Journal of Scientific Computing (2021) 86:38
https://doi.org/10.1007/s10915-021-01409-y

# Complexity of Proximal Augmented Lagrangian for Nonconvex Optimization with Nonlinear Equality Constraints

Yue Xie[1] · Stephen J. Wright[2]

Received: 11 October 2019 / Revised: 1 September 2020 / Accepted: 15 January 2021 /
Published online: 2 February 2021

**Abstract**
We analyze worst-case complexity of a Proximal augmented Lagrangian (Proximal AL) framework for nonconvex optimization with nonlinear equality constraints. When an approximate first-order (second-order) optimal point is obtained in the subproblem, an $\epsilon$ first-order (second-order) optimal point for the original problem can be guaranteed within $\mathcal{O}(1/\epsilon^{2-\eta})$ outer iterations (where $\eta$ is a user-defined parameter with $\eta \in [0, 2]$ for the first-order result and $\eta \in [1, 2]$ for the second-order result) when the proximal term coefficient $\beta$ and penalty parameter $\rho$ satisfy $\beta = \mathcal{O}(\epsilon^{\eta})$ and $\rho = \Omega(1/\epsilon^{\eta})$, respectively. We also investigate the total iteration complexity and operation complexity when a Newton-conjugate-gradient algorithm is used to solve the subproblems. Finally, we discuss an adaptive scheme for determining a value of the parameter $\rho$ that satisfies the requirements of the analysis.

**Keywords** Optimization with nonlinear equality constraints · Nonconvex optimization · Proximal augmented Lagrangian · Complexity analysis · Newton-conjugate-gradient

**Mathematics Subject Classification** 68Q25 · 90C06 · 90C26 · 90C30 · 90C60

## 1 Introduction

Nonconvex optimization problems with nonlinear equality constraints are common in some areas, including matrix optimization and machine learning, where such requirements as normalization, orthogonality, or consensus must be satisfied. Relevant problems include dictionary learning [34], distributed optimization [26], and spherical PCA [28]. We consider

✉ Yue Xie
xieyue1990@gmail.com

Stephen J. Wright
swright@cs.wisc.edu

[1] Wisconsin Institute for Discovery, University of Wisconsin, 330 N. Orchard St., Madison, WI 53715, USA

[2] Computer Sciences Department, University of Wisconsin, 1210 W. Dayton St., Madison, WI 53706, USA

the formulation

$$\min\ f(x) \quad \text{subject to} \quad c(x) = 0, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c(x) \triangleq (c_1(x), \ldots, c_m(x))^T$, $c_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, m$, and all functions are twice continuously differentiable.

We have the following definitions related to points that satisfy approximate first- and second-order optimality coniditions for (1). (Here and throughout, $\|\cdot\|$ denotes the Euclidean norm of a vector.)

**Definition 1** ($\epsilon$-1o) We say that $x$ is an $\epsilon$-1o solution of (1) if there exists $\lambda \in \mathbb{R}^m$ such that

$$\|\nabla f(x) + \nabla c(x)\lambda\| \leq \epsilon, \quad \|c(x)\| \leq \epsilon.$$

**Definition 2** ($\epsilon$-2o) We say that $x$ is an $\epsilon$-2o solution of (1) if there exists $\lambda \in \mathbb{R}^m$ such that:

$$\|\nabla f(x) + \nabla c(x)\lambda\| \leq \epsilon, \quad \|c(x)\| \leq \epsilon, \tag{2a}$$

$$d^T \left( \nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 c_i(x) \right) d \geq -\epsilon \|d\|^2, \tag{2b}$$

for any $d \in S(x) \triangleq \{d \in \mathbb{R}^n \mid \nabla c(x)^T d = 0\}$.

These definitions are consistent with those of $\epsilon$-KKT and $\epsilon$-KKT2 in [11], and similar to those of [23], differing only in choice of norm and use of $\|c(x)\| \leq \epsilon$ rather than $c(x) = 0$. The following theorem is implied by several results in [4,11], which consider a larger class of problem than (1). (A proof tailored to (1) is supplied in the "Appendix".)

**Theorem 1** *If $x^*$ is an local minimizer of (1), then there exists $\epsilon_k \to 0^+$ and $x_k \to x^*$ such that $x_k$ is $\epsilon_k$-2o, thus $\epsilon_k$-1o.*

Theorem 1 states that being the limit of a sequence of points satisfying Definitions 1 or 2 for a decreasing sequence of $\epsilon$ is a necessary condition of a local minimizer. When certain constraint qualifications hold, a converse of this result is also true: $x^*$ satisfies first-order (KKT) conditions when $x_k$ is $\epsilon_k$-1o and second-order conditions when $x_k$ is $\epsilon_k$-2o (See [4,5]). These observations justify our strategy of seeking points that satisfy Definitions 1 or 2.

The augmented Lagrangian (AL) framework is a penalty-type algorithm for solving (1), originating with Hestenes [25] and Powell [31]. Rockafellar proposed in [32] the proximal version of this method, which has both theoretical and practical advantages. The monograph [7] summarizes development of this method during the 1970s, when it was known as the "method of multipliers". Interest in the algorithm has resurfaced in recent years because of its connection to ADMM [13].

The augmented Lagrangian of (1) is defined as:

$$\mathcal{L}_\rho(x, \lambda) \triangleq f(x) + \sum_{i=1}^m \lambda_i c_i(x) + \frac{\rho}{2} \sum_{i=1}^m |c_i(x)|^2 = f(x) + \lambda^T c(x) + \frac{\rho}{2} \|c(x)\|^2,$$

where $\lambda \triangleq (\lambda_1, \ldots, \lambda_m)^T$. The (ordinary) Lagrangian of (1) is $\mathcal{L}_0(x, \lambda)$.

### 1.1 Complexity Measures

In this paper, we discuss measures of worst-case complexity for finding points that satisfy Definitions 1 and 2. Since our method has two nested loops — an outer loop for the Proximal AL procedure, and an inner loop for solving the subproblems — we consider the following measures of complexity.

- *Outer iteration complexity,* which corresponds to the number of outer-loop iterations of Proximal AL or some other framework;
- *Total iteration complexity,* which measures the total number of iterations of the inner-loop procedure that is required to find points satisfying approximate optimality of the subproblems;
- *Operation complexity,* which measures the number of some unit operation (in our case, computation of a matrix-vector product involving the Hessian of the Proximal augmented Lagrangian) required to find approximately optimal points.

We also use the term "total iteration complexity" in connection with algorithms that have only one main loop, such as those whose complexities are shown in Table 1.

We prove results for all three types of complexity for the Proximal AL procedure, where the inner-loop procedure is a Newton-conjugate-gradient (Newton-CG) algorithm for the unconstrained nonconvex subproblems. Details are given in Sect. 1.3.

**Algorithm 1** Augmented Lagrangian (AL)

```
0. Initialize $x_0$, $\lambda_0$ and $\rho_0 > 0$, $\Lambda \triangleq [\Lambda_{\min}, \Lambda_{\max}]$, $\tau \in (0, 1)$, $\gamma > 1$; Set $k \leftarrow 0$;
1. Update $x_k$: Find approximate solution $x_{k+1}$ to $\min_x \mathcal{L}_{\rho_k}(x, \lambda_k)$;
2. Update $\lambda_k$: $\lambda_{k+1} \leftarrow P_\Lambda(\lambda_k + \rho_k c(x_{k+1}))$;
3. Update $\rho_k$: if $k = 0$ or $\|c(x_{k+1})\|_\infty \le \tau\|c(x_k)\|_\infty$, set $\rho_{k+1} = \rho_k$; otherwise, set $\rho_{k+1} = \gamma\rho_k$;
4. If termination criterion is satisfied, STOP; otherwise, $k \leftarrow k + 1$ and return to Step 1.
```

### 1.2 Related Work

*AL for nonconvex optimization.* We consider first the basic augmented Lagrangian framework outlined in Algorithm 1. When $f$ is a nonconvex function, convergence of the augmented Lagrangian framework has been studied in [9,11], with many variants described in [1–3,6,19]. In [11], Algorithm 1 is investigated and generalized for a larger class of problems, showing in particular that if $x_{k+1}$ is a first-order (second-order) approximate solution of the subproblem, with error driven to 0 as $k \to \infty$, then every feasible limit point is an approximate first-order (second-order) KKT point of the original problem. In [9], it is shown that when the subproblem in Algorithm 1 is solved to approximate global optimality with error approaching 0, the limit point is feasible and is a global solution of the original problem.

There are few results in the literature on outer iteration complexity in the nonconvex setting. Some quite recent results appear in [12,22]. In [22], the authors apply a general version of augmented Lagrangian to nonconvex optimization with both equality and inequality constraints. With an aggressive updating rule for the penalty parameter, they show that the algorithm obtains an approximate KKT point (whose exact definition is complicated, but similar to our definition of $\epsilon$-1o optimality when only equality constraints are present) within $\mathcal{O}(\epsilon^{-2/(\alpha-1)})$ outer-loop iterations, where $\alpha > 1$ is an algorithmic parameter. This complexity is improved to $\mathcal{O}(|\log \epsilon|)$ when boundedness of the sequence of penalty parameters is assumed. Total iteration complexity measures are obtained for the case of linear equality constraints when the subproblem is solved with a $p$-order method ($p \geq 2$). In [12], the authors study an augmented Lagrangian framework named ALGENCAN to problems with equality and inequality constraints. An $\epsilon$-accurate first-order point (whose precise definition is again similar to our $\epsilon$-1o optimality in the case of equality constraints only) is obtained in $\mathcal{O}(|\log \epsilon|)$ outer iterations when the penalty parameters are bounded. The practicality of the assumption of bounded penalty parameters in these two works [12,22] is open to question, since the use of an increasing sequence of penalty parameters is critical to both approaches, and there is no clear prior reason why the sequence should be bounded[1].

*Proximal AL for nonconvex optimization: Linear equality constraints.* The Proximal augmented Lagrangian framework, with fixed positive parameters $\rho$ and $\beta$, is shown in Algorithm 2.

**Algorithm 2** Proximal augmented Lagrangian (Proximal AL)

```
0. Initialize $x_0$, $\lambda_0$ and $\rho > 0$, $\beta > 0$; Set $k \leftarrow 0$;
1. Update $x_k$: Find approximate solution $x_{k+1}$ to $\min_x \ \mathcal{L}_\rho(x, \lambda_k) + \frac{\beta}{2}\|x - x_k\|^2$;
2. Update $\lambda_k$: $\lambda_{k+1} \leftarrow \lambda_k + \rho c(x_{k+1})$;
3. If termination criterion is satisfied, STOP; otherwise, $k \leftarrow k + 1$ and return to Step 1.
```

For this proximal version, in the case of *linear* constraints $c(\cdot)$, outer iteration complexity results become accessible in the nonconvex regime [24,26,27,35]. The paper [26] analyzes the outer iteration complexity of this approach (there named "proximal primal dual algorithm (Prox-PDA)") to obtain a first-order optimal point, choosing a special proximal term to make each subproblem strongly convex and suitable for distributed implementation. An outer iteration complexity estimate of $\mathcal{O}(\epsilon^{-1})$ is proved for an $\sqrt{\epsilon}$-1o point. This result is consistent with our results in this paper when the choice of $\beta$ and $\rho$ is independent of $\epsilon$ and $c(x)$ is linear.

The paper [24] proposes a "perturbed proximal primal dual algorithm," a variant of Algorithm 2, to obtain outer iteration complexity results for a problem class where the objective function may be nonconvex and nonsmooth. In particular, an outer iteration complexity of $\mathcal{O}(\epsilon^{-2})$ is required to obtain $\epsilon$-stationary solution, where the latter term is defined in a way that suits that problem class. A modified inexact Proximal AL method is investigated in [35]. Here, an exponentially weighted average of previous updates is used as the anchor point in the proximal term, total iteration complexity of $\mathcal{O}(\epsilon^{-2})$ to locate an $\epsilon$ stationary point similar to $\epsilon$-1o is derived and a certain kind of linear convergence is proved for quadratic programming (QP). The paper [27] derives outer iteration complexity of $\mathcal{O}(\epsilon^{-2})$ for a proximal ADMM procedure to find an $\epsilon$ stationary solution defined for their problem class.

To our knowledge, outer iteration complexity of Proximal AL in the case of *nonlinear* $c(x)$ and its complexity for convergence to second-order optimal points have not yet been studied.

*Complexity for constrained nonconvex optimization.* For constrained nonconvex optimization, worst-case total iteration complexity results of various algorithms to find $\epsilon$-perturbed first-order and second-order optimal points have been obtained in recent years. If only first-derivative information is used, total iteration complexity to obtain an $\epsilon$-accurate first-order

[1] Circumstances under which the penalty parameter sequence of ALGENCAN is bounded are discussed in [1, Section 5].

optimal point may be $\mathcal{O}(\epsilon^{-2})$ [8,23,29]. If Hessian information is used (either explicitly or via Hessian-vector products), total iteration complexity for an $\epsilon$-accurate first-order point can be improved to $\mathcal{O}(\epsilon^{-3/2})$ [8,23,30], while the total iteration complexity to obtain an $\epsilon$-accurate second-order point is typically $\mathcal{O}(\epsilon^{-3})$ [8,23,29,30]. More details about these results can be found in Table 1.

Other approaches focus on nonlinear equality constraints and seek evaluation complexity bounds ("Evaluation complexity" refers to the number of evaluations of $f$ and $c$ and their derivatives required, and corresponds roughly to our "total iteration complexity".) for approximate first-order optimality. An algorithm based on linear approximation of the exact penalty function for (1) is described in [14], and attains a worst-case evaluation complexity of $\mathcal{O}(\epsilon^{-5})$ by using only function and gradient information. Two-phase approaches, which first seek an approximately feasible point by minimizing the nonlinear least-squares objective $\|c(x)\|_2^2$ (or equivalently $\|c(x)\|$), and then apply a target-chasing method to find an approximate first-order point for (1), are described in [16,17]. (See Table 1.) Extensions of these techniques to approximate second-order optimality is not straightforward; most such efforts focus on special cases such as convex constraints. A recent work that tackles the general case is [18], which again considers the two-phase approach and searches for approximate first-, second-, and third-order critical points. Specific definitions of the critical points are less interpretable; we do not show them in Table 1. They are related to scaled KKT conditions for the first order point, and to local optimality with tolerance of a function of $\epsilon$ for second and third order points.

### 1.3 Contributions

We apply the Proximal AL framework of Algorithm 2 to (1) for nonlinear constraints $c(x)$. Recalling Definitions 1 and 2 of approximately optimal points, we show that when approximate first-order (second-order) optimality is attained in the subproblems, the outer iteration complexity to obtain an $\epsilon$-1o ($\epsilon$-2o) point is $\mathcal{O}(1/\epsilon^{2-\eta})$ if we let $\beta = \mathcal{O}(\epsilon^{\eta})$ and $\rho = \Omega(1/\epsilon^{\eta})$, where $\eta$ is a user-defined parameter with $\eta \in [0, 2]$ for the first-order result and $\eta \in [1, 2]$ for the second-order result. We require uniform boundedness and full rank of the constraint Jacobian on a certain bounded level set, and show that the primal and dual sequence of Proximal AL is bounded and the limit point satisfies first-order KKT conditions.

We also derive total iteration complexity of the algorithm when the Newton-CG algorithm of [33] is used to solve the subproblem at each iteration of Algorithm 2. The operation complexity for this overall procedure is also described, taking as unit operation the computation of a Hessian-vector product. When $c(x)$ is linear and $\eta = 2$, the total iteration complexity matches the known results in literature for second-order algorithms: $\mathcal{O}(\epsilon^{-3/2})$ for an $\epsilon$-1o point and $\mathcal{O}(\epsilon^{-3})$ for an $\epsilon$-2o point.

Finally, we present a scheme for determining the algorithmic parameter $\rho$ adaptively, by increasing it until convegence to an approximately-optimal point is identified within the expected number of iterations.

### 1.4 Organization

In Sect. 2, we list the notations and main assumptions used in the paper. We discuss outer iteration complexity of Proximal AL in Sect. 3. Total iteration complexity and operation complexity are derived in Sect. 4. A framework for determining the parameter $\rho$ in Proximal AL is proposed in Sect. 5. We summarize the paper and discuss future work in Sect. 6.

**Table 1** Total iteration or evaluation complexity estimates for constrained nonconvex optimization procedures

| Point type | Complexity | Constraint type | Lit. |
|---|---|---|---|
| $\begin{cases} \lvert[X\nabla f(x)]_i\rvert \le \epsilon, & \text{if } x_i < (1-\epsilon/2)b_i \\ [\nabla f(x)]_i \le \epsilon, & \text{if } x_i \ge (1-\epsilon/2)b_i \end{cases}$ | $\mathcal{O}(\epsilon^{-2})$ (gradient) | $0 \le x \le b$ | [8] |
| $\Vert X\nabla f(x)\Vert_\infty \le \epsilon,\ X\nabla^2 f(x)X \succeq -\sqrt{\epsilon}I_n$ | $\mathcal{O}(\epsilon^{-3/2})$ (Hessian) | $x \ge 0$ | [8] |
| $\begin{cases} Ax = b,\ x > 0,\ \nabla f(x) + A^T\lambda \ge -\epsilon\mathbf{1} \\ \Vert X(\nabla f(x) + A^T\lambda)\Vert_\infty \le \epsilon \end{cases}$ | $\mathcal{O}(\epsilon^{-2})$ (gradient) | $Ax = b, x \ge 0$ | [23] |
| $\begin{cases} Ax = b,\ x > 0,\ \nabla f(x) + A^T\lambda \ge -\epsilon\mathbf{1} \\ \Vert X(\nabla f(x) + A^T\lambda)\Vert_\infty \le \epsilon \\ d^T(X\nabla^2 f(x)X + \sqrt{\epsilon}I)d \ge 0, \\ \forall d \in \{d \mid AXd = 0\} \end{cases}$ | $\mathcal{O}(\epsilon^{-3/2})$ (Hessian) | $Ax = b, x \ge 0$ | [23] |
| $\left\{\begin{array}{l} \left\lvert \left\{\begin{array}{l} \min_s \langle\nabla f(x), s\rangle \\ s.t.\ x + s \in \mathcal{F},\ \Vert s\Vert \le 1 \end{array}\right\} \right\rvert \le \epsilon_g \\ \left\lvert \left\{\begin{array}{l} \min_d d^T\nabla^2 f(x)d \\ s.t.\ x + d \in \mathcal{F},\ \Vert d\Vert \le 1, \\ \quad \langle\nabla f(x), d\rangle \le 0 \end{array}\right\} \right\rvert \le \epsilon_H \end{array}\right.$ | $\mathcal{O}(\max\{\epsilon_g^{-2}, \epsilon_H^{-3}\})$<br>(Hessian) | $x \in \mathcal{F}$,<br>$\mathcal{F}$is closed<br>and convex | [29] |
| $\begin{cases} x > 0,\ \nabla f(x) \ge -\epsilon\mathbf{1},\ \Vert\bar{X}\nabla f(x)\Vert_\infty \le \epsilon, \\ \bar{X}\nabla^2 f(x)\bar{X} \succeq -\sqrt{\epsilon}I \end{cases}$ | $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ (Hessian) | $x \ge 0$ | [30] |
| $\Vert\nabla f(x) + \nabla c(x)\lambda\Vert \le \epsilon, \Vert c(x)\Vert \le \epsilon,$ or<br>$x$is an approximate critical point of$\Vert c(x)\Vert$ | $\mathcal{O}(\epsilon^{-5})$ (gradient) | $c(x) = 0$ | [14] |
| $\Vert c(x)\Vert \le \epsilon_p,\ \Vert\nabla f(x) + \nabla c(x)\lambda\Vert \le \epsilon_d\Vert(\lambda, 1)\Vert$<br>or $\Vert\nabla c(x)c(x)\Vert \le \epsilon_d\Vert c(x)\Vert$ | $\mathcal{O}(\epsilon_d^{-3/2}\epsilon_p^{-1/2})$<br>(Hessian) | c(x) = 0 | [16] |

**Table 1** continued

| Point type | Complexity | Constraint type | Lit. |
|---|---|---|---|
| $\|\nabla f(x) + \nabla c(x)\lambda\| \leq \epsilon$, $\|c(x)\| \leq \epsilon$, or<br>$\|\nabla c(x)\mu\| \leq \epsilon$, $\|c(x)\| \geq \kappa\epsilon$. | $\mathcal{O}(\epsilon^{-2})$ (gradient) | $c(x) = 0$ | [17] |
| x is $\epsilon$approximate first order critical point<br>of the constrained problem or of$\|c(x)\|$ | $\mathcal{O}(\epsilon^{-(p+2)/p})$<br>($p$th derivative) | $c(x) = 0, x \in \mathcal{F}$<br>$\mathcal{F}$ is closed<br>and convex | [18] |
| x is $\epsilon$approximate$q$th order critical point<br>of the constrained problem or of $\|c(x)\|$<br>$q = 1, 2, 3.$ | $\mathcal{O}(\epsilon^{-2q-1})$<br>($q$th derivative) | $c(x) = 0, x \in \mathcal{F}$<br>$\mathcal{F}$ is closed<br>and convex | [18] |

Here $X = \text{diag}(x)$ and $\bar{X} = \text{diag}(\min\{x, \mathbf{1}\})$. $\tilde{\mathcal{O}}$ represents $\mathcal{O}$ with logarithm factors hidden. Gradient or Hessian in parenthesis means that the algorithm uses only gradient or both gradient and Hessian information, respectively. $p$th derivative means that the algorithm needs to evaluate function derivatives up to $p$th order

Most proofs appear in the main body of the paper; some elementary results are proved in the "Appendix".

## 2 Preliminaries

*Notation.* We use $\|\cdot\|$ to denote the Euclidean norm of a vector and $\|\cdot\|_2$ to denote the operator 2-norm of a matrix. For a given matrix $H$, we denote by $\sigma_{\min}(H)$ its minimal singular value and by $\lambda_{\min}(H)$ its minimal eigenvalue. We denote steps in $x$ and $\lambda$ as follows:

$$\Delta x_{k+1} \triangleq x_{k+1} - x_k, \quad \Delta\lambda_{k+1} \triangleq \lambda_{k+1} - \lambda_k. \tag{3}$$

In estimating complexities, we use order notation $\mathcal{O}(\cdot)$ in the usual sense, and $\tilde{\mathcal{O}}$ to hide factors that are logarithmic in the arguments. We use $\beta(\alpha) = \Omega(\gamma(\alpha))$ (where $\beta(\alpha)$ and $\gamma(\alpha)$ are both positive) to indicate that $\beta(\alpha)/\gamma(\alpha)$ is bounded below by a positive real number for all $\alpha$ sufficiently small.
*Assumptions.*

The following assumptions are used throughout this work.

**Assumption 1** Suppose that there exists $\rho_0 \geq 0$ such that $f(x) + \frac{\rho_0}{2}\|c(x)\|^2$ has compact level sets, that is, for all $\alpha \in \mathbb{R}$, the set

$$S_\alpha^0 \triangleq \left\{x \middle| f(x) + \frac{\rho_0}{2}\|c(x)\|^2 \leq \alpha\right\} \tag{4}$$

is empty or compact.

Assumption 1 holds in any of the following cases:

1. $f(x) + \frac{\rho_0}{2}\|c(x)\|^2$ is coercive for some $\rho_0 \geq 0$.
2. $f(x)$ is strongly convex.
3. $f(x)$ is bounded below and $c(x) = x^T x - 1$, as occurs in orthonormal dictionary learning applications.
4. $f(x) \triangleq \frac{1}{2}x^T Qx - p^T x$, $c(x) \triangleq Ax - b$, $Q$ is positive definite on $\text{null}(A) \triangleq \{x \mid Ax = 0\}$.

An immediate consequence of this assumption is the following, proof of which appears in the "Appendix".

**Lemma 1** *Suppose that Assumption 1 holds, then $f(x) + \frac{\rho_0}{2}\|c(x)\|^2$ is lower bounded.*

Therefore, Assumption 1 implies

$$\bar{L} \triangleq \inf_{x\in\mathbb{R}^n} \left\{f(x) + \frac{\rho_0}{2}\|c(x)\|^2\right\} > -\infty. \tag{5}$$

We use this definition of $\bar{L}$ throughout this paper whenever Assumption 1 holds.

The second assumption concerns certain smoothness and nondegeneracy assumptions on $f$ and $c$ over a compact set.

**Assumption 2** Given a compact set $\mathcal{S} \subseteq \mathbb{R}^n$, there exist positive constants $M_f$, $M_c$, $\sigma$, $L_c$ such that the following conditions on functions $f$ and $c$ hold.

(i) $\|\nabla f(x)\| \leq M_f$, $\|\nabla f(x) - \nabla f(y)\| \leq L_f\|x - y\|$, for all $x, y \in \mathcal{S}$.
(ii) $\|\nabla c(x)\|_2 \leq M_c$, $\sigma_{\min}(\nabla c(x)) \geq \sigma > 0$ for all $x \in \mathcal{S}$.
(iii) $\|\nabla c(x) - \nabla c(y)\|_2 \leq L_c\|x - y\|$, for all $x, y \in \mathcal{S}$.

This assumption may allow a general class of problems; in particular, (i) holds if $f(x)$ is smooth and $\nabla f(x)$ is locally Lipschitz continuous on a neighborhood of $\mathcal{S}$. (ii) holds when $c(x)$ is smooth on a neighborhood of $\mathcal{S}$ and satisfies an LICQ condition over $\mathcal{S}$, and (iii) holds if $\nabla c(x)$ is locally Lipschitz continuous on $\mathcal{S}$.

**Assumption 3** Suppose that $f(x) \leq \bar{U}$ for any $x \in \{x \mid \|c(x)\| \leq 1\}$.

A sufficient condition for Assumption 3 to hold is the compactness of $\{x \mid \|c(x)\| \leq 1\}$. This assumption is not needed if $c(x_0) = 0$, that is, the initial point is feasible.

## 3 Outer Iteration Complexity of Proximal AL

In this section, we derive the outer iteration complexity of Proximal AL (Algorithm 2) when the subproblem is solved inexactly. We assume that $x_{k+1}$ in Step 1 of Algorithm 2 satisfies the following approximate first-order optimality condition:

$$\nabla_x \mathcal{L}_\rho(x_{k+1}, \lambda_k) + \beta(x_{k+1} - x_k) = \tilde{r}_{k+1}, \quad \text{for all } k \geq 0, \tag{6}$$

where $\tilde{r}_{k+1}$ is some error vector. We additionally assume that

$$\mathcal{L}_\rho(x_{k+1}, \lambda_k) + \frac{\beta}{2}\|x_{k+1} - x_k\|^2 \leq \mathcal{L}_\rho(x_k, \lambda_k), \quad \text{for all } k \geq 0. \tag{7}$$

This condition can be achieved if we choose $x_k$ as the initial point of the subproblem in Step 1 of Algorithm 2, with subsequent iterates decreasing the objective of this subproblem. To analyze convergence, we use a Lyapunov function defined as follows for any $k \geq 1$, inspired by [26]:

$$P_k \triangleq \mathcal{L}_\rho(x_k, \lambda_k) + \frac{\beta}{4}\|x_k - x_{k-1}\|^2. \tag{8}$$

For any $k \geq 1$, we have that

$$\begin{aligned}
P_{k+1} - P_k &= \mathcal{L}_\rho(x_{k+1}, \lambda_{k+1}) - \mathcal{L}_\rho(x_k, \lambda_k) + \frac{\beta}{4}\|\Delta x_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_k\|^2 \\
&= \mathcal{L}_\rho(x_{k+1}, \lambda_{k+1}) - \mathcal{L}_\rho(x_{k+1}, \lambda_k) + \mathcal{L}_\rho(x_{k+1}, \lambda_k) - \mathcal{L}_\rho(x_k, \lambda_k) \\
&\quad + \frac{\beta}{4}\|\Delta x_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_k\|^2 \\
&= (\lambda_{k+1} - \lambda_k)^T c(x_{k+1}) + \mathcal{L}_\rho(x_{k+1}, \lambda_k) - \mathcal{L}_\rho(x_k, \lambda_k) \\
&\quad + \frac{\beta}{4}\|\Delta x_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_k\|^2 \\
&= \frac{1}{\rho}\|\Delta\lambda_{k+1}\|^2 + \mathcal{L}_\rho(x_{k+1}, \lambda_k) - \mathcal{L}_\rho(x_k, \lambda_k) + \frac{\beta}{4}\|\Delta x_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_k\|^2 \\
&\overset{(7)}{\leq} \frac{1}{\rho}\|\Delta\lambda_{k+1}\|^2 - \frac{\beta}{2}\|\Delta x_{k+1}\|^2 + \frac{\beta}{4}\|\Delta x_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_k\|^2 \\
&= \frac{1}{\rho}\|\Delta\lambda_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_{k+1}\|^2 - \frac{\beta}{4}\|\Delta x_k\|^2,
\end{aligned} \tag{9}$$

where the fourth equality holds because of Step 2 in Algorithm 2. We start with a technical result on bounding $\|\Delta\lambda_{k+1}\|^2 = \|\lambda_{k+1} - \lambda_k\|^2$.

**Lemma 2** (Bound for $\|\lambda_{k+1} - \lambda_k\|^2$) *Consider Algorithm 2 with (6) and (7). Suppose that for a fixed $k \geq 1$, Assumption 2 holds for some set $\mathcal{S}$ and that $x_k, x_{k+1} \in \mathcal{S}$. Then,*

$$\|\lambda_{k+1} - \lambda_k\|^2 \leq C_1 \|\Delta x_{k+1}\|^2 + C_2 \|\Delta x_k\|^2 + \frac{16 M_c^2}{\sigma^4} \|\tilde{r}_k\|^2 + \frac{4}{\sigma^2} \|\tilde{r}_{k+1} - \tilde{r}_k\|^2, \tag{10}$$

*where $C_1$ and $C_2$ are defined by*

$$C_1 \triangleq \frac{4}{\sigma^2} \left( L_f + \frac{L_c M_f}{\sigma} + \beta \right)^2, \quad C_2 \triangleq \frac{4}{\sigma^2} \left( \beta + \frac{2 M_c \beta}{\sigma} \right)^2. \tag{11}$$

***Proof*** The first-order optimality condition (6) for Step 1 implies that for all $t \geq 0$, we have

$$\begin{aligned} &\nabla f(x_{t+1}) + \nabla c(x_{t+1}) \lambda_t + \rho \nabla c(x_{t+1}) c(x_{t+1}) + \beta (x_{t+1} - x_t) = \tilde{r}_{t+1}. \\ &\implies \nabla f(x_{t+1}) + \nabla c(x_{t+1}) \lambda_{t+1} + \beta (x_{t+1} - x_t) = \tilde{r}_{t+1}. \end{aligned} \tag{12}$$

Likewise, by replacing $t$ with $t-1$, for $t \geq 1$, we obtain

$$\nabla f(x_t) + \nabla c(x_t) \lambda_t + \beta (x_t - x_{t-1}) = \tilde{r}_t. \tag{13}$$

By combining (12) and (13) and using the notation (3) along with $\Delta \tilde{r}_{t+1} \triangleq \tilde{r}_{t+1} - \tilde{r}_t$, we have for any $t \geq 1$ that

$$\begin{aligned} &\nabla f(x_{t+1}) - \nabla f(x_t) + \nabla c(x_{t+1}) \Delta \lambda_{t+1} \\ &\qquad + (\nabla c(x_{t+1}) - \nabla c(x_t)) \lambda_t + \beta (\Delta x_{t+1} - \Delta x_t) = \Delta \tilde{r}_{t+1}, \end{aligned}$$

which by rearrangement gives

$$\begin{aligned} -\nabla c(x_{t+1}) \Delta \lambda_{t+1} = {} & \nabla f(x_{t+1}) - \nabla f(x_t) + (\nabla c(x_{t+1}) - \nabla c(x_t)) \lambda_t \\ & + \beta (\Delta x_{t+1} - \Delta x_t) - \Delta \tilde{r}_{t+1}. \end{aligned}$$

For the given $k \geq 1$, since $\sigma$ is a lower bound on the smallest singular value of $\nabla c(x_{k+1})$ by Assumption 2, we have that

$$\begin{aligned} \|\Delta \lambda_{k+1}\| \leq {} & \frac{1}{\sigma} \big( \|\nabla f(x_{k+1}) - \nabla f(x_k)\| + \|\nabla c(x_{k+1}) - \nabla c(x_k)\| \|\lambda_k\| \\ & + \beta (\|\Delta x_{k+1}\| + \|\Delta x_k\|) + \|\Delta \tilde{r}_{k+1}\| \big). \end{aligned} \tag{14}$$

We have from (13) that

$$\nabla c(x_k) \lambda_k = -\nabla f(x_k) - \beta (x_k - x_{k-1}) + \tilde{r}_k,$$

so that

$$\|\lambda_k\| \leq \frac{1}{\sigma} \left( \|\nabla f(x_k)\| + \beta \|\Delta x_k\| + \|\tilde{r}_k\| \right) \leq \frac{1}{\sigma} \left( M_f + \beta \|\Delta x_k\| + \|\tilde{r}_k\| \right). \tag{15}$$

We also have from Assumption 2 that

$$\|\nabla c(x_{k+1}) - \nabla c(x_k)\| \leq L_c \|x_{k+1} - x_k\|, \quad \|\nabla c(x_{k+1}) - \nabla c(x_k)\| \leq 2 M_c. \tag{16}$$

By substituting Assumption 2(i), (15), and (16) into (14), we obtain the following for the given $k \geq 1$.

$$\begin{aligned} \|\Delta \lambda_{k+1}\| \leq \frac{1}{\sigma} \Bigg( & L_f \|\Delta x_{k+1}\| + \beta \|\Delta x_{k+1}\| + \beta \|\Delta x_k\| \\ & + \|\nabla c(x_{k+1}) - \nabla c(x_k)\|_2 \left( \frac{1}{\sigma} M_f + \frac{\beta}{\sigma} \|\Delta x_k\| + \frac{1}{\sigma} \|\tilde{r}_k\| \right) + \|\Delta \tilde{r}_{k+1}\| \Bigg) \end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{\sigma}\bigg( L_f\|\Delta x_{k+1}\| + \beta\|\Delta x_{k+1}\| + \beta\|\Delta x_k\| + \frac{L_c M_f}{\sigma}\|\Delta x_{k+1}\| + \frac{2M_c\beta}{\sigma}\|\Delta x_k\| \\
&\quad + \frac{2M_c}{\sigma}\|\tilde{r}_k\| + \|\Delta\tilde{r}_{k+1}\|\bigg) \\
&\le \frac{1}{\sigma}\left(L_f + \frac{L_c M_f}{\sigma} + \beta\right)\|\Delta x_{k+1}\| + \frac{1}{\sigma}\left(\beta + \frac{2M_c\beta}{\sigma}\right)\|\Delta x_k\| \\
&\quad + \frac{2M_c}{\sigma^2}\|\tilde{r}_k\| + \frac{1}{\sigma}\|\Delta\tilde{r}_{k+1}\|.
\end{aligned}
$$

By using the bound $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ for positive scalars $a, b, c, d$, and using the definition (11), we obtain the result. □

For the rest of this section, we use the following definitions for $c_1$ and $c_2$:

$$
c_1 \triangleq \frac{\beta}{4} - \frac{C_1}{\rho}, \quad c_2 \triangleq \frac{\beta}{4} - \frac{C_2}{\rho}, \tag{17}
$$

where $C_1$ and $C_2$ are defined in (11). Next we show that sequences $\{x_k\}$ and $\{\lambda_k\}$ are bounded and $\{P_k\}_{k\ge1}$ satisfies certain properties under Assumption 1–3, for suitable choices of the algorithmic parameters.

**Lemma 3** *Consider Algorithm 2 with conditions (6) and (7). Choose $\{\tilde{r}_k\}_{k\ge1}$ such that $\sum_{k=1}^{\infty}\|\tilde{r}_k\|^2 \le R < +\infty$ and $\|\tilde{r}_k\| \le 1$, for all $k \ge 1$. Let $\{P_k\}_{k\ge1}$ be defined as in (8). Suppose that Assumptions 1 and 3 hold and define*

$$
\hat{\alpha} \triangleq 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + 2, \tag{18}
$$

*where $C_0 > 0$ is any fixed constant. Suppose that Assumption 2 holds with $\mathcal{S} = S^0_{\hat{\alpha}}$. Choose $\rho$ and $\beta$ such that*

$$
\rho \ge \max\left\{\frac{(M_f + \beta D_S + 1)^2}{2\sigma^2} + \rho_0, \frac{16(M_c^2 + \sigma^2)R}{\sigma^4}, 3\rho_0, 1\right\}, \tag{19}
$$

*where*

$$
D_S \triangleq \max\{\|x - y\| \mid x, y \in S^0_{\hat{\alpha}}\}, \tag{20}
$$

*and that $c_1$ and $c_2$ defined in (17) are both positive. Suppose that $x_0$ satisfies $\|c(x_0)\|^2 \le \min\{C_0/\rho, 1\}$, where $C_0$ is the constant appearing in (18). Then*

$$
\{x_k\}_{k\ge0} \subseteq S^0_{\hat{\alpha}} \quad \text{and} \quad \|\lambda_k\| \le \frac{M_f + \beta D_S + 1}{\sigma}, \quad \text{for all } k \ge 1. \tag{21}
$$

*Furthermore, (10) and the following inequality hold for any $k \ge 1$,*

$$
P_{k+1} - P_k \le -c_1\|\Delta x_{k+1}\|^2 - c_2\|\Delta x_k\|^2 + \frac{16M_c^2}{\rho\sigma^4}\|\tilde{r}_k\|^2 + \frac{4}{\rho\sigma^2}\|\tilde{r}_{k+1} - \tilde{r}_k\|^2. \tag{22}
$$

***Proof*** Note that Assumption 3 implies that

$$
f(x_0) \le \bar{U}, \tag{23}
$$

since $\|c(x_0)\| \le 1$. Therefore,

$$
\mathcal{L}_\rho(x_0, \lambda_0) = f(x_0) + \lambda_0^T c(x_0) + \frac{\rho}{2}\|c(x_0)\|^2
$$

$$\begin{aligned}
&\le f(x_0) + \frac{\|\lambda_0\|^2}{2\rho} + \frac{\rho}{2}\|c(x_0)\|^2 + \frac{\rho}{2}\|c(x_0)\|^2 \\
&\overset{(23)}{\le} \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0.
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
\bar{U} + C_0 - \bar{L} &\overset{(24)}{\ge} f(x_0) + \lambda_0^T c(x_0) + \frac{\rho}{2}\|c(x_0)\|^2 - \frac{\|\lambda_0\|^2}{2\rho} - \bar{L} \\
&\overset{(\rho\ge 3\rho_0)}{\ge} f(x_0) + \frac{\rho_0}{2}\|c(x_0)\|^2 - \bar{L} + \lambda_0^T c(x_0) + \frac{\rho}{3}\|c(x_0)\|^2 - \frac{\|\lambda_0\|^2}{2\rho} \\
&\ge 0 + \frac{\rho}{3}\left\| c(x_0) + \frac{3\lambda_0}{2\rho}\right\|^2 - \frac{3}{4\rho}\|\lambda_0\|^2 - \frac{\|\lambda_0\|^2}{2\rho} \\
&\overset{(\rho\ge 1)}{\ge} -\frac{5}{4}\|\lambda_0\|^2
\end{aligned} \tag{25}$$

We prove the theorem by induction. We show that the following bounds hold for all $i \ge 1$:

$$x_i \in S^0_{\hat{\alpha}}, \tag{26a}$$

$$\|\lambda_i\|^2 \le \frac{(M_f + \beta D_S + 1)^2}{\sigma^2} \le 2(\rho - \rho_0), \tag{26b}$$

$$P_i \le 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + \frac{16M_c^2}{\rho\sigma^4}\sum_{t=1}^{i-1}\|\tilde{r}_t\|^2 + \frac{4}{\rho\sigma^2}\sum_{t=1}^{i-1}\|\tilde{r}_{t+1} - \tilde{r}_t\|^2. \tag{26c}$$

We verify first that (26) holds when $i = 1$. From (7) we have

$$\begin{aligned}
&f(x_1) + \lambda_0^T c(x_1) + \frac{\rho}{2}\|c(x_1)\|^2 + \frac{\beta}{2}\|x_1 - x_0\|^2 \\
&\quad\le f(x_0) + \lambda_0^T c(x_0) + \frac{\rho}{2}\|c(x_0)\|^2 \overset{(24)}{\le} \bar{U} + \frac{\|\lambda_0\|^2}{2\rho} + C_0,
\end{aligned} \tag{27}$$

so that for $i = 0$ and 1, we have

$$\begin{aligned}
f(x_i) + \frac{\rho}{6}\|c(x_i)\|^2 &\overset{(24),(27)}{\le} \bar{U} + \frac{\|\lambda_0\|^2}{2\rho} + C_0 - \lambda_0^T c(x_i) - \frac{\rho}{3}\|c(x_i)\|^2 \\
&= \bar{U} + \frac{\|\lambda_0\|^2}{2\rho} + C_0 - \frac{\rho}{3}\left\|c(x_i) + \frac{3\lambda_0}{2\rho}\right\|^2 + \frac{3\|\lambda_0\|^2}{4\rho} \\
\overset{(\rho\ge 3\rho_0)}{\Longrightarrow} f(x_i) + \frac{\rho_0}{2}\|c(x_i)\|^2 &\le \bar{U} + \frac{5\|\lambda_0\|^2}{4\rho} + C_0 \\
&\overset{((25),\rho\ge 1)}{\le} \bar{U} + C_0 + \frac{5}{4}\|\lambda_0\|^2 + 6\left(\bar{U} + C_0 - \bar{L} + \frac{5}{4}\|\lambda_0\|^2\right) \\
&\le 7\bar{U} + 7C_0 - 6\bar{L} + \frac{35}{4}\|\lambda_0\|^2 \overset{(18)}{<} \hat{\alpha}.
\end{aligned}$$

Thus, $x_0, x_1 \in S^0_{\hat{\alpha}}$, verifying that (26a) holds for $i = 1$.

Approximate first-order optimality (6) indicates that

$$\nabla f(x_1) + \nabla c(x_1)\lambda_1 + \beta(x_1 - x_0) = \tilde{r}_1.$$

Since $x_0, x_1 \in S^0_{\hat{\alpha}}$, we have by Assumption 2 and (20) that

$$\sigma\|\lambda_1\| \le \|\nabla c(x_1)\lambda_1\| = \|\nabla f(x_1) + \beta(x_1 - x_0) - \tilde{r}_1\| \le M_f + \beta D_S + 1.$$
$$\implies \|\lambda_1\|^2 \le \frac{(M_f + \beta D_S + 1)^2}{\sigma^2} \overset{(19)}{\le} 2(\rho - \rho_0).$$

Thus, (26b) holds for $i = 1$.

Next, we verify (26c) when $i = 1$. Note that

$$\begin{aligned}
P_1 &= \mathcal{L}_\rho(x_1, \lambda_1) + \frac{\beta}{4}\|x_1 - x_0\|^2 \\
&= \mathcal{L}_\rho(x_1, \lambda_1) - \mathcal{L}_\rho(x_1, \lambda_0) + \mathcal{L}_\rho(x_1, \lambda_0) - \mathcal{L}_\rho(x_0, \lambda_0) + \mathcal{L}_\rho(x_0, \lambda_0) \\
&\quad + \frac{\beta}{4}\|x_1 - x_0\|^2 \\
&\overset{(7)}{\le} \frac{1}{\rho}\|\lambda_1 - \lambda_0\|^2 - \frac{\beta}{2}\|x_1 - x_0\|^2 + \mathcal{L}_\rho(x_0, \lambda_0) + \frac{\beta}{4}\|x_1 - x_0\|^2 \\
&= \rho\|c(x_1)\|^2 - \frac{\beta}{4}\|x_1 - x_0\|^2 + \mathcal{L}_\rho(x_0, \lambda_0) \\
&\overset{(24)}{\le} \rho\|c(x_1)\|^2 + \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0, \\
&\overset{(\rho\ge 1)}{\le} \rho\|c(x_1)\|^2 + \bar{U} + \frac{1}{2}\|\lambda_0\|^2 + C_0,
\end{aligned} \tag{28}$$

In addition, (27) indicates that

$$\begin{aligned}
&\frac{\rho}{6}\|c(x_1)\|^2 \\
&\le \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0 - \lambda_0^T c(x_1) - \frac{\rho}{6}\|c(x_1)\|^2 - f(x_1) - \frac{\rho}{6}\|c(x_1)\|^2 \\
&= \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0 - \frac{\rho}{6}\|c(x_1) + 3\lambda_0/\rho\|^2 + \frac{3\|\lambda_0\|^2}{2\rho} - f(x_1) - \frac{\rho}{6}\|c(x_1)\|^2 \\
&\overset{(\rho\ge 3\rho_0)}{\le} \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0 + \frac{3\|\lambda_0\|^2}{2\rho} - f(x_1) - \frac{\rho_0}{2}\|c(x_1)\|^2 \\
&\le \bar{U} + \frac{2}{\rho}\|\lambda_0\|^2 + C_0 - \bar{L} \overset{(\rho\ge 1)}{\le} \bar{U} + 2\|\lambda_0\|^2 + C_0 - \bar{L}.
\end{aligned}$$

By substituting this bound into (28), we have that

$$P_1 \le \bar{U} + \frac{\|\lambda_0\|^2}{2} + C_0 + \rho\|c(x_1)\|^2 \le 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2, \tag{29}$$

so (26c) holds for $i = 1$ also.

We now take the inductive step, supposing that (26) holds when $i = k \ge 1$, and proving that these three conditions continue to hold for $i = k + 1$. By inequality (7), we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{k+1}, \lambda_k) \le \mathcal{L}_\rho(x_k, \lambda_k) \le P_k \\
&\implies f(x_{k+1}) + \frac{\rho}{2}\|c(x_{k+1})\|^2 + \lambda_k^T c(x_{k+1}) \le P_k \\
&\implies f(x_{k+1}) + \frac{\rho}{2}\|c(x_{k+1})\|^2 - \frac{\|\lambda_k\|^2}{2(\rho - \rho_0)} - \frac{(\rho - \rho_0)\|c(x_{k+1})\|^2}{2} \le P_k
\end{aligned}$$

$$
\begin{aligned}
&\Longrightarrow\ f(x_{k+1}) + \frac{\rho_0}{2}\|c(x_{k+1})\|^2 \le P_k + \frac{\|\lambda_k\|^2}{2(\rho-\rho_0)} \overset{(26b)}{\le} P_k + 1 \\
&\overset{(26c)}{\le} 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + \frac{16M_c^2}{\rho\sigma^4}\sum_{t=1}^{k-1}\|\tilde{r}_t\|^2 + \frac{4}{\rho\sigma^2}\sum_{t=1}^{k-1}\|\tilde{r}_{t+1}-\tilde{r}_t\|^2 + 1 \\
&\le 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + \frac{16M_c^2}{\rho\sigma^4}\sum_{t=1}^{k-1}\|\tilde{r}_t\|^2 + \frac{8}{\rho\sigma^2}\sum_{t=1}^{k-1}(\|\tilde{r}_{t+1}\|^2 + \|\tilde{r}_t\|^2) + 1 \\
&\le 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + \frac{16M_c^2}{\rho\sigma^4}\sum_{t=1}^{\infty}\|\tilde{r}_t\|^2 + \frac{16}{\rho\sigma^2}\sum_{t=1}^{\infty}\|\tilde{r}_t\|^2 + 1 \\
&\le 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + \frac{16(M_c^2+\sigma^2)R}{\rho\sigma^4} + 1 \\
&\overset{(19)}{\le} 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + 2 = \hat{\alpha},
\end{aligned}
$$

where the inequality on the third line holds because of $-\frac{r}{2}\|a\|^2 - \frac{1}{2r}\|b\|^2 \le a^T b$, for any $r > 0$, $a, b \in \mathbb{R}^m$. Therefore, $x_{k+1} \in S^0_{\hat{\alpha}}$, so we have proved (26a).

By approximate first-order optimality (6) and the hypothesis $x_k \in S^0_{\hat{\alpha}}$, the argument to establish that $\|\lambda_{k+1}\|^2 \le \frac{(M_f+\beta D_S+1)^2}{\sigma^2} \le 2(\rho-\rho_0)$ is the same as for the case of $i = 1$, so (26b) holds for $i = k+1$.

Since $x_k$ and $x_{k+1}$ both belong to $S^0_{\hat{\alpha}}$, Lemma 2 indicates that (10) holds. By combining (10) with (9), we obtain (22). Therefore,

$$
\begin{aligned}
P_{k+1} &\overset{(22)}{\le} P_k + \frac{16M_c^2}{\rho\sigma^4}\|\tilde{r}_k\|^2 + \frac{4}{\rho\sigma^2}\|\tilde{r}_{k+1}-\tilde{r}_k\|^2 \\
&\overset{(26)}{\le} 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + \frac{16M_c^2}{\rho\sigma^4}\sum_{t=1}^{k}\|\tilde{r}_t\|^2 + \frac{4}{\rho\sigma^2}\sum_{t=1}^{k}\|\tilde{r}_{t+1}-\tilde{r}_t\|^2.
\end{aligned}
$$

Thus we have established (26c) for $i = k+1$. Note that (10) and (22) hold for all $k \ge 1$, so we have completed the proof. □

*First-order complexity.* With the properties of $\{P_k\}_{k\ge1}$ established to this point, we can analyze the complexity of obtaining an $\epsilon$-1o solution. For any given $\epsilon > 0$, we define two quantities which will be referred to repeatedly in subsequent sections:

$$T_\epsilon \triangleq \inf\{t \ge 1 \mid \|\nabla_x \mathcal{L}_0(x_t,\lambda_t)\| \le \epsilon,\ \|c(x_t)\| \le \epsilon\}. \tag{30a}$$

$$\hat{T}_\epsilon \triangleq \inf\{t \ge 1 \mid x_t \text{ is an } \epsilon - 1\text{o solution of (1)}\}. \tag{30b}$$

Note that $\hat{T}_\epsilon$ is independent of the Proximal AL method. Meanwhile, by the definition of $\mathcal{L}_0(x,\lambda)$, we know that $x_{T_\epsilon}$ is an $\epsilon$-1o solution and $\lambda_{T_\epsilon}$ is the associated multiplier, indicating that $\hat{T}_\epsilon \le T_\epsilon$. The definition of $T_\epsilon$ also suggests the following stopping criterion for Algorithm 2:

$$\text{If } \|\nabla_x \mathcal{L}_0(x_t,\lambda_t)\| \le \epsilon \text{ and } \|c(x_t)\| \le \epsilon \text{ then STOP.} \tag{31}$$

Under this criterion, Algorithm 2 will stop at iteration $T_\epsilon - 1$ and output $x_{T_\epsilon}$ as an $\epsilon$-1o solution.

Part (i) of the following result shows subsequential convergence of the generated sequence to the first-order optimal point. Part (ii) describes the speed of such convergence by obtaining an estimate of $T_\epsilon$ in terms of $\epsilon$. In this result, we make a specific choice $\beta = \epsilon^\eta/2$ for the proximality parameter. We could choose $\beta$ to be any fixed multiple of this value (the multiple not depending on $\epsilon$) and obtain a similar result with only trivial changes to the analysis.

**Theorem 2** (First-order complexity) *Consider Algorithm 2 with conditions (6) and (7), and let $\{P_k\}_{k\ge 1}$ be defined as in (8). Suppose that Assumptions 1, 3 and 2 hold with $\mathcal{S} = S^0_{\hat\alpha}$ (with $\hat\alpha$ defined in (18)), and that $\epsilon \in (0, 1]$ and $\eta \in [0, 2]$ are given. Suppose that the residual sequence $\{\tilde r_k\}_{k\ge 1}$ is chosen such that $\sum_{k=1}^{\infty} \|\tilde r_k\|^2 \le R \in [1,\infty)$ and $\|\tilde r_k\| \le \epsilon/2$ for all $k \ge 1$.*

*Define $\beta = \epsilon^\eta/2$ and*

$$\rho \ge \max\left\{\frac{16\max\{C_1, C_2\}}{\epsilon^\eta}, \frac{(M_f + \beta D_S + 1)^2}{2\sigma^2} + \rho_0, \frac{16(M_c^2 + \sigma^2)R}{\sigma^4}, 3\rho_0, 1\right\}, \quad (32)$$

*where $C_1$ and $C_2$ are defined as in (11), and $D_S$ is the diameter of $S^0_{\hat\alpha}$, as defined in (20). Suppose that $x_0$ satisfies $\|c(x_0)\|^2 \le \min\{C_0/\rho, 1\}$, where $C_0$ is the constant appearing in (18). Then we have the following.*

(i) *A subsequence of $\{(x_k, \lambda_k)\}_{k\ge 1}$ generated by Algorithm 2 converges to a point $(x^*, \lambda^*)$ satisfying first-order optimality conditions for (1), namely,*

$$\nabla f(x^*) + \nabla c(x^*)\lambda^* = 0, \quad c(x^*) = 0.$$

(ii) *For $T_\epsilon$ and $\hat T_\epsilon$ defined in (30), we have $\hat T_\epsilon \le T_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta})$. In particular, if $\eta = 2$, then $\hat T_\epsilon = \mathcal{O}(1)$.*

***Proof*** We first prove (i). Checking the positivity of $c_1$ and $c_2$, given the parameter assignments, we have

$$c_1 = \frac{\beta}{4} - \frac{C_1}{\rho} \overset{(32)}{\ge} \frac{\epsilon^\eta}{8} - \frac{\epsilon^\eta}{16} > 0, \quad c_2 = \frac{\beta}{4} - \frac{C_2}{\rho} \overset{(32)}{\ge} \frac{\epsilon^\eta}{16} > 0. \quad (33)$$

$$\begin{aligned} P_k &\ge f(x_k) + \frac{\rho}{2}\|c(x_k)\|^2 + \lambda_k^T c(x_k) \\ &\ge f(x_k) + \frac{\rho}{2}\|c(x_k)\|^2 - \frac{\|\lambda_k\|^2}{2(\rho-\rho_0)} - \frac{(\rho-\rho_0)\|c(x_k)\|^2}{2} \\ &= f(x_k) + \frac{\rho_0}{2}\|c(x_k)\|^2 - \frac{\|\lambda_k\|^2}{2(\rho-\rho_0)} \\ &\overset{\text{(Lemma 3)}}{\ge} f(x_k) + \frac{\rho_0}{2}\|c(x_k)\|^2 - \frac{(M_f + \beta D_S + 1)^2}{2\sigma^2(\rho-\rho_0)} \\ &\overset{(5),(32)}{\ge} \bar L - 1. \end{aligned} \quad (34)$$

Therefore, using (22) from Lemma 3, we have the following for any $k \ge 1$:

$$\begin{aligned} &\sum_{i=1}^{k}\left[c_1\|\Delta x_{i+1}\|^2 + c_2\|\Delta x_i\|^2\right] \\ &\le P_1 - P_{k+1} + \frac{16M_c^2}{\rho\sigma^4}\sum_{i=1}^{k}\|\tilde r_i\|^2 + \frac{4}{\rho\sigma^2}\sum_{i=1}^{k}\|\tilde r_{i+1} - \tilde r_i\|^2 \\ &\le P_1 - P_{k+1} + \frac{16M_c^2}{\rho\sigma^4}\sum_{i=1}^{k}\|\tilde r_i\|^2 + \frac{8}{\rho\sigma^2}\sum_{i=1}^{k}(\|\tilde r_{i+1}\|^2 + \|\tilde r_i\|^2) \end{aligned}$$

$$
\begin{aligned}
&\le P_1 - P_{k+1} + \frac{16(M_c^2+\sigma^2)}{\rho\sigma^4}\sum_{i=1}^{\infty}\|\tilde{r}_i\|^2 \\
&\le P_1 - P_{k+1} + \frac{16(M_c^2+\sigma^2)R}{\rho\sigma^4} && (35)\\
&\overset{(34)}{\le} P_1 - \left(\bar{L}-1\right) + \frac{16(M_c^2+\sigma^2)R}{\rho\sigma^4} \\
&\overset{(32)}{\le} P_1 - \bar{L} + 2 \\
&\overset{(29)}{\le} 7\bar{U} + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 - \bar{L} + 2 = \hat{\alpha} - \bar{L}. && (36)
\end{aligned}
$$

Because of (21) in Lemma 3 and compactness of $S_{\hat{\alpha}}^0$, the sequence $\{(x_k,\lambda_k)\}_{k\ge 1}$ is bounded, so there exists a convergent subsequence $\{(x_k,\lambda_k)\}_{k\in\mathcal{K}}$ with limit $(x^*,\lambda^*)$. Since (36) holds for any $k\ge 1$ and $c_1>0$, $c_2>0$, we have that $\lim\limits_{k\to\infty}\|\Delta x_k\|=0$. Moreover, finiteness of $\sum_{k=1}^{\infty}\|\tilde{r}_k\|^2$ implies that $\lim\limits_{k\to\infty}\|\tilde{r}_k\|=0$. Therefore, we have

$$
\begin{aligned}
\nabla f(x^*) + \nabla c(x^*)\lambda^* &= \lim_{k\in\mathcal{K}}(\nabla f(x_k) + \nabla c(x_k)\lambda_k) \\
&= \lim_{k\in\mathcal{K}}(\nabla f(x_k) + \nabla c(x_k)(\lambda_{k-1} + \rho c(x_k))) = \lim_{k\in\mathcal{K}} \nabla_x \mathcal{L}_\rho(x_k,\lambda_{k-1}) \\
&\overset{(6)}{=} \lim_{k\in\mathcal{K}}(-\beta\Delta x_k + \tilde{r}_k) = 0.
\end{aligned}
$$

Since (10) holds for any $k\ge 1$ by Lemma 3, we have

$$
\begin{aligned}
\|c(x^*)\|^2 &= \lim_{k\in\mathcal{K}}\|c(x_k)\|^2 = \lim_{k\in\mathcal{K}}\|\lambda_k-\lambda_{k-1}\|^2/\rho^2 \\
&\overset{(10)}{\le} \lim_{k\in\mathcal{K}} \frac{C_1}{\rho^2}\|\Delta x_k\|^2 + \frac{C_2}{\rho^2}\|\Delta x_{k-1}\|^2 + \frac{16M_c^2}{\rho^2\sigma^4}\|\tilde{r}_{k-1}\|^2 + \frac{4}{\rho^2\sigma^2}\|\tilde{r}_k-\tilde{r}_{k-1}\|^2 = 0,
\end{aligned}
$$

completing the proof of (i).

We now prove (ii). Define

$$
C_1^o \triangleq \frac{4}{\sigma^2}\left(L_f + \frac{L_c M_f}{\sigma}\right)^2 \le C_1, \tag{37a}
$$

$$
\hat{\Delta} \triangleq (\hat{\alpha}-\bar{L})\max\left\{16,\, 1/(8C_1^o)\right\}. \tag{37b}
$$

We want to show that $T_\epsilon \le \lceil \hat{\Delta}/\epsilon^{2-\eta}\rceil + 1$. Let $K \triangleq \lceil \hat{\Delta}/\epsilon^{2-\eta}\rceil$, and note that since (36) holds for $k=K$, we have that there exists some $k^*\in\{1,2,\ldots,K\}$ such that

$$
c_1\|x_{k^*+1}-x_{k^*}\|^2 + c_2\|x_{k^*}-x_{k^*-1}\|^2 \le (\hat{\alpha}-\bar{L})/K. \tag{38}
$$

Thus, we have

$$
\begin{aligned}
&\|\nabla\mathcal{L}_0(x_{k^*+1},\lambda_{k^*+1})\| = \|\nabla\mathcal{L}_\rho(x_{k^*+1},\lambda_{k^*})\| \overset{(6.)}{\le} \beta\|x_{k^*+1}-x_{k^*}\| + \|\tilde{r}_{k^*+1}\| \\
&\overset{(38)}{\le} \beta\sqrt{\frac{(\hat{\alpha}-\bar{L})/c_1}{K}} + \frac{\epsilon}{2} \overset{(33)}{\le} \frac{\epsilon^\eta}{2}\sqrt{\frac{(\hat{\alpha}-\bar{L})/(\epsilon^\eta/16)}{K}} + \frac{\epsilon}{2} \\
&\le \frac{\epsilon^\eta}{2}\sqrt{\frac{16(\hat{\alpha}-\bar{L})/(\epsilon^\eta)}{\hat{\Delta}\epsilon^{\eta-2}}} + \frac{\epsilon}{2} \overset{(37b)}{\le} \frac{\epsilon^\eta}{2}\sqrt{\frac{16(\hat{\alpha}-\bar{L})}{16(\hat{\alpha}-\bar{L})\epsilon^{2\eta-2}}} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}
$$

For the constraint norm, we have

$$\begin{aligned}
\|c(x_{k^*+1})\|^2 &= \|\Delta\lambda_{k^*+1}\|^2/\rho^2 \\
&\overset{(10)}{\le} \frac{C_1}{\rho^2}\|\Delta x_{k^*+1}\|^2 + \frac{C_2}{\rho^2}\|\Delta x_{k^*}\|^2 + \frac{16M_c^2}{\rho^2\sigma^4}\|\tilde{r}_{k^*}\|^2 + \frac{4}{\rho^2\sigma^2}\|\tilde{r}_{k^*+1} - \tilde{r}_{k^*}\|^2 \\
&\le \frac{C_1}{\rho^2}\|\Delta x_{k^*+1}\|^2 + \frac{C_2}{\rho^2}\|\Delta x_{k^*}\|^2 + \frac{16M_c^2}{\rho^2\sigma^4}\|\tilde{r}_{k^*}\|^2 + \frac{8}{\rho^2\sigma^2}(\|\tilde{r}_{k^*}\|^2 + \|\tilde{r}_{k^*+1}\|^2) \\
&\le \frac{C_1}{\rho^2}\|\Delta x_{k^*+1}\|^2 + \frac{C_2}{\rho^2}\|\Delta x_{k^*}\|^2 + \frac{16(M_c^2+\sigma^2)}{\rho^2\sigma^4}\cdot\frac{\epsilon^2}{4} \\
&\le \frac{1}{\rho^2}\max\left\{\frac{C_1}{c_1}, \frac{C_2}{c_2}\right\}(c_1\|\Delta x_{k^*+1}\|^2 + c_2\|\Delta x_{k^*}\|^2) + \frac{4(M_c^2+\sigma^2)\epsilon^2}{\rho^2\sigma^4} \\
&\overset{(32),(38)}{\le} \frac{\max\{C_1, C_2\}/(\epsilon^\eta/16)}{(16\max\{C_1, C_2\}/\epsilon^\eta))^2}\cdot\frac{\hat{\alpha} - \bar{L}}{K} + \frac{4(M_c^2+\sigma^2)\epsilon^2}{\rho^2\sigma^4} \\
&\le \frac{(\hat{\alpha} - \bar{L})\epsilon^\eta}{16\max\{C_1, C_2\}K} + \frac{4(M_c^2+\sigma^2)}{\rho^2\sigma^4}\cdot\epsilon^2 \\
&\overset{(37a)}{\le} \frac{(\hat{\alpha} - \bar{L})\epsilon^\eta}{16C_1^oK} + \frac{4(M_c^2+\sigma^2)}{\rho^2\sigma^4}\cdot\epsilon^2 \\
&\overset{(32)}{\le} \frac{(\hat{\alpha} - \bar{L})\epsilon^\eta}{16C_1^o\hat{\Delta}\epsilon^{\eta-2}} + \frac{\epsilon^2}{4\rho R} \\
&\overset{(\rho\ge1, R\ge1)}{\le} \frac{(\hat{\alpha} - \bar{L})\epsilon^\eta}{16C_1^o\hat{\Delta}\epsilon^{\eta-2}} + \frac{\epsilon^2}{4} \\
&\overset{(37b)}{\le} \frac{\epsilon^2}{2} + \frac{\epsilon^2}{4} < \epsilon^2.
\end{aligned}$$

Therefore, we have

$$T_\epsilon \le k^* + 1 \le K + 1 = \lceil\hat{\Delta}/\epsilon^{2-\eta}\rceil + 1. \tag{39}$$

It follows that $\hat{T}_\epsilon \le T_\epsilon \le \lceil\hat{\Delta}/\epsilon^{2-\eta}\rceil + 1$, completing the proof. □

***Remark 1*** (i) The condition $\|c(x_0)\|^2 \le \min\{C_0/\rho, 1\}$ is obviously satisfied by a feasible point, for which $c(x_0) = 0$. In this case, we do not need Assumption 3 and can prove a result with $\hat{\alpha} = 7f(x_0) + 7C_0 - 6\bar{L} + 13\|\lambda_0\|^2 + 2$. An initial phase can be applied, if necessary, to find a point with small $\|c(x_0)\|$; we discuss this point in a later section.

(ii) When $\eta = 0$, the complexity result is consistent with that of [26]. However, our parameter choices $\beta = \epsilon^\eta$ for $\eta > 0$ allows us to choose $\beta$ to be small because, unlike [26], we are not concerned with maintaining strong convexity of the subproblem in Step 1 of Algorithm 2. Another benefit of small $\beta$ is that it allows complexity results to be proved for $\epsilon$-2o points, which follows from part (ii) of Theorem 2, as we see next.

*Second-order complexity.* We further assume that in Step 1 of Algorithm 2, $x_{k+1}$ satisfies the following approximate second-order optimality conditions:

$$\nabla_{xx}^2\mathcal{L}_\rho(x_{k+1}, \lambda_k) + \beta I \succeq -\epsilon_{k+1}^H I, \quad \text{for all } k \ge 0, \tag{40}$$

where $\{\epsilon_{k+1}^H\}_{k\ge0}$ is a chosen error sequence.

In corresponding fashion to the definition of $\hat{T}_\epsilon$ in (30b), we define $\widetilde{T}_\epsilon$ as follows:

$$\widetilde{T}_\epsilon \triangleq \inf\{t \geq 1 \mid x_t \text{ is an } \epsilon - \text{ 2o solution of}(1)\}. \tag{41}$$

We have the following result for complexity of obtaining an $\epsilon$-2o stationary point of (1) through Algorithm 2.

**Corollary 1** (Second-order complexity) *Suppose that all assumptions and settings in Theorem 2 hold. Assume that, in addition, Step 1 of Algorithm 2 satisfies (40), with $\epsilon_k^H \equiv \epsilon/2$ for all $k \geq 1$. Let $\eta \in [1, 2]$. Then for $\widetilde{T}_\epsilon$ defined in (41), we have $\widetilde{T}_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta})$.*

***Proof*** Since $\beta = \epsilon^\eta/2 \leq \epsilon/2$ and $\epsilon_{k+1}^H \equiv \epsilon/2$, for any $k \geq 0$, we have from (40) that

$$\nabla_{xx}^2 \mathcal{L}_\rho(x_{k+1}, \lambda_k) \succeq -(\beta + \epsilon_{k+1}^H)I \succeq -\epsilon I.$$

This fact indicates that

$$\nabla^2 f(x_{k+1}) + \sum_{i=1}^m [\lambda_{k+1}]_i \nabla^2 c_i(x_{k+1}) + \rho \nabla c(x_{k+1})[\nabla c(x_{k+1})]^T \succeq -\epsilon I,$$

which implies that

$$d^T(\nabla^2 f(x_{k+1}) + \sum_{i=1}^m [\lambda_{k+1}]_i \nabla^2 c_i(x_{k+1}))d \geq -\epsilon\|d\|^2,$$

for any $d \in S(x_{k+1}) \triangleq \{d \in \mathbb{R}^n \mid [\nabla c(x_{k+1})]^T d = 0\}$. This is exactly condition (2b) of Definition 2. Therefore, we have

$$\begin{aligned}
\widetilde{T}_\epsilon &= \inf\{t \geq 1 \mid \exists \lambda \in \mathbb{R}^m, \|\nabla f(x_t) + \nabla c(x_t)\lambda\| \leq \epsilon, \|c(x_t)\| \leq \epsilon, \\
&\qquad d^T(\nabla^2 f(x_t) + \sum_{i=1}^m \lambda_i \nabla^2 c_i(x_t))d \geq -\epsilon\|d\|^2, \quad \text{for all } d \in S(x_t)\} \\
&\leq \inf\{t \geq 1 \mid \|\nabla f(x_t) + \nabla c(x_t)\lambda_t\| \leq \epsilon, \|c(x_t)\| \leq \epsilon, \\
&\qquad d^T(\nabla^2 f(x_t) + \sum_{i=1}^m [\lambda_t]_i \nabla^2 c_i(x_t))d \geq -\epsilon\|d\|^2, \quad \text{for all } d \in S(x_t)\} \\
&= \inf\{t \geq 1 \mid \|\nabla f(x_t) + \nabla c(x_t)\lambda_t\| \leq \epsilon, \|c(x_t)\| \leq \epsilon\} = T_\epsilon.
\end{aligned}$$

The result now follows from Theorem 2. □

## 4 Total Iteration/Operation Complexity of Proximal AL

In this section, we will choose an appropriate method to solve the subproblem and estimate the total iteration and operation complexity of our Proximal AL approach to find an $\epsilon$-1o or $\epsilon$-2o solution. To solve the subproblem at each major iteration of Algorithm 2, we can use methods for unconstrained smooth nonconvex optimization that allow the decrease condition (7) to hold, and approximate optimality conditions (6) or (40) to be enforced in a natural way, finding points that satisfy such conditions within a certain number of iterations that depends on the tolerances [10,15,20,21,33]. Among these, the Newton-CG method described in [33] has good complexity guarantees as well as good practical performance.

To review the properties of the algorithm in [33], we consider the following unconstrained problem:

$$\min_{z \in \mathbb{R}^n} \quad F(z) \tag{42}$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is a twice Lipschitz continuously differentiable function. The Newton-CG approach makes use of the following assumption.

**Assumption 4** (a) The set $\{z \mid F(z) \le F(z_0)\}$ is compact, where $z_0$ is the initial point.
(b) $F$ is twice uniformly Lipschitz continuously differentiable on a neighborhood of $\{z \mid F(z) \le F(z_0)\}$ that includes the trial points generated by the algorithm.
(c) Given $\epsilon_H > 0$ and $0 < \delta \ll 1$, a procedure called by the algorithm to verify approximate positive definiteness of $\nabla^2 F(z)$ either certifies that $\nabla^2 F(z) \succeq -\epsilon_H I$ or finds a direction along which curvature of $\nabla^2 F(z)$ is smaller than $-\epsilon_H/2$ in at most $N_{\text{meo}} \triangleq \min\{n, 1 + \lceil \mathcal{C}_{\text{meo}} \epsilon_H^{-1/2} \rceil\}$ Hessian-vector products, with probability $1 - \delta$, where $\mathcal{C}_{\text{meo}}$ depends at most logarithmically on $\delta$ and $\epsilon_H$.

Based on the above assumption, the following iteration complexity is indicated by [33, Theorem 4].

**Theorem 3** *Suppose that Assumption 4 holds. The Newton-CG terminates at a point satisfying*

$$\|\nabla F(z)\| \le \epsilon_g, \quad \lambda_{\min}(\nabla^2 F(z)) \ge -\epsilon_H, \tag{43}$$

*in at most $\bar{K}$ iterations with probability at least $(1-\delta)^{\bar{K}}$, where*

$$\bar{K} \triangleq \left\lceil C_{\text{NCG}} \max\{L_{F,H}^3, 1\}(F(z_0) - F_{\text{low}}) \max\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\} \right\rceil + 2. \tag{44}$$

*(With probability at most $1 - (1-\delta)^{\bar{K}}$, it terminates incorrectly within $\bar{K}$ iterations at a point at which $\|\nabla F(z)\| \le \epsilon_g$ but $\lambda_{\min}(\nabla^2 F(z)) < -\epsilon_H$.) Here, $C_{\text{NCG}}$ is a constant that depends on user-defined algorithmic parameters, $L_{F,H}$ is the Lipschitz constant for $\nabla^2 F$ on the neighborhood defined in Assumption 4(b), and $F_{\text{low}}$ is the lower bound of $F(z)$.*

Since in the Newton-CG approach, Hessian-vector products are the fundamental operations, [33] also derives operation complexity results, in which the operations are either evaluations of $\nabla F(z)$ or evaluations of matrix-vector products involving $\nabla^2 F(z)$ and an arbitrary vector (which can be computed without actually evaluating the Hessian itself).

**Corollary 2** *Suppose that Assumption 4 holds. Let $\bar{K}$ be defined as in (44). Then with probability at least $(1-\delta)^{\bar{K}}$, Newton-CG terminates at a point satisfying (43) after at most*

$$(\max\{2\min\{n, J(U_{F,H}, \epsilon_H)\} + 2, N_{\text{meo}}\})\bar{K}$$

*Hessian-vector products, where $U_{F,H}$ is the upper bound for $\nabla^2 F(z)$ on the neighborhood defined in Assumption 4(b) and $J(\cdot,\cdot)$ satisfies*

$$J(U_{F,H}, \epsilon_H) \le \min\left\{n, \left\lceil \left(\sqrt{\kappa} + \frac{1}{2}\right) \log\left(\frac{144(\sqrt{\kappa}+1)^2\kappa^6}{\zeta^2}\right) \right\rceil \right\}, \tag{45}$$

*where $\kappa \triangleq \frac{U_{F,H}+2\epsilon_H}{\epsilon_H}$ and $\zeta$ is a user-defined algorithmic parameter. (With probability at most $1-(1-\delta)^{\bar{K}}$, it terminates incorrectly within such complexity at a point for which $\|\nabla F(z)\| \le \epsilon_g$ but $\lambda_{\min}(\nabla^2 F(z)) < -\epsilon_H$.)*

To get total iteration and operation complexity we can aggregate the cost of applying Newton-CG to each subproblem in Algorithm 2. We present a critical lemma before deriving the total iteration complexity and operation complexity (Theorem 4 and Corollary 3). For these purposes, we denote the objective to be minimized at iteration $k$ of Algorithm 2 as follows:

$$\psi_k(x) \triangleq \mathcal{L}_\rho(x, \lambda_k) + \frac{\beta}{2}\|x - x_k\|^2. \tag{46}$$

Additionally, we recall from Assumption 1 that $S_\alpha^0 \triangleq \{f(x) + \frac{\rho_0}{2}\|c(x)\|^2 \le \alpha\}$ is either empty or compact for all $\alpha$.

**Lemma 4** *Suppose that assumptions and parameter settings in Theorem 2 hold. In addition, suppose that $\rho \ge \frac{1}{2}\|\lambda_0\|^2 + \rho_0$. Then we have*

$$\{x \mid \psi_k(x) \le \psi_k(x_k)\} \subseteq S_{\hat{\alpha}}^0,$$

*and*

$$\psi_k(x_k) - \psi_k^{low} \le \hat{\alpha} - \bar{L}, \tag{47}$$

*for all $k \ge 0$, where $\psi_k^{low} \triangleq \inf_{x \in \mathbb{R}^n} \psi_k(x)$ and $\hat{\alpha}$ is defined in (18). Hence $\{x \mid \psi_k(x) \le \psi_k(x_k)\}$ is compact for all $k \ge 0$.*

***Proof*** Because of $c_1 > 0$ and $c_2 > 0$ and (35), we have for any $k \ge 1$ that

$$P_{k+1} \le P_1 + \frac{16(M_c^2 + \sigma^2)R}{\rho\sigma^4} \overset{(32)}{\le} P_1 + 1. \tag{48}$$

Thus for any $k \ge 1$, we have

$$\psi_k(x_k) = \mathcal{L}_\rho(x_k, \lambda_k) \le P_k \le P_1 + 1,$$

which, using (29) and (18), implies that

$$\psi_k(x_k) \le 7\bar{U} + 7C_0 + 13\|\lambda_0\|^2 - 6\bar{L} + 1 = \hat{\alpha} - 1. \tag{49}$$

Note that (49) also holds when $k = 0$ since

$$\begin{aligned}
\psi_0(x_0) = \mathcal{L}_\rho(x_0, \lambda_0) &\overset{(24)}{\le} \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0 \\
&\overset{(25)}{\le} \bar{U} + \frac{1}{2\rho}\|\lambda_0\|^2 + C_0 + 6(\bar{U} + C_0 - \bar{L} + (5/4)\|\lambda_0\|^2) \\
&\overset{(\rho \ge 1)}{\le} 7\bar{U} + 7C_0 - 6\bar{L} + 8\|\lambda_0\|^2 < \hat{\alpha} - 1
\end{aligned}$$

Further, for any $k \ge 0$, we have

$$\begin{aligned}
\psi_k(x) &= \mathcal{L}_\rho(x, \lambda_k) + \frac{\beta}{2}\|x - x_k\|^2 \\
&= f(x) + \frac{\rho}{2}\|c(x)\|^2 + \lambda_k^T c(x) + \frac{\beta}{2}\|x - x_k\|^2 \\
&\ge f(x) + \frac{\rho}{2}\|c(x)\|^2 - \frac{\|\lambda_k\|^2}{2(\rho - \rho_0)} - \frac{(\rho - \rho_0)\|c(x)\|^2}{2} \\
&\ge f(x) + \frac{\rho_0}{2}\|c(x)\|^2 - \frac{1}{2(\rho - \rho_0)} \max\left\{\|\lambda_0\|^2, \frac{(M_f + \beta D_S + 1)^2}{\sigma^2}\right\}
\end{aligned}$$

$$\begin{aligned}&\overset{(32)}{\geq} f(x)+\frac{\rho_0}{2}\|c(x)\|^2-\max\left\{\frac{\|\lambda_0\|^2}{2(\rho-\rho_0)},1\right\},\\&\overset{(2(\rho-\rho_0)\geq\|\lambda_0\|^2)}{=} f(x)+\frac{\rho_0}{2}\|c(x)\|^2-1.\end{aligned}\tag{50}$$

The second inequality holds because $\|\lambda_k\|\leq(M_f+\beta D_S+1)^2/\sigma^2, \forall k\geq 1$ from Lemma 3. Then, for any $k\geq 0$, we have by combining (49) and (50) that

$$\psi_k(x_k)-\psi_k(x)\leq\hat{\alpha}-\left(f(x)+\frac{\rho_0}{2}\|c(x)\|^2\right).\tag{51}$$

Thus, for any $k\geq 0$, we have

$$\psi_k(x)\leq\psi_k(x_k)\implies\psi_k(x_k)-\psi_k(x)\geq 0\overset{(51)}{\implies}f(x)+\frac{\rho_0}{2}\|c(x)\|^2\leq\hat{\alpha}.$$

Therefore $\{x\mid\psi_k(x)\leq\psi_k(x_k)\}\subseteq S^0_{\hat{\alpha}}$ for all $k\geq 0$. For the claim (47), note that

$$\begin{aligned}\psi_k(x_k)-\psi_k^{low}&=\sup_{x\in\mathbb{R}^n}(\psi_k(x_k)-\psi_k(x))\\&\overset{(51)}{\leq}\sup_{x\in\mathbb{R}^n}\left(\hat{\alpha}-\left(f(x)+\frac{\rho_0}{2}\|c(x)\|^2\right)\right)\overset{(5)}{=}\hat{\alpha}-\bar{L}.\end{aligned}$$

□

By Lemma 4, we know that if the Newton-CG method of [33] is used to minimize $\psi_k(x)$ at iteration $k$ of Algorithm 2, Assumption 4(a) is satisfied at the initial point $x_k$. It also shows that the amount $\psi_k(x)$ can decrease at iteration $k$ is uniformly bounded for any $k\geq 0$. This is important in estimating iteration complexity of Newton-CG to solve the subproblem.

The following assumption is needed to prove complexity results about the Newton-CG method. Recall from definition (46) that

$$\begin{aligned}&\nabla^2\psi_k(x)\\&=\nabla^2 f(x)+\sum_{i=1}^m[\lambda_k]_i\nabla^2 c_i(x)+\rho\sum_{i=1}^m c_i(x)\nabla^2 c_i(x)+\rho\nabla c(x)\nabla c(x)^T+\beta I.\end{aligned}\tag{52}$$

**Assumption 5** (a) There exists a bounded open convex neighborhood $\mathcal{N}_{\hat{\alpha}}$ of $S^0_{\hat{\alpha}}$, where $\hat{\alpha}$ is defined as in (18), such that for any $k\geq 0$, the trial points of Newton-CG in iteration $k$ of Algorithm 2 lie in $\mathcal{N}_{\hat{\alpha}}$. Suppose that on $\mathcal{N}_{\hat{\alpha}}$, the functions $f(x)$ and $c_i(x), i=1,2,\ldots,m$ are twice uniformly Lipschitz continuously differentiable.

(b) Given $\epsilon^H_{k+1}>0$ and $0<\delta\ll 1$ at iteration $k\geq 0$, the procedure called by Newton-CG to verify sufficient positive definiteness of $\nabla^2\psi_k$ either certifies that $\nabla^2\psi_k(x)\succeq-\epsilon^H_{k+1}I$ or else finds a vector of curvature smaller than $-\epsilon^H_{k+1}/2$ in at most

$$N_{\text{meo}}\triangleq\min\{n,1+\lceil\mathcal{C}_{\text{meo}}(\epsilon^H_{k+1})^{-1/2}\rceil\}\tag{53}$$

Hessian-vector products, with probability $1-\delta$, where $\mathcal{C}_{\text{meo}}$ depends at most logarithmically on $\delta$ and $\epsilon^H_{k+1}$.

Boundedness and convexity of $\mathcal{N}_{\hat{\alpha}}$ and Assumption 5(a) imply that $\nabla^2\psi_k(x)$ is Lipschitz continuous on $\mathcal{N}_{\hat{\alpha}}$. Thus, Assumption 4(b) holds for each subproblem. Further, if we denote the Lipschitz constant for $\nabla^2\psi_k$ by $L_{k,H}$, then there exist $U_1$ and $U_2$ such that

$$L_{k,H}\leq U_1\rho+U_2,\tag{54}$$

where $U_1$ and $U_2$ depend only on $f$ and $c, \mathcal{N}_{\hat{\alpha}}$, and the upper bound for $\|\lambda_k\|$ from Lemma 3. Moreover, if $c(x)$ is linear, then $L_{k,H} = L_H$, where $L_H$ is the Lipschitz constant for $\nabla^2 f$.

The next theorem analyzes the total iteration complexity, given the parameter settings in Theorem 2 (with some additional requirements).

**Theorem 4** *Consider Algorithm 2 with stopping criterion (31), and suppose that the subproblem in Step 1 is solved with the Newton-CG procedure such that $x_{k+1}$ satisfies (6), (7) and with high probability satisfies (40). Suppose that Assumptions 1, 2 with $\mathcal{S} = S^0_{\hat{\alpha}}$ (with $\hat{\alpha}$ defined in (18)), Assumptions 3 and 5 hold. $\epsilon \in (0, 1]$ and $\eta \in [1, 2]$ are given. In addition, let $\|\tilde{r}_k\| \le \epsilon^g_k \triangleq \min\{1/k, \epsilon/2\}$, for all $k \ge 1$ (so that $R = \sum_{k=1}^{\infty} 1/k^2 = \pi^2/6$). Let $\beta = \epsilon^\eta/2$ and assume that $\rho \in [\rho_\eta, C_\rho \rho_\eta]$, where $C_\rho > 1$ is constant and*

$$\begin{aligned}\rho_\eta := \max\Bigg\{&\frac{16\max\{C_1, C_2\}}{\epsilon^\eta}, \frac{(M_f + \beta D_S + 1)^2}{2\sigma^2} + \rho_0,\\ &\frac{\|\lambda_0\|^2}{2} + \rho_0, \frac{16(M_c^2 + \sigma^2)R}{\sigma^4}, 3\rho_0, 1\Bigg\},\end{aligned} \tag{55}$$

*where $C_1$ and $C_2$ are defined in (11) and $D_S$ is the diameter of $S^0_{\hat{\alpha}}$ (see (20)). Suppose that $x_0$ satisfies $\|c(x_0)\|^2 \le \min\{C_0/\rho, 1\}$, where $C_0$ is the constant appearing in (18). Then we have the following.*

(i) *If we set $\epsilon^H_k \equiv \sqrt{\epsilon}/2$, then the total number of iterations of Newton-CG before Algorithm 2 stops and outputs an $\epsilon$-1o solution is $\mathcal{O}(\epsilon^{-2\eta-7/2})$, which is optimized when $\eta = 1$. When $c(x)$ is linear, this total iteration complexity is $\mathcal{O}(\epsilon^{\eta-7/2})$, which is optimized when $\eta = 2$.*

(ii) *If we let $\epsilon^H_k \equiv \epsilon/2$, then the total iteration number before Algorithm 2 stops and outputs an $\epsilon$-1o solution with probability 1 and an $\epsilon$-2o solution with probability at least $(1-\delta)^{\bar{K}_{T_\epsilon}}$ is $\mathcal{O}(\epsilon^{-2\eta-5})$, and $\bar{K}_{T_\epsilon} = \mathcal{O}(\epsilon^{-3\eta-3})$, where $T_\epsilon$ is defined in (30a) and $\bar{K}_{T_\epsilon}$ is the iteration complexity at iteration $T_\epsilon - 1$, defined below in (56). This bound is optimized when $\eta = 1$. When $c(x)$ is linear, this complexity is $\mathcal{O}(\epsilon^{\eta-5})$, and $\bar{K}_{T_\epsilon} = \mathcal{O}(\epsilon^{-3})$, so the optimal setting for $\eta$ is $\eta = 2$ in this case.*

***Proof*** We first prove (i). Note that if we use $x_k$ as the initial point for Newton-CG at iteration $k$, then (7) will be automatically satisfied because Newton-CG decreases the objective $\psi_k$ at each iteration. Due to Lemma 4 and Assumption 5, we know that Assumption 4 is satisfied for each subproblem. Thus, at iteration $k$, according to Theorem 3, given positive tolerances $\epsilon_g = \epsilon^g_{k+1}$ and $\epsilon_H = \epsilon^H_{k+1}$, Newton-CG will terminate at a point $x_{k+1}$ that satisfies (6) such that $\|\tilde{r}_{k+1}\| \le \epsilon^g_{k+1}$ with probability 1, and that satisfies (40) with probability $(1-\delta)^{\bar{K}_{k+1}}$, within

$$\begin{aligned}&\bar{K}_{k+1}\\ &\triangleq \left\lceil C_{\text{NCG}} \max\{L^3_{k,H}, 1\}(\psi_k(x_k) - \psi^{low}_k) \max\{(\epsilon^g_{k+1})^{-3}(\epsilon^H_{k+1})^3, (\epsilon^H_{k+1})^{-3}\}\right\rceil + 2.\end{aligned} \tag{56}$$

iterations, where $L_{k,H}$ is the Lipschitz constant for $\nabla^2 \psi_k(x)$. By substituting (47) from Lemma 4 into (56), we obtain

$$\bar{K}_{k+1} \le \left\lceil C_{\text{NCG}} \max\{L^3_{k,H}, 1\}(\hat{\alpha} - \bar{L}) \max\{(\epsilon^g_{k+1})^{-3}(\epsilon^H_{k+1})^3, (\epsilon^H_{k+1})^{-3}\}\right\rceil + 2, \tag{57}$$

for any $k \ge 0$. From (54) and the conditions on $\rho$, we have

$$L_{k,H} \le U_1\rho + U_2 = \mathcal{O}(\epsilon^{-\eta}), \tag{58}$$

where $U_1$ and $U_2$ depend only on $f$ and $c$, $\mathcal{N}_{\hat{\alpha}}$, and the upper bound for $\|\lambda_k\|$ from Lemma 3. When $c(x)$ is linear, we have $L_{k,H} \equiv L_H$.

Since $\|\tilde{r}_k\| \le \epsilon_k^g = \min\{1/k, \epsilon/2\}$ for all $k \ge 1$, the definition of $T_\epsilon$ in (30a) and the result of Theorem 2 imply that $T_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta})$. Therefore, for any $k \le T_\epsilon$ and $\eta \in [1, 2]$, we have

$$1/k \ge 1/T_\epsilon = \Omega(\epsilon^{2-\eta}) \implies \epsilon_k^g = \Omega(\epsilon) \implies (\epsilon_k^g)^{-1} = \mathcal{O}(\epsilon^{-1}). \tag{59}$$

Thus, when $\epsilon_k^H \equiv \sqrt{\epsilon}/2$, the term involving $\epsilon_{k+1}^g$ and $\epsilon_{k+1}^H$ on the right-hand sides of (56) and (57) are $\mathcal{O}(\epsilon^{-3/2})$. Therefore, we have from the bound for $\bar{K}_k$, the estimate (58), and $T_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta})$ that the total iteration complexity to obtain an $\epsilon$-1o solution is

$$\sum_{k=1}^{T_\epsilon} \bar{K}_k = \sum_{k=1}^{T_\epsilon} \max\{L_{k-1,H}^3, 1\}\mathcal{O}(\epsilon^{-3/2}) = T_\epsilon \mathcal{O}(\epsilon^{-3\eta})\mathcal{O}(\epsilon^{-3/2}) = \mathcal{O}(\epsilon^{-2\eta-7/2}).$$

This bound is optimized when $\eta = 1$. When $c(x)$ is linear, we have from $L_{k,H} = L_H = \mathcal{O}(1)$ that the complexity is

$$\sum_{k=1}^{T_\epsilon} \bar{K}_k = \sum_{k=1}^{T_\epsilon} \max\{L_H^3, 1\}\mathcal{O}(\epsilon^{-3/2}) = T_\epsilon \mathcal{O}(\epsilon^{-3/2}) = \mathcal{O}(\epsilon^{\eta-7/2}).$$

This bound is optimized when $\eta = 2$.

We turn now to (ii). Since Algorithm 2 stops at iteration $T_\epsilon - 1$, Newton-CG will stop at the point $x_{T_\epsilon}$ satisfying (6) with probability 1 and (40) with probability at least $(1-\delta)^{\bar{K}_{T_\epsilon}}$. Since $\epsilon_{T_\epsilon}^H = \epsilon/2$, $\eta \in [1, 2]$, and $\beta = \epsilon^\eta/2 \le \epsilon/2$, the following conditions are satisfied with probability at least $(1-\delta)^{\bar{K}_{T_\epsilon}}$:

$$\begin{aligned}
&\nabla_{xx}^2 \mathcal{L}_\rho(x_{T_\epsilon}, \lambda_{T_\epsilon - 1}) \overset{(40)}{\succeq} -(\beta + \epsilon_{T_\epsilon}^H) I \succeq -\epsilon I, \\
&\quad \implies \nabla^2 f(x_{T_\epsilon}) + \sum_{i=1}^m [\lambda_{T_\epsilon}]_i \nabla^2 c_i(x_{T_\epsilon}) + \rho \nabla c(x_{T_\epsilon}) \nabla c(x_{T_\epsilon})^T \succeq -\epsilon I, \\
&\quad \implies d^T \left( \nabla^2 f(x_{T_\epsilon}) + \sum_{i=1}^m [\lambda_{T_\epsilon}]_i \nabla^2 c_i(x_{T_\epsilon}) \right) d \ge -\epsilon \|d\|^2, \\
&\text{for any } d \in S(x_{T_\epsilon}) \triangleq \{d \in \mathbb{R}^n \mid [\nabla c(x_{T_\epsilon})]^T d = 0\},
\end{aligned}$$

which matches condition (2b) of Definition 2. Therefore, $x_{T_\epsilon}$ is an $\epsilon$-1o solution with probability 1 and an $\epsilon$-2o solution with probability at least $(1-\delta)^{\bar{K}_{T_\epsilon}}$. Since we have $(\epsilon_k^g)^{-1} = \mathcal{O}(\epsilon^{-1})$ for $k \le T_\epsilon$ as in (59), and $\epsilon_k^H = \epsilon/2$, the term involving $\epsilon_{k+1}^g$ and $\epsilon_{k+1}^H$ on the right-hand side of (56) and (57) is $\mathcal{O}(\epsilon^{-3})$. Recalling that $T_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta})$, the total iteration complexity to obtain $x_{T_\epsilon}$

$$\sum_{k=1}^{T_\epsilon} \bar{K}_k \overset{(57)}{=} \sum_{k=1}^{T_\epsilon} \max\{L_{k-1,H}^3, 1\}\mathcal{O}(\epsilon^{-3}) \overset{(58)}{=} T_\epsilon \mathcal{O}(\epsilon^{-3\eta})\mathcal{O}(\epsilon^{-3}) = \mathcal{O}(\epsilon^{-2\eta-5}).$$

This bound is optimized when $\eta = 1$. Note that

$$\bar{K}_{T_\epsilon} \overset{(57)}{=} \max\{L_{T_\epsilon-1,H}^3, 1\}\mathcal{O}(\epsilon^{-3}) \overset{(58)}{=} \mathcal{O}(\epsilon^{-3\eta-3}).$$

When $c(x)$ is linear, $L_{k,H} = L_H = \mathcal{O}(1)$ and the complexity to get $x_{T_\epsilon}$ is

$$\sum_{k=1}^{T_\epsilon} \bar{K}_k \overset{(57)}{=} \sum_{k=1}^{T_\epsilon} \max\{L_H^3, 1\}\mathcal{O}(\epsilon^{-3}) = T_\epsilon \mathcal{O}(\epsilon^{-3}) = \mathcal{O}(\epsilon^{\eta-5}),$$

which is optimized when $\eta = 2$. Note that in this case

$$\bar{K}_{T_\epsilon} \overset{(57)}{=} \max\{L_H^3, 1\}\mathcal{O}(\epsilon^{-3}) = \mathcal{O}(\epsilon^{-3}).$$

□

***Remark 2*** (i). A feasible point, if available, will satisfy $\|c(x_0)\|^2 \leq \min\{C_0/\rho, 1\}$ for all $\rho$. Otherwise, a "Phase I" procedure may be applied to the problem of minimizing $\|c(x)\|^2$. Since $\rho = \mathcal{O}(\epsilon^{-\eta})$, we have that $C_0/\rho = \Omega(\epsilon^\eta)$, $\eta \in [1, 2]$. Thus the Newton-CG algorithm could be use to find an approximate first-order point $\bar{x}$ such that $\|\nabla_x(\|c(x)\|^2)\ |_{x=\bar{x}}\ \| = \|2\nabla c(\bar{x})c(\bar{x})\| \leq \min\{\epsilon, \sqrt{C_0/\rho}, 1\} = \Omega(\epsilon)$. If $\|c(\bar{x})\| \leq \min\{\sqrt{C_0/\rho}, 1\}$, we can set $x_0 = \bar{x}$. Otherwise, $\|c(\bar{x})\| > \min\{\sqrt{C_0/\rho}, 1\} = \Omega(\epsilon)$, and we can terminate at the approximate infeasible critical point of $\|c(x)\|^2$, as in [17]. When Assumption 4 holds for $F(z) = \|c(z)\|^2$, Theorem 3 indicates that the iteration complexity of Newton-CG to find $\bar{x}$ is $\mathcal{O}(\epsilon^{-3/2})$ (where $\epsilon_g = \min\{\epsilon, \sqrt{C_0/\rho}, 1\}$, $\epsilon_H = \sqrt{\epsilon_g}$). Thus, the total iteration complexity of Proximal AL is not affected when we account for Phase 1.

(ii). Note that when $c(x)$ is linear, the optimized total iteration complexity bounds to obtain $\epsilon$-1o and $\epsilon$-2o point are $\mathcal{O}(\epsilon^{-3/2})$ and $\mathcal{O}(\epsilon^{-3})$, respectively. These bounds match the best known ones in literature for linear constraints (see Table 1 and corresponding discussion in Sect. 1.2). When $c(x)$ is nonlinear, the optimized total iteration complexity bounds to locate $\epsilon$-1o and $\epsilon$-2o point are $\mathcal{O}(\epsilon^{-11/2})$ and $\mathcal{O}(\epsilon^{-7})$, respectively. These bound are not competitive with the evaluation complexity bounds derived for two-phase second-order methods in [16,18] (see Table 1). These methods require solving a cubic regularization subproblem or minimizing a nonconvex program to global optimality per evaluation, which are potentially expensive computational tasks. The Newton-CG algorithm used for our inner loop has standard iterative linear algebra subproblems, and is equipped with worst-case operation complexity guarantees. We take up the issue of total operation complexity next.

Recalling the formula for $\nabla^2\psi_k$ in (52), we define a constant $U_H$ such that

$$\|\nabla^2\psi_k(x)\| \leq U_H,\ \forall k \geq 0,\ \forall x \in \mathcal{S}_{\hat{\alpha}}. \tag{60}$$

Since $f(x)$ and $c_i(x)$, $i = 1, 2, \ldots, m$ are twice continuously differentiable on a neighborhood $\mathcal{N}_{\hat{\alpha}} \supseteq \mathcal{S}_{\hat{\alpha}}$ (by Assumption 5), and $\mathcal{S}_{\hat{\alpha}}$ is compact and $\lambda_k$ is upper bounded (by Lemma 3), then such a $U_H > 0$ exists. Moreover, there exist quantities $\tilde{U}_1$, $\tilde{U}_2$ such that

$$U_H \leq \tilde{U}_1\rho + \tilde{U}_2, \tag{61}$$

where $\tilde{U}_1$, $\tilde{U}_2$ depend only $f(\cdot)$, $c(\cdot)$, $\mathcal{S}_{\hat{\alpha}}$, $\beta$ (which is bounded if equals to $\epsilon^\eta/2$ for all $\epsilon \leq 1$ and $\eta \geq 0$), and the upper bound for $\|\lambda_k\|$ in Lemma 3.

We conclude this section with the result concerning operation complexity of Algorithm 2 in which the subproblems are solved inexactly with Newton-CG.

**Corollary 3** *Suppose that the setup and assumptions of Theorem 4 are satisfied. $U_H$ is a constant satisfying (60) and (61). $J(\cdot,\cdot)$ and $N_{\text{meo}}$ are specified in Corollary 2 and Assumption 5(b), respectively. Let $\bar{K}_{\text{total}} \triangleq \sum_{k=1}^{T_\epsilon} \bar{K}_k$ denote the total iteration complexity for Algorithm 2 with Newton-CG applied to the subproblems, where $\bar{K}_k$ is defined as in (56). Then the following claims are true.*

(i) *When* $\epsilon_k^H \equiv \sqrt{\epsilon}/2$, *then the total number of Hessian-vector products before Algorithm* 2 *stops and outputs an* $\epsilon$*-1o solution is bounded by*

$$\max\{2\min\{n, J(U_H, \sqrt{\epsilon}/2)\} + 2, N_{\text{meo}}\}\bar{K}_{\text{total}}.$$

*For all* $n$ *sufficiently large, this bound is* $\tilde{\mathcal{O}}(\epsilon^{-5\eta/2-15/4})$ *(which reduces to* $\tilde{\mathcal{O}}(\epsilon^{\eta/2-15/4})$ *when* $c(x)$ *is linear).*

(ii) *If we let* $\epsilon_k^H \equiv \epsilon/2$, *then the total number of Hessian-vector products before Algorithm* 2 *stops and outputs an* $\epsilon$*-1o solution with probability* 1 *and* $\epsilon$*-2o with probability at least* $(1-\delta)^{\bar{K}_{T_\epsilon}}$ *is bounded by*

$$\max\{2\min\{n, J(U_H, \epsilon/2)\} + 2, N_{\text{meo}}\}\bar{K}_{\text{total}}.$$

*For all* $n$ *sufficiently large, this bound is* $\tilde{\mathcal{O}}(\epsilon^{-5\eta/2-11/2})$ *(which reduces to* $\tilde{\mathcal{O}}(\epsilon^{\eta/2-11/2})$ *when* $c(x)$ *is linear).*

***Proof*** Since $\{\psi_k(x) \le \psi_k(x_k)\} \subseteq S_{\hat{\alpha}}^0$ (Lemma 4), then $\|\nabla^2\psi_k(x)\| \le U_H$ on $\{\psi_k(x) \le \psi_k(x_k)\}$ for each $k \ge 0$. Therefore, from Corollary 2, to solve the subproblem in iteration $k-1$ of Algorithm 2 (for $k \ge 1$), Newton-CG requires at most

$$(\max\{2\min\{n, J(U_H, \epsilon_k^H)\} + 2, N_{\text{meo}}\})\bar{K}_k \tag{62}$$

Hessian-vector products, where $\bar{K}_k$ is defined in (56), and $J(\cdot,\cdot)$ is bounded as in (45). From the latter definition and the fact that $U_H = \mathcal{O}(\rho) = \mathcal{O}(\epsilon^{-\eta})$, we have for sufficiently large $n$ that

$$J(U_H, \epsilon_k^H) \le \min\left(n, \tilde{\mathcal{O}}((U_H/\epsilon_k^H)^{1/2})\right) = \tilde{\mathcal{O}}\left((\epsilon_k^H)^{-1/2}\epsilon^{-\eta/2}\right). \tag{63}$$

From (53), we have at iteration $k-1$, for sufficiently large $n$, that

$$N_{\text{meo}} = \min\left(n, \tilde{\mathcal{O}}((\epsilon_k^H)^{-1/2})\right) = \tilde{\mathcal{O}}((\epsilon_k^H)^{-1/2}). \tag{64}$$

By noting that the bound in (63) dominates that of (64), we have from (62) that the number of Hessian-vector products needed at iteration $k-1$ is bounded by

$$\tilde{\mathcal{O}}\left((\epsilon_k^H)^{-1/2}\epsilon^{-\eta/2}\right)\bar{K}_k. \tag{65}$$

To prove (i), we have $\epsilon_k^H = \sqrt{\epsilon}/2$, so by substituting into (65) and summing over $k = 1, 2, \ldots, T_\epsilon$, we obtain the following bound on the total number of Hessian-vector products before termination:

$$\tilde{\mathcal{O}}(\epsilon^{-\eta/2-1/4})\bar{K}_{\text{total}}, \tag{66}$$

where $\bar{K}_{\text{total}} = \mathcal{O}(\epsilon^{-2\eta-7/2})$ from Theorem 4(i). By substituting into (66), we prove the result. When $c(x)$ is linear, we obtain the tighter bound by using the estimate $\bar{K}_{\text{total}} = \mathcal{O}(\epsilon^{\eta-7/2})$ that pertains to this case.

For (ii), we have from Theorem 4(ii) that $x_{T_\epsilon}$ is an $\epsilon$-1o solution with probability 1 and an $\epsilon$-2o solution with probability at least $(1-\delta)^{\bar{K}_{T_\epsilon}}$. By substituting $\epsilon_k^H = \epsilon/2$ into (65) and summing over $k = 1, \ldots, T_\epsilon$, we have that the total number of Hessian-vector products before termination is bounded by

$$\tilde{\mathcal{O}}(\epsilon^{-\eta/2-1/2})\bar{K}_{\text{total}}, \tag{67}$$

where $\bar{K}_{\text{total}} = \mathcal{O}(\epsilon^{-2\eta-5})$ from Theorem 4(ii), so the result is obtained by substituting into (67). When $c(x)$ is linear, we obtain the tighter bound by using the estimate $\bar{K}_{\text{total}} = \mathcal{O}(\epsilon^{\eta-5})$ that pertains to this case. □

## 5 Determining $\rho$

Our results above on outer iteration complexity, total iteration complexity, and operation complexity for Algorithm 2 are derived under the assumption that $\rho$ is larger than a certain threshold. However, this threshold cannot be determined a priori without knowledge of many parameters related to the functions and the algorithm. In this section, we sketch a framework for determining a sufficiently large value of $\rho$ without knowledge of these parameters. This framework executes Algorithm 2 as an inner loop and increases $\rho$ by a constant multiple in an outer loop whenever convergence of Algorithm 2 has not been attained in a number of iterations set for this outer loop. The framework is specified as Algorithm 3. The next theorem shows that $\{\rho_\tau\}_{\tau\geq 1}$ and $\{T_\tau\}_{\tau\geq 1}$ can be defined as geometrically increasing sequences without any dependence on problem-related parameter, and that this choice of sequences leads to an iteration complexity for Algorithm 3 that matches that of Algorithm 2 (from Theorem 2) to within a logarithm factor.

**Algorithm 3** Proximal AL with trial value of $\rho$

```
0. Choose initial multiplier $\Lambda_0$, positive sequences $\{\rho_\tau\}_{\tau\geq 1}$ and $\{T_\tau\}_{\tau\geq 1}$; set $\tau \leftarrow 1$.
1. Call Algorithm 2 with $x_0 = z_\tau$, $\lambda_0 = \Lambda_0$, $\rho = \rho_\tau$ and run Algorithm 2 for $T_\tau$ number of iterations, or
   until the stopping criteria are satisfied.
2. If the stopping criterion of Algorithm 2 are satisfied, STOP the entire algorithm and output solutions
   given by Algorithm 2; otherwise, $\tau \leftarrow \tau + 1$ and return to Step 1.
```

**Theorem 5** *Suppose that all the assumptions and settings in Theorem 2 for Algorithm 2 hold except for the choice of $\rho$. In particular, the values of $\epsilon \in (0, 1)$, $\eta \in [0, 2]$, $\beta$ and $R$ are the same in each loop of Algorithm 3, and $z_\tau$ satisfies $\|c(z_\tau)\|^2 \leq \min\{C_0/\rho_\tau, 1\}$, where $C_0$ is the constant appearing in (18). Suppose that Algorithm 3 terminates when the conditions (31) are satisfied. For user-defined parameters $q > 1$ and $T_0 \in \mathbb{Z}_{++}$, we define the sequences $\{\rho_\tau\}_{\tau\geq 1}$ and $\{T_\tau\}_{\tau\geq 1}$ as follows:*

$$\begin{cases} \rho_\tau = \max\{q^\tau \epsilon^{2-2\eta}, 1\},\ T_\tau = \lceil T_0 q^\tau \rceil + 1, & \text{if } \eta \in [0, 1), \\ \rho_\tau = q^\tau,\ T_\tau = \lceil T_0 q^\tau \rceil + 1, & \text{if } \eta = 1, \\ \rho_\tau = q^\tau,\ T_\tau = \max\{\lceil T_0 q^\tau \epsilon^{2\eta-2} \rceil + 1, T_0\}, & \text{if } \eta \in (1, 2]. \end{cases}$$

*Then Algorithm 3 stops within $\log_q\left(\epsilon^{\min\{\eta-2,-\eta\}}\right) + \mathcal{O}(1)$ iterations. The number of iterations of Algorithm 2 that are performed before Algorithm 3 stops is $\tilde{\mathcal{O}}(\epsilon^{\eta-2})$.*

***Proof*** According to Theorem 2, at iteration $\tau$, the stopping criterion must be satisfied within $T_\tau$ number of iterations if $\rho$ satisfies (32) and $T_\tau$ is greater than the upper bound for $T_\epsilon$ estimated in Theorem 2 (see (39)). Therefore, Algorithm 3 is guaranteed to stop when $\rho_\tau$ and $T_\tau$ are large enough.

When $\eta \in (0, 1)$, $\rho_\tau$ will satisfy (32) if

$$\begin{aligned} q^\tau \epsilon^{2-2\eta} &\geq \max\left\{\frac{16\max\{C_1, C_2\}}{\epsilon^\eta}, \frac{(M_f + \beta D_S + 1)^2}{2\sigma^2} + \rho_0, \frac{16(M_c^2 + \sigma^2)R}{\sigma^4}, 3\rho_0, 1\right\} \\ \Leftrightarrow \tau &\geq \max\left\{\log_q\left(16\max\{C_1, C_2\}\epsilon^{\eta-2}\right),\right. \\ &\qquad \left.\log_q\left(\max\left\{\frac{(M_f + \beta D_S + 1)^2}{2\sigma^2} + \rho_0, \frac{16(M_c^2 + \sigma^2)R}{\sigma^4}, 3\rho_0, 1\right\}\epsilon^{2\eta-2}\right)\right\} \end{aligned}$$

$$= \log_q(\epsilon^{\eta-2}) + \mathcal{O}(1).$$

$T_\tau$ will be larger than the upper bound for $T_\epsilon$, that is, $\lceil \hat{\Delta}\epsilon^{\eta-2} \rceil + 1$ (for $\hat{\Delta}$ defined in (37b)) if

$$T_0 q^\tau \geq \hat{\Delta}\epsilon^{\eta-2} \Leftrightarrow \tau \geq \log_q\left(\hat{\Delta}\epsilon^{\eta-2}/T_0\right) = \log_q(\epsilon^{\eta-2}) + \mathcal{O}(1).$$

Therefore, Algorithm 3 will stop in $\log_q(\epsilon^{\eta-2}) + \mathcal{O}(1)$ number of iterations. Note that $\rho_\tau = \mathcal{O}(\epsilon^{-\eta})$, $T_\tau = \mathcal{O}(\epsilon^{\eta-2})$ for any $\tau$ before the algorithm stops. Therefore the total number of iteration of Algorithm 2 is $\tilde{\mathcal{O}}(\epsilon^{\eta-2})$. The same result holds when $\eta \in [1, 2]$ and the proof is similar (thus omitted). □

***Remark 3*** In Theorem 5, we almost recover the iteration complexity of Algorithm 2 derived in Theorem 2, except for a factor of $\log(1/\epsilon)$. The iteration complexity required to obtain $\epsilon$-2o (Corollary 1) is immediate by Algorithm 3 if Step 1 of Algorithm 2 satisfies (40). To recover the iteration complexity of subproblem solver (Newton-CG) derived in Theorem 4, we could use similar approach by setting a limit on the iteration of Newton-CG and increasing this limit geometrically with respect to $\tau$. The approach and analysis are quite similar to that presented above, so we omit the details.

## 6 Conclusion

We have analyzed complexity of a Proximal AL algorithm to solve smooth nonlinear optimization problems with nonlinear equality constraints. Three types of complexity are discussed: outer iteration complexity, total iteration complexity and operation complexity. In particular, we showed that if the first-order (second-order) stationary point is computed inexactly in each subproblem, then the algorithm outputs an $\epsilon$-1o ($\epsilon$-2o) solution within $\mathcal{O}(1/\epsilon^{2-\eta})$ outer iterations ($\beta = \mathcal{O}(\epsilon^\eta)$, $\rho = \Omega(1/\epsilon^\eta)$; $\eta \in [0, 2]$ for first-order case and $\eta \in [1, 2]$ for second-order case). We also investigate total iteration complexity and operation complexity when the Newton-CG method of [33] is used to solve the subproblems. A framework for determining the appropriate value of algorithmic parameter $\rho$ is presented, and we show that the iteration complexity increases by only a logarithmic factor for this approach by comparison with the version in which $\rho$ is known in advance.

There are several possible extensions of this work. First, we may consider a framework in which $\rho$ is varied within Algorithm 2. Second, extensions to nonconvex optimization problems with nonlinear *inequality* constraints remain to be studied.

**Acknowledgements** Research supported by Award N660011824020 from the DARPA Lagrange Program, NSF Awards 1628384, 1634597, and 1740707; and Subcontract 8F-30039 from Argonne National Laboratory.

## Appendix: Proofs of Elementary Results

***Proof of Theorem 1*** Since $x^*$ is a local minimizer of (1), it is the unique global solution of

$$\min\ f(x) + \frac{1}{4}\|x - x^*\|^4 \quad \text{subject to} \quad c(x) = 0, \quad \|x - x^*\| \leq \delta, \tag{68}$$

for $\delta > 0$ sufficiently small. For the same $\delta$, we define $x_k$ to be the global solution of

$$\min \quad f(x) + \frac{\rho_k}{2}\|c(x)\|^2 + \frac{1}{4}\|x - x^*\|^4 \quad \text{subject to} \quad \|x - x^*\| \leq \delta, \tag{69}$$

for a given $\rho_k$, where $\{\rho_k\}_{k\geq 1}$ is a positive sequence such that $\rho_k \to +\infty$. Note that $x_k$ is well defined because the feasible region is compact and the objective is continuous. Suppose that $z$ is any accumulation point of $\{x_k\}_{k\geq 1}$, that is, $x_k \to z$ for $k \in \mathcal{K}$, for some subsequence $\mathcal{K}$. Such a $z$ exists because $\{x_k\}_{k\geq 1}$ lies in a compact set, and moreover, $\|z - x^*\| \leq \delta$. We want to show that $z = x^*$. By the definition of $x_k$, we have for any $k \geq 1$ that

$$\begin{aligned} f(x^*) &= f(x^*) + \frac{\rho_k}{2}\|c(x^*)\|^2 + \frac{1}{4}\|x^* - x^*\|^4 \\ &\geq f(x_k) + \frac{\rho_k}{2}\|c(x_k)\|^2 + \frac{1}{4}\|x_k - x^*\|^4 \geq f(x_k) + \frac{1}{4}\|x_k - x^*\|^4. \end{aligned} \tag{70}$$

By taking the limit over $\mathcal{K}$, we have $f(x^*) \geq f(z) + \frac{1}{4}\|z - x^*\|^4$. From (70), we have

$$\frac{\rho_k}{2}\|c(x_k)\|^2 \leq f(x^*) - f(x_k) \leq f(x^*) - \inf_{k\geq 1} f(x_k) < +\infty. \tag{71}$$

By taking limits over $\mathcal{K}$, we have that $c(z) = 0$. Therefore, $z$ is the global solution of (68), so that $z = x^*$.

Without loss of generality, suppose that $x_k \to x^*$ and $\|x_k - x^*\| < \delta$. By first and second-order optimality conditions for (69), we have

$$\nabla f(x_k) + \rho_k \nabla c(x_k) c(x_k) + \|x_k - x^*\|^2 (x_k - x^*) = 0,$$

$$\nabla^2 f(x_k) + \rho_k \sum_{i=1}^{m} c_i(x_k)\nabla^2 c_i(x_k) + \rho_k \nabla c(x_k)[\nabla c(x_k)]^T \tag{72}$$

$$+ 2(x_k - x^*)(x_k - x^*)^T + \|x_k - x^*\|^2 I \succeq 0. \tag{73}$$

Define $\lambda_k \triangleq \rho_k c(x_k)$ and $\epsilon_k \triangleq \max\{\|x_k - x^*\|^3, 3\|x_k - x^*\|^2, \sqrt{2(f(x^*) - \inf_{k\geq 1} f(x_k))/\rho_k}\}$. Then by (71), (72), (73) and Definition 2, $x_k$ is $\epsilon_k$-2o. Note that $x_k \to x^*$ and $\rho_k \to +\infty$, so $\epsilon_k \to 0^+$. □

***Proof of Lemma 1*** We prove by contradiction. Otherwise for any $\alpha$ we could select sequence $\{x_k\}_{k\geq 1} \subseteq S_\alpha^0$ such that $f(x_k) + \frac{\rho_0}{2}\|c(x_k)\|^2 < -k$. Let $x^*$ be an accumulation point of $\{x_k\}_{k\geq 1}$ (which exists by compactness of $S_\alpha^0$). Then there exists index $K$ such that $f(x^*) + \frac{\rho_0}{2}\|c(x^*)\|^2 \geq -K + 1 > f(x_k) + \frac{\rho_0}{2}\|c(x_k)\|^2 + 1$ for all $k \geq K$, which contradicts the continuity of $f(x) + \frac{\rho_0}{2}\|c(x)\|^2$. □

## References

1. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: On augmented Lagrangian methods with general lower-level constraints. SIAM J. Optim. **18**(4), 1286–1309 (2008). https://doi.org/10.1137/060654797
2. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: Second-order negative-curvature methods for box-constrained and general constrained optimization. Comput. Optim. Appl. **45**(2), 209–236 (2010). https://doi.org/10.1007/s10589-009-9240-y
3. Andreani, R., Fazzio, N., Schuverdt, M., Secchin, L.: A sequential optimality condition related to the quasi-normality constraint qualification and its algorithmic consequences. SIAM J. Optim. **29**(1), 743–766 (2019). https://doi.org/10.1137/17M1147330
4. Andreani, R., Haeser, G., Ramos, A., Silva, P.J.S.: A second-order sequential optimality condition associated to the convergence of optimization algorithms. IMA J. Numer. Anal. **37**(4), 1902–1929 (2017)
5. Andreani, R., Martínez, J.M., Ramos, A., Silva, P.J.S.: A cone-continuity constraint qualification and algorithmic consequences. SIAM J. Optim. **26**(1), 96–110 (2016). https://doi.org/10.1137/15M1008488

6. Andreani, R., Secchin, L., Silva, P.: Convergence properties of a second order augmented Lagrangian method for mathematical programs with complementarity constraints. SIAM J. Optim. **28**(3), 2574–2600 (2018). https://doi.org/10.1137/17M1125698
7. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, Cambridge (2014)
8. Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. Math. Program. **149**(1), 301–327 (2015). https://doi.org/10.1007/s10107-014-0753-5
9. Birgin, E.G., Floudas, C.A., Martínez, J.M.: Global minimization using an augmented Lagrangian method with variable lower-level constraints. Math. Program. **125**(1), 139–162 (2010). https://doi.org/10.1007/s10107-009-0264-y
10. Birgin, E.G., Gardenghi, J., Martínez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**(1–2), 359–368 (2017)
11. Birgin, E.G., Haeser, G., Ramos, A.: Augmented Lagrangians with constrained subproblems and convergence to second-order stationary points. Comput. Optim. Appl. **69**(1), 51–75 (2018). https://doi.org/10.1007/s10589-017-9937-2
12. Birgin, E.G., Martínez, J.M.: Complexity and performance of an augmented Lagrangian algorithm. Optim. Methods Softw. (2020). https://doi.org/10.1080/10556788.2020.1746962
13. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011). https://doi.org/10.1561/2200000016
14. Cartis, C., Gould, N., Toint, P.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. **21**(4), 1721–1739 (2011)
15. Cartis, C., Gould, N., Toint, P.: Complexity bounds for second-order optimality in unconstrained optimization. J. Complex. **28**(1), 93–108 (2012)
16. Cartis, C., Gould, N.I.M., Toint, P.L.: On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization. SIAM J. Optim. **23**(3), 1553–1574 (2013). https://doi.org/10.1137/120869687
17. Cartis, C., Gould, N.I.M., Toint, P.L.: On the complexity of finding first-order critical points in constrained nonlinear optimization. Math. Program. Ser. A **144**, 93–106 (2014)
18. Cartis, C., Gould, N.I.M., Toint, P.L.: Optimization of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. J. Complex. **53**, 68–94 (2019)
19. Curtis, F.E., Jiang, H., Robinson, D.P.: An adaptive augmented Lagrangian method for large-scale constrained optimization. Math. Program. **152**(1), 201–245 (2015). https://doi.org/10.1007/s10107-014-0784-y
20. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**(1), 59–99 (2016). https://doi.org/10.1007/s10107-015-0871-8
21. Grapiglia, G.N., Nesterov, Y.: Regularized Newton methods for minimizing functions with Hölder continuous Hessians. SIAM J. Optim. **27**(1), 478–506 (2017). https://doi.org/10.1137/16M1087801
22. Grapiglia, G.N., Yuan, Y.X.: On the complexity of an augmented Lagrangian method for nonconvex optimization. arXiv e-prints arXiv:1906.05622 (2019)
23. Haeser, G., Liu, H., Ye, Y.: Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. Math. Program. (2018). https://doi.org/10.1007/s10107-018-1290-4
24. Hajinezhad, D., Hong, M.: Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. Math. Program. (2019). https://doi.org/10.1007/s10107-019-01365-4
25. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969). https://doi.org/10.1007/BF00927673
26. Hong, M., Hajinezhad, D., Zhao, M.M.: Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In: D. Precup, Y.W. Teh (eds.) Proceedings of the 34th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 70, pp. 1529–1538. PMLR (2017). http://proceedings.mlr.press/v70/hong17a.html
27. Jiang, B., Lin, T., Ma, S., Zhang, S.: Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. Comput. Optim. Appl. **72**(1), 115–157 (2019). https://doi.org/10.1007/s10589-018-0034-y
28. Liu, K., Li, Q., Wang, H., Tang, G.: Spherical principal component analysis. In: Proceedings of the 2019 SIAM International Conference on Data Mining, pp. 387–395 (2019). https://doi.org/10.1137/1.9781611975673.44

29. Nouiehed, M., Lee, J.D., Razaviyayn, M.: Convergence to second-order stationarity for constrained nonconvex optimization. arXiv e-prints arXiv:1810.02024 (2018)
30. O'Neill, M., Wright, S.J.: A log-barrier Newton-CG method for bound constrained optimization with complexity guarantees. IMA J. Numer. Anal. (2020). https://doi.org/10.1093/imanum/drz074
31. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Optimization (Sympos., Univ. Keele, Keele, 1968), pp. 283–298. Academic Press, London (1969)
32. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976). https://doi.org/10.1287/moor.1.2.97
33. Royer, C.W., O'Neill, M., Wright, S.J.: A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. Math. Program. (2019)
34. Sun, J., Qu, Q., Wright, J.: Complete dictionary recovery over the sphere. In: 2015 International Conference on Sampling Theory and Applications (SampTA), pp. 407–410 (2015)
35. Zhang, J., Luo, Z.Q.: A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. SIAM J. Optim. **30**(3), 2272–2302 (2020). https://doi.org/10.1137/19M1242276