is used to bound primal infeasibility of Algorithm 1 and follows directly from the proof of [45, Lemma 7].

LEMMA A.1. *Let* $\mathbf{x}^* = \arg\min_{\mathbf{x}} \left\{ \widehat{f}(\mathbf{x}),\ s.t.\ \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{g}(\mathbf{x}) \le \mathbf{0} \right\}$, *where* $\widehat{f}$ *is strongly convex, and each component function in* $\mathbf{g}$ *is convex. Let* $\mathcal{L}_\beta(\mathbf{x}; \mathbf{p})$ *be the AL function with a multiplier* $\mathbf{p} = (\mathbf{y}, \mathbf{z})$ *and a penalty parameter* $\beta > 0$. *Suppose* $\mathbf{x}^*$ *is a KKT point of the problem with a corresponding multiplier* $\mathbf{p}^* = (\mathbf{y}^*, \mathbf{z}^*)$. *Start from any* $\mathbf{p}$ *with* $\mathbf{z} \ge \mathbf{0}$; *let* $\widehat{\mathbf{x}}$ *satisfy* $\mathcal{L}_\beta(\widehat{\mathbf{x}}; \mathbf{p}) \le \min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}; \mathbf{p}) + \delta$ *for some* $\delta \ge 0$; *set* $\mathbf{y}^+ = \mathbf{y} + \beta(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b})$, $\mathbf{z}^+ = [\mathbf{z} + \beta\mathbf{g}(\widehat{\mathbf{x}})]_+$. *Then* $\|\mathbf{p}^+ - \mathbf{p}^*\|^2 \le \|\mathbf{p} - \mathbf{p}^*\|^2 + 2\beta\delta$.

Proof of Lemma 2.3. Recall the definitions of J_1^k and J_2^k given in (2.3) and further define

$$J_+^k := \left\{i : g_i\left(\mathbf{x}^{k+1}\right) \geq 0\right\}, \quad J_-^k := \left\{i : g_i\left(\mathbf{x}^{k+1}\right) < 0\right\}. \tag{A.1}$$

From the update of $\mathbf{z}^{k+1}$, we arrive at

$$\begin{cases} z_i^{k+1} = \left(1 - \dfrac{\gamma_k}{\beta_k}\right) z_i^k \text{ and } g_i\left(\mathbf{x}^{k+1}\right) < 0, & \text{if } i \in J_1^k, \\ z_i^{k+1} = z_i^k + \gamma_k g_i\left(\mathbf{x}^{k+1}\right) \text{ and } g_i\left(\mathbf{x}^{k+1}\right) < 0, & \text{if } i \in J_2^k \cap J_-^k, \\ z_i^{k+1} = z_i^k + \gamma_k g_i\left(\mathbf{x}^{k+1}\right) \text{ and } g_i\left(\mathbf{x}^{k+1}\right) \geq 0, & \text{if } i \in J_2^k \cap J_+^k. \end{cases} \tag{A.2}$$

Below we look at the term $\frac{\beta_{k+1}}{2}\|[\mathbf{g}(\mathbf{x}^{k+1}) + \frac{\mathbf{z}^{k+1}}{\beta_{k+1}}]_+\|^2 - \frac{\beta_k}{2}\|[\mathbf{g}(\mathbf{x}^{k+1}) + \frac{\mathbf{z}^k}{\beta_k}]_+\|^2$ for these three cases.

Case I: $i \in J_1^k$: We have $g_i\left(\mathbf{x}^{k+1}\right) + \frac{1}{\beta_{k+1}}\left(1 - \frac{\gamma_k}{\beta_k}\right) z_i^k \leq g_i\left(\mathbf{x}^{k+1}\right) + \frac{1}{\beta_k} z_i^k \leq 0$, since $0 < \beta_k \leq \beta_{k+1}$, $0 \leq \gamma_k \leq \beta_k$, and $\mathbf{z}^k \geq \mathbf{0}$ by Lemma 2.1. Hence, $\frac{\beta_{k+1}}{2}[g_i(\mathbf{x}^{k+1}) + \frac{z_i^{k+1}}{\beta_{k+1}}]_+^2 - \frac{\beta_k}{2}[g_i(\mathbf{x}^{k+1}) + \frac{z_i^k}{\beta_k}]_+^2 \leq 0$.

Case II: $i \in J_2^k \cap J_-^k$: We have $\beta_k g_i\left(\mathbf{x}^{k+1}\right) + z_i^k > 0$ and $g_i\left(\mathbf{x}^{k+1}\right) < 0$. Thus $g_i(\mathbf{x}^{k+1}) + \frac{z_i^k}{\beta_k} \geq 0$. Below we discuss two subcases based on the sign of $(\beta_{k+1} + \gamma_k) g_i\left(\mathbf{x}^{k+1}\right) + z_i^k$.

When $(\beta_{k+1} + \gamma_k) g_i\left(\mathbf{x}^{k+1}\right) + z_i^k \geq 0$, we have

$$\begin{aligned} &\tfrac{\beta_{k+1}}{2}\left[g_i\left(\mathbf{x}^{k+1}\right) + \tfrac{1}{\beta_{k+1}}\left(z_i^k + \gamma_k g_i\left(\mathbf{x}^{k+1}\right)\right)\right]_+^2 - \tfrac{\beta_k}{2}\left[g_i\left(\mathbf{x}^{k+1}\right) + \tfrac{z_i^k}{\beta_k}\right]_+^2 \\ &= \tfrac{1}{2\beta_{k+1}}\left[(\beta_{k+1} + \gamma_k)^2 g_i^2(\mathbf{x}^{k+1}) + 2(\beta_{k+1} + \gamma_k) z_i^k g_i\left(\mathbf{x}^{k+1}\right) + \left(z_i^k\right)^2\right] - \tfrac{\beta_k}{2}\left(g_i\left(\mathbf{x}^{k+1}\right) + \tfrac{z_i^k}{\beta_k}\right)^2 \\ &= \tfrac{\gamma_k}{\beta_{k+1}} g_i\left(\mathbf{x}^{k+1}\right)\left(z_i^k + (\beta_{k+1} + \tfrac{\gamma_k}{2}) g_i\left(\mathbf{x}^{k+1}\right)\right) + \tfrac{\beta_{k+1} - \beta_k}{2}(g_i(\mathbf{x}^{k+1}))^2 + \left(\tfrac{1}{2\beta_{k+1}} - \tfrac{1}{2\beta_k}\right)\left(z_i^k\right)^2 \\ &\leq \tfrac{\beta_{k+1} - \beta_k}{2} g_i^2(\mathbf{x}^{k+1}) + \left(\tfrac{1}{2\beta_{k+1}} - \tfrac{1}{2\beta_k}\right)\left(z_i^k\right)^2 \\ &\leq \tfrac{\beta_{k+1} - \beta_k}{2}\tfrac{\left(z_i^k\right)^2}{\beta_k^2} + \left(\tfrac{1}{2\beta_{k+1}} - \tfrac{1}{2\beta_k}\right)\left(z_i^k\right)^2 = \tfrac{(\beta_k - \beta_{k+1})^2}{2\beta_k^2\beta_{k+1}}\left(z_i^k\right)^2, \end{aligned} \tag{A.3}$$

where the first inequality holds because the first term in (A.3) is negative, and the last inequality follows from $g_i^2(\mathbf{x}^{k+1}) \leq \left(z_i^k\right)^2/\beta_k^2$. When $(\beta_{k+1} + \gamma_k) g_i\left(\mathbf{x}^{k+1}\right) + z_i^k < 0$, the above inequality holds trivially as the LHS is non-positive.

Case III: $i \in J_2^k \cap J_+^k$: In this case, (A.3) still holds as $g_i(\mathbf{x}^{k+1}) + \frac{z_i^k}{\beta_k} \geq 0$ and together with $\beta_k \leq \beta_{k+1}$ gives

$$\begin{aligned} &\frac{\beta_{k+1}}{2}\left[g_i\left(\mathbf{x}^{k+1}\right) + \frac{1}{\beta_{k+1}}\left(z_i^k + \gamma_k g_i\left(\mathbf{x}^{k+1}\right)\right)\right]_+^2 - \frac{\beta_k}{2}\left[g_i\left(\mathbf{x}^{k+1}\right) + \frac{z_i^k}{\beta_k}\right]_+^2 \\ &\leq \left(\frac{\beta_{k+1} - \beta_k}{2} + \gamma_k + \frac{\gamma_k^2}{2\beta_{k+1}}\right)\left[g_i(\mathbf{x}^{k+1})\right]_+^2 + \frac{\gamma_k}{\beta_{k+1}} z_i^k g_i\left(\mathbf{x}^{k+1}\right). \end{aligned}$$

Combining the above three cases, we get

$$\begin{aligned} &\frac{\beta_{k+1}}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right) + \frac{\mathbf{z}^{k+1}}{\beta_{k+1}}\right]_+\right\|^2 - \frac{\beta_k}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right) + \frac{\mathbf{z}^k}{\beta_k}\right]_+\right\|^2 \\ &\leq \left(\frac{\beta_{k+1} - \beta_k}{2} + \gamma_k + \frac{\gamma_k^2}{2\beta_{k+1}}\right)\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2 + \frac{\gamma_k}{\beta_{k+1}}\|\mathbf{z}^k\|\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\| + \frac{(\beta_k - \beta_{k+1})^2}{2\beta_k^2\beta_{k+1}}\|\mathbf{z}^k\|^2 \end{aligned} \tag{A.4}$$

Summing up the inequality in (A.4) from $k=0$ to $K-1$ yields

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\left(\frac{\beta_{k+1}}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)+\frac{\mathbf{z}^{k+1}}{\beta_{k+1}}\right]_+\right\|^2-\frac{\beta_k}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)+\frac{\mathbf{z}^{k}}{\beta_{k}}\right]_+\right\|^2\right)\\
&\leq\sum_{k=0}^{K-1}\left(\left(\frac{\beta_{k+1}-\beta_k}{2}+\gamma_k+\frac{\gamma_k^2}{2\beta_{k+1}}\right)\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2+\frac{\gamma_k}{\beta_{k+1}}\|\mathbf{z}^k\|\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|+\frac{(\beta_k-\beta_{k+1})^2}{2\beta_k^2\beta_{k+1}}\left\|\mathbf{z}^k\right\|^2\right)\\
&=\sum_{k=0}^{K-1}\left(\frac{\beta_{k+1}-\beta_k}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2+\gamma_k\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2+\frac{\gamma_k}{2\beta_{k+1}}\gamma_k\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2\right)\\
&\quad+\sum_{k=0}^{K-1}\left(\frac{\|\mathbf{z}^k\|}{\beta_{k+1}}\gamma_k\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|\right)+\sum_{k=0}^{K-1}\frac{(\beta_k-\beta_{k+1})^2}{2\beta_k^2\beta_{k+1}}\left\|\mathbf{z}^k\right\|^2\\
&\leq\sum_{k=0}^{K-1}\frac{\beta_{k+1}-\beta_k}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2+\sum_{k=0}^{K-1}\left(w_k\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|+\frac{\beta_k w_k\|[\mathbf{g}\left(\mathbf{x}^{k+1}\right)]_+\|}{2\beta_{k+1}}\right)+\sum_{k=0}^{K-1}\frac{\|\mathbf{z}^k\|}{\beta_{k+1}}w_k\\
&\quad+\sum_{k=0}^{K-1}\frac{(\beta_k-\beta_{k+1})^2}{2\beta_k^2\beta_{k+1}}\left\|\mathbf{z}^k\right\|^2\\
&\leq\sum_{k=0}^{K-1}\frac{\beta_{k+1}-\beta_k}{2}\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|^2+\sum_{k=0}^{K-1}\frac{3}{2}w_k\left\|\left[\mathbf{g}\left(\mathbf{x}^{k+1}\right)\right]_+\right\|+\sum_{k=0}^{K-1}\frac{w_k}{\beta_{k+1}}\|\mathbf{z}^k\|+\sum_{k=0}^{K-1}\frac{(\beta_k-\beta_{k+1})^2}{2\beta_k^2\beta_{k+1}}\left\|\mathbf{z}^k\right\|^2\\
&\leq\sum_{k=0}^{K-1}\frac{C_P^2(\beta_{k+1}-\beta_k)}{2\beta_k^2}+\frac{3}{2}\sum_{k=0}^{K-1}w_k\frac{C_P}{\beta_k}+\sum_{k=0}^{K-1}\frac{w_k}{\beta_{k+1}}\|\mathbf{z}^k\|+\sum_{k=0}^{K-1}\frac{\beta_{k+1}-\beta_k}{2\beta_k^2}\left\|\mathbf{z}^k\right\|^2\\
&\leq\frac{3C_P^2}{4\beta_0}+\frac{3C_{\mathbf{z}}C_P}{2\beta_0}+\frac{C_{\mathbf{z}}}{\beta_0}C_{\mathbf{z}}+\frac{3C_{\mathbf{z}}^2}{4\beta_0}=\frac{1}{4\beta_0}(3C_P^2+6C_{\mathbf{z}}C_P+7C_{\mathbf{z}}^2),
\end{aligned}
$$

where the second inequality holds from $\gamma_k\|[\mathbf{g}(\mathbf{x}^{k+1})]_+\|\leq w_k$ and $\gamma_k\leq\beta_k$, the third one is due to $\beta_k\leq\beta_{k+1}$, the fourth one uses Lemma 2.1, and the last inequality follows from Lemma 2.1, the bound

$$
\sum_{k=0}^{K-1}\frac{\beta_{k+1}-\beta_k}{\beta_k^2}=\sum_{k=0}^{K-1}\frac{1}{\beta_0(k+1)(\sqrt{k+2}+\sqrt{k+1})}\leq\frac{1}{2\beta_0}+\int_1^K\frac{1}{2\beta_0x^{\frac{3}{2}}}dx\leq\frac{3}{2\beta_0},\tag{A.5}
$$

$\beta_0\leq\beta_k$ for all $k\geq0$, and $\sum_{k=0}^{\infty}w_k\leq C_{\mathbf{z}}$. The proof of (2.8) is then completed.

To prove (2.9), we start by noticing $\left\langle\mathbf{y}^{k+1}-\mathbf{y}^k,\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\right\rangle=\alpha_k\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|^2\leq v_k\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|$ from the update of $\mathbf{y}^k$ and definition of α_k. Hence, by Lemma 2.1, the definitions of v_k,β_k, and (2.4), it follows that $\sum_{k=0}^{K-1}\langle\mathbf{y}^{k+1}-\mathbf{y}^k,\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\rangle\leq\sum_{k=0}^{K-1}C_Pv_k/\beta_k\leq C_{\mathbf{y}}C_P/\beta_0$, where we have used $\sum_{k=0}^{K-1}v_k\leq C_{\mathbf{y}}$, and $\beta_k\geq\beta_0$ for all $k\geq0$. In addition, we have

$$
\textstyle\sum_{k=0}^{K-1}\frac{\beta_{k+1}-\beta_k}{2}\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|^2\overset{(2.4)}{\leq}\sum_{k=0}^{K-1}\frac{\beta_{k+1}-\beta_k}{2\beta_k^2}C_P^2\leq\frac{3C_P^2}{4\beta_0},
$$

where the second inequality is obtained from (A.5). Therefore, we complete the proof.

Appendix B. Nesterov's Accelerated Proximal Gradient (APG) Method. In this section, we review Nesterov's APG method in [35] for solving composite convex problems in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^d} \phi(\mathbf{x}) := \widetilde{f}(\mathbf{x}) + \widetilde{h}(\mathbf{x}), \tag{B.1}$$

where $\widetilde{f}$ is convex and $L_{\widetilde{f}}$-smooth, and $\widetilde{h}$ is closed and μ-strongly convex with $\mu > 0$.

The algorithm is shown in Algorithm 2, where $\phi'(\mathcal{M}_L(\mathbf{y})) := L(\mathbf{y} - \mathcal{M}_L(\mathbf{y})) + \nabla\widetilde{f}(\mathcal{M}_L(\mathbf{y})) - \nabla\widetilde{f}(\mathbf{y})$, $\mathcal{M}_L(\mathbf{y}) := \mathbf{prox}_{\widetilde{h}/L}(\mathbf{y} - \nabla\widetilde{f}(\mathbf{y})/L)$ for some $L > 0$, and $\psi_{t+1}(\mathbf{x}) := \psi_t(\mathbf{x}) + a_{t+1}[\widetilde{f}(\mathbf{x}^{t+1}) + \langle\nabla\widetilde{f}(\mathbf{x}^{t+1}), \mathbf{x} - \mathbf{x}^{t+1}\rangle + \widetilde{h}(\mathbf{x})]$ with a positive sequence $\{a_t\}$ and $\psi_0(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^0\|^2$.

Algorithm 2: Nesterov's Accelerated Proximal Gradient (APG) Method for (B.1)

Initialization: choose $\mathbf{x}^0, L := L_0, \gamma_d \geq \gamma_u > 1, \Delta > 0$ and set $A_0 = 0, \mathbf{v}^0 = \mathbf{y}^0 = \mathbf{y} = \mathbf{x}^0$.
for $t = 0, 1, \ldots$ **do**
  while $\langle\phi'(\mathcal{M}_L(\mathbf{y})), \mathbf{y} - \mathcal{M}_L(\mathbf{y})\rangle < \frac{1}{L}\|\phi'(\mathcal{M}_L(\mathbf{y}))\|^2$ **do**
    $L \leftarrow L\gamma_u$; let $a > 0$ and satisfy $\frac{a^2}{A_t + a} = 2\frac{1+\mu A_t}{L}$; $\mathbf{y} = \frac{A_t\mathbf{x}^t + a\mathbf{v}^t}{A_t + a}$.
  Set $L_{t+1} \leftarrow L$, $\mathbf{y}^t \leftarrow \mathbf{y}, a_{t+1} = a$, $L \leftarrow L_{t+1}/\gamma_d$, $\mathbf{x}^{t+1} \leftarrow \mathcal{M}_{L_{t+1}}(\mathbf{y}^t)$, $A_{t+1} = A_t + a_{t+1}$. Let $\mathbf{v}^{t+1} := \arg\min_{\mathbf{x}} \psi_{t+1}(\mathbf{x}) = \mathrm{prox}_{A_{t+1}\widetilde{h}(\cdot)}(\mathbf{x}^0 - \sum_{i=1}^{t+1} a_i\nabla\widetilde{f}(\mathbf{x}^i))$.
  if $\mathrm{dist}(\mathbf{0}, \partial\phi(\mathbf{x}^{t+1})) \leq \Delta$ **then** output $\mathbf{x}^{t+1}$ and stop.

The next theorem gives the number of iterations for Algorithm 2 to output the desired solution.

THEOREM B.1. *Suppose that $\{\mathbf{x}^t\}$ is the sequence generated by Algorithm 2 and $\mathrm{dom}(\widetilde{h})$ is bounded with a diameter $D = \max_{\mathbf{x},\mathbf{x}'\in\mathrm{dom}(\widetilde{h})}\|\mathbf{x} - \mathbf{x}'\| < \infty$. Then $\mathrm{dist}(\mathbf{0}, \partial\phi(\mathbf{x}^{t+1})) \leq \Delta, \forall\, t \geq T$, where*

$$T = \left\lceil \max\left\{1/\log 2, 2\sqrt{(\gamma_u L_{\widetilde{f}})/(2\mu)}\right\} \log\frac{3(1+\gamma_u)DL_{\widetilde{f}}\sqrt{2\gamma_u L_{\widetilde{f}}/\mu}}{2\Delta}\right\rceil + 1.$$

Proof. Let $\mathbf{x}^*$ be the minimizer of problem (B.1). Then from [35, Theorem 6], it holds

$$\phi(\mathbf{x}^{t+1}) - \phi(\mathbf{x}^*) \leq \frac{\gamma_u L_{\widetilde{f}}}{4}\|\mathbf{x}^* - \mathbf{x}^0\|^2\left[1 + \sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{-2t}. \tag{B.2}$$

By the optimality condition in the definition of $\mathcal{M}_{L_{t+1}}(\mathbf{y}^t)$, we have $\nabla\widetilde{f}(\mathbf{x}^{t+1}) - \nabla\widetilde{f}(\mathbf{y}^t) - L_{t+1}\left(\mathbf{x}^{t+1} - \mathbf{y}^t\right) \in \partial\phi(\mathbf{x}^{t+1})$. Also, from [35, Eqn. (4.11)], it holds $L_{t+1} \leq \gamma_u L_{\widetilde{f}}$. Using these, we obtain

$$\begin{aligned}\mathrm{dist}\left(0, \partial\phi\left(\mathbf{x}^{t+1}\right)\right) &\leq \left\|\nabla\widetilde{f}(\mathbf{x}^{t+1}) - \nabla\widetilde{f}(\mathbf{y}^t) - L_{t+1}\left(\mathbf{x}^{t+1} - \mathbf{y}^t\right)\right\| \leq L_{\widetilde{f}}\left\|\mathbf{x}^{t+1} - \mathbf{y}^t\right\| + L_{t+1}\left\|\mathbf{x}^{t+1} - \mathbf{y}^t\right\| \\ &\leq L_{\widetilde{f}}(1+\gamma_u)\left\|\mathbf{x}^{t+1} - \mathbf{y}^t\right\| \leq L_{\widetilde{f}}(1+\gamma_u)\left(\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\| + \left\|\mathbf{x}^t - \mathbf{y}^t\right\|\right).\end{aligned} \tag{B.3}$$

Notice that ψ_t is a $(\mu A_t + 1)$-strongly convex function. Hence,

$$\begin{aligned}\left\|\mathbf{x}^t - \mathbf{v}^t\right\|^2 &\leq \frac{2}{\mu A_t + 1}\left(\psi_t(\mathbf{x}^t) - \psi_t^*\right) \leq \frac{2}{\mu A_t + 1}\left(A_t\phi(\mathbf{x}^t) - \psi_t^* + \frac{1}{2}\left\|\mathbf{x}^t - \mathbf{x}^0\right\|^2\right) \\ &\leq \frac{1}{\mu A_t + 1}\left\|\mathbf{x}^t - \mathbf{x}^0\right\|^2 \leq \frac{D^2}{\mu A_t + 1},\end{aligned} \tag{B.4}$$

where $\psi_t^* := \min_{\mathbf{x}} \psi_t(\mathbf{x})$, and the second/third inequality is from [35, Eqn. (4.4)].

Combining the μ-strong convexity of $\phi(\mathbf{x})$ and (B.2) gives

$$\frac{\mu}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2 \leq \phi(\mathbf{x}^{t+1})-\phi(\mathbf{x}^*) \leq \frac{\gamma_u L_{\widetilde{f}}}{4}\|\mathbf{x}^*-\mathbf{x}^0\|^2\left[1+\sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{-2t}, \tag{B.5}$$

which implies

$$\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\| \leq \sqrt{\gamma_u L_{\widetilde{f}}/(2\mu)}\left\|\mathbf{x}^*-\mathbf{x}^0\right\|\left[1+\sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{-t} \leq D\sqrt{\gamma_u L_{\widetilde{f}}/(2\mu)}\left[1+\sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{-t}. \tag{B.6}$$

Since $\mathbf{y}^t = \frac{A_t\mathbf{x}^t + a_{t+1}\mathbf{v}^t}{A_t+a_{t+1}}$ and $a_{t+1}>0$, it must hold $\|\mathbf{y}^t-\mathbf{x}^t\| \leq \|\mathbf{v}^t-\mathbf{x}^t\|$. Using this in (B.3) and combining it with (B.4), we get

$$\operatorname{dist}\left(0,\partial\phi\left(\mathbf{x}^{t+1}\right)\right) \leq (1+\gamma_u)L_{\widetilde{f}}\left(\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\| + D/\sqrt{\mu A_t+1}\right). \tag{B.7}$$

By the triangle inequality, we have $\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\| \leq \left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\| + \|\mathbf{x}^t-\mathbf{x}^*\|$, which together with (B.6) gives

$$\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\| \leq D\sqrt{2\gamma_u L_{\widetilde{f}}/\mu}\left[1+\sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{-t+1}. \tag{B.8}$$

Now by (B.7), (B.8), and the fact that $A_t \geq 2/(\gamma_u L_{\widetilde{f}})\left[1+\sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{2(t-1)}$ from [35, Lemma 8], we have

$$\operatorname{dist}\left(0,\partial\phi\left(\mathbf{x}^{t+1}\right)\right) \leq \frac{3}{2}(1+\gamma_u)DL_{\widetilde{f}}\left(\sqrt{2\gamma_u L_{\widetilde{f}}/\mu}\right)\left[1+\sqrt{\mu/(2\gamma_u L_{\widetilde{f}})}\right]^{-t+1}, \tag{B.9}$$

which indicates $\operatorname{dist}(\mathbf{0},\partial\phi(\mathbf{x}^{t+1})) \leq \Delta, \forall\, t \geq T$ from the definition of T and $\frac{1}{\log(1+x)} \leq \frac{2}{x}, \forall\, x \in (0,1)$. □

REFERENCES

[1] Z. Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *International Conference on Machine Learning*, pages 89–97. PMLR, 2017. 2

[2] N. S. Aybat and G. Iyengar. A first-order smoothed penalty method for compressed sensing. *SIAM Journal on Optimization*, 21(1):287–313, 2011. 2

[3] A. Bayandina, P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. *Large-scale and distributed optimization*, pages 181–213, 2018. 2

[4] Y. Chen, T.-Z. Huang, W. He, X.-L. Zhao, H. Zhang, and J. Zeng. Hyperspectral image denoising using factor group sparsity-regularized non-convex low-rank approximation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 1

[5] F. H. Clarke. *Optimization and non-smooth Analysis*, volume 5. SIAM, Philadelphia, 1990. 5

[6] F. E. Curtis and M. L. Overton. A sequential quadratic programming algorithm for non-convex, non-smooth constrained optimization. *SIAM Journal on Optimization*, 22(2):474–500, 2012. 3

[7] M. Danilova, P. Dvurechensky, A. Gasnikov, E. Gorbunov, S. Guminov, D. Kamzolov, and I. Shibaev. Recent theoretical advances in non-convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*, pages 79–163. Springer, 2022. 1

[8] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $\mathcal{O}(k^{-1/4})$ on weakly convex functions. *ArXiv, preprint:1802.02988*, 2018. 2

[9] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. 2

[10] D. Davis and B. Grimmer. Proximally guided stochastic subgradient method for non-smooth, non-convex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019. 2

[11] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019. 2, 3, 4, 5, 12, 13, 14, 20, 21

[12] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for non-convex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. 2

[13] S. Ghadimi and G. Lan. Accelerated gradient methods for non-convex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016. 2

[14] M. L. Goncalves, J. G. Melo, and R. D. Monteiro. Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving non-convex linearly constrained problems. *ArXiv, preprint:1702.01850*, 2017. 3

[15] D. Hajinezhad and M. Hong. Perturbed proximal primal–dual algorithm for non-convex non-smooth optimization. *Mathematical Programming*, 176(1-2):207–245, 2019. 3

[16] M. Hong. Decomposing linearly constrained non-convex problems by a proximal primal-dual approach: algorithms, convergence, and applications. *ArXiv, preprint:1604.00543*, 2016. 3

[17] Y. Huang and Q. Lin. Single-loop switching subgradient methods for non-smooth weakly convex optimization with non-smooth convex constraints. *ArXiv, preprint:2301.13314*, 2023. 1, 3

[18] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured non-convex and non-smooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 72(1):115–157, 2019. 3

[19] W. Kong, J. G. Melo, and R. D. Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained non-convex composite programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019. 3

[20] W. Kong, J. G. Melo, and R. D. Monteiro. Iteration complexity of a proximal augmented Lagrangian method for solving non-convex composite optimization problems with nonlinear convex constraints. *Mathematics of Operations Research*, 2022. 3, 5, 17, 18, 19, 20

[21] G. Lan and R. D. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1-2):115–139, 2013. 2

[22] G. Lan and R. D. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016. 2

[23] G. Lan and Y. Yang. Accelerated stochastic algorithms for non-convex finite-sum and multi-block optimization. *SIAM Journal on Optimization*, 29(4):2753–2784, 2019. 2

[24] G. Lan and Z. Zhou. Algorithms for stochastic optimization with functional or expectation constraints. *ArXiv, preprint:1604.03887*, 2016. 2

[25] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Rate-improved inexact augmented Lagrangian method for constrained non-convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2170–2178. PMLR, 2021. 3

[26] Z. Li and Y. Xu. Augmented Lagrangian–based first-order methods for convex-constrained programs with weakly convex objective. *INFORMS Journal on Optimization*, 3(4):373–397, 2021. 3, 5, 9, 11, 13, 15, 17, 18, 19, 20

[27] Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational Optimization and Applications*, 82(1):175–224, 2022. 2, 3, 5, 10

[28] Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *International conference on machine learning*, pages 3112–3121. PMLR, 2018. 2

[29] Y.-F. Liu, X. Liu, and S. Ma. On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Mathematics of Operations Research*, 44(2):632–650, 2019. 2

[30] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *ArXiv, preprint:1803.09941*, 2018. 2

[31] J. G. Melo and R. D. Monteiro. Iteration-complexity of a Jacobi-type non-Euclidean ADMM for multi-block linearly constrained non-convex programs. *ArXiv, preprint:1705.07229*, 2017. 3

[32] J. G. Melo, R. D. Monteiro, and H. Wang. Iteration-complexity of an inexact proximal accelerated augmented Lagrangian method for solving linearly constrained smooth non-convex composite optimization problems. *ArXiv, preprint:2006.08048*, 2020. 3

[33] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optimization Methods and Software*, 34(2):305–335, 2019. 2

[34] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003. 17

[35] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. 8, 24, 25

[36] A. Patrascu, I. Necoara, and Q. Tran-Dinh. Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. *Optimization Letters*, 11:609–626, 2017. 2

[37] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for non-convex optimization. In

International conference on machine learning, pages 314–323. PMLR, 2016. 2

[38] P. Rigollet and X. Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12(Oct):2831–2855, 2011. 1

[39] M. F. Sahin, A. Alacaoglu, F. Latorre, V. Cevher, et al. An inexact augmented Lagrangian framework for non-convex optimization with nonlinear constraints. In *Advances in Neural Information Processing Systems*, pages 13965–13977, 2019. 3

[40] Q. Tran-Dinh and V. Cevher. A primal-dual algorithmic framework for constrained convex minimization. *ArXiv, preprint:1406.5403*, 2014. 2

[41] X. Wei and M. J. Neely. Primal-dual frank-wolfe for constrained stochastic programs with convex and non-convex objectives. *ArXiv, preprint:1806.00709*, 2018. 2

[42] X. Wei, H. Yu, Q. Ling, and M. Neely. Solving non-smooth constrained programs with lower complexity than $\mathcal{O}(\frac{1}{\varepsilon})$: A primal-dual homotopy smoothing approach. *Advances in Neural Information Processing Systems*, 31, 2018. 2

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

[44] Y. Xu. First-order methods for constrained convex programming based on linearized augmented Lagrangian function. *INFORMS Journal on Optimization*, 3(1):89–117, 2021. 2

[45] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Mathematical Programming*, 185(185):199–244, 2021. 2, 21

[46] Y. Xu. First-order methods for problems with $O(1)$ functional constraints can have almost the same convergence rate as for unconstrained problems. *SIAM Journal on Optimization*, 32(3):1759–1790, 2022. 17

[47] Y. Xu, W. Yin, Z. Wen, and Y. Zhang. An alternating direction algorithm for matrix completion with nonnegative factors. *Frontiers of Mathematics in China*, 7:365–384, 2012. 1

[48] H. Yu and M. J. Neely. A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017. 2

[49] J. Zeng, W. Yin, and D.-X. Zhou. Moreau envelope augmented Lagrangian method for non-convex optimization with linear constraints. *Journal of Scientific Computing*, 91(2):61, 2022. 3, 4, 17, 18, 19

[50] J. Zhang and Z.-Q. Luo. A proximal alternating direction method of multiplier for linearly constrained non-convex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020. 3

[51] J. Zhang and Z.-Q. Luo. A global dual error bound and its application to the analysis of linearly constrained non-convex optimization. *SIAM Journal on Optimization*, 32(3):2319–2346, 2022. 3

[52] S. Zhang and N. He. On the convergence rate of stochastic mirror descent for non-smooth non-convex optimization. *ArXiv, preprint:1806.04781*, 2018. 2