GLOBAL CONVERGENCE RATES OF AUGMENTED LAGRANGIAN METHODS FOR CONSTRAINED CONVEX PROGRAMMING

YANGYANG XU

Abstract. Augmented Lagrangian method (ALM) has been popularly used for solving constrained optimization problems. Its convergence and local convergence speed have been extensively studied. However, its global convergence rate is still open for problems with nonlinear inequality constraints. In this paper, we work on general constrained convex programs. For these problems, we establish the global convergence rate of ALM and its inexact variants.

We first assume exact solution to each subproblem in the ALM framework and establish an $O(1/k)$ ergodic convergence result, where $k$ is the number of iterations. Then we analyze an inexact ALM that approximately solves the subproblems. Assuming summable errors, we prove that the inexact ALM also enjoys $O(1/k)$ convergence if smaller stepizes are used in the multiplier updates. Furthermore, we apply the inexact ALM to a constrained composite convex problem with each subproblem solved by Nesterov’s optimal first-order method. We show that $O(\varepsilon^{-3/2} - \delta)$ gradient evaluations are sufficient to guarantee an $\varepsilon$-optimal solution in terms of both primal objective and feasibility violation, where $\delta$ is an arbitrary positive number. Finally, for constrained smooth problems, we modify the inexact ALM by adding a proximal term to each subproblem and improve the iteration complexity to $O(\varepsilon^{-1} |\log \varepsilon|)$.

Keywords: augmented Lagrangian method (ALM), nonlinearly constrained problem, first-order method, global convergence rate, iteration complexity

Mathematics Subject Classification: 90C06, 90C25, 68W40, 49M27.

1. Introduction. In this paper, we consider the constrained convex programming

$$
\min_{x \in X} f_0(x), \text{ s.t. } Ax = b, f_i(x) \leq 0, \quad i = 1, \ldots, m,
$$

where $X$ is a closed convex set, and $f_i$ is a convex function for every $i = 0, 1, \ldots, m$. Any convex optimization problem can be written in the standard form of (1.1).

Note that the constraint $x \in X$ can be equivalently represented by using an inequality constraint $i_X(x) \leq 0$ or adding $i_X(x)$ to the objective, where $i_X$ denotes the indicator function on $X$. However, we explicitly use it for technical reason. In addition, every affine constraint $a_j^T x = b_j$ can be equivalently represented by two inequality constraints: $a_j^T x - b_j \leq 0$ and $-a_j^T x + b_j \leq 0$. That way does not change theoretical results of an algorithm but will make the problem computationally more difficult.

Problems formulated in the form of (1.1) appear in many areas including statistics, machine learning, data mining, engineering, signal processing, finance, operations research, and so on.

One popular method for solving (1.1) is the augmented Lagrangian method (ALM), which first appeared in [14, 25]. Its global convergence and local convergence rate have been extensively studied; see the books [4, 5]. Several recent works (e.g., [12, 13]) establish the global convergence rate of ALM and/or its variants for affinely constrained
problems. However, to the best of our knowledge, its global convergence rate for problems with nonlinear inequality constraints still remains open\footnote{Although the global convergence rate in terms of augmented dual objective can be easily shown from existing works, that does not indicate the convergence speed from the perspective of the primal objective and feasibility.}. We will address this open question and also analyze iteration complexity of its inexact variants.

1.1. Augmented Lagrangian function. In the literature, there are several different penalty terms used in an augmented Lagrangian function, such as the classic one \cite{26,27}, the quadratic penalty on constraint violation \cite{3}, and the exponential penalty \cite{30}. The work \cite{2} gives a general class of augmented penalty functions that satisfy certain properties. In this paper, we use the classic one. As discussed below, it can be derived from a quadratic penalty on an equivalent equality constrained problem.

Introducing nonnegative slack variable $s_i$’s, one can write (1.1) equivalently to

\begin{equation}
\min_{x \in X, s \geq 0} f_0(x), \text{ s.t. } Ax = b, f_i(x) + s_i = 0, i = 1, \ldots, m.
\end{equation}

With quadratic penalty on the equality constraints, the augmented Lagrangian function of (1.2) is

\begin{equation}
\tilde{L}_\beta(x, y, z) = f_0(x) + y^T (Ax - b) + \sum_{i=1}^{m} z_i (f_i(x) + s_i) + \frac{\beta}{2} \|Ax - b\|^2 + \frac{\beta}{2} \sum_{i=1}^{m} (f_i(x) + s_i)^2,
\end{equation}

where $y$ and $z$ are multipliers, and $\beta$ is the augmented penalty parameter. Minimizing $\tilde{L}_\beta$ with respect to $s \geq 0$ while fixing $x, y$ and $z$, we have the optimal $s$ given by

\begin{equation}
s_i = \left[ -\frac{z_i}{\beta} - f_i(x) \right]_+, i = 1, \ldots, m.
\end{equation}

Plugging the above $s$ into $\tilde{L}_\beta$ gives

\begin{equation}
\tilde{L}_\beta(x, y, z) = f_0(x) + y^T (Ax - b) + \frac{\beta}{2} \|Ax - b\|^2 + \sum_{i=1}^{m} \psi_\beta(f_i(x), z_i),
\end{equation}

where

\begin{equation}
\psi_\beta(u, v) = \begin{cases} 
uv + \frac{\beta}{2}u^2, & \text{if } \beta u + v \geq 0, \\
-\frac{v^2}{2\beta}, & \text{if } \beta u + v < 0.
\end{cases}
\end{equation}

Let

\begin{equation}
\Psi_\beta(x, z) = \sum_{i=1}^{m} \psi_\beta(f_i(x), z_i),
\end{equation}

and we obtain the classical augmented Lagrangian function of (1.1):

\begin{equation}
L_\beta(x, y, z) = f_0(x) + y^T (Ax - b) + \frac{\beta}{2} \|Ax - b\|^2 + \Psi_\beta(x, z).
\end{equation}
It is shown in [26] that the augmented dual function\(^2\)
\[
d_\beta(y, z) = \min_{x \in X} \mathcal{L}_\beta(x, y, z)
\]
is continuously differentiable, and \(\nabla d_\beta\) is Lipschitz continuous with constant \(\frac{1}{\beta}\). In addition, it turns out that the (inexact) ALM is an (inexact) augmented dual gradient ascent [27], and thus convergence rate of the (inexact) ALM can be shown from analysis on (inexact) gradient method [29]. Our analysis will be different from this line, and the assumptions we make are weaker.

1.2. Contributions. The contributions of this paper are mainly on establishing global convergence rate results of ALM and its inexact variants for solving (1.1) and listed as follows.

- We establish the global convergence of ALM (see Algorithm 2) that employs the augmented Lagrangian function in (1.3). We show that the first-order optimality conditions hold asymptotically and the multiplier sequence converges to an optimal solution of the dual problem. In addition, \(O(1/k)\) global convergence rate is established in terms of both primal objective and feasibility violation, where \(k\) is the total number of iterations.
- We also analyze an inexact ALM (see Algorithm 3) that approximately solves every \(x\)-subproblem to a specified error tolerance. If the errors are summable, the inexact ALM is guaranteed to converge in terms of first-order optimality condition by using smaller stepsizes for the multiplier updates. In addition, if the accumulated error grows slower than \(k\), we establish a sublinear convergence rate of the inexact ALM in terms of both primal objective and feasibility violation.
- We then apply the inexact ALM to a constrained composite convex program and also modify it to solve a constrained smooth convex program. Exploring the structure of the problem, we use Nesterov’s optimal first-order method (see Algorithm 1) to find an approximate solution of each \(x\)-subproblem. We show that to reach an \(\varepsilon\)-optimal solution (see Definition 1.1), we only need evaluate gradients of \(f_i\)'s for \(O(\varepsilon^{-\frac{3}{2}-\delta})\) times on solving the constrained composite convex program and \(O(\varepsilon^{-1}|\log \varepsilon|)\) times on the constrained smooth convex program, where \(\delta\) is an arbitrarily small positive number.

1.3. Notation. For simplicity, throughout the paper, we focus on a finite-dimensional Euclidean space, but our analysis can be directly extended to a general Hilbert space.

We denote \([m]\) as the set \(\{1, 2, \ldots, m\}\) for any positive integer \(m\). Given a real number \(a\), we let \([a]_+ = \max(0, a)\) and \([a]_- = \min(0, a)\). By \(p(k) \lesssim q(k)\), we mean \(\limsup_k \frac{p(k)}{q(k)} \leq 1\), i.e., \(p(k)\) is dominated by \(q(k)\). Given a symmetric positive definite (SPD) matrix \(P\), we denote \(\|\cdot\|_P\) as \(P\)-weighted norm, i.e., \(\|w\|_P = \sqrt{w^\top P w}\). If \(P = I\), we simply write it as \(\|\cdot\|\).

Given a convex function \(f\), \(\nabla f(x)\) represents one subgradient of \(f\) at \(x\), namely,
\[
f(\hat{x}) \geq f(x) + \langle \nabla f(x), \hat{x} - x \rangle, \quad \forall \hat{x},
\]

\(^2\)Although [26] only considers the inequality constrained case, the results derived there apply to the case with both equality and inequality constraints.
and \( \partial f(x) \) denotes its subdifferential, i.e., the set of all subgradients. When \( f \) is differentiable, we simply write its subgradient as \( \nabla f(x) \). For a convex set \( \mathcal{X} \), we use \( \iota_{\mathcal{X}} \) as its indicator function, i.e.,
\[
\iota_{\mathcal{X}}(x) = \begin{cases} 
0, & \text{if } x \in \mathcal{X}; \\
+\infty, & \text{if } x \notin \mathcal{X}, 
\end{cases}
\]
and \( \mathcal{N}_{\mathcal{X}}(x) = \partial \iota_{\mathcal{X}}(x) \) as its normal cone at \( x \in \mathcal{X} \).

**Definition 1.1 (\( \varepsilon \)-optimal solution).** Let \( f^* \) be the optimal value of (1.1). Given \( \varepsilon \geq 0 \), the point \( x \in \mathcal{X} \) is called an \( \varepsilon \)-optimal solution to (1.1) if
\[
|f_0(x) - f^*| \leq \varepsilon \quad \text{and also} \quad \|Ax - b\| + \sum_{i=1}^{m} |f_i(x)| \leq \varepsilon.
\]

**1.4. Outline.** The rest of the paper is organized as follows. In section 2, we give a few preparatory results and review Nesterov’s optimal first-order method for solving a composite convex program. The convergence result and rate estimate of the ALM are shown in section 3. In section 4, we first analyze an inexact ALM for general convex optimization problems and then apply it to constrained composite convex programs and its modified version to constrained smooth convex problems. Iteration complexity in terms of the number of gradient evaluations is shown. Related works are reviewed and compared in section 5, and finally section 6 concludes the paper.

**2. Preliminary results and accelerated proximal gradient.** In this section, we give a few preliminary results and also review Nesterov’s optimal first-order method for composite convex programs.

**2.1. Basic facts.** A point \((x, y, z)\) satisfies the Karush-Kuhn-Tucker (KKT) conditions for (1.1) if
\[
\begin{align*}
0 & \in \partial f_0(x) + \mathcal{N}_{\mathcal{X}}(x) + A^\top y + \sum_{i=1}^{m} z_i \partial f_i(x), \\
Ax & = b, \quad x \in \mathcal{X}, \\
z_i & \geq 0, \quad f_i(x) \leq 0, \quad z_i f_i(x) = 0, \forall i \in [m].
\end{align*}
\]

From the convexity of \( f_i \)’s, if \((x^*, y^*, z^*)\) is a KKT point, then
\[
(f_0(x) - f_0(x^*) + \langle y^*, Ax - b \rangle + \sum_{i=1}^{m} z_i^* f_i(x) \geq 0, \forall x \in \mathcal{X}.
\]

The following result is well-known (c.f., [8, Prop. 2.3]). It will be used for showing iterate convergence of the analyzed algorithms.

**Lemma 2.1.** Assume that \( \{\eta_k\} \) is a nonnegative sequence and \( \sum_{k=1}^{\infty} \eta_k < \infty \). Let \( P \) be an SPD matrix and \( W \) a nonempty set. If the sequence \( \{w^k\} \) satisfies
\[
\|w^{k+1} - w\|_P^2 \leq \|w^k - w\|_P^2 + \eta_k, \forall w \in W,
\]
and \( \{w^k\} \) has a cluster point \( \bar{w} \) in \( W \), then \( w^k \) converges to \( \bar{w} \).

The result below will be used to establish convergence rate of the algorithms.
Lemma 2.2. Assume \((x^*,y^*,z^*)\) is a KKT point of (1.1). Let \(\bar{x}\) be a point such that for any \(y\) and any \(z \geq 0\),

\[
f_0(\bar{x}) - f_0(x^*) + y^\top (A\bar{x} - b) + \sum_{i=1}^{m} z_i f_i(\bar{x}) \leq \alpha + c_1 \|y\|^2 + c_2 \|z\|^2,
\]

where \(\alpha\) and \(c_1, c_2\) are nonnegative constants independent of \(y\) and \(z\). Then

\[
-\left(\alpha + 4c_1 \|y^*\|^2 + 4c_2 \sum_{i=1}^{m} (z_i^*)^2\right) \leq f_0(\bar{x}) - f_0(x^*) \leq \alpha,
\]

\[
\|A\bar{x} - b\| + \sum_{i=1}^{m} |f_i(\bar{x})|_+ \leq \alpha + c_1 \|y^*\|^2 + c_2 \sum_{i=1}^{m} (1 + z_i^*)^2.
\]

Proof. Letting \(y = 0\) and \(z = 0\) in (2.3) gives the second inequality in (2.4). For any nonnegative \(\gamma_y\) and \(\gamma_i, i \in [m]\), we let

\[
y = \gamma_y \frac{A\bar{x} - b}{\|A\bar{x} - b\|}, \quad z_i = \gamma_i \frac{|f_i(\bar{x})|_+}{|f_i(\bar{x})|}, \forall i \in [m]
\]

and have from (2.3) by using the convention \(\frac{0}{0} = 0\) that

\[
f_0(\bar{x}) - f_0(x^*) + \gamma_y \|A\bar{x} - b\| + \sum_{i=1}^{m} \gamma_i |f_i(\bar{x})|_+ \leq \alpha + c_1 \gamma_y^2 + c_2 \sum_{i=1}^{m} \gamma_i^2.
\]

Noting

\[
-\langle y^*, A\bar{x} - b \rangle \geq -\|y^*\| \cdot \|A\bar{x} - b\|, \quad -z_i^* f_i(\bar{x}) \geq -z_i^* |f_i(\bar{x})|_+, \forall i \in [m],
\]

we have from (2.2) and (2.6) that

\[
(\gamma_y - \|y^*\|) \|A\bar{x} - b\| + \sum_{i=1}^{m} (\gamma_i - z_i^*) |f_i(\bar{x})|_+ \leq \alpha + c_1 \gamma_y^2 + c_2 \sum_{i=1}^{m} \gamma_i^2.
\]

In the above inequality, letting \(\gamma_y = 1 + \|y^*\|\) and \(\gamma_i = 1 + z_i^*, \forall i \in [m]\) gives (2.5), and letting \(\gamma_y = 2\|y^*\|\) and \(\gamma_i = 2z_i^*, \forall i \in [m]\) gives the first inequality in (2.4) by (2.2) and (2.7). \(\square\)

2.2. Nesterov’s optimal first-order method. In this subsection, we review Nesterov’s optimal first-order method for composite convex programs. The method will be used to approximately solve subproblems in our inexact ALMs. It aims at finding a solution of the following problem

\[
\min_x \phi(x) + \psi(x),
\]

where \(\phi\) is a Lipschitz differentiable and strongly convex function with gradient Lipschitz constant \(L_\phi\) and strong convexity modulus \(\mu \geq 0\), and \(\psi\) is a simple (possibly nondifferentiable) closed convex function. Algorithm 1 summarizes the method. Here, for simplicity, we assume \(L_\phi\) and \(\mu\) are known. The method does not require the value
Algorithm 1: Nesterov’s optimal first-order method for (2.8)

1. Initialization: choose \(\hat{x}^0 = x^0\), \(α_0 \in (0, 1]\), and let \(q = \frac{\sqrt{2}}{L_φ}\).

2. for \(k = 0, 1, \ldots\), do

3. Let

\[
x^{k+1} = \text{arg min}_x \langle \nabla \phi(\hat{x}^k), x \rangle + \frac{L_φ}{2} \| x - \hat{x}^k \|^2 + ψ(x).
\]

4. Set

\[
α_{k+1} = \frac{q - α_k^2 + \sqrt{(q - α_k^2)^2 + 4α_k^2}}{2},
\]

and

\[
\hat{x}^{k+1} = x^{k+1} + \frac{α_k(1 - α_k)}{α_k^2 + α_{k+1}}(x^{k+1} - x^k).
\]

of \(L_φ\) but can estimate it by backtracking. In addition, it only requires a lower estimate of \(μ\). For our purpose, we will have \(μ = 1\); see the problem (4.16) in Algorithm 4.

The theorem below gives the convergence rate of Algorithm 1 for both convex (i.e., \(μ = 0\)) and strongly convex (i.e., \(μ > 0\)) cases; see [1, 24]. We will use the results to estimate iteration complexity of inexact ALMs.

**Theorem 2.3.** Let \(\{x^k\}\) be the sequence generated from Algorithm 1. Assume \(x^*\) to be a minimizer of (2.8). The following results holds:

1. If \(μ = 0\) and \(α_0 = 1\), then

\[
φ(x^k) + ψ(x^k) - φ(x^*) - ψ(x^*) \leq \frac{2L_φ\|x^0 - x^*\|^2}{k^2}, \forall k ≥ 1.
\]

2. If \(μ > 0\), \(α_0 = \sqrt{\frac{μ}{L_φ}}\), and \(ψ = i_X\) for a convex set \(X\), then

\[
φ(x^k) - φ(x^*) \leq \frac{(L_φ + μ)\|x^0 - x^*\|^2}{2} \left(1 - \sqrt{\frac{μ}{L_φ}}\right), \forall k ≥ 1.
\]

3. Augmented Lagrangian method. In this section, we analyze the global convergence rate of an ALM based on \(L_β(x, y, z)\) given in (1.3). The method is summarized in Algorithm 2. It is more general than that in [26], which does not explicitly include the equality constraint and simply set \(ρ_z = β\). Although Algorithm 2 can be regarded as a special case of the inexact ALM in Algorithm 3 by setting \(ε_k = 0, \forall k\), the analysis on the exact ALM brings us inspirations on analyzing the inexact one, and in addition, stronger convergence results can be shown for the exact ALM.

3.1. Technical assumptions. Throughout this section, we make the following assumptions.
Algorithm 2: Augmented Lagrangian method for (1.1)

1 Initialization: choose $x^0, y^0, z^0$ and $\beta, \rho_y, \rho_z$
2 for $k = 0, 1, \ldots$ do
3 Perform the updates
   (3.1a) $x^{k+1} = \arg \min_{x \in X} L_\beta(x, y^k, z^k),$
   (3.1b) $y^{k+1} = y^k + \rho_y (Ax^{k+1} - b),$
   (3.1c) $z_{i}^{k+1} = z_i^k + \rho_z \cdot \max \left( -\frac{z_i^k}{\beta}, f_i(x^{k+1}) \right), i = 1, \ldots, m.$

Assumption 1. There exists a point $(x^*, y^*, z^*)$ satisfying the KKT conditions in (2.1).

Assumption 2. For any $k \geq 0$, the problem (3.1a) has at least one solution.

The first assumption holds if a certain regularity condition is satisfied, such as the Slater condition (namely, there is an interior point $x$ of $X$ such that $Ax = b$ and $f_i(x) < 0, \forall i \in [m]$). The second assumption is for the well-definedness of the algorithm. It holds if $X$ is compact and $f_i$’s are continuous.

3.2. Convergence analysis. To show the convergence results of Algorithm 2, we first establish a few lemmas.

Lemma 3.1. Let $y$ and $z$ be updated by (3.1b) and (3.1c) respectively. Then for any $k$, it holds

(3.2) $\frac{1}{2\rho_y} \left[ ||y^{k+1} - y||^2 - ||y^k - y||^2 + ||y^{k+1} - y^k||^2 \right] - \langle y^{k+1} - y, r^{k+1} \rangle = 0,$

(3.3) $\frac{1}{2\rho_z} \left[ ||z_i^{k+1} - z_i||^2 - ||z_i^k - z_i||^2 + ||z_i^{k+1} - z_i^k||^2 \right] - \sum_{i=1}^{m} (z_i^{k+1} - z_i) \cdot \max \left( -\frac{z_i^k}{\beta}, f_i(x^{k+1}) \right) = 0,$

where $r^k = Ax^k - b$.

Proof. Using the equality $2u^Tv = ||u||^2 - ||u - v||^2 + ||v||^2$, we have the results from the updates (3.1b) and (3.1c).

Lemma 3.2. For any $z \geq 0$, we have

(3.4) $\sum_{i=1}^{m} \left( (z_i^k + \beta f_i(x^{k+1}))_+ - z_i \right) f_i(x^{k+1}) - \sum_{i=1}^{m} (z_i^{k+1} - z_i) \cdot \max \left( -\frac{z_i^k}{\beta}, f_i(x^{k+1}) \right) \geq \frac{1}{\rho_z} (\beta - \rho_z) ||z^{k+1} - z^k||^2.$

Proof. Denote

(3.5) $I_+^k = \{ i \in [m] : z_i^k + \beta f_i(x^{k+1}) \geq 0 \}, \quad I_-^k = [m] \setminus I_+^k.$
Then

the left hand side of (3.4)

\[
= \sum_{i \in I_k^+} \left( (z_i^k - z_i) f_i(x^{k+1}) + \beta [f_i(x^{k+1})]^2 - (z_i^k + \rho_z f_i(x^{k+1}) - z_i) f_i(x^{k+1}) \right)
+ \sum_{i \in I_k^+} \left[ -z_i f_i(x^{k+1}) - (z_i^k - \rho_z z_i^k) (z_i - z_i^k) \right] = (\beta - \rho_z) \sum_{i \in I_k^+} [f_i(x^{k+1})]^2 + \sum_{i \in I_k^+} \left[ -z_i (f_i(x^{k+1}) + \frac{z_i^k}{\beta}) + \frac{1}{\beta^2} (\beta - \rho_z) (z_i^k)^2 \right] \\
\geq (\beta - \rho_z) \sum_{i \in I_k^+} [f_i(x^{k+1})]^2 + \frac{1}{\beta^2} (\beta - \rho_z) \sum_{i \in I_k^+} (z_i^k)^2
\]

\[= \frac{1}{\rho_z^2} (\beta - \rho_z) \| z^{k+1} - z^k \|^2, \]

where the inequality follows from \( z_i \geq 0 \) and \( f_i(x^{k+1}) + \frac{z_i^k}{\beta} \leq 0, \forall i \in I_k^+ \), and the last equality holds due to the update (3.1c).

\[\square\]

Using the above two lemmas, we establish a fundamental result on Algorithm 2.

**Theorem 3.3 (One-iteration progress of ALM).** Let \( \{x^k, y^k, z^k\} \) be the sequence generated from Algorithm 2. Then for any \( x \in \mathcal{X} \) such that \( Ax = b \) and \( f_i(x) \leq 0, \forall i \in [m] \), any \( y \), and any \( z \geq 0 \), it holds that

\[
f_0(x^{k+1}) - f_0(x) + y^\top r^{k+1} + \sum_{i=1}^{m} z_i f_i(x^{k+1}) + (\beta - \frac{\rho_y}{2}) \| r^{k+1} \|^2
\]

\[
+ \frac{1}{\rho_z^2} (\beta - \rho_z) \| z^{k+1} - z^k \|^2 + \frac{1}{2 \rho_y} \| y^{k+1} - y \|^2 + \frac{1}{2 \rho_z} \| z^{k+1} - z \|^2
\]

\[
\leq \frac{1}{2 \rho_y} \| y^k - y \|^2 + \frac{1}{2 \rho_z} \| z^k - z \|^2. \quad (3.6)
\]

**Proof.** From the optimality of \( x^{k+1} \), it follows that

\[
\langle x^{k+1} - x, \nabla_x \mathcal{L}(x^{k+1}, y^k, z^k) \rangle \leq 0, \forall x \in \mathcal{X},
\]

namely,

\[
\langle x^{k+1} - x, \nabla f_0(x^{k+1}) + A^\top y^k + \beta A^\top r^{k+1} + \sum_{i=1}^{m} [z_i^k + \beta f_i(x^{k+1})] + \nabla f_i(x^{k+1}) \rangle \leq 0, \forall x \in \mathcal{X}. \quad (3.7)
\]
From the convexity of $f_i, i = 0, 1, \ldots, m$, we have
\begin{equation*}
\begin{aligned}
&\left\langle x^{k+1} - x, \nabla f_0(x^{k+1}) + \sum_{i=1}^{m}[z_i^k + \beta f_i(x^{k+1})] + \nabla f_i(x^{k+1}) \right\rangle \\
\geq &\, f_0(x^{k+1}) - f_0(x) + \sum_{i=1}^{m}[z_i^k + \beta f_i(x^{k+1})] + (f_i(x^{k+1}) - f_i(x)) \\
= &\, f_0(x^{k+1}) - f_0(x) + \sum_{i=1}^{m}z_i f_i(x^{k+1}) + \sum_{i=1}^{m}[z_i^k + \beta f_i(x^{k+1})] + (f_i(x^{k+1}) - f_i(x)) - \sum_{i=1}^{m} z_i f_i(x^{k+1}).
\end{aligned}
\end{equation*}

In addition, for any $x$ such that $Ax = b$, it holds that
\begin{equation*}
\begin{aligned}
&\langle x^{k+1} - x, A^\top y^k + \beta A^\top r^{k+1} \rangle = \langle x^{k+1}, y^k \rangle + \beta \| r^{k+1} \|^2 \\
= &\, \langle x^{k+1}, y^{k+1} \rangle + (\beta - \rho_y) \| r^{k+1} \|^2 \\
= &\, y^\top r^{k+1} + \langle y^{k+1} - y, r^{k+1} \rangle + (\beta - \rho_y) \| r^{k+1} \|^2.
\end{aligned}
\end{equation*}

Plugging the above two equations into (3.7) and using Lemma 3.1, we have
\begin{equation*}
\begin{aligned}
f_0(x^{k+1}) - f_0(x) + \sum_{i=1}^{m} z_i f_i(x^{k+1}) + \sum_{i=1}^{m}[z_i^k + \beta f_i(x^{k+1})] + (f_i(x^{k+1}) - f_i(x)) - \sum_{i=1}^{m} z_i f_i(x^{k+1}) \\
+ y^\top r^{k+1} + (\beta - \rho_y) \| r^{k+1} \|^2 + \frac{1}{2\rho_y} \left[ \| y^{k+1} - y \|^2 - \| y^k - y \|^2 + \| y^{k+1} - y \|^2 \right] \\
+ \frac{1}{2\rho_z} \left[ \| z^{k+1} - z \|^2 - \| z^k - z \|^2 + \| z^{k+1} - z \|^2 \right] - \sum_{i=1}^{m} (z_i^{k+1} - z_i) \cdot \max \left( -\frac{z_i^k}{\beta}, f_i(x^{k+1}) \right)
\end{aligned}
\end{equation*}

(3.8)
\begin{equation*}
\leq 0.
\end{equation*}

Using Lemma 3.2 and also noting $[z_i^k + \beta f_i(x^{k+1})] + f_i(x) \leq 0, \forall i$ yield the desired result.

Now we are ready to state and show the convergence and rate results of Algorithm 2.

**Theorem 3.4** (Global convergence of ALM). Under Assumptions 1 and 2, let $\{(x^k, y^k, z^k)\}$ be the sequence generated from Algorithm 2 with $0 < \rho_y < 2\beta$ and $0 < \rho_z < 2\beta$. Then we have:

1. The sequence $\{x^k\} \subset X$ is a minimizing sequence, namely,
\begin{equation*}
\limsup_{k \to \infty} f_0(x^k) \leq f_0(x^\ast), \quad \lim_{k \to \infty} f_i(x^k) = 0, \forall i \in [m].
\end{equation*}

2. If $f_i(x)$ is lower bounded and $\partial f_i(x)$ bounded for any $x \in X$ and $i \in [m]$, then the sequence $\{(x^k, y^k, z^k)\}$ asymptotically satisfies the KKT conditions in (2.1), namely,
\begin{align*}
(3.10a) &\quad \lim_{k \to \infty} \text{dist}(0, \partial f_0(x^k) + \mathcal{N}_X(x^k) + A^\top y^k + \sum_{i=1}^{m} z_i^k \partial f_i(x^k)) = 0, \\
(3.10b) &\quad \lim_{k \to \infty} Ax^k - b = 0, \\
(3.10c) &\quad \lim_{k \to \infty} [z_i^k]_+ = 0, \quad \lim_{k \to \infty} [f_i(x^k)]_+ = 0, \quad \lim_{k \to \infty} z_i^k f_i(x^k) = 0, \forall i \in [m].
\end{align*}
3. In addition, if \( \{x^k\} \) has a cluster point \( \bar{x} \) and \( f_i(x) \) is continuous on \( \mathcal{X} \) for every \( i = 0, 1, \ldots, m \), then \( (y^k, z^k) \) converges to a point \((\bar{y}, \bar{z})\), and \((\bar{x}, \bar{y}, \bar{z})\) satisfies the KKT conditions in (2.1).

**Remark 3.1.** In general, we do not have the convergence of \( x^k \) to \( \bar{x} \). However, it can be guaranteed if we add a proximal term \( \frac{\mu}{2}\|x - x^k\|^2 \) to the subproblem in (3.1a) for any \( \mu > 0 \); see Algorithm 4 and Theorem 4.10.

**Proof.** Letting \((x, y, z) = (x^k, y^k, z^k)\) in (3.6) and using (2.2), we have

\[
(\beta - \frac{\rho_y}{2})\|y^{k+1}\|^2 + \frac{1}{\rho_z^2}(\beta - \frac{\rho_z}{2})\|z^{k+1} - z^k\|^2 + \frac{1}{2\rho_y}\|y^{k+1} - y^k\|^2 + \frac{1}{2\rho_z}\|z^{k+1} - z^k\|^2
\]

\[
\leq \frac{1}{2\rho_y}\|y^k - y^*\|^2 + \frac{1}{2\rho_z}\|z^k - z^*\|^2,
\]

which implies the boundedness of \( y^k \) and \( z^k \) since \( 0 < \rho_y < 2\beta \) and \( 0 < \rho_z < 2\beta \). Summing up (3.11) over \( k \) yields

\[
\lim_{k \to \infty} y^{k+1} - y^k = \lim_{k \to \infty} \rho_y r^{k+1} = 0, \quad \lim_{k \to \infty} z^{k+1} - z^k = 0.
\]

Hence, (3.10b) holds.

Note that if \( z^k_i < 0 \), then \( 0 \leq \frac{z^k_i}{\beta} \leq z^{k+1}_i - z^k_i \to 0 \). Hence, \( [z^k_i]_+ \to 0 \) as \( k \to \infty \).

Similarly, one can show \( f_i(x^{k+1}) \to 0 \).

Since \( z^k \) is bounded, there is \( Z > 0 \) such that \( |z^k| \leq Z, \forall i, k \). For any \( \varepsilon > 0 \), since \( z^{k+1}_i - z^k_i \to 0 \), there is a \( k \) such that

\[
\max \left( - \frac{z^k_i}{\beta}, f_i(x^{k+1}) \right) < \varepsilon, \forall k \geq \tilde{k}.
\]

For \( k \geq \tilde{k} \), when \( f_i(x^{k+1}) \geq -\frac{z^k_i}{\beta} \), it follows from (3.13) that \( |f_i(x^{k+1})| < \varepsilon \) and thus

\[
|z^{k+1}_i f_i(x^{k+1})| \leq |z^{k+1}_i| \varepsilon \leq Z \varepsilon,
\]

\[
||z^{k+1}_i + \beta f_i(x^{k+1})| - z^{k+1}_i| = |\beta - \rho_z| \cdot |f_i(x^{k+1})| \leq |\beta - \rho_z| \varepsilon.
\]

Hence, \( \limsup \Psi_\beta(x^{k+1}, z^k) \geq 0 \) from the arguments below

\[
\Psi_\beta(x^{k+1}, z^k) = \sum_{i \in I_\beta^k} z_i^k f_i(x^{k+1}) + \frac{\beta}{2} |f_i(x^{k+1})|^2 - \sum_{i \in I_\mu^k} \left( \frac{(z_i^k)^2}{2\beta} \right)
\]

\[
\geq \sum_{i \in I_\beta^k} z_i^k f_i(x^{k+1}) - \frac{\beta}{2} |f_i(x^{k+1})|^2 - \sum_{i \in I_\mu^k} \left( \frac{(z_i^k)^2}{2\beta} \right)
\]

\[
= \sum_{i \in I_\beta^k} z_i^k f_i(x^{k+1}) - \frac{\beta}{2\rho_z^2} \|z^{k+1} - z^k\|^2 \to 0, \text{ as } k \to \infty,
\]

where \( I_\beta^k \) and \( I_\mu^k \) are given in (3.5), and the last convergence to 0 is from (3.12) and (3.14a). Therefore, we obtain the results in (3.9) by noting \( \Psi_\beta(x^*, z^k) \leq 0 \).
From the lower boundedness of $f_i$ on $\mathcal{X}$, there is $F_i > 0$ such that $f_i(x) \geq -F_i, \forall x \in \mathcal{X}$. For $k \geq k_i$, when $f_i(x^{k+1}) < -\frac{\varepsilon}{\beta}$, it follows from (3.13) that $|z_i^k| < \beta \varepsilon$ and thus

$$|z_i^{k+1} - f_i(x^{k+1})| = |1 - \frac{\rho}{\beta} z_i^k - |z_i^k| \cdot |f_i(x^{k+1})| = |\beta - \rho z_i^k| \cdot |f_i(x^{k+1})| \leq |\beta - \rho z_i^k| \max(Z/\beta, F_i) \varepsilon,$$

and thus

$$|z_i^{k+1} + \beta f_i(x^{k+1})| = |1 - \frac{\rho z_i^k}{\beta} \cdot |z_i^k| \leq |\beta - \rho z_i^k| \varepsilon.$$

Therefore, (3.10c) follows from the above two equations and (3.14), and

$$\lim_{k \to \infty} \left[|z_i^k + \beta f_i(x^{k+1})| - z_i^{k+1}\right] = 0, \forall i \in [m].$$

In addition, the optimality of $x^{k+1}$ indicates

$$0 \in \partial f_0(x^{k+1}) + N_{\mathcal{X}}(x^{k+1}) + A^t y^k + \beta A^t r^{k+1} + \sum_{i=1}^m [z_i^k + \beta f_i(x^{k+1})], \partial f_i(x^{k+1}),$$

which together with (3.12), (3.15), and the boundedness of $\partial f_i(x^{k+1})$ implies (3.10a).

If $\{x_i^k\}$ has a cluster point $\bar{x}$, then from the boundedness of $(y^k, z^k)$, it follows that $\{x_i^k, y^k, z^k\}$ must also have a cluster point $(\bar{x}, \bar{y}, \bar{z})$ which satisfies KKT conditions in (2.1) from (3.10) and the continuity of $f_i$'s. Hence, (3.11) holds with $(\bar{y}, \bar{z}) = (\bar{y}, \bar{z})$, and thus $(y^k, z^k)$ converges to $(\bar{y}, \bar{z})$ from Lemma 2.1.

**Theorem 3.5 (Global sublinear convergence rate of ALM).** Under Assumptions 1 and 2, let $\{x^k, y^k, z^k\}$ be the sequence generated from Algorithm 2 with $y^0 = 0, z^0 = 0$ and $0 < \rho_y < 2/\beta, \rho_z < 2\beta$. Then

$$\lim_{k \to \infty} \left[|z_i^k + \beta f_i(x^{k+1})| - z_i^{k+1}\right] = 0, \forall i \in [m].$$

Proof. Summing up (3.6) with $x = x^*$ over $k$ gives

$$\sum_{i=0}^k \left[ f_0(x^{t+1}) - f_0(x^*) + g^T r^{t+1} + \sum_{i=1}^m z_i f_i(x^{t+1}) + (\beta - \frac{\rho_y}{2}) ||r^{t+1}||^2 \right.$$

$$\left. + \frac{1}{\rho_z^2} (\beta - \rho z) ||r^{t+1}||^2 \right] + \frac{1}{\rho_y} ||y^{k+1} - y||^2 + \frac{1}{\rho_z} ||z^{k+1} - z||^2$$

$$\leq \frac{1}{2\rho_y} ||y^0 - y||^2 + \frac{1}{2\rho_z} ||z^0 - z||^2.$$
where we have used $y^0 = 0$ and $z^0 = 0$. The desired results are obtained from Lemma 2.2 with $\bar{x} = \bar{x}^{k+1}$, $\alpha = 0$, $c_1 = \frac{1}{2\rho_y(k+1)}$, and $c_2 = \frac{1}{2\rho_z(k+1)}$. □

Before finishing this section, we make the two remarks on the results in Theorem 3.5.

**Remark 3.2.** The convergence rate results in Theorem 3.5 imply that we should choose large $\rho_y$ and $\rho_z$. However, this could cause difficulty for solving (3.1a) to a high accuracy, as discussed in Remark 4.3.

**Remark 3.3.** From [26], it follows that the augmented dual function $d_\beta(y, z)$ is differentiable and $\nabla d_\beta(y, z)$ is Lipschitz continuous with constant $\frac{1}{\bar{x}}$. Therefore, if $\rho_y, \rho_z \in (0, \beta]$, then $d_\beta(y^k, z^k)$ converges to the optimal value in $O(1/k)$ rate. However, this result does not indicate the convergence speed in terms of primal objective and feasibility violation. In addition, the results in Theorem 3.5 allow larger stepsizes that can in potential make the algorithm numerically converge faster.

4. Inexact augmented Lagrangian methods. Algorithm 2 requires an exact solution to (3.1a). In general, it is very expensive or even impossible to solve the subproblems exactly or to a high accuracy. In this section, we first present and analyze an inexact augmented Lagrangian method (iALM) for (1.1). It updates the dual variables $y$ and $z$ in the same way as that in Algorithm 2. But instead of exact solution to each primal subproblem, iALM allows a certain level of error specified by $\varepsilon_k$. The method is summarized in Algorithm 3. Then we assume certain special structures on (1.1) and estimate the iteration complexity of the iALM and/or its variant.

**Algorithm 3:** Inexact augmented Lagrangian method for (1.1)

1. **Initialization:** choose $x^0, y^0, z^0$ and $\beta, \rho_y, \rho_z$
2. for $k = 0, 1, \ldots$ do
3. Find $x^{k+1} \in \mathcal{X}$ such that
   \[ L_\beta(x^{k+1}, y^k, z^k) \leq \min_{x \in \mathcal{X}} L_\beta(x, y^k, z^k) + \varepsilon_k. \] (4.1)
4. Let $y^{k+1} \leftarrow (3.1b)$ and $z^{k+1} \leftarrow (3.1c)$.

4.1. Convergence analysis of iALM. In this subsection, we analyze the convergence behavior of Algorithm 3 under Assumption 1 and the following assumption that ensures well-definedness of $x^{k+1}$.

**Assumption 3.** For every $k$, there is $x^{k+1}$ satisfying (4.1).

We first show a fundamental result that is similar to Theorem 3.3.

**Theorem 4.1** (One-iteration progress of iALM). Let $\{(x^k, y^k, z^k)\}$ be the sequence generated from Algorithm 3. Then for any $x \in \mathcal{X}$ such that $Ax = b$ and
\( f_i(x) \leq 0, \forall i \in [m], \) any \( y, \) and any \( z \geq 0, \) it holds that

\[
\begin{align*}
f_0(x^{k+1}) - f_0(x) + y^\top r^{k+1} + \sum_{i=1}^{m} z_i f_i(x^{k+1}) & + \sum_{i=1}^{m} (|z_i^k + \beta f_i(x^{k+1})| + z_i) f_i(x^{k+1}) \\
+ \left( \frac{\beta}{2} - \rho_y \right) \|r^{k+1}\|^2 + \Psi_\beta(x^{k+1}, z^k) - f_0(x) - \Psi_\beta(x, z^k) & \leq \varepsilon_k.
\end{align*}
\]

Proof. From (4.1), it follows that for any \( x \) such that \( Ax = b, \)

\[
\begin{align*}
f_0(x^{k+1}) + \langle y^k, r^{k+1} \rangle + \frac{\beta}{2} \|r^{k+1}\|^2 + \Psi_\beta(x^{k+1}, z^k) - f_0(x) - \Psi_\beta(x, z^k) & \leq \varepsilon_k.
\end{align*}
\]

Since \( \langle y^k, r^{k+1} \rangle = \langle y^{k+1} - y, r^{k+1} \rangle + \langle y, r^{k+1} \rangle - \rho_y \|r^{k+1}\|^2, \) by adding (3.2) and (3.3) to the above inequality, we have

\[
\begin{align*}
f_0(x^{k+1}) - f_0(x) + y^\top r^{k+1} + \sum_{i=1}^{m} z_i f_i(x^{k+1}) & + \sum_{i=1}^{m} (|z_i^k + \beta f_i(x^{k+1})| + z_i) f_i(x^{k+1}) \\
+ \left( \frac{\beta}{2} - \rho_y \right) \|r^{k+1}\|^2 + \Psi_\beta(x^{k+1}, z^k) - \Psi_\beta(x, z^k) & + \frac{1}{2\rho_y} \left[ \|y^{k+1} - y\|^2 - \|y^k - y\|^2 + \|y^{k+1} - y\|^2 \right] \\
+ \frac{1}{2\rho_z} \left[ \|z^{k+1} - z\|^2 - \|z^k - z\|^2 + \|z^{k+1} - z\|^2 \right] & - \sum_{i=1}^{m} (z_i^{k+1} - z_i) \cdot \max \left( -\frac{z_i^k}{\beta}, f_i(x^{k+1}) \right)
\end{align*}
\]

\( \leq \varepsilon_k. \)

Note that

\[
\begin{align*}
\Psi_\beta(x^{k+1}, z^k) - \sum_{i=1}^{m} [z_i^k + \beta f_i(x^{k+1})] & = \sum_{i \in I^k} \left[ z_i^k f_i(x^{k+1}) + \frac{\beta}{2} [f_i(x^{k+1})]^2 - [z_i^k + \beta f_i(x^{k+1})] f_i(x^{k+1}) \right] + \sum_{i \in I^k} \left[ \frac{(z_i^k)^2}{2\beta} \right] \\
& = - \sum_{i \in I^k} \frac{\beta}{2} [f_i(x^{k+1})]^2 - \sum_{i \in I^k} \frac{(z_i^k)^2}{2\beta} \\
& \leq - \frac{\beta}{2\rho_z^2} \|z^{k+1} - z^k\|^2,
\end{align*}
\]

where the sets \( I^k \) and \( I^k \) are defined in (3.5). In addition, if \( f_i(x) \leq 0, \forall i \in [m], \) then \( \Psi_\beta(x, z^k) \leq 0. \) Hence, plugging (3.4) and (3.5) into (4.4) yields (4.2).

By Theorem 4.1, we establish convergence results of Algorithm 3 as follows.
Theorem 4.2 (Global convergence of iALM). Under Assumptions 1 and 3, let \( \{(x^k, y^k, z^k)\} \) be the sequence generated from Algorithm 3 with \( 0 < \rho_y < \beta \), \( 0 < \rho_z < \beta \), and \( \varepsilon_k \) satisfying \( \sum_{k=1}^{\infty} \varepsilon_k < \infty \). Then the sequence \( \{x^k\} \) is a minimizing sequence, namely (3.9) holds. In addition, assume that \( f_i(x) \) is lower bounded and continuous on \( \mathcal{X} \) for all \( i \in [m] \). If \( \{x^k\} \) has a cluster point \( \bar{x} \), then \( (y^k, z^k) \) converges to a point \( (\bar{y}, \bar{z}) \), and \( (\bar{x}, \bar{y}, \bar{z}) \) satisfies the KKT conditions in (2.1).

Proof. Letting \( (x, y, z) = (x^*, y^*, z^*) \) in (4.2) and using (2.2) give

\[
\frac{\beta - \rho_y}{2} \|v^{k+1}\|^2 + \frac{\beta - \rho_z}{2\rho_z^2} \|z^{k+1} - z^*\|^2 + \frac{1}{2\rho_y} \|y^{k+1} - y^*\|^2 + \frac{1}{2\rho_z} \|z^{k+1} - z^*\|^2 + \varepsilon_k.
\]

(4.6) Since \( 0 < \rho_y < \beta \), \( 0 < \rho_z < \beta \), and \( \sum_{k=1}^{\infty} \varepsilon_k < \infty \), summing up (4.6), we have (3.12) and the boundedness of \( (y^k, z^k) \). Hence, from the same arguments as those in the proof of Theorem 3.4, it follows that (3.9) holds, and we can show that \( (\bar{x}, \bar{y}, \bar{z}) \) satisfies (2.1b) and (2.1c). From (4.1), the continuity of \( f_i \)'s, and \( \varepsilon_k \to 0 \), it follows that

\[ \mathcal{L}_0(\bar{x}, \bar{y}, \bar{z}) \leq \mathcal{L}_0(x, \bar{y}, \bar{z}), \forall x \in \mathcal{X}, \]

so the first-order optimality condition holds:

\[ 0 \in \partial f_0(\bar{x}) + \mathcal{N}_\mathcal{X}(\bar{x}) + A^T \bar{y} + \sum_{i=1}^{m} [z_i + \beta f_i(\bar{x})]_+ \partial f_i(\bar{x}). \]

Note that (3.15) implies \( [z_i + \beta f_i(\bar{x})]_+ = \bar{z}_i \). Therefore, \( (\bar{x}, \bar{y}, \bar{z}) \) also satisfies (2.1a). Hence, (4.6) holds with \( (y^*, z^*) \) replaced by \( (\bar{y}, \bar{z}) \), and \( (y^k, z^k) \) converges to \( (\bar{y}, \bar{z}) \) from Lemma 2.1. \( \square \)

Remark 4.1. The work [27] has also analyzed the convergence of Algorithm 3 through the augmented dual function \( d_\beta \). It shows the results in (3.9) by assuming \( \sum_{k=1}^{\infty} \sqrt{\varepsilon_k} < \infty \), which is strictly stronger than our assumption \( \sum_{k=1}^{\infty} \varepsilon_k < \infty \).

Compared to the results in Theorem 3.4, we require smaller \( \rho_y \) and \( \rho_z \) to have the convergence of Algorithm 3. In addition, we are unable to show the result in (3.10a). These are because in general, (4.1) does not guarantee a point \( v^k \) such that \( \|v^k\| = O(\varepsilon_k) \) and \( v^k \in \partial \mathcal{L}_0(x^{k+1}, y^k, z^k) + \mathcal{N}_\mathcal{X}(x^{k+1}) \).

By Lemma 2.2 and Theorem 4.1, we have the following convergence rate estimate of Algorithm 3.

Theorem 4.3 (Global convergence rate of iALM). Under Assumptions 1 and 3, let \( \{(x^k, y^k, z^k)\} \) be the sequence generated from Algorithm 3 with \( y^0 = 0, z^0 = 0 \) and
0 < \rho_y < \beta, 0 < \rho_z < \beta. Then

(4.7a)
\[ -\frac{1}{k+1} \left( \frac{2\|y^*\|^2}{\rho_y} + 2\rho_z \sum_{i=1}^{m} (z_i^*)^2 + \sum_{j=0}^{k} \varepsilon_j \right) \leq f_0(\bar{x}^{k+1}) - f_0(x^*) \leq \frac{1}{k+1} \sum_{j=0}^{k} \varepsilon_j, \]

(4.7b)
\[ \|A\bar{x}^{k+1} - b\| + \sum_{i=1}^{m} [f_i(\bar{x}^{k+1})]_+ \leq \frac{1}{k+1} \left( \frac{1 + \|y^*\|^2}{2\rho_y} + \frac{1}{2\rho_z} \sum_{i=1}^{m} (1 + z_i^*)^2 + \sum_{j=0}^{k} \varepsilon_j \right), \]

where \( \bar{x}^{k+1} = \frac{\sum_{i=1}^{k+1} x^i}{k+1} \).

Proof. Summing up (4.2) with \( x = x^* \) and using the convexity of \( f_i \)'s give

\[ f_0(\bar{x}^{k+1}) - f_0(x^*) + y^\top (A\bar{x}^{k+1} - b) + \sum_{i=1}^{m} z_i f_i(\bar{x}^{k+1}) \leq \frac{1}{k+1} \left( \frac{1}{2\rho_y} \|y\|^2 + \frac{1}{2\rho_z} \|z\|^2 + \sum_{j=0}^{k} \varepsilon_j \right). \]

The results thus follow from Lemma 2.2 with \( \alpha = \frac{\sum_{j=0}^{k} \varepsilon_j}{k+1} \), \( c_1 = \frac{1}{2\rho_y(k+1)} \), and \( c_2 = \frac{1}{2\rho_z(k+1)} \).

\[ \square \]

Remark 4.2. The results in (4.7) imply that if \( \sum_{j=0}^{k} \varepsilon_j = o(k) \), we have sublinear convergence.

4.2. Iteration complexity of iALM for constrained composite convex programs. In this subsection, we assume composite convex structure on (1.1). More specifically, we assume

(4.8)
\[ f_0(x) = g(x) + h(x), \]

where \( g \) is a Lipschitz differentiable convex function on \( \mathcal{X} \), and \( h \) is a simple\(^3\) (possibly nondifferentiable) convex function. Also, \( f_i \) is convex and Lipschitz differentiable on \( \mathcal{X} \) for every \( i \in [m] \), namely, there are constants \( L_0, L_1, \ldots, L_m \) such that

(4.9)
\[ \|\nabla g(\hat{x}) - \nabla g(\tilde{x})\| \leq L_0 \|\hat{x} - \tilde{x}\|, \forall \hat{x}, \tilde{x} \in \mathcal{X}, \]

(4.10)
\[ \|\nabla f_i(\hat{x}) - \nabla f_i(\tilde{x})\| \leq L_i \|\hat{x} - \tilde{x}\|, \forall \hat{x}, \tilde{x} \in \mathcal{X}, \forall i \in [m]. \]

In addition, we assume the boundedness of \( \mathcal{X} \) and denote its diameter as

\[ D_\mathcal{X} = \max_{\hat{x}, \tilde{x} \in \mathcal{X}} \|\hat{x} - \tilde{x}\|. \]

We will explore the above structure to find \( x^{k+1} \) in (4.1) and estimate the total number of gradient evaluations to produce a solution with a specified accuracy.

The following results are easy to show from the Lipschitz differentiability of \( g \) and \( f_i, i \in [m] \).

\(^3\)By “simple”, we mean the proximal mapping of \( h \) is easy to evaluate, i.e., it is easy to find a solution to \( \min_{\hat{x}} h(x) + \frac{1}{2\gamma} \|x - \hat{x}\|^2 \) for any \( \hat{x} \) and \( \gamma > 0 \).
Proposition 4.4. Assume (4.9) and (4.10). If $\mathcal{X}$ is bounded, then there exist constants $B_1, \ldots, B_m$ such that

\begin{align}
(4.11a) & \quad \max \left( |f_i(x)|, \|\nabla f_i(x)\| \right) \leq B_i, \forall x \in \mathcal{X}, \forall i \in [m], \\
(4.11b) & \quad |f_i(\hat{x}) - f_i(\bar{x})| \leq B_i \|\hat{x} - \bar{x}\|, \forall \hat{x}, \bar{x} \in \mathcal{X}, \forall i \in [m].
\end{align}

Let the smooth part of $L_\beta$ be denoted as

$$F_\beta(x, y, z) = L_\beta(x, y, z) - h(x).$$

Based on (4.11), we are able to show Lipschitz continuity of $\nabla_x F_\beta(x, y, z)$ for every $(y, z)$.

Lemma 4.5. If (4.9) through (4.11) hold, then for any $(y, z)$, $\nabla_x F_\beta(x, y, z)$ is Lipschitz continuous on $\mathcal{X}$ in terms of $x$ with constant

$$L(z) = L_0 + \beta \|A^T A\| + \sum_{i=1}^{m} (\beta B_i (B_i + L_i) + L_i |z_i|).$$

Proof. First we notice that $\frac{\partial}{\partial u} \psi_\beta(u, v) = [\beta u + v]_+$, and thus for any $v$,

$$\left| \frac{\partial}{\partial u} \psi_\beta(\tilde{u}, v) - \frac{\partial}{\partial \tilde{u}} \psi_\beta(\tilde{u}, v) \right| \leq \beta |\tilde{u} - \tilde{u}|, \forall \tilde{u}, \tilde{u}.$$

Let $h_i(x, z_i) = \psi_\beta(f_i(x), z_i)$, $i = 1, \ldots, m$. Then

$$\begin{align*}
\| \nabla_x h_i(\hat{x}, z_i) - \nabla_x h_i(\bar{x}, z_i) \| & = \left\| \frac{\partial}{\partial u} \psi_\beta(f_i(\hat{x}), z_i) \nabla f_i(\hat{x}) - \frac{\partial}{\partial u} \psi_\beta(f_i(\bar{x}), z_i) \nabla f_i(\bar{x}) \right\| \\
& \leq \left\| \frac{\partial}{\partial u} \psi_\beta(f_i(\hat{x}), z_i) \nabla f_i(\hat{x}) - \frac{\partial}{\partial u} \psi_\beta(f_i(\bar{x}), z_i) \nabla f_i(\bar{x}) \right\| \\
& + \left\| \frac{\partial}{\partial u} \psi_\beta(f_i(\hat{x}), z_i) \nabla f_i(\hat{x}) - \frac{\partial}{\partial u} \psi_\beta(f_i(\bar{x}), z_i) \nabla f_i(\bar{x}) \right\| \\
& \leq \beta |f_i(\hat{x}) - f_i(\bar{x})| \cdot \| \nabla f_i(\hat{x}) \| + \left\| \frac{\partial}{\partial u} \psi_\beta(f_i(\hat{x}), z_i) \right\| \cdot \| \nabla f_i(\hat{x}) - \nabla f_i(\bar{x}) \| \\
& \leq \beta B_i^2 \|\hat{x} - \bar{x}\| + L_i (\beta B_i + |z_i|) \|\hat{x} - \bar{x}\|.
\end{align*}$$

Hence,

$$\begin{align*}
\| \nabla_x F_\beta(\hat{x}, y, z) - \nabla_x F_\beta(\bar{x}, y, z) \| & \leq \| \nabla g(\hat{x}) - \nabla g(\bar{x}) \| + \beta \|A^T A\| \|\hat{x} - \bar{x}\| + \sum_{i=1}^{m} \| \nabla_x h_i(\hat{x}, z_i) - \nabla_x h_i(\bar{x}, z_i) \| \\
& \leq \left( L_0 + \beta \|A^T A\| + \sum_{i=1}^{m} [\beta B_i^2 + L_i (\beta B_i + |z_i|)] \right) \|\hat{x} - \bar{x}\|,
\end{align*}$$

which completes the proof. \(\square\)
Therefore, we can apply Nesterov’s optimal first-order method in Algorithm 1 to find $x^{k+1}$ in (4.1). From Theorem 2.3, if starting from $x^k$ and running Algorithm 1 for $t_k$ iterations to produce $x^{k+1}$ with

$$\phi(x) = F_\beta(x, y^k, z^k), \quad \psi(x) = h(x) + i_X(x),$$

then we have

$$L_\beta(x^{k+1}, y^k, z^k) - \min_{x \in X} L_\beta(x, y^k, z^k) \leq \frac{2L(z^k)[\text{dist}(x^k, X^*_k)]^2}{t_k},$$

(4.13)

where $X^*_k$ denotes the set of optimal solutions to $\min_{x \in X} L_\beta(x, y^k, z^k)$.

From (4.13), we have the following result.

**Lemma 4.6.** Let $\{x^k, y^k, z^k\}$ be the sequence from Algorithm 3 with $y^0 = 0, z^0 = 0, 0 \leq \rho_y < \beta, 0 \leq \rho_z < \beta$, and $x^{k+1}$ given by running Algorithm 1 for $t_k$ iterations on $\min_{x \in X} L_\beta(x, y^k, z^k)$ with initial point set to $x^k$, where

$$t_k = \left[\sqrt{2L(z^k)D_X(k + 1)^{\frac{1 + \delta}{2}}}\right], \forall k.$$

Here, $L(z^k)$ is defined in (4.12), and $\delta > 0$ is a constant. Then

$$\sum_{j=0}^{k} t_j \leq \sum_{j=0}^{k} \left(\sqrt{2L(z^k)(j + 1)^{\frac{1 + \delta}{2}}} + 1\right) \leq \frac{2\sqrt{2L(z^k)D_X}}{3 + \delta} (k + 2)^{\frac{1 + \delta}{2}},$$

where

$$L_z = L_0 + \beta\|A^\top A\| + \sum_{i=1}^{m} \beta B_i(B_i + L_i) + \sum_{i=1}^{m} L_i^2 \left(\frac{\rho_z}{\rho_y}\|y^*\| + 2\|z^*\| + \sqrt{2\rho_z(1 + \frac{1}{\delta})}\right).$$

**Proof.** Summing up (4.6) and using $y^0 = 0, z^0 = 0$ give

$$\frac{1}{2\rho_y}\|y^k - y^*\|^2 + \frac{1}{2\rho_z}\|z^k - z^*\|^2 \leq \frac{1}{2\rho_y}\|y^*\|^2 + \frac{1}{2\rho_z}\|z^*\|^2 + \sum_{j=0}^{k-1} \varepsilon_j.$$

(4.14)

Note $\varepsilon_k \leq \frac{1}{(k+1)^{1+\delta}}$ from (4.13) and the choice of $t_k$, and thus

$$\sum_{j=0}^{\infty} \varepsilon_j \leq \sum_{j=0}^{\infty} \frac{1}{(k + 1)^{1+\delta}} \leq 1 + \int_{0}^{\infty} \frac{1}{(u + 1)^{1+\delta}} du = 1 + \frac{1}{\delta}.$$

From the above inequality and (4.14), it follows

$$\|z^k\| \leq \|z^*\| + \|z^k - z^*\|$$

$$\leq \|z^*\| + \sqrt{\frac{\rho_z}{\rho_y}\|y^*\|^2 + \|z^*\|^2 + 2\rho_z \sum_{j=0}^{k-1} \varepsilon_j}$$

$$\leq \sqrt{\frac{\rho_z}{\rho_y}\|y^*\|^2 + 2\|z^*\|^2} + \sqrt{2\rho_z(1 + \frac{1}{\delta})}.$$
Hence,
\[ \sum_{i=1}^{m} L_i |z_i^k| \leq \sqrt{\sum_{i=1}^{m} L_i^2 \left( \frac{D_x}{p_y} \|y^*\| + 2\|z^*\| + \sqrt{2\rho_z(1 + \frac{1}{\delta})} \right)}, \]
and thus \( L(z^k) \leq L_z, \forall k \). Therefore,
\[ \sum_{j=0}^{k} t_j \leq \sum_{j=0}^{k} \left( \sqrt{2L_z D_X (j + 1)\frac{3+\eta}{\delta}} + 1 \right) \leq \sqrt{2L_z D_X} \left( \frac{2}{3 + \delta} (k + 1)\frac{3+\eta}{\delta} + 1 \right) + k + 1, \]
where the last inequality uses
\[ \sum_{j=0}^{k} (j + 1)\frac{3+\eta}{\delta} \leq \int_{0}^{k+1} (u + 1)\frac{3+\eta}{\delta} du = \frac{2}{3 + \delta} \left( (k + 2)\frac{3+\eta}{\delta} - 1 \right). \]

This completes the proof. \( \Box \)

From Theorem 4.3 and Lemma 4.6, we have the following estimate on the number of gradient evaluations to reach an \( \varepsilon \)-optimal solution.

**Theorem 4.7 (Iteration complexity of iALM).** Under Assumption 1, let \( \{(x^k, y^k, z^k)\} \) be the sequence from Algorithm 3 with \( y^0 = 0, z^0 = 0 \), and \( 0 < \rho_y < \beta, 0 < \rho_z < \beta \). Then given any \( \varepsilon > 0 \) and \( \delta > 0 \), we can have an \( \varepsilon \)-optimal solution by evaluating the gradients of \( g \) and \( f_i, i \in [m] \) for \( T \) times in total, where
\[ T \lesssim \frac{2\sqrt{2L_z D_X}}{3 + \delta} \left( \frac{\max \{4\|y^*\|^2, (1 + \|y^*\|^2)\}}{2\rho_y} + \sum_{i=1}^{m} \frac{\max \{4(z^*_i)^2, (1 + |z^*_i|^2)\}}{2\rho_z} + 1 + \frac{1}{\delta} \right)^{\frac{3+\eta}{\delta} \frac{3+\eta}{\delta}} \varepsilon^{-\frac{3+\eta}{\delta}}. \]

**Remark 4.3.** From the above iteration complexity estimate, it follows that the best penalty parameter \( \beta \) and multiplier update stepsizes \( \rho_y, \rho_z \) depend on \( (y^*, z^*) \). Simply setting \( \rho_y = \rho_z = \gamma \beta \) for a certain \( \gamma \in (0, 1) \), we can optimize \( T \) with respect to \( \beta \). Since \( (y^k, z^k) \) converges to an optimal dual solution, we can estimate \( (y^*, z^*) \) from the iterates.

### 4.3. Inexact ALM for smooth constrained convex programs.

In this subsection, we assume that \( f_i \) is smooth for every \( i = 0, 1, \ldots, m \), i.e., \( h = 0 \) in (4.8). From Lemma 4.5, it follows that \( \min_{x \in X} L_\beta(x, y^k, z^k) \) becomes a smooth optimization problem. Based on the smooth structure, we modify Algorithm 3 to improve the iteration complexity given in Theorem 4.7. The modification is inspired from the linear convergence of Algorithm 1 for strongly convex problems and the following result, which can be shown from Prop. 3 and Lemma 4 in [17]. The method, named iPALM, is summarized in Algorithm 4.

**Lemma 4.8.** Let \( X \) be a closed convex set and \( f \) a convex and Lipschitz differentiable function with gradient Lipschitz constant \( L_f \). For any \( \eta \geq 0 \), if there is a point \( x \in X \) such that
\[ f(x) - \min_{x \in X} f(x) \leq \eta, \]
then there is a vector \( v \) such that \( \|v\| \leq 2\sqrt{2L_f\eta} \) and

\[
v \in \nabla f(x^+) + N_X(x^+),
\]

where \( x^+ = \text{Proj}_X \left( x - \frac{1}{L_f} \nabla f(x) \right) \).

**Algorithm 4: Inexact proximal augmented Lagrangian method (iPALM)**

1. **Initialization:** choose \( x^0, y^0, z^0 \) and \( \beta, \rho_y, \rho_z \)
2. **for** \( k = 0, 1, \ldots \) **do**
   
   3. Find \( x^{k+1} \in X \) such that
      
      \[
      L_{\beta}(\tilde{x}^{k+1}, y^k, z^k) + \frac{1}{2}\|\tilde{x}^{k+1} - x^k\|^2 \leq \min_{x \in X} \left\{ L_{\beta}(x, y^k, z^k) + \frac{1}{2}\|x - x^k\|^2 \right\} + \varepsilon_k
      \]
      
      Let
      
      \[
      x^{k+1} = \text{Proj}_X \left( \tilde{x}^{k+1} - \frac{1}{L(z^k) + 1} (\nabla_x L_{\beta}(\tilde{x}^{k+1}, y^k, z^k) + \tilde{x}^{k+1} - x^k) \right).
      \]

      Let \( y^{k+1} \leftarrow (3.1b) \) and \( z^{k+1} \leftarrow (3.1c) \).

To analyze the convergence behavior of Algorithm 4, we first establish a fundamental result that is similar to Theorems 3.3 and 4.1.

**Theorem 4.9** (One iteration progress of iPALM). Let \( \{(x^k, y^k, z^k)\} \) be the sequence generated from Algorithm 4. Then

\[
f_0(x^{k+1}) - f_0(x) + y^\top r^{k+1} + \sum_{i=1}^m z_i f_i(x^{k+1}) + (\beta - \frac{\rho_y}{2})\|x^{k+1}\|^2 + \frac{1}{2}\|x^{k+1} - x\|^2
\]

\[
+ \frac{1}{2}\|x^{k+1} - x\|^2 + \frac{1}{\rho_z} (\beta - \frac{\rho_y}{2})\|z^{k+1} - z\|^2 + \frac{1}{2\rho_y}\|y^{k+1} - y\|^2 + \frac{1}{2\rho_z}\|z^{k+1} - z\|^2
\]

\[
\leq \frac{1}{2}\|x^k - x\|^2 + \frac{1}{2\rho_y}\|y^k - y\|^2 + \frac{1}{2\rho_z}\|z^k - z\|^2 + 2D_X \sqrt{2(L(z^k) + 1)\varepsilon_k}.
\]

**Proof.** From Lemma 4.8, it follows that there is \( v^k \) such that \( \|v^k\| \leq 2\sqrt{2L(z^k) + 1}\varepsilon_k \) and

\[
v^k \in \nabla_x L_{\beta}(x^{k+1}, y^k, z^k) + x^{k+1} - x^k + N_X(x^{k+1}).
\]

Hence,

\[
\langle x^{k+1} - x, \nabla_x L_{\beta}(x^{k+1}, y^k, z^k) + x^{k+1} - x^k \rangle \leq \langle x^{k+1} - x, v^k \rangle \leq 2D_X \sqrt{2(L(z^k) + 1)\varepsilon_k}.
\]

Following the arguments in the proof of Theorem 3.3 and noting

\[
\langle x^{k+1} - x, x^{k+1} - x^k \rangle = \frac{1}{2} \left[ \|x^{k+1} - x\|^2 - \|x^k - x\|^2 + \|x^{k+1} - x^k\|^2 \right],
\]
we have the desired result from the above inequality. □

Using Theorem 4.9, we show the iterate convergence of Algorithm 4 below.

**Theorem 4.10** (Global iterate convergence of iPALM). Under Assumption 1, let \( \{(x^k, y^k, z^k)\} \) be the sequence generated from Algorithm 4 with \( 0 < \rho_y < 2\beta \), and \( 0 < \rho_z < 2\beta \). If

\[
\sum_{k=0}^{\infty} \sqrt{2(L(z^k) + 1)} \varepsilon_k < \infty,
\]

then \( (x^k, y^k, z^k) \) converges to a point \((\bar{x}, \bar{y}, \bar{z})\) that satisfies the KKT conditions in (2.1).

**Proof.** Letting \((x, y, z) = (x^*, y^*, z^*)\) in (4.17) and using (2.2), we have

\[
\left(\beta - \frac{\rho_y}{2}\right)\|r^{k+1}\|^2 + \frac{1}{2}\|x^{k+1} - x^*\|^2 + \frac{1}{2}\|x^k - x^*\|^2
\]
\[
+ \frac{1}{\rho_z^2} (\beta - \frac{\rho_z}{2})\|z^{k+1} - z^*\|^2 + \frac{1}{2\rho_y}\|y^{k+1} - y^*\|^2 + \frac{1}{2\rho_z}\|z^{k+1} - z^*\|^2
\]

(4.19) \[ \leq \frac{1}{2}\|x^k - x^*\|^2 + \frac{1}{2\rho_y}\|y^k - y^*\|^2 + \frac{1}{2\rho_z}\|z^k - z^*\|^2 + 2Dx\sqrt{2(L(z^k) + 1)}\varepsilon_k. \]

Summing up the above inequality gives

\[
\sum_{j=0}^{k} \left(\left(\beta - \frac{\rho_y}{2}\right)\|r^{j+1}\|^2 + \frac{1}{2}\|x^{j+1} - x^*\|^2 + \frac{1}{2}\|x^j - x^*\|^2 \right)
\]
\[
+ \frac{1}{\rho_z^2} (\beta - \frac{\rho_z}{2})\|z^{j+1} - z^*\|^2 + \frac{1}{2\rho_y}\|y^{j+1} - y^*\|^2 + \frac{1}{2\rho_z}\|z^{j+1} - z^*\|^2
\]

(4.20) \[ \leq \frac{1}{2}\|x^0 - x^*\|^2 + \frac{1}{2\rho_y}\|y^0 - y^*\|^2 + \frac{1}{2\rho_z}\|z^0 - z^*\|^2 + \sum_{j=0}^{k} 2Dx\sqrt{2(L(z^j) + 1)}\varepsilon_j. \]

Since \( 0 < \rho_y < 2\beta \), \( 0 < \rho_z < 2\beta \), (4.20) implies

\[
\lim_{k \to \infty} x^{k+1} - x^k = 0, \quad \lim_{k \to \infty} y^{k+1} - y^k = 0, \quad \lim_{k \to \infty} z^{k+1} - z^k = 0.
\]

Then by the same arguments as those in the proof of Theorem 3.4, any cluster point of \( \{(x^k, y^k, z^k)\} \) satisfies the KKT conditions in (2.1). Let \((\bar{x}, \bar{y}, \bar{z})\) be a cluster point. Then (4.19) holds with \((x^*, y^*, z^*)\) replaced by \((\bar{x}, \bar{y}, \bar{z})\). Therefore, \((x^k, y^k, z^k)\) converges to \((\bar{x}, \bar{y}, \bar{z})\) from Lemma 2.1. □

In addition, we have the convergence rate result of Algorithm 4 as follows.

**Theorem 4.11** (Global convergence rate of iPALM). Under Assumption 1, let \( \{(x^k, y^k, z^k)\} \) be the sequence generated from Algorithm 4 with \( y^0 = 0, z^0 = 0 \) and \( 0 < \rho_y < 2\beta \), \( 0 < \rho_z < 2\beta \). Then

\[
\frac{-1}{k+1} \left( C_0 + \frac{2\|y^*\|^2}{\rho_y} + \frac{2}{\rho_z} \sum_{i=1}^{m} (z^*_i)^2 \right) \leq f_0(x^{k+1}) - f_0(x^*) \leq \frac{C_0}{k+1},
\]

(4.22a) \[ \|Ax^{k+1} - b\| + \sum_{i=1}^{m} |f_i(x^{k+1})| \leq \frac{1}{k+1} \left( C_0 + \frac{(1 + \|y^*\|^2}{2\rho_y} + \frac{1}{2\rho_z} \sum_{i=1}^{m} (1 + z^*_i)^2 \right), \]

(4.22b)
where $\bar{x}^{k+1} = \frac{\sum_{i=1}^{k+1} x_i}{k+1}$, and
\[
C_0 = \frac{1}{2} \| x^0 - x^* \|^2 + \sum_{j=0}^{k} D_X \sqrt{2(L(z^j) + 1) \varepsilon_j}.
\]

Proof. Letting $x = x^*$ in (4.17) and summing it up give
\[
f_0(\bar{x}^{k+1}) - f_0(x^*) + y^T (A \bar{x}^{k+1} - b) + \sum_{i=1}^{m} z_i f_i(\bar{x}^{k+1})
\leq \frac{1}{k+1} \sum_{t=0}^{k} \left[ f_0(x^{t+1}) - f_0(x^*) + y^T r^{t+1} + \sum_{i=1}^{m} z_i f_i(x^{t+1}) \right]
\leq \frac{1}{k+1} \left( \frac{1}{2} \| x^0 - x^* \|^2 + \frac{1}{2\rho_y} \| y \|^2 + \frac{1}{2\rho_z} \| z \|^2 + \sum_{t=0}^{k} 2D_X \sqrt{2(L(z^t) + 1) \varepsilon_t} \right).
\]

The desired results follow from Lemma 2.2 with $\alpha = \frac{C_0}{k+1}$, $c_1 = \frac{1}{2\rho_y (k+1)}$, and $c_2 = \frac{1}{2\rho_z (k+1)}$. $\square$

The rest of this subsection aims at estimating the iteration complexity of Algorithm 4. Note that if $x^{k+1}$ in (4.16) is produced by running Algorithm 1 for $t_k$ iterations on $\min_{x \in X} \{ L_\beta(x, y^k, z^k) + \frac{1}{2} \| x - x^k \| \}$, then
\[
(4.23) \quad \varepsilon_k \leq \frac{D^2_A (L(z^k) + 2)}{2} \left( 1 - \sqrt{\frac{1}{L(z^k) + 1}} \right)^{t_k}.
\]

Therefore, we have the next result.

Lemma 4.12. Let $\{ (x^k, y^k, z^k) \}$ be the sequence from Algorithm 4 with $y^0 = 0$, $z^0 = 0$, $0 < \rho_y < 2\beta$, $0 < \rho_z < 2\beta$, and $x^{k+1}$ produced by running Algorithm 1 for $t_k$ iterations on $\min_{x \in X} \{ L_\beta(x, y^k, z^k) + \frac{1}{2} \| x - x^k \| \}$ with
\[
t_k = \left\lfloor \frac{2 \log 2D_A^2 (L(z^k) + 2) + 2(1 + \delta) \log (k+1)}{\log \left( 1 + \frac{1}{\sqrt{L(z^k) + 1}} \right)} \right\rfloor, \forall k,
\]
where $L(z^k)$ is defined in (4.12), and $\delta > 0$ is a constant. Then
\[
\sum_{j=0}^{k} t_j \leq \frac{2(1 + \delta)}{\log \left( 1 + \frac{1}{\sqrt{L(z^{k+1})}} \right)} (k + 2) \log (k + 2),
\]
where
\[
L_z = L_0 + 2\beta \| A^T A \| + \sum_{i=1}^{m} \beta B_i (B_i + L_i) + \sqrt{\sum_{i=1}^{m} L_i^2} \left( \sqrt{\rho_z} \| x^0 - x^* \| + \sqrt{\frac{\rho_z}{\rho_y}} \| y^* \| + 2 \| z^* \| + 2 \rho_z (1 + \frac{1}{\delta}) \right).
\]
Proof. From (4.23) and the choice of $t_k$, it follows that

$$2D_X \sqrt{2(L(z^k) + 1)\varepsilon_k} \leq \frac{1}{(k+1)^{1+\delta}}, \forall k.$$ 

Hence, similar to (4.15), we have

$$\|z^k\| \leq \sqrt{\rho_z} \|x^0 - x^*\| + \sqrt{\frac{\rho_z}{\rho_y}} \|y^*\| + 2\|z^*\| + \sqrt{2\rho_z(1 + \frac{1}{\delta})}.$$ 

Hence,

$$\sum_{i=1}^{m} L_i |z^k_i| \leq \sqrt{\sum_{i=1}^{m} L_i^2} \left(\sqrt{\rho_z} \|x^0 - x^*\| + \sqrt{\frac{\rho_z}{\rho_y}} \|y^*\| + 2\|z^*\| + \sqrt{2\rho_z(1 + \frac{1}{\delta})}\right),$$

and thus $L(z^k) \leq L_z$, $\forall k$. Therefore,

$$\sum_{j=0}^{k} t_j \leq \sum_{j=0}^{k} \left(\frac{2\log 2D_X^2(L_z + 2) + 2(1 + \delta)\log(k + 1)}{\log \left(1 + \frac{1}{\sqrt{L(z^*_j)} + 1 - 1}\right)} + 1\right) \leq \frac{2(1 + \delta)}{\log \left(1 + \frac{1}{\sqrt{L_z} + 1 - 1}\right)} \int_0^{k+1} \log(u + 1) du = \frac{2(1 + \delta)}{\log \left(1 + \frac{1}{\sqrt{L_z} + 1 - 1}\right)} ((k + 2) \log(k + 2) - (k + 1)), $$

which completes the proof. \qed

Lemma 4.12 together with Theorem 4.11 implies the following result.

**Theorem 4.13.** Under Assumption 1, let $\{(x^k, y^k, z^k)\}$ be the sequence from Algorithm 4 with $y^0 = 0, z^0 = 0$, and $0 < \rho_y < 2\beta, 0 < \rho_z < 2\beta$. Then given any $\varepsilon > 0$ and $\delta > 0$, to reach an $\varepsilon$-optimal solution, we only need to evaluate the gradients of $f_i, i = 0, 1, \ldots, m$ for $T$ times in total with

$$T \lesssim \frac{2(1 + \delta)}{\log \left(1 + \frac{1}{\sqrt{L_z} + 1 - 1}\right)} \frac{C_0}{\varepsilon \log C_0},$$

where

$$C_0 = \frac{1}{2} \|x^0 - x^*\|^2 + \max \left(\frac{4\|y^*\|^2}{2\rho_y}, (1 + \|y^*\|)^2\right) + \sum_{i=1}^{m} \max \left(\frac{4(z^*_i)^2}{2\rho_z}, (1 + |z^*_i|)^2\right) + 1 + \frac{1}{\delta}. $$

5. Related works. In this section, we review related works and compare them to our results. Our review and comparison focus on convex optimization, but note that ALM has also been popularly applied to nonconvex optimization problems; see [4–6] and the references therein.

*Affinely constrained convex problems.* Several recent works have established the convergence rate of ALM and its inexact version for affinely constrained convex problems. Assuming exact solution to every $x$-subproblem, [12] first shows $O(1/k)$ convergence of ALM for smooth problems in terms of dual objective and then accelerates the rate to $O(1/k^2)$ by applying Nesterov’s extrapolation technique to the
multiplier update. The results are extended to nonsmooth problems in [16] that uses similar technique. By adapting parameters, [31] establishes $O(1/k^2)$ convergence of a linearized ALM in terms of primal objective and feasibility violation. The linearized ALM allows linearization to smooth part in the objective but still assumes exact solvability of $x$-subproblems.

When the objective is strongly convex, [15] proves $O(1/k^2)$ convergence of an inexact ALM with extrapolation technique applied to the multiplier update. It requires summable error and subproblems to be solved more and more accurately. However, it does not give an estimate of iteration complexity on solving all subproblems to the required accuracies. For smooth convex problems, [18] analyzes the iteration complexity of the inexact ALM. It applies Nesterov’s optimal first-order method to every $x$-subproblem and shows that $O(\varepsilon^{-\frac{1}{2}})$ gradient evaluations are required to reach an $\varepsilon$-optimal solution. Compared to this complexity, our result in Theorem 4.7 is better by an order nearly $O(\varepsilon^{-\frac{1}{4}})$. In addition, [18] modifies the inexact ALM by adding a proximal term to each $x$-subproblem. This modification is similar to that in Algorithm 4. Motivated by the model predictive control, [21] also analyzes the iteration complexity of inexact dual gradient methods (iDGM) that are essentially inexact ALMs. It shows that to reach an $\varepsilon$-optimal solution, a nonaccelerated iDGM requires $O(\varepsilon^{-1})$ outer iterations and every $x$-subproblem solved to an accuracy $O(\varepsilon^2)$, and an accelerated iDGM requires $O(\varepsilon^{-\frac{1}{2}})$ outer iterations and every $x$-subproblem solved to an accuracy $O(\varepsilon^3)$. While the iteration complexity in [18] is estimated based on the best iterate, and that in [21] is ergodic, [19] establishes non-ergodic convergence of inexact ALM. It requires $O(\varepsilon^{-2})$ outer iterations and every subproblem solved to an accuracy $O(\varepsilon^2)$ to reach an $\varepsilon$-optimal solution in the non-ergodic sense.

Another line of existing works on inexact ALM assume two or multiple block structure on the problem and simply perform one cycle of Gauss-Seidel update to the block variables or update one randomly selected block. Global sublinear convergence of these methods has also been established. Exhausting all such works is impossible and out of scope of this paper. We refer interested readers to [7, 9–11, 13, 32] and the references therein.

**General convex problems.** As there are nonlinear inequality constraints, we do not find any work in the literature showing the global convergence rate of ALM or its inexact versions, though its local convergence rate has been extensively studied (e.g., [3, 26, 28]). Many existing works on nonlinearly constrained convex problems employ Lagrangian function instead of the augmented one and establish global convergence rate through dual subgradient approach (e.g., [20, 22, 23]). For general convex problems, these methods enjoy $O(1/\sqrt{k})$ convergence, and for strongly convex case, the rate can be improved to $O(1/k)$. Assuming Lipschitz continuity of $f_i$ for every $i \in [m]$, [34] proposes a new algorithm for nonlinearly constrained convex programs.

---

4 [21] assumes every subproblem solved to the condition $\langle \nabla L_\beta(x^{k+1}, y^k), x - x^{k+1} \rangle \geq -O(\varepsilon), \forall x \in \mathcal{X}$, which is implied by $L_\beta(x^{k+1}, y^k) - \min_{x \in \mathcal{X}} L_\beta(x, y^k) \leq O(\varepsilon^2)$ if $L_\beta$ is Lipschitz differentiable with respect to $x$. 

23
Every iteration, it minimizes a proximal Lagrangian function and updates the multiplier in a novel way. With sufficiently large proximal parameter that depends on the Lipschitz constants of $f_i$’s, the algorithm converges in $O(1/k)$ ergodic rate. The follow-up paper [33] focuses on smooth constrained convex problems and proposes a linearized variant of the algorithm in [34]. Assuming compactness of the set $\mathcal{X}$, it also establishes $O(1/k)$ ergodic convergence of the linearized method.

6. Concluding remarks. We have established $O(1/k)$ convergence of ALM and its inexact version for general constrained convex programs. Furthermore, we have shown that to reach an $\varepsilon$-optimal solution, it is sufficient to evaluate gradients of $f_i$’s for $O(\varepsilon^{-\frac{2}{3} - \delta})$ times on solving a constrained composite convex problem and $O(\varepsilon^{-1} |\log \varepsilon|)$ on a constrained smooth problem, where $\delta$ is an arbitrary positive number.

From (4.13), we see that the iteration complexity of the inexact ALM for constrained composite convex problems may be improved if $\text{dist}(x^k, \mathcal{X}_k^k)$ approaches to zero sublinearly. In addition, for constrained smooth convex problems, as in [33], simply performing one projected gradient step at each iteration, we may be able to remove the logarithmic term in the iteration complexity result. We will explore these issues in the future work.

REFERENCES

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
[2] A. Ben-Tal and M. Zibulevsky. Penalty/barrier multiplier methods for convex programming problems. *SIAM Journal on Optimization*, 7(2):347–366, 1997.
[3] D. P. Bertsekas. Convergence rate of penalty and multiplier methods. In *Decision and Control including the 12th Symposium on Adaptive Processes, 1973 IEEE Conference on*, volume 12, pages 260–264. IEEE, 1973.
[4] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.
[5] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.
[6] E. G. Birgin, R. Castillo, and J. M. Martínez. Numerical comparison of augmented lagrangian algorithms for nonconvex problems. *Computational Optimization and Applications*, 31(1):31–55, 2005.
[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
[8] P. L. Combettes and J.-C. Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.
[9] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.
[10] X. Gao, Y. Xu, and S. Zhang. Randomized primal-dual proximal block coordinate updates. *arXiv preprint arXiv:1605.05969*, 2016.
[11] R. Glowinski. On alternating direction methods of multipliers: a historical perspective. In *Modeling, simulation and optimization for science and technology*, pages 59–82. Springer, 2014.
[12] B. He and X. Yuan. On the acceleration of augmented lagrangian method for linearly constrained optimization. *Optimization online*, 2010.
[13] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.
1. M. R. Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

2. M. Kang, M. Kang, and M. Jung. Inexact accelerated augmented lagrangian methods. *Computational Optimization and Applications*, 62(2):373–404, 2015.

3. M. Kang, S. Yun, H. Woo, and M. Kang. Accelerated bregman method for linearly constrained $\ell_1-\ell_2$ minimization. *Journal of Scientific Computing*, 56(3):515–534, 2013.

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

5. G. Lan, D. Renato, and C. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016.

6. Y.-F. Liu, X. Liu, and S. Ma. On the non-ergodic convergence rate of an inexact augmented lagrangian framework for composite convex programming. *arXiv preprint arXiv:1603.05738*, 2016.

7. I. Necoara and V. Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5):1232–1243, 2014.

8. V. Nedelcu, I. Necoara, and Q. Tran-Dinh. Computational complexity of inexact gradient augmented lagrangian methods: application to constrained mpc. *SIAM Journal on Control and Optimization*, 52(5):3109–3134, 2014.

9. A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.

10. A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

11. Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publisher, 2004.

12. M. J. Powell. A method for non-linear constraints in minimization problems. in *Optimization*, R. Fletcher Ed., Academic Press, New York, NY, 1969.

13. R. T. Rockafellar. The multiplier method of hestenes and powell applied to convex programming. *Journal of Optimization Theory and applications*, 12(6):555–562, 1973.

14. R. T. Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976.

15. M. Schmidt, N. L. Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

16. P. Tseng and D. P. Bertsekas. On the convergence of the exponential multiplier method for convex programming. *Mathematical Programming*, 60(1):1–19, 1993.

17. Y. Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *arXiv preprint arXiv:1606.09155*, 2016.

18. Y. Xu. Asynchronous parallel primal-dual block update methods. *arXiv preprint arXiv:1705.06391*, 2017.

19. H. Yu and M. J. Neely. A primal-dual type algorithm with the $O(1/t)$ convergence rate for large scale constrained convex programs. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 1900–1905. IEEE, 2016.

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