Abstract

We analyze the iteration complexity of a Proximal Augmented Lagrangian (Proximal AL) framework for nonconvex optimization with nonlinear equality constraints. When a 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 $O(1/\epsilon^{2-\eta})$ major iterations ($1 \leq \eta \leq 2$) when the proximal term coefficient $\beta$ and penalty parameter $\rho$ satisfy $\beta = O(\epsilon^\eta)$ and $\rho = O(1/\epsilon^\eta)$, respectively. Further, when the subproblems are solved inexactly, the same order of complexity can be recovered by imposing certain verifiable conditions on the error sequence. Preliminary numerical results support our findings and demonstrate efficiency of this traditional method on dictionary learning.

1 Introduction

Nonconvex optimization captures a host of applications in machine learning. When such requirements as normalization, orthogonality, or consensus are imposed on the optimizer, the model may include nonlinear equality constraints. Relevant problems include dictionary learning [25], distributed optimization [17], and spherical PCA [19].

We consider the following problem:

$$\min \ f(x) \ \text{subject to} \ c_i(x) = 0, \ i = 1, \ldots, m, \quad (1)$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $c_i : \mathbb{R}^n \to \mathbb{R}, i \in 1, \ldots, m$ are continuous over their domains and second-order continuously differentiable over the interiors of their domains.

The Augmented Lagrangian (AL) framework is a penalty-type algorithm for solving (1), originating with Hestenes [16] and Powell [22]. Rockafellar proposed in [23] the proximal version of this method, which has both theoretical and practical advantages. The monograph [10] summarizes development of this method (then known as the method of multipliers) during the 1970s. Interest in the algorithm has resurfaced in recent years because of its connection to ADMM [10], which is based on AL.

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 $c(x) \triangleq (c_1(x), \ldots, c_m(x))^T$ and $\lambda \triangleq (\lambda_1, \ldots, \lambda_m)^T$. The (ordinary) Lagrangian of (1) is $\mathcal{L}_0(x, \lambda)$. 
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 \( k := 0 \);
1. Update \( x_k \): find approximate solution \( x_{k+1} \) to argmin \( L_{\rho_k}(x, \lambda_k) \);
2. Update \( \lambda_k \): \( \lambda_{k+1} := P_{\Lambda}(\lambda_k + \rho_k c(x_{k+1})) \);
3. Update \( \rho_k \): if \( k = 0 \) or \( \|c(x_{k+1})\|_{\infty} \leq \tau \|c(x_k)\|_{\infty}, \rho_{k+1} = \rho_k \); otherwise, \( \rho_{k+1} = \gamma \rho_k \);
4. If termination criteria is satisfied, STOP; otherwise, \( k := k + 1 \) and return to Step 1.

1.1 Related work

**AL for nonconvex optimization.** We consider first the basic augmented Lagrangian framework outlined in Algorithm \( \text{I} \). When \( f \) is a nonconvex function, convergence of the augmented Lagrangian framework has been studied in \( \text{[6,14,20]} \), with many other variants studied in \( \text{[1,4,12]} \). In \( \text{[6]} \), Algorithm \( \text{I} \) is investigated and generalized for a larger class of problems. In particular, this paper shows 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 \( \text{[8]} \), it is shown that when the subproblem in Algorithm \( \text{I} \) is solved to approximate global optimality with error approaching 0, the limit point is feasible and is a global solution of the original problem. However, none of this literature discusses the complexity, that is, a bound on the number of iterations required to achieve approximate optimality. In fact, there is little literature that addresses complexity for AL frameworks in the nonconvex setting.

The proximal augmented Lagrangian framework is presented in Algorithm \( \text{II} \).

Algorithm 2 Proximal Augmented Lagrangian (Proximal AL)

0. Initialize \( x_0, \lambda_0 \) and \( \rho > 0, \beta > 0 \); Set \( k := 0 \);
1. Update \( x_k \): Find approximate solution \( x_{k+1} \) to argmin \( L_{\rho}(x, \lambda_k) + \frac{\beta}{2}\|x - x_k\|^2 \);
2. Update \( \lambda_k \): \( \lambda_{k+1} := \lambda_k + \rho c(x_{k+1}) \);
3. If termination criteria is satisfied, STOP; otherwise, \( k := k + 1 \) and return to Step 1.

For this proximal version, complexity results become accessible in the nonconvex regime \( \text{[15,17,18,20]} \). The paper \( \text{[17]} \) analyzes the complexity of this approach (there named “proximal primal dual”) to obtain a first-order optimal point, choosing a special proximal term to make each subproblem strongly convex. Later, \( \text{[15]} \) proposes a “perturbed proximal primal dual algorithm,” a variant of Algorithm \( \text{II} \) to obtain complexity results for a problem class more general than \( \text{[I]} \). A proximal inexact augmented Lagrangian multiplier method is investigated in \( \text{[20]} \). This paper uses an exponentially weighted average of previous updates as the anchor point in the proximal term, and proves linear convergence in a certain measure on quadratic programming (QP). The paper \( \text{[18]} \) shows complexity of a proximal ADMM for obtaining a first-order optimal point. In all these works, \( c(x) \) is assumed to be linear. To our knowledge, complexity in the case of nonlinear \( c(x) \) and complexity for convergence to second-order optimal points have not yet been studied.

**Complexity for nonconvex optimization.** For constrained nonconvex optimization, worst case complexity of algorithms to obtain \( \epsilon \)-perturbed first-order and second-order optimal points has been studied in recent years. In particular, if only first-derivative information is used, iteration complexity to obtain an \( \epsilon \)-accurate first-order optimal point may be \( \mathcal{O}(\epsilon^{-2}) \) \( \text{[6,14,20]} \). If Hessian information is used, iteration complexity for an \( \epsilon \)-accurate first-order point can be improved to \( \mathcal{O}(\epsilon^{-3/2}) \) \( \text{[6,14,21]} \), while the complexity to obtain an \( \epsilon \)-accurate second-order point is typically \( \mathcal{O}(\epsilon^{-3}) \) \( \text{[6,14,21,22]} \).

The major iteration complexity of Proximal AL in \( \text{[17]} \) to obtain \( \epsilon \)-accurate first-order point (corresponding to our Definition \( \text{II} \) of \( \sqrt{\epsilon} - 1 \)) for nonconvex optimization with linear equality constraints is \( \mathcal{O}(\epsilon^{-1}) \). This is consistent with our result when choice of \( \beta \) and \( \rho \) is independent of \( \epsilon \). We could improve this complexity and derive the one to get \( \epsilon \)-2o by allowing \( \beta \) and \( \rho \) to be dependent on \( \epsilon \).

\footnote{Some very recent results on complexity analysis of AL in the nonconvex setting can be found in Grapiglia and Yuan \( \text{[13]} \), and Birgin and Martínez \( \text{[7]} \).}
1.2 Contributions

We apply the proximal augmented Lagrangian framework, Algorithm 2, to (1) where \( c(x) \) is non-linear. We define \( \epsilon \) first and second order optimal points (\( \epsilon\)-1o and \( \epsilon\)-2o) and show the following.

(i) When first-order (second-order) optimality is attained in the subproblems, the complexity of major iterations to obtain an \( \epsilon\)-1o (\( \epsilon\)-2o) point is \( O(1/\epsilon^{2-\eta}) \) if we let \( \beta = O(\epsilon^{\eta}) \) and \( \rho = O(1/\epsilon^{\eta}) \), \( 1 \leq \eta \leq 2 \). We show that the assumption of uniform boundedness and full rank of the constraint Jacobian can be restricted to a bounded level set, and that the primal and dual sequence of Proximal AL is bounded and the limit point satisfies first-order KKT conditions. (ii) If the subproblems are solved inexactly, the same order of complexity can be recovered by assuming appropriate checkable conditions on the sequence of errors. The preliminary numerical experiments, reported in Section 5, are consistent with the theoretical findings and show the efficiency of Proximal AL on dictionary learning.

Organization. In Section 2 all the definitions and assumptions used in subsequent analysis are listed. We discuss complexity of Proximal AL in Section 3 and the case with inexact subproblem solutions in Section 4. Preliminary numerical experiments are presented in Section 5 and we summarize and discuss future work in Section 6. Proofs of results in the main paper can be found in Section 4.

2 Preliminaries

Notation. \( \| \cdot \| \) denotes the Euclidean norm. \( \| \cdot \|_2 \) and \( \| \cdot \|_\infty \) denote the operator 2-norm and Frobenius norm of a matrix, respectively. For a given symmetric matrix \( H \), we denote \( \sigma_{\text{min}}(H) \) and \( \sigma_{\text{max}}(H) \) as its minimal and maximal eigenvalues, respectively. Denote \( \Delta x_{k+1} \equiv x_{k+1} - x_k \) and \( \Delta \lambda_{k+1} \equiv \lambda_{k+1} - \lambda_k \). Let \( \mathcal{D} \equiv \{ x \mid f(x) < +\infty \} \).

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:
\[
\| \nabla f(x) + \nabla c(x) \lambda \| \leq \epsilon, \quad \|c(x)\| \leq \epsilon,
\]
\[
d^T(\nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 c_i(x))d \geq -\epsilon \|d\|^2,
\]
for any \( d \in S(x) \equiv \{ d \in \mathbb{R}^n \mid |\nabla c(x)|^T d = 0 \} \).

These definitions are consistent with (9) and are suggested by the optimality conditions. In particular, we have the following theorem from (9).

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

We assume throughout the rest of discussion that function \( f \) is Lipschitz smooth over its domain \( \mathcal{D} \), that is, there exists a constant \( L_f \) such that
\[
\| \nabla f(x) - \nabla f(y) \| \leq L_f \| x - y \|, \quad \forall x, y \in \mathcal{D}.
\]

The following assumptions are used in the subsequent analysis.

Assumption 1. The following conditions on functions \( f \) and \( c \) hold:
(i) \( \| \nabla f(x) \| \leq M_f, \forall x \in \mathcal{D} \).
(ii) \( \| \nabla c(x) \|_2 \leq M_c, \sigma_{\text{min}}(\|\nabla c(x)\|^T \nabla c(x)) \geq \sigma^2 > 0, \forall x \in \mathcal{D} \).
(iii) \( \| \nabla c(x) - \nabla c(y) \|_2 \leq L_c \| x - y \|, \forall x, y \in \mathcal{D} \).

Assumption 2. \( \exists \rho_0 \in \mathbb{R} \) such that \( \inf_{x \in \mathbb{R}^n} \{ f(x) + \frac{\rho_0}{2} \|c(x)\|^2 \} \geq \bar{L} > -\infty \).

Remark. Assumption 2 holds in any of the following circumstances:
1. \( f \) is lower bounded over its domain.
2. \( f \equiv \frac{1}{2} x^T Q x - p^T x \) and \( c(x) \equiv Ax - b \). \( Q \) is positive definite on \( \text{null}(A) \equiv \{ x \mid Ax = 0 \} \).
3. \( f(x) + \frac{\rho_0}{2} \|c(x)\|^2 \) is lower semi-continuous and coercive. Moreover, \( \bar{L} \) is easy to see that for any \( \rho \geq \rho_0 \), we have
\[
\inf_{x \in \mathbb{R}^n} \{ f(x) + \frac{\rho}{2} \|c(x)\|^2 \} \geq \inf_{x \in \mathbb{R}^n} \{ f(x) + \frac{\rho_0}{2} \|c(x)\|^2 \} = \bar{L}.
\]
Our first results require Assumption 1 to hold. We then discuss a weakened version of this assumption, which requires the conditions (i), (ii), (iii) to hold only in a compact level set of the form $S^0_\alpha = \{ x \mid f(x) + \frac{\rho}{2} \| c(x) \|^2 \leq \alpha \}$, for some $\rho > 0$.

## 3 Complexity Analysis of Proximal AL

Throughout this section, we assume that the choice of $x_{k+1}$ used in Step 1 of Algorithm 2 satisfies the following conditions:

$$\nabla_x L_\rho(x_{k+1}, \lambda_k) + \beta (x_{k+1} - x_k) = 0. \tag{5}$$

We assume that (5) can be satisfied exactly for the time being. (We consider a relaxation of this condition in Section 4.) We additionally assume the following:

$$L_\rho(x_{k+1}, \lambda_k) + \beta \frac{2}{\rho} \| x_{k+1} - x_k \|^2 \leq L_\rho(x_k, \lambda_k). \tag{6}$$

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 $k \geq 1, \gamma > 0$, inspired by [17]:

$$P_k \triangleq L_\rho(x_k, \lambda_k) + \frac{\gamma}{2} \| x_k - x_{k-1} \|^2. \tag{7}$$

Then, for any $k \geq 1$, we have that

$$P_{k+1} - P_k = L_\rho(x_{k+1}, \lambda_{k+1}) - L_\rho(x_k, \lambda_k) + \frac{\gamma}{2} \| x_{k+1} - x_k \|^2 - \frac{\gamma}{2} \| x_k - x_{k-1} \|^2$$

$$= L_\rho(x_{k+1}, \lambda_{k+1}) - L_\rho(x_k, \lambda_k) + L_\rho(x_{k+1}, \lambda_k) - L_\rho(x_k, \lambda_k) + \frac{\gamma}{2} \| \Delta x_{k+1} \|^2 - \frac{\gamma}{2} \| \Delta x_k \|^2$$

$$\leq \frac{1}{\rho} \| \lambda_{k+1} - \lambda_k \|^2 - \beta \| x_{k+1} - x_k \|^2 + \gamma \| x_{k+1} - x_k \|^2 - \frac{\gamma}{2} \| x_k - x_{k-1} \|^2$$

$$= \frac{1}{\rho} \| \lambda_{k+1} - \lambda_k \|^2 - \beta \| x_{k+1} - x_k \|^2 - \frac{\gamma}{2} \| x_k - x_{k-1} \|^2. \tag{8}$$

We want to show that $\{ P_k \}_{k \geq 1}$ is a nonincreasing sequence, which requires bounding the term $\| \lambda_{k+1} - \lambda_k \|^2$.

### Lemma 2 (Bound for $\| \lambda_{k+1} - \lambda_k \|^2$).

Consider Algorithm 2 with (5) and (6), and suppose that Assumption 1 holds. Then for any $k \geq 1$, we have

$$\| \lambda_{k+1} - \lambda_k \|^2 \leq C_1 \| \Delta x_{k+1} \|^2 + C_2 \| \Delta x_k \|^2, \tag{9}$$

where

$$C_1 \triangleq \frac{2}{\sigma^2} \left( L_f + \frac{L_c M_f}{\sigma} + \beta \right)^2, \quad C_2 \triangleq \frac{2}{\rho} \frac{\gamma}{\sigma^2} \left( \beta + \frac{2M_f \gamma}{\sigma} \right)^2. \tag{10}$$

We now define two constants using the parameters from Algorithm 2 and Assumption 1

$$c_1 \triangleq \frac{\beta - \gamma}{2} - \frac{C_1}{\rho}, \quad c_2 \triangleq \frac{\gamma}{2} - \frac{C_2}{\rho}. \tag{11}$$

We show next that if certain parameters are chosen appropriately, then the sequence $\{ P_k \}_{k \geq 1}$ is nonincreasing and lower bounded.

### Lemma 3.

Consider Algorithm 2 with (5) and (6), with $\{ P_k \}_{k \geq 1}$ is defined as in (7). Suppose that $\beta > \gamma$ and $\rho$ is chosen large enough such that $c_1 > 0, c_2 > 0$ (defined in (11)). Then we have

$$P_{k+1} - P_k \leq -c_1 \| x_{k+1} - x_k \|^2 - c_2 \| x_k - x_{k-1} \|^2, \quad \text{for all } k \geq 1, \tag{12}$$

so that $\{ P_k \}_{k \geq 1}$ is a nonincreasing sequence.

**Proof.** (12) follows from (8) and (9). Since $c_1 > 0$ and $c_2 > 0, P_{k+1} \leq P_k$, for all $k \geq 1$. \[Q.E.D.\]

### Lemma 4.

Consider Algorithm 2 with (5) and (6), with $\{ P_k \}_{k \geq 1}$ is defined as in (7). Suppose that Assumption 1 and Assumption 2 hold. In addition, $c_1 > 0, c_2 > 0$, and $\rho \geq \rho_0$. Then $\{ P_k \}_{k \geq 1}$ is lower bounded by $L$, where $L$ is defined in Assumption 2.
**First-order complexity.** With these properties of \( \{P_k\}_{k \geq 1} \) shown, we are able to analyze the complexity of obtaining an \( \epsilon \)-1o solution. Part (ii) of the following result shows \( O(\epsilon^{-2}) \) complexity for fixed choices of parameters \( \beta, \rho, \) and \( \gamma \). Part (iii) shows that for specific choices of these parameters, depending on \( \epsilon \), we can improve the complexity to \( O(\epsilon^{-1}) \).

**Theorem 5** (First-order complexity - exact case). Consider Algorithm 2 with \( 5 \) and \( 9 \), and let \( \{P_k\}_{k \geq 1} \) be defined as in \( 7 \). Suppose that Assumption 7 and Assumption 2 hold. In addition, \( c_1 > 0, c_2 > 0 \) (cf. \( 11 \)) and \( \rho \geq \rho_0 \). Define \( r_k \triangleq \min_{1 \leq t \leq k} \{ P_i - P_{i+1} \}, \forall k \geq 1 \). Then the following holds:

(i) \( r_k = O(\frac{1}{\epsilon}) \);

(ii) Fix parameters \( \beta, \rho \) and \( \gamma \). For any \( \epsilon > 0 \), define

\[
T_\epsilon \triangleq \inf\{ t \geq 1 \mid x_t \text{ is an } \epsilon \text{-1o solution of } 1 \}.
\]

Also define \( \Delta \triangleq C \max \left\{ \frac{x^2}{c_1^2}, \frac{c_1}{c_2 \rho^2} \right\} \), where \( C \triangleq P_1 - \bar{L} \), with \( \bar{L} \) defined in Assumption 2 and \( C_1 \) and \( C_2 \) defined in \( 10 \). Then \( T_\epsilon \leq \lceil \Delta/\epsilon^2 \rceil + 1 \).

(iii) Choose \( x_0 \) such that \( c(x_0) = 0 \). For any \( \epsilon > 0 \) and some \( \eta \in [0, 2] \), suppose that

\[
\beta = \epsilon^\eta, \quad \gamma = \epsilon^{\eta/2}, \quad \rho = \max\{8/\epsilon^\eta, \max\{C_1, C_2\}, 3\rho_0, 1\},
\]

where \( C_1 \) and \( C_2 \) are defined as in \( 10 \) and \( T_\epsilon \) is defined as in (ii). Then \( T_\epsilon = O(1/\epsilon^{2-\eta}) \). In particular, if \( \eta = 2 \), then \( T_\epsilon = O(1) \).

**Proof.**

(i). According to Lemma 4, \( P_k \geq \bar{L}, \forall k \geq 1 \). Therefore, \( \sum_{i=1}^k (P_i - P_{i+1}) = P_1 - P_{k+1} \leq P_1 - \bar{L} < +\infty, \forall k \geq 1 \). Thus, the sequence \( \{P_i - P_{i+1}\}_{i \geq 1} \) is summable. This fact implies that \( r_k = O(\frac{1}{\epsilon}) \) (Proposition 3.4, 24).

(ii). Let \( K \triangleq \lceil \Delta/\epsilon^2 \rceil \). Since \( \sum_{i=1}^K (P_i - P_{i+1}) = P_1 - P_{K+1} \leq P_1 - \bar{L} = C, \) and \( P_i - P_{i+1} \geq 0, \forall i \geq 1 \) according to Lemma 3, \( \exists k \in [1, K] \), s.t., \( P_k - P_{k+1} \leq C/K \leq C\epsilon^2/\Delta \). By 12 from Lemma 3, \( \|x_{k+1} - x_k\|^2 \leq C\epsilon^2/(c_1 \Delta) \). Further, the first-order optimality condition 5 indicates that

\[
\|\nabla_x L_0(x_{k+1}, \lambda_{k+1})\|^2 \geq \beta^2 \|x_{k+1} - x_k\|^2 \leq \beta^2 C\epsilon^2/(c_1 \Delta) \leq \epsilon^2.
\]

Meanwhile, by 9 from Lemma 2

\[
\|c(x_{k+1})\|^2 \leq \|\lambda_{k+1} - \lambda_k\|^2/\rho^2 \leq (C_1/\rho^2)\|x_{k+1} - x_k\|^2 + (C_2/\rho^2)\|x_k - x_{k-1}\|^2
\]

\[
\leq \max \left\{ \frac{C_1}{c_1}, \frac{C_2}{c_2} \right\} \cdot \frac{1}{\rho^2} c_1 \|x_{k+1} - x_k\|^2 + c_2 \|x_k - x_{k-1}\|^2
\]

\[
\leq \max \left\{ \frac{C_1}{c_1}, \frac{C_2}{c_2} \right\} \cdot \frac{1}{\rho^2} (P_k - P_{k+1}) \leq \max \left\{ \frac{C_1}{c_1}, \frac{C_2}{c_2} \right\} \cdot \frac{C\epsilon^2}{\rho^2 \Delta} \leq \epsilon^2.
\]

According to the definition of \( T_\epsilon \),

\[
T_\epsilon = \inf\{ t \geq 1 \mid \exists \lambda \in \mathbb{R}^n, \|\nabla_x L_0(x_t, \lambda)\| \leq \epsilon, \|c(x_t)\| \leq \epsilon \}
\]

\[
\leq \inf\{ t \geq 1 \mid \|\nabla_x L_0(x_t, \lambda_t)\| \leq \epsilon, \|c(x_t)\| \leq \epsilon \}
\]

\[
\leq k + 1 \leq K + 1 = \lceil \Delta/\epsilon^2 \rceil + 1.
\]

(iii). We would like to show that

\[
T_\epsilon \leq \left[ \frac{(7f(x_0) + 9\|\lambda_0\|^2 - 7\bar{L}) \max\{8, 1/(8\epsilon^\eta)\}}{\epsilon^{2-\eta}} \right] + 1,
\]

where

\[
C_1^\circ \triangleq \frac{2}{\sigma^2} \left( L_f + \frac{L_c M_\ell}{\sigma} \right)^2.
\]
Recall the definitions of $C_1$ and $C_2$ in (10), of $c_1$ and $c_2$ in (11), and of $\beta, \gamma, \rho$ in (13). Then we have that:

$$
c_1 = \frac{\beta - \gamma}{2} - \frac{C_1}{\rho} \geq \epsilon^\eta, \quad c_2 = \frac{\gamma - C_2}{\rho} \geq \epsilon^\eta.
$$

Therefore, $c_1 > 0$, $c_2 > 0$, and $\rho \geq \rho_0$ are satisfied and the parameter assignment is legitimate. We now apply the result from part (ii), noting that the value of $\Delta$ defined there is now a function of $e$, because of how we define the parameters $\beta$, $\gamma$, and $\rho$. In fact, we show in the remainder of the proof that $\Delta = O(\epsilon^\eta)$.

We show first that $C = P_1 - \bar{L} = O(1)$. Note that

$$
P_1 = \mathcal{L}_\rho(x_1, \lambda_1) + \frac{\gamma}{2} \|x_1 - x_0\|^2
\leq \mathcal{L}_\rho(x_1, \lambda_1) - \mathcal{L}_\rho(x_1, \lambda_0) + \mathcal{L}_\rho(x_0, \lambda_0) - \mathcal{L}_\rho(x_0, \lambda_0) + \frac{\gamma}{2} \|x_1 - x_0\|^2
\leq \frac{1}{\rho} \|\lambda_1 - \lambda_0\|^2 - \frac{\beta}{2} \|x_1 - x_0\|^2 + \mathcal{L}_\rho(x_0, \lambda_0) + \frac{\gamma}{2} \|x_1 - x_0\|^2
= \rho \|c(x_1)\|^2 - \left(\frac{\beta - \gamma}{2}\right) \|x_1 - x_0\|^2 + f(x_0) + \lambda_0^T c(x_0) + \frac{\rho}{2} \|c(x_0)\|^2
= \rho \|c(x_1)\|^2 - \frac{\epsilon^\eta}{4} \|x_1 - x_0\|^2 + f(x_0) \leq \rho \|c(x_1)\|^2 + f(x_0),
$$

(17)

where the last equality follows from the definitions of $\beta$ and $\gamma$ together with $c(x_0) = 0$. In addition, we have

$$
f(x_1) + \lambda_0^T c(x_1) + \frac{\rho}{2} \|c(x_1)\|^2 + \frac{\beta}{2} \|x_1 - x_0\|^2
\leq f(x_0) + \lambda_0^T c(x_0) + \frac{\rho}{2} \|c(x_0)\|^2 = f(x_0),
$$

which indicates that

$$
\frac{\rho}{6} \|c(x_1)\|^2 \leq f(x_0) - \lambda_0^T c(x_1) - \frac{\rho}{6} \|c(x_1)\|^2 - f(x_1) - \frac{\rho}{6} \|c(x_1)\|^2
\leq f(x_0) - \frac{\rho}{6} \|c(x_1)\|^2 + 3\lambda_0^T c(x_1) + \frac{3}{2} \|\lambda_0\|^2 + \frac{3}{2} \|x_1\|^2 - f(x_1) - \frac{\rho}{6} \|c(x_1)\|^2
\leq f(x_0) + \frac{3}{2} \|\lambda_0\|^2 - f(x_1) - \frac{\rho}{2} \|c(x_1)\|^2.
$$

(18)

Therefore,

$$
C = P_1 - \bar{L} \leq \rho \|c(x_1)\|^2 + f(x_0) - \bar{L}
\leq 6f(x_0) + 9\|\lambda_0\|^2/\rho - 6\bar{L} + f(x_0) - \bar{L}
= 7f(x_0) + 9\|\lambda_0\|^2/\rho - 7\bar{L}
\leq 7f(x_0) + 9\|\lambda_0\|^2 - 7\bar{L},
$$

(proving that $C = O(1)$).

Next, we examine the terms $\frac{\beta^2}{c_1}$, $\frac{\gamma^2}{c_1 \rho^2}$, and $\frac{\lambda_0^2}{c_1 \rho^2}$, which together with $C$ make up the definition of $\Delta$ in part (ii). For the first of these terms, we have

$$
\frac{\beta^2}{c_1} \leq \frac{\epsilon^\eta}{8} \leq \frac{\epsilon^\eta}{8} = \epsilon^\eta.
$$

(19)

For $i = 1, 2$, we have

$$
\frac{C_i}{c_i \rho^2} \leq \frac{C_i}{\epsilon^\eta/8 \max\{C_1, C_2\}^2} \leq \frac{\epsilon^\eta}{8 \max\{C_1, C_2\}} \leq \frac{\epsilon^\eta}{8 C_1^2}.
$$

(19)
We have the following result for complexity of obtaining an \( \epsilon \). Apply inequality (14) from Theorem 5 (iii) and the result follows.

\[
\Delta = C \max \left\{ \frac{\beta^2}{c_1}, \frac{C_1}{c_1 \rho^2}, \frac{C_2}{c_2 \rho^2} \right\} \leq (7f(x_0) + 9\|\lambda_0\|^2 - 7L) \max \left\{ 8\eta, \frac{\epsilon^n}{8C_1^2} \right\}
\]

\[
= \epsilon^n (7f(x_0) + 9\|\lambda_0\|^2 - 7L) \max \left\{ 8, \frac{1}{8C_1^2} \right\}
\]

Then,

\[
T_e \leq \left[ \frac{\Delta}{\epsilon^2} \right] + 1 \leq \left[ \frac{(7f(x_0) + 9\|\lambda_0\|^2 - 7L) \max \{ 8, 1/(8C_1^2) \}}{\epsilon^2} \right] + 1,
\]

completing the proof.

**Remark.** The complexity result in (ii) is consistent with that of \([17]\). But part (iii) yields an improved complexity result, due to the special choice of the parameters: \( \beta = \epsilon^n \) and \( \rho = O(1/\epsilon^n) \). We can choose \( \beta \) to be small because, unlike \([17]\), we do not need the subproblem in Step 1 of Algorithm 2 to be strongly convex. Another benefit of small \( \beta \) is that it enables complexity analysis to obtain \( \epsilon \)-2o, which is a by-product of (iii), as we will see in the next corollary.

**Second-order complexity.** Let us further assume that \( x_{k+1} \) is a second-order stationary point of its subproblem, that is,

\[
\nabla^2_{xx} \mathcal{L}(x_{k+1}, \lambda_k) + \beta I \succeq 0.
\]

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

**Corollary 1 (Second-order complexity - exact case).** Consider Algorithm 2 with \( \{P_k\}_{k \geq 1} \) defined as in (7). In particular, the subproblem in Step 1 is solved such that second-order optimality conditions (5) (19) hold along with the decrease condition (6). Suppose that Assumptions 7 and 2 hold. Choose \( x_0 \) such that \( c(x_0) = 0 \). For any \( \epsilon > 0 \), define

\[
\bar{T}_e = \inf \{ t \geq 1 \mid x_t \text{ is an } \epsilon \text{-} 2O \text{ solution of } [1] \},
\]

and choose the parameters as follows:

\[
\beta = \epsilon^n, \quad \gamma = \epsilon^n/2, \quad \rho = \max \{ (8/\epsilon^n) \max \{C_1, C_2 \}, 3\rho_0, 1 \}, \quad 1 \leq \eta \leq 2, \quad \epsilon \leq 1.
\]

where \( C_1, C_2 \) are defined as in Theorem 5 (ii). Then \( \bar{T}_e = O(1/\epsilon^{2-\eta}) \). In particular, if \( \eta = 2 \), then \( \bar{T}_e = O(1) \).

**Proof.** Since \( \beta = \epsilon^n \), by (19), for any \( k \geq 0 \), \( \nabla^2_{xx} \mathcal{L}_\mu(x_{k+1}, \lambda_k) \succeq -\epsilon^n I \). This fact indicates that

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

which implies that \( d^T (\nabla^2_{xx} f(x_{k+1}) + \sum_{i=1}^m [\lambda_{k+1}], \nabla^2_{xx} c_i(x_{k+1})) d \geq -\epsilon^n \|d\|^2 \geq -\epsilon \|d\|^2 \), for any \( d \in S(x_{k+1}) \equiv \{ d \in \mathbb{R}^n \mid \|\nabla c(x_{k+1})\|^T d = 0 \}. \) This is exactly condition (2b) of Definition 2. Therefore,

\[
\bar{T}_e = \inf \{ t \geq 1 \mid \exists \lambda \in \mathbb{R}^m, \|\nabla f(x_t) + \nabla c(x_t) \lambda\| \leq \epsilon, \|c(x_t)\| \leq \epsilon, \}
\]

\[
d^T (\nabla^2_{xx} f(x_t) + \sum_{i=1}^m \lambda_i \nabla^2_{xx} c_i(x_t)) d \geq -\epsilon \|d\|^2, \forall d \in S(x_t) \}
\]

\[
\leq \inf \{ t \geq 1 \mid \|\nabla f(x_t) + \nabla c(x_t) \lambda_t\| \leq \epsilon, \|c(x_t)\| \leq \epsilon, \}
\]

\[
d^T (\nabla^2_{xx} f(x_t) + \sum_{i=1}^m \lambda_i \nabla^2_{xx} c_i(x_t)) d \geq -\epsilon \|d\|^2, \forall d \in S(x_t) \}
\]

\[
= \inf \{ t \geq 1 \mid \|\nabla f(x_t) + \nabla c(x_t) \lambda_t\| \leq \epsilon, \|c(x_t)\| \leq \epsilon \}.
\]

Apply inequality (14) from Theorem 5 (iii) and the result follows.
Weakening Assumption\cite{weakening}. If the domain $D$ in Assumption\cite{weakening} is infinite, the assumption may be violated even by quadratic functions. Instead, we may require the conditions of the assumption to hold only in some compact set that includes all the iterates. We start by assuming the following.

**Assumption 3.** Suppose that $\exists \rho_0 \geq 0$ such that $f(x) + \frac{\rho_0}{2} \|c(x)\|^2$ has compact sublevel sets, i.e., for any $\alpha \in \mathbb{R}$, $S^0_\alpha \triangleq \{f(x) + \frac{\rho_0}{2} \|c(x)\|^2 \leq \alpha\}$ is empty or compact.

This assumption holds when $f$ is strongly convex. It also holds when $f$ is bounded below and $c(x) = x^T x - 1$, as occurs in dictionary learning \cite{weakening}. It holds too when $f \triangleq \frac{1}{2} x^T Q x - p^T x$, $c(x) \triangleq A x - b$, $Q$ is positive definite on null($A$) \triangleq \{x \mid Ax = 0\}$. Assumption\cite{weakening} indicates lower boundedness of $f(x) + \frac{\rho_0}{2} \|c(x)\|^2$.

**Lemma 6.** Suppose that Assumption\cite{weakening} holds, then $f(x) + \frac{\rho_0}{2} \|c(x)\|^2$ is lower bounded.

*Proof.* Otherwise for any $\alpha$ we could select sequence $\{x_k\}_{k \geq 1} \subseteq S^0_\alpha$ and $f(x_k) + \frac{\rho_0}{2} \|c(x_k)\|^2 < -k$. Let $x^*$ be a cluster point of $\{x_k\}_{k \geq 1}$ ($x^*$ exists since $S^0_\alpha$ is compact). Then $\exists K$ such that $f(x^*) + \frac{\rho_0}{2} \|c(x^*)\|^2 \geq -K + 1 > f(x_k) + \frac{\rho_0}{2} \|c(x_k)\|^2 + 1, \forall k \geq K$, which contradicts the continuity of $f(x) + \frac{\rho_0}{2} \|c(x)\|^2$. \hfill $\blacksquare$

The weakened form of Assumption\cite{weakening} is as follows.

**Assumption 4.** Given a compact set $S \subseteq \mathbb{R}^n$, there exists positive constants $M_f, M_c, \sigma, L_c$ such that the following conditions on functions $f$ and $c$ hold.

(i). $\|\nabla f(x)\| \leq M_f, \forall x \in S$.

(ii). $\|\nabla c(x)\|_2 \leq M_c, \sigma_{\min}(\nabla c(x)^T \nabla c(x)) \geq \sigma^2 > 0, \forall x \in S$.

(iii). $\|\nabla c(x) - \nabla c(y)\|_2 \leq L_c \|x - y\|, \forall x, y \in S$.

We show now that under Assumption\cite{weakening} and weakened Assumption\cite{weakening} the results of Lemma\cite{lemma6} and Lemma\cite{lemma7} continue to hold.

**Lemma 7.** Consider Algorithm\cite{algorithm} with conditions $\cite{conditions}$ and $\cite{conditions}$. Let $\{P_k\}_{k \geq 1}$ be defined in $\cite{algorithm}$. Suppose that Assumption\cite{weakening} holds, that $c(x_0) = 0$, and define

$$\alpha \triangleq 7 f(x_0) - 6 \rho_0 + 9 \|\lambda_0\|^2 + 1, \text{ where } \rho_0 \triangleq \inf_{x \in \mathbb{R}^n} \left\{f(x) + \frac{\rho_0}{2} \|c(x)\|^2\right\}. \quad (20)$$

Suppose too that Assumption\cite{weakening} holds with $S = S^0_\alpha$. Choose $\rho, \beta, \gamma$ such that

$$\rho \geq \max \left\{\frac{(M_f + \beta D_S)^2}{2 \sigma^2} + \rho_0, 3 \rho_0, 1\right\},$$

and also that $c_1 > 0$ and $c_2 > 0$, where $c_1$ and $c_2$ are both defined in $\cite{conditions}$, with $C_1$ and $C_2$ defined in $\cite{conditions}$, where $D_S \triangleq \max\{|x - y| \mid x, y \in S^0_\alpha\}$. Then $\{P_k\}_{k \geq 1}$ is a nonincreasing sequence, and the following inequalities hold for any $k \geq 1$,

$$\|\lambda_{k+1} - \lambda_k\|^2 \leq C_1 \|\Delta x_{k+1}\|^2 + C_2 \|\Delta x_k\|^2,$$

$$P_{k+1} - P_k \leq -c_1 \|\Delta x_{k+1}\|^2 - c_2 \|\Delta x_k\|^2.$$

Furthermore, $\{x_k\}_{k \geq 1} \subseteq S^0_\alpha$ and $\|\lambda_k\|^2 \leq \frac{(M_f + \beta D_S)^2}{\sigma^2}$, $\forall k \geq 1$.

*Proof.* We prove the result by induction. We want to show that the following three bounds hold for all $i \geq 1$:

$$x_i \in S^0_\alpha, \quad \|\lambda_i\|^2 \leq \frac{(M_f + \beta D_S)^2}{\sigma^2} \leq 2(\rho - \rho_0), \quad P_i \leq 7 f(x_0) - 6 \rho_0 + 9 \|\lambda_0\|^2. \quad (21)$$

We verify first that $\cite{bounds}$ holds when $i = 1$. By inequality $\cite{bounds}$, we have

$$f(x_1) + \lambda_1^T c(x_1) + \frac{\rho}{2} \|c(x_1)\|^2 + \frac{\beta}{2} \|x_1 - x_0\|^2.$$
We now take the inductive step, supposing that (21) holds when
\[\rho \geq 3p_0,\]
where the last inequality follows from the definition of \(\rho\). This verifies that the second condition in (21) for \(i = 1\), so the third condition in (21) holds for \(i = 1\) also.

We now take the inductive step, supposing that (21) holds when \(i = k \geq 1\), and proving that these three conditions continue to hold for \(i = k + 1\). By inequality (6), we have

\[
f(x_{k+1}) + \frac{\rho_0}{2}||c(x_{k+1})||^2 + \frac{\beta}{2}||\Delta x_{k+1}||^2 \leq f(x_k) + \frac{\rho}{3}||c(x_k)||^2 \leq P_k
\]

\[
\Rightarrow f(x_{k+1}) + \frac{\rho_0}{2}||c(x_{k+1})||^2 + \frac{\beta}{2}||\Delta x_{k+1}||^2 \leq P_k
\]

\[
\Rightarrow f(x_{k+1}) + \frac{\rho_0}{2}||c(x_{k+1})||^2 \leq P_k + \frac{\lambda_k^2}{2(\rho - \rho_0)} \leq 7f(x_0) - 6l_0 + 9||\lambda_0||^2 + \alpha.
\]

(The inequality on the third line holds because of \(-R||a||^2 - 1/2||b||^2 \leq a^Tb\), for any \(R > 0\), \(a, b \in \mathbb{R}^m\).) Therefore, \(x_{k+1} \in S_0^0\), so we have proved the first condition in (21).

By the first order optimality (5) and the hypothesis \(x_k \in S_0^0\), the argument to establish that \(\|\lambda_{k+1}\|^2 \leq (M_1 + \beta D_S)^2 \leq 2(\rho - \rho_0)\) is the same as for the case of \(i = 1\). This establishes the second condition in (21) for \(i = k + 1\).

Since \(x_k, x_{k+1} \in S_0^0\), we can show in the same fashion as in the proof of Lemma [4] that

\[
\|\lambda_{k+1} - \lambda_k\|^2 \leq C_1\|\Delta x_{k+1}\|^2 + C_2\|\Delta x_k\|^2.
\]

By combining (22) with (8), we obtain

\[
P_{k+1} - P_k \leq -c_1\|\Delta x_{k+1}\|^2 - c_2\|\Delta x_k\|^2 \leq 0 \Rightarrow P_{k+1} \leq P_k.
\]

Thus \(P_{k+1} \leq 7f(x_0) - 6l_0 + 9||\lambda_0||^2\) and we have established the third condition in (21) for \(i = k + 1\). Note that (22) and (23) hold for all \(k \geq 1\), so we have completed the proof.

Remark. For dictionary learning (31) with equality constraints \(q^Tq - 1 = 0\), the assumptions in Lemma [4] are satisfied when \(\rho_0\) is large enough that \(S_0^0 \subseteq \{q \mid 0 < l \leq ||q|| \leq u\}\) for certain positive numbers \(l\) and \(u\).
Theorem 8. Consider Algorithm 2 with conditions (5) and (6). Suppose that \( \{P_k\}_{k \geq 1} \) is defined as in (7), that Assumption 3 holds, and that \( c(x_0) = 0 \). Let \( \alpha \) and \( \lambda_0 \) be defined as in (20). Suppose that Assumption 4 holds with \( S = S_0^0 \). For any \( \epsilon > 0 \) and \( \eta \in [0, 2] \), choose \( \rho, \beta, \gamma \) such that
\[
\beta = \epsilon^3, \quad \gamma = \epsilon^3 / 2, \quad \rho \geq \max \left\{ \frac{(M_f + \beta D_S)^2}{2\sigma^2} + \rho_0, \left(\frac{8}{\epsilon^3}\right) \max\{C_1, C_2\}, 3\rho_0, 1 \right\},
\]
where \( D_S \triangleq \max\{\|x - y\| \mid x, y \in S_0^0\} \) and \( C_1, C_2 \) are defined as in Theorem 5(i). Then the following statements are true:

(i). The sequence \( \{(x_k; \lambda_k)\}_{k \geq 1} \) generated by Algorithm 2 is bounded, and any accumulation point \((x^*, \lambda^*)\) of this sequence satisfies
\[
\nabla f(x^*) + \nabla c(x^*) \lambda^* = 0, c(x^*) = 0.
\]

(ii). Defining \( T_e \triangleq \inf \{t \geq 1 \mid x_t \text{ is an }\epsilon\text{-}1\text{o solution of } (1) \} \), we have \( T_e = \mathcal{O}(1/\epsilon^2 - \eta) \).

(iii). Suppose that \( \eta \in [1, 2] \) and \( \epsilon \in (0, 1] \). Defining \( \hat{T}_e \triangleq \inf \{t \geq 1 \mid x_t \text{ is an }\epsilon\text{-}2\text{o solution of } (1) \} \), we have that \( \hat{T}_e = \mathcal{O}(1/\epsilon^2 - \eta) \).

Proof. (i). Lemma 7 ensures that \( \{x_k\}_{k \geq 1} \subseteq S_0^0 \) where \( S_0^0 \) is compact, and \( \|\lambda_k\| \leq \frac{M_f + \beta D_S}{\sigma} \) for all \( k \geq 1 \). Therefore, sequence \( \{(x_k; \lambda_k)\}_{k \geq 1} \) is bounded. Since \( \{P_k\}_{k \geq 1} \) is a nonincreasing sequence as indicated in Lemma 7 and we have that
\[
\inf_{x \in \mathbb{R}^n} \{ f(x) + \frac{\rho_0}{2} \| c(x) \|^2 \} \geq \inf_{x \in \mathbb{R}^n} \{ f(x) + \frac{\rho_0}{2} \| c(x) \|^2 \} \geq l_0,
\]
we can show that \( P_k \geq l_0, \forall k \geq 1 \), following the proof of Lemma 4. Therefore, by (23) in the proof of Lemma 4, we have that
\[
\sum_{k=1}^{K} \| dx_{k+1} \|^2 + \sum_{k=1}^{K} \| dx_k \|^2 = P_1 - P_{K+1} \leq P_1 - l_0 < +\infty, \quad \text{for all } K \geq 1.
\]
Recall the definition of \( c_1 \) and \( c_2 \) from (11). Then
\[
c_1 = \frac{\beta - \gamma}{2} - \frac{C_1}{\rho} \geq \frac{\epsilon^3}{4} - \frac{\epsilon^3}{8} = \frac{\epsilon^3}{8} > 0,
\] and
\[
c_2 = \frac{\gamma}{2} - \frac{C_2}{\rho} \geq \frac{\epsilon^3}{4} - \frac{\epsilon^3}{8} = \frac{\epsilon^3}{8} > 0.
\]
Thus, \( \lim_{k \to \infty} \| dx_k \| = 0 \). Further, by (22), \( \lim_{k \to \infty} \| c(x_{k+1}) \| = \| \lambda_{k+1} - \lambda_k \| / \rho = 0 \). These facts indicate that for any cluster point \((x^*; \lambda^*)\), we have
\[
\nabla f(x^*) + \nabla c(x^*) \lambda^* = \lim_{k \to K} \nabla f(x_k) + \nabla c(x_k) \lambda_k \triangleq \lim_{k \to K} -\beta dx_k = 0,
\]
and \( c(x^*) = \lim_{k \to K} c(x_k) = 0 \), where \( K \) is a infinite subset of index such that \( \lim_{k \to K} x_k = x^* \), \( \lim_{k \to K} \lambda_k = \lambda^* \).

Proofs of (ii) and (iii) are similar to Theorem 5 and Corollary 1 thus omitted.

4 Proximal AL with inexact subproblems

In this section, we examine the case in which the subproblems are solved inexactly for \( x_{k+1} \) at each iteration \( k \). Specifically, consider Algorithm 2 and assume that in Step 1, the condition (6) holds along with
\[
\nabla_x \mathcal{L}_\rho(x_{k+1}, \lambda_k) + \beta (x_{k+1} - x_k) = \tilde{r}_{k+1},
\]
for some error vector \( \tilde{r}_{k+1} \). We continue to use the definition (7) of the Lyapunov function and note that (5) still holds despite of the inexactness. Also note that we continue to use Assumption 1 for main results in this section, but it can be weakened in a similar fashion to the second part of Section 3.

The inexactness leads to a modified bound on \( \| \lambda_{k+1} - \lambda_k \|^2 \) as we show now. (We continue to make use of the definitions (10) of constants \( C_1 \) and \( C_2 \).)

Lemma 9 (Bound for \( \| \lambda_{k+1} - \lambda_k \|^2 \) - Inexact Case). Consider Algorithm 2 with (6) and (24), and suppose that Assumption 1 holds. Then for any \( k \geq 1 \), we have that
\[
\| \lambda_{k+1} - \lambda_k \|^2 \leq 2C_1 \| dx_{k+1} \|^2 + 2C_2 \| dx_k \|^2 + \frac{16M^2}{\sigma^4} \| \tilde{r}_k \|^2 + \frac{4}{\sigma^2} \| \tilde{r}_{k+1} - \tilde{r}_k \|^2,
\]
where \( C_1 \) and \( C_2 \) are defined in (10).
Condition on the error sequence. In the inexact case, we are able to recover the complexity of the exact case, but need to control the error sequence \( \{ \hat{r}_k \}_{k \geq 1} \). In particular, a sufficient condition to achieve this is: \( \sum_{k=1}^{\infty} \| \hat{r}_k \|^2 < \infty, \| \hat{r}_k \| \leq \epsilon/2, \forall k \geq 1 \). For the rest of this subsection, we use the following definitions for \( \hat{c}_1 \) and \( \hat{c}_2 \):

\[
\hat{c}_1 \triangleq \frac{\beta - \gamma}{2} - \frac{2}{\rho} C_1, \quad \hat{c}_2 \triangleq \frac{\gamma}{2} - \frac{2}{\rho} C_1, \tag{26}
\]

where \( C_1 \) and \( C_2 \) are defined in (10). Analogously to Lemma 3 and Lemma 4, we derive the following properties of \( \{ P_k \}_{k \geq 1} \).

**Lemma 10.** Consider Algorithm 2 with (6) and (24), and let \( \{ P_k \}_{k \geq 1} \) be defined as in (7). Then for any \( k \geq 1 \), we have

\[
P_{k+1} - P_k \leq -\hat{c}_1 \| x_{k+1} - x_k \|^2 - \hat{c}_2 \| x_k - x_{k-1} \|^2 + \frac{16M^2}{\rho \sigma^4} \| \hat{r}_k \|^2 + \frac{4}{\rho \sigma^2} \| \hat{r}_{k+1} - \hat{r}_k \|^2. \tag{27}
\]

**Proof.** Use inequalities (8) and (25) and the result follows. \( \blacksquare \)

**Lemma 11.** Consider Algorithm 2 with (6) and (24), and let \( \{ P_k \}_{k \geq 1} \) be defined as in (7). Suppose that Assumption 7 and Assumption 2 hold. Further, let \( \hat{c}_1 > 0, \hat{c}_2 > 0 \) be defined as in (26), and let \( \rho \geq \rho_0 \), where \( \rho_0 \) is defined in Assumption 2. In addition, suppose that the residual sequence \( \{ \hat{r}_k \}_{k \geq 1} \) is chosen such that \( \sum_{k=1}^{\infty} \| \hat{r}_k \|^2 \leq R < \infty \). Then

\[
P_k \geq \tilde{L} - \frac{16(M^2 + \sigma^2)R}{\rho \sigma^4}, \quad \text{for all } k \geq 1.
\]

The next theorem claims that we are able to recover the complexity of exact case by imposing the checkable condition on \( \{ \hat{r}_k \}_{k \geq 1} \).

**Theorem 12** (First-order complexity - Inexact case). Consider Algorithm 2 with (6) and (24), and let \( \{ P_k \}_{k \geq 1} \) be defined as in (7). Suppose that Assumption 7 and Assumption 2 hold, and that \( \epsilon > 0 \) and \( \epsilon \in [0, 2] \) are given. Suppose that the residual sequence \( \{ \hat{r}_k \}_{k \geq 1} \) is chosen such that \( \sum_{k=1}^{\infty} \| \hat{r}_k \|^2 \leq R < \infty \) and \( \| \hat{r}_k \| \leq \epsilon/2 \) for all \( k \geq 1 \). Suppose that \( c(x_0) = 0 \). Then if we define

\[
T_\epsilon \triangleq \inf \{ t \geq 1 \mid \| \nabla_x L_0(x_t, \lambda_t) \| \leq \epsilon, \| c(x_t) \| \leq \epsilon \},
\]

and let

\[
\beta = \epsilon^3/2, \quad \gamma = \epsilon^3/4, \quad \rho = \max \{ 32 \max \{ C_1, C_2 \}/\epsilon^9, \sqrt{8(M^2 + \sigma^2)}/\sigma^2, 3\rho_0, 1 \}, \tag{28}
\]

where \( C_1 \) and \( C_2 \) are defined as in (10), then \( T_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta}) \). In particular, if \( \eta = 2 \), then \( T_\epsilon = \mathcal{O}(1) \). Therefore, \( \inf \{ t \geq 1 \mid x_t \text{ is an } \epsilon\text{-1o solution of (1)} \} = \mathcal{O}(1/\epsilon^{2-\eta}) \).

We further assume that in Step 1 of Algorithm 2, \( x_{k+1} \) can be computed such that the following condition is satisfied:

\[
\nabla^2_{xx} L_0(x_{k+1}, \lambda_k) + \beta I \geq -\epsilon^H_{k+1} I, \tag{29}
\]

where \( \{ \epsilon^H_{k+1} \}_{k \geq 1} \) is a chosen error sequence. Then second-order complexity can be obtained as a corollary of Theorem 12.

**Corollary 2** (Second-order complexity - inexact case). Consider Algorithm 2 with the \( x_{k+1} \) in Step 1 satisfying (24), (29), and (6). Suppose that Assumption 7 and Assumption 2 hold, and that \( \epsilon \in (0, 1) \) and \( \eta \in [0, 2] \) are given. In addition, assume that the error sequence \( \{ \hat{r}_k \}_{k \geq 1} \) is selected such that \( \sum_{k=1}^{\infty} \| \hat{r}_k \|^2 \leq R < \infty \) and \( \| \hat{r}_k \| \leq \epsilon/2 \) for all \( k \geq 1 \). Let \( c(x_0) = 0 \) and suppose that \( \epsilon^H_k = \epsilon/2 \) for all \( k \). Then if we define \( T_\epsilon \triangleq \inf \{ t \geq 1 \mid x_t \text{ is an } \epsilon\text{-2o solution of (1)} \} \) and choose the parameters as follows:

\[
\beta = \epsilon^3/2, \quad \gamma = \epsilon^3/4, \quad \rho = \max \left\{ (32/\epsilon^9) \max \{ C_1, C_2 \}, \sqrt{8(M^2 + \sigma^2)}/\sigma^2, 3\rho_0, 1 \right\}, \tag{30}
\]

where \( C_1, C_2 \) are defined as in (10), then \( T_\epsilon = \mathcal{O}(1/\epsilon^{2-\eta}) \).
5 Numerical experiment

We apply Proximal AL to dictionary learning (DL) ([8, 23]), collecting some preliminary numerical results that support our theoretical findings and showcase the efficiency of Proximal AL against an efficient technique proposed recently for this application.

**Problem description.** Let the data matrix $Y$ be created by $Y = A_0 X_0$, where $Y \in \mathbb{R}^{n \times p}$, $A_0 \in \mathbb{R}^{n \times n}$, $X_0 \in \mathbb{R}^{n \times p}$, $A_0$ is an orthogonal matrix and $X_0$ is sparse. We want to reconstruct the complete dictionary $A_0$ by solving the following optimization problem:

$$\min f(q^T Y) \quad \text{subject to } \|q\|_2 - 1 = 0,$$

where $f(\cdot)$ is a regularization function that enforces sparsity of $q^T Y$. The intuition is that based on statistical models, $q^T Y = q^T A_0 X_0$ is most sparse when $q$ is a column of $A_0$ up to sign (therefore $q^T A$ has only one nonzero element). This approach is also used in [5] where $f(z) \triangleq \frac{1}{p} \sum_{i=1}^{p} h_\mu(z_i)$, where $h_\mu(x) \triangleq \mu \log \left( \frac{\exp(x/\mu) + \exp(-x/\mu)}{2} \right) = \mu \log \cosh(x/\mu)$, as suggested in [23].

**Setup.** We use Matlab R2018b and Mac Air with 1.3 GHz Intel Core i5 CPU and 8GB Memory for experiments. We used the values $n = 30$ and $n = 50$. For each $n$, we define $p = 30n^2$; choose dictionary $A_0$ to be a randomly generated orthogonal matrix; choose $X_0$ from a Bernoulli-Gaussian distribution, (that is, $[X_0]_{ij} = B_{ij} G_{ij}$, where $B_{ij} \sim Ber(\theta = 0.3)$ and $G_{ij} \sim N(0, 1)$). For each data matrix $Y = A_0 X_0$, we run the algorithms from the same initial point $q_0$ chosen randomly from the unit sphere ($\|q_0\| = 1$), repeating this choice several times. We tested two methods.

(i). Proximal AL (the method of this paper) with three parameter settings: $(\beta, \rho) \in \{(1, 1), (0.1, 10), (0.01, 100)\}$. Also, set the smoothing parameter $\mu = 0.01$ and choose $\lambda_0 = 0$. We use gradient descent with backtracking linesearch for the subproblem, terminating when $\|\Delta x_k\| \leq \min \{1, \epsilon/\sqrt{k}\}$. We stop the algorithm if $\max \{\beta \|\Delta x_{k+1}\|, \|c(x_{k+1})\|\} \leq \epsilon$. Therefore, the algorithm outputs $x_{k+1}$ as a $2\epsilon$-1o solution. We fix $\epsilon = 10^{-3}$ and define error $\overset{\Delta}{=} \min_{1 \leq i \leq n} \{\max \{\|a_i - q_{\text{output}}\|, \|a_i + q_{\text{output}}\|\}\} = \|a_i\|$. We set 300 seconds as the maximum runtime allowed.

(ii). Subgradient descent described in [5] for (31) when $f$ is $\ell_1$-norm. We use the same algorithm setting as in [5, Section 5]. In particular, we terminate when $\|a_i - q_{\text{best}}\| \leq \epsilon = 10^{-3}$, where $q_{\text{best}}$ is the solution with best function value. $a_i$ is a column of $A_0$. Recall that in Theorem 5 (iii) and Theorem 12 we are able to obtain better complexity of $O(1/\epsilon)$ by assigning small $\beta$ and large $\rho$. The numerical results are consistent with this theory. Note that the computation time may not drop all the way with the iteration number, because when $\rho$ is large, solving the subproblem becomes slow using first-order methods. In addition, we find that Proximal AL may outperform subgradient descent method; the latter gives impressive results when compared with other methods in [5].

**Result.** Table 1 shows that as we increase $\rho$ and decrease $\beta$, the number of iterations decreases. We have analyzed complexity of Proximal AL to solve smooth nonlinear optimization problems with nonlinear equality constraints. We showed that if the first-order (second-order) stationary point is computed exactly or inexactly in each subproblem, then the algorithm outputs an $\epsilon$-1o ($\epsilon$-2o) solution within $O(1/\epsilon^{2-\eta})$ number of iterations ($1 \leq \eta \leq 2; \beta = O(\epsilon^\eta), \rho = O(1/\epsilon^\eta)$). Numerical experiments are presented to support the theoretical findings and prove the good performance of Proximal AL on dictionary learning.

---

Source code and simulation environment are available online: [https://github.com/exybrh/PALDL](https://github.com/exybrh/PALDL).
Table 1: Comparison between Proximal AL and Subgradient method on dictionary learning. 300* = maximum runtime reached.

| n = 30, p = 27000 | Proximal AL | Subgradient Descent |
|-------------------|-------------|---------------------|
| β = 1, ρ = 1      | β = 0.1, ρ = 10 | β = 0.01, ρ = 100 |
| error (t(s) iter.) | error (t(s) iter.) | error (t(s) iter.) |
| 9.4e-4 3.0 25 7.2e-4 1.0 | 3 4.5e-4 11 2 1.0e-3 183 | |
| 7.5e-4 2.4 23 7.3e-4 0.9 | 3 4.7e-4 8.5 2 1.0e-3 160 | |
| 8.0e-4 3.0 28 7.6e-4 1.0 | 3 5.0e-4 11 2 1.0e-3 178 | |

| n = 50, p = 75000 |
|-------------------|-------------|---------------------|
| β = 1, ρ = 1      | β = 0.1, ρ = 10 | β = 0.01, ρ = 100 |
| error (t(s) iter.) | error (t(s) iter.) | error (t(s) iter.) |
| 6.8e-4 9.9 28 6.4e-4 5.0 | 3 3.4e-4 57 2 1.0e-3 295 | |
| 7.4e-4 11 29 6.8e-4 4.0 | 3 3.3e-4 44 2 1.1e-3 300* | |
| 5.5e-4 10 27 7.1e-4 5.0 | 3 3.8e-4 49 2 1.2e-3 300* | |

There are several possible extensions of this work. First, we may investigate the overall computational complexity, taking into account the cost of solving the subproblems. Second, we may consider a framework in which β and ρ are varied during the algorithm, an approach which has more appeal in practice. Third, we will investigate extensions to nonconvex optimization with nonlinear inequality constraints.

Acknowledgments

Research supported by Award N660011824020 from the DARPA Lagrange Program, NSF Awards IIS-1447449, 1628384, 1634597, and 1740707; and Subcontract 8F-30039 from Argonne National Laboratory.

References

[1] R. Andreani, E. Birgin, J. Martínez, and M. Schuverdt. On augmented lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18(4):1286–1309, 2008. doi: 10.1137/060654797. URL https://doi.org/10.1137/060654797.

[2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. Second-order negative-curvature methods for box-constrained and general constrained optimization. *Computational Optimization and Applications*, 45(2):209–236, Mar 2010. ISSN 1573-2894. doi: 10.1007/s10589-009-9240-y. URL https://doi.org/10.1007/s10589-009-9240-y.

[3] R. Andreani, L. Secchin, and P. Silva. Convergence properties of a second order augmented lagrangian method for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 28(3):2574–2600, 2018. doi: 10.1137/17M1125698. URL https://doi.org/10.1137/17M1125698.

[4] R. Andreani, N. Fazzio, M. Schuverdt, and L. Secchin. A sequential optimality condition related to the quasi-normality constraint qualification and its algorithmic consequences. *SIAM Journal on Optimization*, 29(1):743–766, 2019. doi: 10.1137/17M1147330. URL https://doi.org/10.1137/17M1147330.

[5] Y. Bai, Q. Jiang, and J. Sun. Subgradient Descent Learns Orthogonal Dictionaries. *arXiv e-prints*, art. arXiv:1810.10702, Oct 2018.

[6] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Mathematical Programming*, 149(1):301–327, Feb 2015. ISSN 1436-4646. doi: 10.1007/s10107-014-0753-5. URL https://doi.org/10.1007/s10107-014-0753-5.
[7] E. Birgin and J. Martínez. Complexity and performance of an augmented lagrangian algorithm. *arXiv preprint arXiv:1907.02401*, Jul 2019.

[8] E. G. Birgin, C. A. Floudas, and J. M. Martínez. Global minimization using an augmented lagrangian method with variable lower-level constraints. *Mathematical Programming*, 125(1):139–162, Sep 2010. ISSN 1436-4646. doi: 10.1007/s10107-009-0264-y. URL https://doi.org/10.1007/s10107-009-0264-y

[9] E. G. Birgin, G. Haeser, and A. Ramos. Augmented lagrangians with constrained subproblems and convergence to second-order stationary points. *Computational Optimization and Applications*, 69(1):51–75, 2018.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[11] C. Cartis, N. Gould, and P. Toint. On the evaluation complexity of constrained nonlinear least-squares and general constrained nonlinear optimization using second-order methods. *SIAM Journal on Numerical Analysis*, 53(2):836–851, 2015. doi: 10.1137/130915546. URL https://doi.org/10.1137/130915546

[12] F. E. Curtis, H. Jiang, and D. P. Robinson. An adaptive augmented lagrangian method for large-scale constrained optimization. *Mathematical Programming*, 152(1):201–245, Aug 2015. ISSN 1436-4646. doi: 10.1007/s10107-014-0784-y. URL https://doi.org/10.1007/s10107-014-0784-y

[13] G. N. Grapiglia and Y.-x. Yuan. On the complexity of an augmented lagrangian method for nonconvex optimization. *arXiv: 1906.05622*, Jun 2019.

[14] G. Haeser, H. Liu, and Y. Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Mathematical Programming*, May 2018. ISSN 1436-4646. doi: 10.1007/s10107-018-1290-4. URL https://doi.org/10.1007/s10107-018-1290-4

[15] D. Hajinezhad and M. Hong. Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming*, Feb 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01365-4. URL https://doi.org/10.1007/s10107-019-01365-4

[16] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, Nov 1969. ISSN 1573-2878. doi: 10.1007/BF00927673. URL https://doi.org/10.1007/BF00927673

[17] M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *arXiv preprint arXiv:1604.00543*, 2016.

[18] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 72(1):115–157, Jan 2019. ISSN 1573-2894. doi: 10.1007/s10589-018-0034-y. URL https://doi.org/10.1007/s10589-018-0034-y

[19] K. Liu, Q. Li, H. Wang, and G. Tang. Spherical Principal Component Analysis. *arXiv e-prints*, art. arXiv:1903.06877, Mar 2019.

[20] M. Nouiehed, J. D. Lee, and M. Razaviyayn. Convergence to Second-Order Stationarity for Constrained Non-Convex Optimization. *arXiv e-prints*, art. arXiv:1810.02024, Oct 2018.

[21] M. O’Neill and S. J. Wright. A Log-Barrier Newton-CG Method for Bound Constrained Optimization with Complexity Guarantees. *arXiv e-prints*, art. arXiv:1904.03563, Apr 2019.

[22] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In *Optimization (Sympos., Univ. Keele, Keele, 1968)*, pages 283–298. Academic Press, London, 1969.
Appendix

Proof of Lemma 2

Proof. The first-order optimality condition for Step 1 implies that for all \( k \geq 0 \),
\[
\nabla f(x_{k+1}) + \nabla c(x_{k+1})\lambda_k + \rho \nabla c(x_{k+1})c(x_{k+1}) + \beta(x_{k+1} - x_k) = 0.
\]
\[
\Rightarrow \nabla f(x_{k+1}) + \nabla c(x_{k+1})\lambda_{k+1} + \beta(x_{k+1} - x_k) = 0.
\]
Likewise, by replacing \( k \) with \( k - 1 \), we obtain
\[
\nabla f(x_k) + \nabla c(x_k)\lambda_k + \beta(x_k - x_{k-1}) = 0.
\]

By combining (32) and (33) and using the notation \( \Delta \lambda_{k+1} \equiv \lambda_{k+1} - \lambda_k \) and \( \Delta x_{k+1} \equiv x_{k+1} - x_k \), we have
\[
\nabla f(x_{k+1}) - \nabla f(x_k) + \nabla c(x_{k+1})\Delta \lambda_{k+1} + (\nabla c(x_{k+1}) - \nabla c(x_k))\lambda_k + \beta(\Delta x_{k+1} - \Delta x_k) = 0,
\]
which by rearrangement gives
\[
\nabla c(x_{k+1})\Delta \lambda_{k+1} = -(\nabla f(x_{k+1}) - \nabla f(x_k) + (\nabla c(x_{k+1}) - \nabla c(x_k))\lambda_k + \beta(\Delta x_{k+1} - \Delta x_k)).
\]

Since \( \sigma \) is a lower bound on the smallest singular value of \( \nabla c(x_{k+1}) \), we have
\[
\|
\Delta \lambda_{k+1} \|
\leq \frac{1}{\sigma} \left( \|
\nabla f(x_{k+1}) - \nabla f(x_k)\| + \|
\nabla c(x_{k+1}) - \nabla c(x_k)\|_2\|
\lambda_k\| + \beta(\|
\Delta x_{k+1}\| + \|
\Delta x_k\|) \right).
\]
we have from (33) that
\[
\nabla c(x_k)\lambda_k = -\nabla f(x_k) - \beta(x_k - x_{k-1}),
\]
so that
\[
\|
\lambda_k\| \leq \frac{1}{\sigma} (\|
\nabla f(x_k)\| + \beta\|
\Delta x_k\|) \leq \frac{1}{\sigma} (M_f + \beta\|
\Delta x_k\|).
\]

We also have
\[
\|
\nabla c(x_{k+1}) - \nabla c(x_k)\|_2 \leq L_c\|
\Delta x_{k+1} - x_k\|.
\]

By substituting (3), (36), and (37) into (35), we obtain
\[
\|
\Delta \lambda_{k+1}\|
\leq \frac{1}{\sigma} \left( L_f\|
\Delta x_{k+1}\| + \beta\|
\Delta x_{k+1}\| + \beta\|
\Delta x_k\| + \|
\nabla c(x_{k+1}) - \nabla c(x_k)\|_2 \left( \frac{1}{\sigma}M_f + \frac{\beta}{\sigma}\|
\Delta x_k\| \right) \right)
\]
\[
\leq \frac{1}{\sigma} \left( L_f\|
\Delta x_{k+1}\| + \beta\|
\Delta x_{k+1}\| + \beta\|
\Delta x_k\| + \frac{L_cM_f}{\sigma}\|
\Delta x_{k+1}\| + \frac{2M_c\beta}{\sigma}\|
\Delta x_k\| \right)
\]
\[
\leq \frac{1}{\sigma} \left( L_f + \frac{L_cM_f}{\sigma} + \beta \right)\|
\Delta x_{k+1}\| + \frac{1}{\sigma} \left( \beta + \frac{2M_c\beta}{\sigma} \right)\|
\Delta x_k\|.
\]

By using the bound \( a \leq b + c \Rightarrow a^2 \leq 2b^2 + 2c^2 \) for positive scalars \( a, b, c \), and using the definition (10), we obtain the result.
Proof of Lemma 4

Proof. Note that for all $k \geq 1$,
\[ \lambda_k^T c(x_k) = \lambda_k^T (\lambda_k - \lambda_{k-1})/\rho = \frac{1}{2\rho} (\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2 + \|\lambda_k - \lambda_{k-1}\|^2). \]
According to Assumption 2 and the fact that $\rho \geq \rho_0$, (4) holds. Therefore, for any $k \geq 1$,
\[
\sum_{j=1}^{k} P_j = \sum_{j=1}^{k} \{ f(x_j) + \frac{\rho}{2} \sum_{i=1}^{m} \|c_i(x_j)\|^2 + \frac{\rho}{2} \|x_j - x_{j-1}\|^2 + \lambda_j^T c(x_j) \}
= \sum_{j=1}^{k} \{ f(x_j) + \frac{\rho}{2} \sum_{i=1}^{m} \|c_i(x_j)\|^2 + \frac{\rho}{2} \|x_j - x_{j-1}\|^2 \} + \sum_{j=1}^{k} \lambda_j^T c(x_j)
\geq \sum_{j=1}^{k} \bar{L} + \frac{1}{2\rho} \sum_{j=1}^{k} (\|\lambda_j\|^2 - \|\lambda_{j-1}\|^2 + \|\lambda_j - \lambda_{j-1}\|^2)
\geq \sum_{j=1}^{k} \bar{L} + \frac{1}{2\rho} \sum_{j=1}^{k} (\|\lambda_j\|^2 - \|\lambda_{j-1}\|^2)
= \sum_{j=1}^{k} \bar{L} + \frac{1}{2\rho} (\|\lambda_k\|^2 - \|\lambda_0\|^2) \geq \sum_{j=1}^{k} \bar{L} - \frac{1}{2\rho} \|\lambda_0\|^2 \Rightarrow \sum_{j=1}^{k} (P_j - \bar{L}) \geq \frac{1}{2\rho} \|\lambda_0\|^2.
\]
Note that the above inequality holds for all $k \geq 1$. Thus, nonincreasing property of $\{P_k - \bar{L}\}_{k \geq 1}$ (because $c_1, c_2 > 0$ and Lemma 3 indicates its nonnegativity, that is, $P_k \geq \bar{L}, \forall k \geq 1$.

Proof of Lemma 9

Proof. The first-order optimality condition (24) for Step 1 implies that for all $k \geq 0$, we have
\[
\nabla f(x_{k+1}) + \nabla c(x_{k+1})\lambda_k + \rho \nabla c(x_{k+1})c(x_{k+1}) + \beta(x_{k+1} - x_k) = \tilde{r}_{k+1}.
\]
Likewise, by replacing $k$ with $k - 1$, we obtain
\[
\nabla f(x_k) + \nabla c(x_k)\lambda_k + \beta(x_k - x_{k-1}) = \tilde{r}_k.
\]
By combining (38) and (39) and using the notation $\Delta \lambda_{k+1} \triangleq \lambda_{k+1} - \lambda_k, \Delta x_{k+1} \triangleq x_{k+1} - x_k$ and $\Delta \tilde{r}_{k+1} \triangleq \tilde{r}_{k+1} - \tilde{r}_k$, we have
\[
\nabla f(x_{k+1}) - \nabla f(x_k) + \nabla c(x_{k+1})\Delta \lambda_{k+1} + \nabla c(x_{k+1}) - \nabla c(x_k))\lambda_k + \beta(\Delta x_{k+1} - \Delta x_k) = \Delta \tilde{r}_{k+1},
\]
which by rearrangement gives
\[
\nabla c(x_{k+1})\Delta \lambda_{k+1}
= - (\nabla f(x_{k+1}) - \nabla f(x_k) + \nabla c(x_{k+1}) - \nabla c(x_k))\lambda_k + \beta(\Delta x_{k+1} - \Delta x_k) - \Delta \tilde{r}_{k+1}).
\]
Since $\sigma$ is a lower bound on the smallest singular value of $\nabla c(x_{k+1})$, we have
\[
\|\Delta \lambda_{k+1}\| \leq \frac{1}{\sigma} (\|\nabla f(x_{k+1}) - \nabla f(x_k)\| + \|\nabla c(x_{k+1}) - \nabla c(x_k)\|\|\lambda_k\| + \beta(\|\Delta x_{k+1}\| + \|\Delta x_k\|) + \|\Delta \tilde{r}_{k+1}\|).
\]
we have from (39) that
\[
\nabla c(x_k)\lambda_k = -\nabla f(x_k) - \beta(x_k - x_{k-1}) + \tilde{r}_k,
\]
so that
\[
\|\lambda_k\| \leq \frac{1}{\sigma} (\|\nabla f(x_k)\| + \beta\|\Delta x_k\| + \|\tilde{r}_k\|) \leq \frac{1}{\sigma} (\|\nabla f(x_k)\| + \beta\|\Delta x_k\| + \|\tilde{r}_k\|).
\]
We also have
\[ \| \nabla c(x_{k+1}) - \nabla c(x_k) \|_2 \leq L_c \| x_{k+1} - x_k \|. \] (43)

By substituting (3), (42), and (43) into (41), we obtain
\[
\begin{align*}
\| \Delta \lambda_{k+1} \| & \leq \frac{1}{\sigma} \left( L_f \| \Delta x_{k+1} \| + \beta \| \Delta x_{k+1} \| + \beta \| \Delta x_k \| + \frac{1}{\sigma} \| \hat{r}_k \| + \| \Delta \hat{r}_{k+1} \| \right) \\
& + \| \nabla c(x_{k+1}) - \nabla c(x_k) \|_2 \left( \frac{1}{\sigma} M_f + \frac{\beta}{\sigma} \| \Delta x_k \| + \frac{1}{\sigma} \| \hat{r}_k \| + \| \Delta \hat{r}_{k+1} \| \right) \\
& \leq \frac{1}{\sigma} \left( L_f \| \Delta x_{k+1} \| + \beta \| \Delta x_{k+1} \| + \frac{L_c M_f}{\sigma} \| \Delta x_{k+1} \| + \frac{2 M_c \beta}{\sigma} \| \Delta x_k \| \\
& + \frac{2 M_c \beta}{\sigma} \| \hat{r}_k \| + \| \Delta \hat{r}_{k+1} \| \right) \\
& \leq \frac{1}{\sigma} \left( L_f + \frac{L_c M_f}{\sigma} + \beta \right) \| \Delta x_{k+1} \| + \frac{1}{\sigma} \left( \beta + \frac{2 M_c \beta}{\sigma} \right) \| \Delta x_k \| + \frac{2 M_c \beta}{\sigma^2} \| \hat{r}_k \| + \frac{1}{\sigma} \| \Delta \hat{r}_{k+1} \|. 
\end{align*}
\]

By using the bound \((a + b + c + d)^2 \leq 4(a^2 + b^2 + c^2 + d^2)\) for positive scalars \(a, b, c, d\), and using the definition (10), we obtain the result. \(\blacksquare\)

**Proof of Lemma 11**

**Proof.** Since \(\rho \geq \rho_0\), according to Assumption 2 we have that \(\inf_{x \in \mathbb{R}^n} \{ f(x) + \frac{\rho}{2} \| c(x) \|^2 \} \geq \bar{L}\). By an argument similar to the proof of Lemma 4, we have that \(\sum_{i=1}^{k}(P_i - \bar{L}) \geq -\frac{1}{2\rho} \| \lambda_0 \|^2\), for any \(k \geq 1\). We prove the claim that \(P_k \geq \bar{L} - \frac{16(M_2^2 + \sigma^2)\rho}{\rho_0^4}\) for any \(k \geq 1\), by contradiction. Otherwise, assume that \(\exists K \geq 1\) such that \(P_K = \bar{L} - \frac{16(M_2^2 + \sigma^2)\rho}{\rho_0^4} - \delta\) for some \(\delta > 0\). According to Lemma 10, we have for any \(k \geq 1\) that
\[
P_{k+1} - P_k \leq -\tilde{c}_1 \| x_{k+1} - x_k \|^2 - \tilde{c}_2 \| x_k - x_{k-1} \|^2 + \frac{16 M_2^2}{\rho_0^4} \| \tilde{r}_k \|^2 + \frac{4}{\rho_0^2} \| \tilde{r}_{k+1} - \hat{r}_k \|^2 \\
\leq \frac{16 M_2^2 + 8 \sigma^2}{\rho_0^4} \| \tilde{r}_k \|^2 + \frac{8}{\rho_0^2} \| \tilde{r}_{k+1} \|^2.
\]

Then for any \(k \geq K + 1\), we have
\[
P_k \leq P_K + \frac{16 M_2^2 + 8 \sigma^2}{\rho_0^4} \sum_{i=K}^{k-1} \| \tilde{r}_i \|^2 + \frac{8}{\rho_0^2} \sum_{i=K}^{k-1} \| \tilde{r}_{i+1} \|^2 \leq P_K + \frac{16(M_2^2 + \sigma^2)}{\rho_0^4} \sum_{i=1}^{\infty} \| \tilde{r}_i \|^2 \\
\leq P_K + \frac{16(M_2^2 + \sigma^2)\rho}{\rho_0^4} = \bar{L} - \delta,
\]
so that \(P_k - \bar{L} \leq -\delta\) for all \(k \geq K + 1\). Thus, \(\sum_{i=1}^{k}(P_i - \bar{L}) \to -\infty\) as \(k \to \infty\), a contradiction. \(\blacksquare\)

**Proof of Theorem 12**

**Proof.** Define \(C_{\eta}^1\) as in (15), and set
\[
C \triangleq 7 f(x_0) + 9 \| \lambda_0 \|^2 - 7 \bar{L} + \frac{(M_2^2 + \sigma^2)\rho \eta}{\sigma^4 C_{\eta}^1}, \quad \Delta \triangleq C \max \{ 16, 1/(16C_{\eta}^1) \}.
\]

We want to show that \(T_{\epsilon} \leq [\Delta/\epsilon^2 - \eta] + 1\). First, let us check the positivity of \(\hat{c}_1\) and \(\hat{c}_2\), given the parameter assignments:
\[
\hat{c}_1 = \frac{\beta - \gamma}{2} - \frac{2 C_1}{\rho} \geq \frac{\epsilon^\eta}{8} - \frac{\epsilon^\eta}{16} = \frac{\epsilon^\eta}{16} > 0, \quad \hat{c}_2 = \frac{\gamma}{2} - \frac{2 C_2}{\rho} \geq \frac{\epsilon^\eta}{16} > 0.
\]
(44)
By Lemma [10], we have for any $k \geq 1$ that
\[
\begin{align*}
P_{k+1} - P_k &\leq -\hat{c}_1 \|x_{k+1} - x_k\|^2 - \hat{c}_2 \|x_k - x_{k-1}\|^2 + \frac{16M_c^2}{\rho \sigma^4} \|\tilde{r}_k\|^2 + \frac{4}{\rho \sigma^2} \|\tilde{r}_{k+1} - \tilde{r}_k\|^2 \\
&\leq -\hat{c}_1 \|x_{k+1} - x_k\|^2 - \hat{c}_2 \|x_k - x_{k-1}\|^2 + \frac{16M_c^2 + 8\sigma^2}{\rho \sigma^4} \|\tilde{r}_k\|^2 + \frac{8}{\rho \sigma^2} \|\tilde{r}_{k+1}\|^2.
\end{align*}
\]

Therefore, for any $k \geq 1$, we have
\[
\begin{align*}
\sum_{i=1}^{k} \left[ \hat{c}_1 \|x_{i+1} - x_i\|^2 + \hat{c}_2 \|x_i - x_{i-1}\|^2 \right] \\
\leq P_1 - P_{k+1} + \frac{16M_c^2 + 8\sigma^2}{\rho \sigma^4} \sum_{i=1}^{k} \|\tilde{r}_i\|^2 + \frac{8}{\rho \sigma^2} \sum_{i=1}^{k} \|\tilde{r}_{i+1}\|^2 \\
&\leq P_1 - P_{k+1} + \frac{16(M_c^2 + \sigma^2)}{\rho \sigma^4} \sum_{i=1}^{\infty} \|\tilde{r}_i\|^2 \leq P_1 - P_{k+1} + \frac{16(M_c^2 + \sigma^2)R}{\rho \sigma^4} \\
\overset{(\text{Lemma }[11])}{\leq} P_1 - \left( L - \frac{16(M_c^2 + \sigma^2)R}{\rho \sigma^4} \right) + \frac{16(M_c^2 + \sigma^2)R}{\rho \sigma^4} = P_1 - L + \frac{32(M_c^2 + \sigma^2)R}{\rho \sigma^4} \\
&\leq P_1 - L + \frac{32(M_c^2 + \sigma^2)R}{\rho \sigma^4} \frac{1}{\sigma^4(32 \max\{C_1, C_2\}/\epsilon^3)} \\
&\leq P_1 - L + \frac{(M_c^2 + \sigma^2)R \epsilon^3}{\sigma^4 C_1^0}. \tag{45}
\end{align*}
\]

By analysis similar to the proof of Theorem [5], we have
\[
P_1 - L \leq 7f(x_0) + 9\|\lambda_0\|^2 - 7L. \tag{46}
\]

By combining (45) with (46), we obtain
\[
\begin{align*}
\sum_{i=1}^{k} \left[ \hat{c}_1 \|x_{i+1} - x_i\|^2 + \hat{c}_2 \|x_i - x_{i-1}\|^2 \right] &\leq 7f(x_0) + 9\|\lambda_0\|^2 - 7L + \frac{(M_c^2 + \sigma^2)R \epsilon^3}{\sigma^4 C_1^0} = C. \tag{47}
\end{align*}
\]

Let $K \triangleq \lceil \Delta/\epsilon^2 \rceil$, and note that (47) holds for $k = K$, we have that there exists $k^* \in [1, K]$ such that
\[
\hat{c}_1 \|x_{k^*+1} - x_{k^*}\|^2 + \hat{c}_2 \|x_{k^*} - x_{k^*-1}\|^2 \leq C/K. \tag{48}
\]

Thus, we have
\[
\begin{align*}
\|\nabla L_0(x_{k^*+1}, \lambda_{k^*+1})\| &= \|\nabla L_0(x_{k^*+1}, \lambda_{k^*})\| \leq \beta(x_{k^*+1} - x_{k^*}) + \tilde{r}_{k^*+1} \\
&\leq \beta \|x_{k^*+1} - x_{k^*}\| + \|\tilde{r}_{k^*+1}\| \leq \beta \sqrt{\|x_{k^*+1} - x_{k^*}\|^2 + \epsilon^2/2} \\
&\overset{(49)}{\leq} \beta \sqrt{\frac{C/\epsilon^2}{K} + \frac{\epsilon^2}{2}} \leq \frac{\epsilon^2}{2} \sqrt{\frac{C}{K} + \frac{16C}{16C \epsilon^2}} + \frac{\epsilon^2}{2} = \epsilon.
\end{align*}
\]

For the constraint norm, we have
\[
\begin{align*}
\|\epsilon(x_{k^*+1})\|^2 &= \|\lambda_{k^*+1} - \lambda_{k^*}\|^2 / \rho^2 \\
&\overset{(45)}{\leq} \frac{2C_1}{\rho^2} \|x_{k^*+1} - x_{k^*}\|^2 + \frac{2C_2}{\rho^2} \|x_{k^*} - x_{k^*-1}\|^2 + \frac{16M_c^2}{\rho^2 \sigma^4} \|\tilde{r}_{k^*}\|^2 + \frac{4}{\rho^2 \sigma^2} \|\tilde{r}_{k^*+1} - \tilde{r}_{k^*}\|^2 \\
&\leq \frac{2C_1}{\rho^2} \|x_{k^*+1} - x_{k^*}\|^2 + \frac{2C_2}{\rho^2} \|x_{k^*} - x_{k^*-1}\|^2 + \frac{16M_c^2 + 8\sigma^2}{\rho^2 \sigma^4} \|\tilde{r}_{k^*}\|^2 + \frac{8}{\rho^2 \sigma^2} \|\tilde{r}_{k^*+1}\|^2 \\
&\leq \frac{2C_1}{\rho^2} \|x_{k^*+1} - x_{k^*}\|^2 + \frac{2C_2}{\rho^2} \|x_{k^*} - x_{k^*-1}\|^2 + \frac{16(M_c^2 + \sigma^2)}{\rho^2 \sigma^4} \epsilon^2/4.
\end{align*}
\]
Therefore, \( T \) follows from Theorem 12.

This fact indicates that \( \frac{\partial^2}{\partial x^2} \epsilon \left( x_{k^*+1} \right) \), completing the proof.

\( \text{Proof of Corollary 2} \)

\( \text{Proof.} \) Since \( \beta = \epsilon^a / 2 \leq \epsilon / 2 \) and \( \epsilon_{k+1} = \epsilon / 2 \), for any \( k \geq 0 \), we have from (29) that

\[ \nabla^2 \mathcal{L}(x_{k+1}, \lambda_k) \geq -\beta + \epsilon_{k+1} \mathbb{I} \geq -\epsilon \mathbb{I}. \]

This fact indicates that

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

which implies that

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

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

\[ \tilde{T}_\epsilon = \inf \{ t \geq 1 \mid \exists \lambda \in \mathbb{R}^m, \| \nabla f(x_t) + \nabla c(x_t) \lambda \| \leq \epsilon, \| c(x_t) \| \leq \epsilon \}, \]

\[ \inf \{ t \geq 1 \mid \| \nabla f(x_t) + \nabla c(x_t) \lambda_t \| \leq \epsilon, \| c(x_t) \| \leq \epsilon \}, \]

The result now follows from Theorem 12. \( \blacksquare \)