Mathematical Programming
https://doi.org/10.1007/s10107-019-01425-9

FULL LENGTH PAPER

Series A

Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming

Yangyang Xu[1]

Received: 24 September 2018 / Accepted: 13 August 2019

Abstract

Augmented Lagrangian method (ALM) has been popularly used for solving constrained optimization problems. Practically, subproblems for updating primal variables in the framework of ALM usually can only be solved inexactly. The convergence and local convergence speed of ALM have been extensively studied. However, the global convergence rate of the inexact ALM is still open for problems with nonlinear inequality constraints. In this paper, we work on general convex programs with both equality and inequality constraints. For these problems, we establish the global convergence rate of the inexact ALM and estimate its iteration complexity in terms of the number of gradient evaluations to produce a primal and/or primal-dual solution with a specified accuracy. We first establish an ergodic convergence rate result of the inexact ALM that uses constant penalty parameters or geometrically increasing penalty parameters. Based on the convergence rate result, we then apply Nesterov's optimal first-order method on each primal subproblem and estimate the iteration complexity of the inexact ALM. We show that if the objective is convex, then $O(\varepsilon^{-1})$ gradient evaluations are sufficient to guarantee a primal ε-solution in terms of both primal objective and feasibility violation. If the objective is strongly convex, the result can be improved to $O(\varepsilon^{-\frac{1}{2}}|\log \varepsilon|)$. To produce a primal-dual ε-solution, more gradient evaluations are needed for convex case, and the number is $O(\varepsilon^{-\frac{4}{3}})$, while for strongly convex case, the number is still $O(\varepsilon^{-\frac{1}{2}}|\log \varepsilon|)$. Finally, we establish a nonergodic convergence rate result of the inexact ALM that uses geometrically increasing penalty parameters. This result is established only for the primal problem. We show that the nonergodic iteration complexity result is in the same order as that for the ergodic result. Numerical experiments on quadratically constrained quadratic programming are conducted to compare the performance of the inexact ALM with different settings.

Keywords Augmented Lagrangian method (ALM) · Nonlinearly constrained problem · First-order method · Global convergence rate · Iteration complexity

This work is partly supported by NSF Grant DMS-1719549.

Extended author information available on the last page of the article

Published online: 21 August 2019

Mathematics Subject Classification 90C06 · 90C25 · 68W40 · 49M27

1 Introduction

In this paper, we consider the constrained convex programming

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\ f_0(\mathbf{x}),\ \text{ s.t. } \mathbf{A}\mathbf{x}=\mathbf{b},\ f_i(\mathbf{x})\le 0, i=1,\ldots,m, \tag{1.1}$$

where $\mathcal{X}\subseteq\mathbb{R}^n$ is a closed convex set, $\mathbf{A}$ and $\mathbf{b}$ are respectively given matrix and vector, 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). It appears in many areas including statistics, machine learning, data mining, engineering, signal processing, finance, operations research, and so on.

Note that the constraint $\mathbf{x}\in\mathcal{X}$ can be equivalently represented by using an inequality constraint $\iota_{\mathcal{X}}(\mathbf{x})\le 0$ or adding $\iota_{\mathcal{X}}(\mathbf{x})$ to the objective, where $\iota_{\mathcal{X}}$ is the indicator function on $\mathcal{X}$ defined in (1.15) below. However, we explicitly use it for technical reason. In addition, every affine constraint $\mathbf{a}_j^\top\mathbf{x}=b_j$ can be equivalently represented by two inequality constraints: $\mathbf{a}_j^\top\mathbf{x}-b_j\le 0$ and $-\mathbf{a}_j^\top\mathbf{x}+b_j\le 0$. That way does not change theoretical results of an algorithm but will make the problem computationally more difficult.

One popular method for solving (1.1) is the augmented Lagrangian method (ALM), which first appeared in [19,36]. ALM alternatingly updates the primal variable and the Lagrangian multipliers. At each update, the primal variable is renewed by minimizing the augmented Lagrangian (AL) function and the multipliers by a dual gradient ascent. The global convergence and local convergence rate of ALM have been extensively studied; see the books [5,6]. Several recent works (e.g., [17,27]) establish the global convergence rate of ALM and/or its variants for affinely constrained problems. In the framework of ALM, the primal subproblem usually can only be solved inexactly, and thus practically inexact ALM (iALM) is often used. However, to the best of our knowledge, the global convergence rate of iALM for problems with nonlinear inequality constraints still remains open.[1] We address this open question in this work and also establish the iteration complexity of iALM in terms of the number of gradient evaluations.

We will assume composite convex structure on (1.1). More specifically, we assume

$$f_0(\mathbf{x})=g(\mathbf{x})+h(\mathbf{x}), \tag{1.2}$$

where g is a differentiable convex function with Lipschitz continuous gradient, and h is a simple[2] (possibly nondifferentiable) closed convex function. Also, f_i is convex

[1] Although the global convergence rate in terms of augmented dual objective can be easily shown from existing works (e.g., see our discussion in Sect. 5), that does not indicate the convergence speed from the perspective of the primal objective and feasibility.

[2] By "simple", we mean the proximal mapping of h is easy to evaluate, i.e., it is easy to find a solution to $\min_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x})+\frac{1}{2\gamma}\|\mathbf{x}-\hat{\mathbf{x}}\|^2$ for any $\hat{\mathbf{x}}$ and $\gamma>0$.

and differentiable with Lipschitz continuous gradient for each $i = 1, \ldots, m$, namely, there are constants $L_0, L_1, \ldots, L_m$ such that

$$\|\nabla g(\hat{\mathbf{x}}) - \nabla g(\tilde{\mathbf{x}})\| \le L_0 \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|, \forall\, \hat{\mathbf{x}}, \tilde{\mathbf{x}} \in \mathrm{dom}(h) \cap \mathcal{X}, \tag{1.3a}$$

$$\|\nabla f_i(\hat{\mathbf{x}}) - \nabla f_i(\tilde{\mathbf{x}})\| \le L_i \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|, \forall\, \hat{\mathbf{x}}, \tilde{\mathbf{x}} \in \mathrm{dom}(h) \cap \mathcal{X}, \forall\, i = 1, \ldots, m. \tag{1.3b}$$

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

$$D = \underset{\hat{\mathbf{x}}, \tilde{\mathbf{x}} \in \mathrm{dom}(h) \cap \mathcal{X}}{\text{maximize}} \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|. \tag{1.4}$$

1.1 Augmented Lagrangian function

In the literature, there are several different penalty terms used in an augmented Lagrangian (AL) function, such as the classic one [37,38], the quadratic penalty on constraint violation [4], and the exponential penalty [41]. The work [3] 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) to an equivalent form:

$$\underset{\mathbf{x} \in \mathcal{X}, \mathbf{s} \ge \mathbf{0}}{\text{minimize}}\ f_0(\mathbf{x}),\ \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{b},\ f_i(\mathbf{x}) + s_i = 0, i = 1, \ldots, m. \tag{1.5}$$

With quadratic penalty on the equality constraints, the AL function of (1.5) is

$$\begin{aligned} \tilde{\mathcal{L}}_\beta(\mathbf{x}, \mathbf{s}, \mathbf{y}, \mathbf{z}) = {} & f_0(\mathbf{x}) + \mathbf{y}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m z_i \big(f_i(\mathbf{x}) + s_i\big) \\ & + \frac{\beta}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \frac{\beta}{2} \sum_{i=1}^m \big(f_i(\mathbf{x}) + s_i\big)^2, \end{aligned} \tag{1.6}$$

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

$$s_i = \max\left(0, -\frac{z_i}{\beta} - f_i(\mathbf{x})\right),\ i = 1, \ldots, m.$$

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

$$\tilde{\mathcal{L}}_\beta(\mathbf{x}, \mathbf{s}, \mathbf{y}, \mathbf{z}) = f_0(\mathbf{x}) + \mathbf{y}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) + \frac{\beta}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \sum_{i=1}^m \psi_\beta(f_i(\mathbf{x}), z_i),$$

where

$$\psi_\beta(u,v)=\begin{cases} uv+\frac{\beta}{2}u^2, & \text{if } \beta u+v\ge 0,\\ -\frac{v^2}{2\beta}, & \text{if } \beta u+v<0.\end{cases} \tag{1.7}$$

Let

$$\Psi_\beta(\mathbf{x},\mathbf{z})=\sum_{i=1}^m \psi_\beta(f_i(\mathbf{x}),z_i),$$

and we obtain the classic AL function of (1.1):

$$\mathcal{L}_\beta(\mathbf{x},\mathbf{y},\mathbf{z})=f_0(\mathbf{x})+\mathbf{y}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})+\frac{\beta}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\Psi_\beta(\mathbf{x},\mathbf{z}). \tag{1.8}$$

The AL function in (1.8) has an important advantage over that in (1.6). The former AL function is convex about the primal variable and concave about the dual variable while that in (1.6) may not be convex about the primal variable. [37, Theorem 3.1] shows that $\mathcal{L}_\beta$ given in (1.8) is convex about $\mathbf{x}$ and concave about $\mathbf{z}$. For completeness, we include a different and short proof here.

Lemma 1 *Assume f_i to be convex for each $i=0,1,\ldots,m$. Then the AL function $\mathcal{L}_\beta$ in* (1.8) *is convex about* $\mathbf{x}$ *and concave about* $(\mathbf{y},\mathbf{z})$.

Proof We only need to show the convexity-concavity of $\Psi_\beta(\mathbf{x},\mathbf{z})$ in $\mathbf{x}$ and $\mathbf{z}$. It is easy to see that $\psi_\beta(u,v)$ in (1.7) is nondecreasing and convex about u and concave about v. Hence, given $\mathbf{x}$, the function $\psi_\beta(f_i(\mathbf{x}),z_i)$ is concave about z_i for each $i=1,\ldots,m$, and thus $\Psi_\beta(\mathbf{x},\mathbf{z})$ is concave about $\mathbf{z}$. To show the convexity of Ψ_β about $\mathbf{x}$, we note that the composition of a nondecreasing convex function with a convex function is still convex; cf. [9, Eq. (3.11)]. Therefore, given $\mathbf{z}$, $\psi_\beta(f_i(\mathbf{x}),z_i)$ is convex about $\mathbf{x}$ for each $i=1,\ldots,m$, and thus $\Psi_\beta(\mathbf{x},\mathbf{z})$ is convex about $\mathbf{x}$. This completes the proof. □

1.2 Inexact augmented Lagrangian method

The augmented Lagrangian method (ALM) was proposed in [19,36]. Within each iteration, ALM first updates the $\mathbf{x}$ variable by minimizing the AL function with respect to $\mathbf{x}$ while fixing $\mathbf{y}$ and $\mathbf{z}$, and then it performs a dual gradient ascent update to $\mathbf{y}$ and $\mathbf{z}$. In general, it is difficult to exactly minimize the AL function about $\mathbf{x}$. A more realistic way is to solve the $\mathbf{x}$-subproblem within a tolerance error, which leads to the inexact ALM. Its pseudocode is given in Algorithm 1 below. If $\varepsilon_k=0$, $\forall k$, it reduces to the ALM.

Note that Algorithm 1 is a framework of iALM since it does not specify how to find $\mathbf{x}^{k+1}$. For problems that have the structure given in (1.2) and (1.3), we will apply an optimal first-order method as a subroutine to inexactly solve each subproblem. In addition, the inequality in (1.9) generally cannot be directly verified. However, it can be guaranteed by setting appropriate stopping conditions such as running the subroutine to a theoretically derived maximum number of iterations or until

Algorithm 1: Inexact augmented Lagrangian method for (1.1)

```
Initialization: choose $\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0$, a positive integer $K$, and $\{\beta_k, \rho_k, \varepsilon_k\}$
for $k = 0, 1, \ldots, K-1$ do
    Find $\mathbf{x}^{k+1} \in \mathcal{X}$ such that
        $\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{y}^k, \mathbf{z}^k) \leq \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \varepsilon_k.$   (1.9)
    Update $\mathbf{y}$ and $\mathbf{z}$ by
        $\mathbf{y}^{k+1} = \mathbf{y}^k + \rho_k(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}),$   (1.10)
        $z_i^{k+1} = z_i^k + \rho_k \cdot \max\left(-\frac{z_i^k}{\beta_k}, f_i(\mathbf{x}^{k+1})\right), i = 1, \ldots, m.$   (1.11)
```

$$\text{dist}\left(\mathbf{0}, \partial_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{y}^k, \mathbf{z}^k) + \mathcal{N}_{\mathcal{X}}(\mathbf{x}^{k+1})\right) \leq \frac{\varepsilon_k}{D},$$

where D is given in (1.4), and $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ is the normal cone of $\mathcal{X}$ at $\mathbf{x}$.

It is shown in [37] that the augmented dual function[3]

$$d_\beta(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \mathbf{z}) \tag{1.12}$$

is continuously differentiable, and ∇d_β is Lipschitz continuous with constant $\frac{1}{\beta}$. In addition, it turns out that the (inexact) ALM is an (inexact) augmented dual gradient ascent [38], and thus convergence rate of the (inexact) ALM in term of d_β can be shown from existing results about (inexact) gradient method [40]. However, directly applying these existing results would require $\sum_{k \geq 0} \sqrt{\varepsilon_k} < \infty$. Our analysis will be different from this line and only needs $\sum_{k \geq 0} \varepsilon_k < \infty$. Our results will be based on both the primal and augmented dual problems.

1.3 Main results

The main results we establish in this paper are summarized as follows. Both ergodic and nonergodic convergence rate results are established. Here, ergodic convergence rate is based on averaged iterates while nonergodic one is about the actual iterates.

Theorem 1 (Summary of main results) *For a given $\varepsilon > 0$, choose a positive integer K and numbers $C_1 > 0$, $C_2 > 0$. Let $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^K$ be the iterates generated from Algorithm 1 with parameters set according to one of the follows:*

(i) $\rho_k = \beta_k = \frac{C_1}{K\varepsilon}$, $\varepsilon_k = \frac{\varepsilon}{2} \frac{C_2}{C_1}$, $\forall k$.

(ii) $\rho_k = \beta_k = \beta_0 \sigma^k$, $\forall k$ *for certain* $\beta_0 > 0$ *and* $\sigma > 1$ *such that* $\sum_{k=0}^{K-1} \beta_k = \frac{C_1}{\varepsilon}$, *and* $\varepsilon_k = \frac{\varepsilon}{2} \frac{C_2}{C_1}$, $\forall k$.

[3] Although [37] only considers the inequality constrained case, the results derived there apply to the case with both equality and inequality constraints.

(iii) *$\rho_k = \beta_k = \beta_0\sigma^k, \forall k$ for certain $\beta_0 > 0$ and $\sigma > 1$ such that $\sum_{k=0}^{K-1}\beta_k = \frac{C_1}{\varepsilon}$. If f_0 is convex, let $\varepsilon_k = \frac{C_2}{2\beta_k^{\frac{1}{3}}}\frac{1}{\sum_{t=0}^{K-1}\beta_t^{\frac{2}{3}}}, \forall k$, and if f_0 is strongly convex, let $\varepsilon_k = \frac{C_2}{2\beta_k^{\frac{1}{2}}}\frac{1}{\sum_{t=0}^{K-1}\beta_t^{\frac{1}{2}}}, \forall k$.*

Then we have the following results:

(a) *For each setting, the averaged point $\bar{\mathbf{x}}^K = \sum_{k=0}^{K-1}\frac{\rho_k\mathbf{x}^{k+1}}{\sum_{t=0}^{K-1}\rho_t}$ is a primal $O(\varepsilon)$-solution (see Definition 1), where the hidden constant depends on C_1, C_2 and dual solution $(\mathbf{y}^*, \mathbf{z}^*)$.*

(b) *For the second and third settings, the actual point $\mathbf{x}^K$ is also a primal $O(\varepsilon)$-solution.*

(c) *For each setting, to obtain the iterates, the total number of evaluations on ∇g and $\nabla f_i, i = 1, \ldots, m$ is $O(\sqrt{K}\varepsilon^{-1} + K\varepsilon^{-\frac{1}{2}})$ if f_0 is convex and $O(K + \sqrt{K}\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ if f_0 is strongly convex.*

(d) *For the first setting, without linear equality constraint, additional t_K gradient evaluations can guarantee to produce $\bar{\mathbf{z}}^{K+}$ such that $(\bar{\mathbf{x}}^K, \bar{\mathbf{z}}^{K+})$ is a primal-dual $O(\varepsilon)$-solution (see Definition 2), where $t_K = O(\frac{1}{K\varepsilon^2})$ if f_0 is convex and $t_K = O(\sqrt{K}\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ if f_0 is strongly convex.*

For the primal ε-solution, the formal statements and the hidden constants are shown in Theorem 5 for the first setting, in Theorems 6 and 10 for the second setting, and in Theorems 7 and 10 for the third setting. The formal statement for the primal-dual ε-solution is given in Theorem 8. We make a few remarks here. First, the integer K could be independent of ε. When $K = 1$, Algorithm 1 solves a single penalized problem and reduces to a penalty method if $\mathbf{y}^0 = \mathbf{0}$ and $\mathbf{z}^0 = \mathbf{0}$. Although the number of gradient evaluations is smallest in item (c) if $K = 1$, numerically we observe better performance by choosing a larger K. Second, as K is independent of ε, the iteration complexity to obtain a primal ε-solution is $O(\varepsilon^{-1})$ for the convex case and $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ for the strongly convex case. The order for the convex case matches with the lower complexity bound established in [35] and thus is optimal. For the strongly convex case, Ouyang and Xu [35] gives a lower bound in the order of $\varepsilon^{-\frac{1}{2}}$, and thus our result is nearly optimal. Third, in item (d), we set $K = O(\varepsilon^{-\frac{2}{3}})$ if f_0 is convex and K independent of ε if f_0 is strongly convex. Therefore, to have a primal-dual ε-solution, the iteration complexity is $O(\varepsilon^{-\frac{4}{3}})$ for the convex case and $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ for the strongly convex case.

1.4 Literature review

In this section, we review related works. Our review focuses on convex optimization, but note that ALM has also been popularly applied to nonconvex optimization problems; see [5–7] 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:

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\ f_0(\mathbf{x}),\ \text{ s.t. } \mathbf{A}\mathbf{x}=\mathbf{b}. \tag{1.13}$$

Assuming exact solution to every $\mathbf{x}$-subproblem, He and Yuan [17] 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 Kang et al. [21] that uses similar technique. By adapting parameters, Xu [42] 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 $\mathbf{x}$-subproblems.

When the objective is strongly convex, Kang et al. [20] proves $O(1/k^2)$ convergence of iALM 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 on the total number of gradient evaluations on solving all subproblems to the required accuracies.

For smooth linearly constrained convex problems, Lan and Monteiro [22] analyzes the iteration complexity of the iALM. It applies Nesterov's optimal first-order method to every $\mathbf{x}$-subproblem and shows that $O(\varepsilon^{-\frac{7}{4}})$ gradient evaluations are required to reach a primal-dual ε-solution $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ in the sense that $\|\mathbf{A}\bar{\mathbf{x}}-\mathbf{b}\| \le \varepsilon$ and $\nabla f_0(\bar{\mathbf{x}}) + \mathbf{A}^\top\bar{\mathbf{y}} \in -\mathcal{N}_{\mathcal{X}}(\bar{\mathbf{x}}) + \mathcal{B}_\varepsilon$, where $\mathcal{B}_\varepsilon$ denotes an ε-ball centered at origin. In addition, Lan and Monteiro [22] modifies the iALM by solving a perturbed problem. The modified iALM requires $O(\varepsilon^{-1}|\log\varepsilon|^{\frac{3}{4}})$ gradient evaluations to produce a primal-dual ε-solution. Motivated by the model predictive control, Nedelcu et al. [29] also analyzes the iteration complexity of inexact dual gradient methods (iDGM) that are essentially iALMs. While the iteration complexity in Lan and Monteiro [22] is estimated based on the best iterate, and that in Nedelcu et al. [29] is ergodic, the recent work [25] establishes non-ergodic convergence of iALM.

Another line of existing works on iALM 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 [8,10–12,18,34,44,46] and the references therein.

General convex problems As there are nonlinear inequality constraints, the local convergence rate of iALM has been extensively studied (e.g., [4,37,39]). However, at the time of our first submission, we did not find any work in the literature showing its global convergence rate. 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., [28,30,31]). For general convex problems, these methods enjoy $O(1/\sqrt{k})$ convergence, and for strongly convex case, the rate can be improved to $O(1/k)$. To achieve a primal-dual ε-solution, compared to our results, their iteration complexity is $O(\varepsilon^{-\frac{2}{3}})$ times worse for the convex problems and $O(\varepsilon^{-\frac{1}{2}})$ worse for the strongly convex problems. Assuming Lipschitz continuity of f_i for every $i \in [m]$, [48] proposes a new primal-dual type algorithm for nonlinearly constrained convex programs. 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 [47] focuses on smooth constrained convex problems and proposes a linearized variant of the algorithm in [48]. Assuming compactness of the set $\mathcal{X}$, it also establishes $O(1/k)$ ergodic convergence of the linearized method. Since our first submission, a few works have been done on first-order methods for solving nonlinear functional constrained problems. For example, [26] analyzes the iteration complexity of first-order iALM and a modified version for convex conic programming, and [16] proposed a first-order primal-dual method for general convex-concave saddle-point problems.

1.5 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 use italic letters $a, c, B, L, \ldots$, for scalars, bold lower-case letters $\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$ for vectors, and bold upper-case letters $\mathbf{A}, \mathbf{B}, \ldots$ for matrices. z_i denotes the i-th entry of a vector $\mathbf{z}$. We use $\mathbf{0}$ to denote a vector or matrix of all zeros, and its size is clear from the context. $[m]$ denotes the set $\{1, 2, \ldots, m\}$ for any positive integer m. Given a real number a, we let $[a]_+ = \max(0, a)$ and $\lceil a \rceil$ be the smallest integer that is no less than a. For a vector $\mathbf{a}$, $[\mathbf{a}]_+$ takes the positive part of $\mathbf{a}$ in a component-wise manner. $\|\mathbf{a}\|$ denotes the Euclidean norm of a vector $\mathbf{a}$ and $\|\mathbf{A}\|$ the spectral norm of a matrix $\mathbf{A}$.

We denote $\boldsymbol{\ell}$ as the vector consisting of $L_i, i \in [m]$, where L_i is the Lipschitz constant of ∇f_i in (1.3b). Also we let $\mathbf{f}$ be the vector function with f_i as the i-th component scalar function. That is

$$\boldsymbol{\ell} = [L_1, \ldots, L_m], \quad \mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})]. \tag{1.14}$$

Given a convex function f, $\tilde{\nabla} f(\mathbf{x})$ represents one subgradient of f at $\mathbf{x}$, namely,

$$f(\hat{\mathbf{x}}) \ge f(\mathbf{x}) + \langle \tilde{\nabla} f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x} \rangle, \ \forall \hat{\mathbf{x}},$$

and $\partial f(\mathbf{x})$ denotes its subdifferential, i.e., the set of all subgradients. When f is differentiable, we simply write its subgradient as $\nabla f(\mathbf{x})$. For a convex set $\mathcal{X}$, we use $\iota_{\mathcal{X}}$ as its indicator function, i.e.,

$$\iota_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{X}, \\ +\infty, & \text{if } \mathbf{x} \notin \mathcal{X}, \end{cases} \tag{1.15}$$

and $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \partial \iota_{\mathcal{X}}(\mathbf{x})$ as its normal cone at $\mathbf{x} \in \mathcal{X}$.

1.6 Outline

The rest of the paper is organized as follows. In Sect. 2, we give a few preparatory results and review Nesterov's optimal first-order method for solving a composite convex program. An ergodic convergence rate result of iALM is given in Sect. 3, and a nonergodic convergence rate result is shown in Sect. 4. Iteration complexity results in terms of the number of gradient evaluations are established for both ergodic and nonergodic cases. Comparison to several existing works is given in Sect. 5, and numerical results are provided in Sect. 6. Finally Sect. 7 concludes the paper.

2 Preliminary results and Nesterov's optimal first-order method

In this section, we give a few preliminary results and also review Nesterov's optimal first-order method for composite convex programs.

2.1 ε-Solutions and basic facts

Given an $\varepsilon > 0$, the primal ε-solution of (1.1) is defined as follows.

Definition 1 (*primal ε-solution*) Let f_0^* be the optimal value of (1.1). Given $\varepsilon \geq 0$, a point $\mathbf{x} \in \mathcal{X}$ is called a primal ε-solution to (1.1) if

$$|f_0(\mathbf{x}) - f_0^*| \leq \varepsilon, \text{ and } \|\mathbf{Ax} - \mathbf{b}\| + \left\|[\mathbf{f}(\mathbf{x})]_+\right\| \leq \varepsilon.$$

The above definition is not new. For linearly constrained problems, Lin et al. [24] adopts a similar definition, and for general nonlinearly constrained problems, Rockafellar, Yu and Neely [39,48] also use the objective distance and feasibility violation to measure the solution quality.

A point $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ satisfies the Karush-Kuhn-Tucker (KKT) conditions for (1.1) if

$$\mathbf{0} \in \partial f_0(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x}) + \mathbf{A}^\top \mathbf{y} + \sum_{i=1}^m z_i \nabla f_i(\mathbf{x}), \tag{2.1a}$$

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{x} \in \mathcal{X}, \tag{2.1b}$$

$$z_i \geq 0, \quad f_i(\mathbf{x}) \leq 0, \quad z_i f_i(\mathbf{x}) = 0, \forall i \in [m]. \tag{2.1c}$$

From the convexity of f_i's, if $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ satisfies the conditions in (2.1), then [43]

$$f_0(\mathbf{x}) - f_0(\mathbf{x}^*) + \langle \mathbf{y}^*, \mathbf{Ax} - \mathbf{b} \rangle + \sum_{i=1}^m z_i^* f_i(\mathbf{x}) \geq 0, \ \forall \mathbf{x} \in \mathcal{X}. \tag{2.2}$$

For any primal feasible point $\mathbf{x}$ of (1.1) and any $(\mathbf{y}, \mathbf{z})$ with $\mathbf{z} \geq \mathbf{0}$, one can easily show the weak duality inequality $d_0(\mathbf{y}, \mathbf{z}) \leq f_0(\mathbf{x})$, where

$$d_0(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in \mathcal{X}} f_0(\mathbf{x}) + \langle \mathbf{y}, \mathbf{Ax} - \mathbf{b} \rangle + \sum_{i=1}^m z_i f_i(\mathbf{x})$$

is the Lagrangian dual function. As a KKT point $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ exists, we have $d_0(\mathbf{y}^*, \mathbf{z}^*) = f_0(\mathbf{x}^*)$, i.e., the strong duality holds. In this case, we define the primal-dual ε-solution of (1.1) as follows.

Definition 2 (*primal-dual ε-solution*) Given $\varepsilon \geq 0$, a point $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ with $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \geq \mathbf{0}$ is called a primal-dual ε-solution to (1.1) if $\mathbf{x}$ is a primal ε-solution and in addition $f_0^* \leq d_0(\mathbf{y}, \mathbf{z}) + \varepsilon$, where f_0^* is the optimal value of (1.1).

The result below will be used to establish convergence rate results of Algorithm 1.

Lemma 2 *Assume $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ satisfies the KKT conditions in (2.1). Let $\bar{\mathbf{x}}$ be a point such that for any $\mathbf{y}$ and any $\mathbf{z} \geq \mathbf{0}$,*

$$f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) + \mathbf{y}^\top(\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}) + \sum_{i=1}^m z_i f_i(\bar{\mathbf{x}}) \leq \alpha + c_1\|\mathbf{y}\|^2 + c_2\|\mathbf{z}\|^2, \tag{2.3}$$

where α and c_1, c_2 are nonnegative constants independent of $\mathbf{y}$ and $\mathbf{z}$. Then

$$-\left(\alpha + 4c_1\|\mathbf{y}^*\|^2 + 4c_2\|\mathbf{z}^*\|^2\right) \leq f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) \leq \alpha, \tag{2.4}$$

$$\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| + \left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\| \leq \alpha + c_1\big(1 + \|\mathbf{y}^*\|\big)^2 + c_2\big(1 + \|\mathbf{z}^*\|\big)^2. \tag{2.5}$$

Proof Letting $\mathbf{y} = \mathbf{0}$ and $\mathbf{z} = \mathbf{0}$ in (2.3) gives the second inequality in (2.4). For any nonnegative γ_y and γ_z, we let

$$\mathbf{y} = \gamma_y \frac{\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}}{\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\|}, \quad \mathbf{z} = \gamma_z \frac{[\mathbf{f}(\bar{\mathbf{x}})]_+}{\left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\|}$$

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

$$f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) + \gamma_y\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| + \gamma_z\left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\| \leq \alpha + c_1\gamma_y^2 + c_2\gamma_z^2. \tag{2.6}$$

Noting

$$-\langle \mathbf{y}^*, \mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\rangle \geq -\|\mathbf{y}^*\| \cdot \|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\|, \quad -\sum_{i=1}^m z_i^* f_i(\bar{\mathbf{x}}) \geq -\|\mathbf{z}^*\| \cdot \left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\|, \tag{2.7}$$

we have from (2.2) and (2.6) that

$$(\gamma_y - \|\mathbf{y}^*\|)\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| + (\gamma_z - \|\mathbf{z}^*\|)\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\| \leq \alpha + c_1\gamma_y^2 + c_2\gamma_z^2$$

In the above inequality, letting $\gamma_y = 1 + \|\mathbf{y}^*\|$ and $\gamma_z = 1 + \|\mathbf{z}^*\|$ gives (2.5), and letting $\gamma_y = 2\|\mathbf{y}^*\|$ and $\gamma_z = 2\|\mathbf{z}^*\|$ gives the first inequality in (2.4) by (2.2) and (2.7). □

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 **x**-subproblems in Algorithm 1. It aims at finding a solution of the following problem

$$\underset{\mathbf{x}}{\text{minimize}}\, \phi(\mathbf{x}) + \psi(\mathbf{x}). \tag{2.8}$$

Here, ϕ is L_ϕ-smooth, i.e., $\nabla\phi$ is Lipschitz continuous with constant L_ϕ, and ϕ is also strongly convex with modulus $\mu \geq 0$. In addition, ψ is a simple (possibly nondifferentiable) closed convex function. Algorithm 2 summarizes the method. Here, for simplicity, we assume L_ϕ and μ are known. The method does not require the value of L_ϕ but can estimate a local Lipschitz constant by backtracking. In addition, it only requires a lower estimate of μ; see [33] for example.

Algorithm 2: Nesterov's optimal first-order method for (2.8)

```
Initialization: choose $\hat{\mathbf{x}}^0 = \mathbf{x}^0$, $\alpha_0 \in (0, 1]$, and let $q = \frac{\mu}{L_\phi}$;
for $k = 0, 1, \ldots,$ do
    Let
        $$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \langle \nabla\phi(\hat{\mathbf{x}}^k), \mathbf{x}\rangle + \frac{L_\phi}{2}\|\mathbf{x} - \hat{\mathbf{x}}^k\|^2 + \psi(\mathbf{x}).$$
    Set
        $$\alpha_{k+1} = \frac{q - \alpha_k^2 + \sqrt{(q - \alpha_k^2)^2 + 4\alpha_k^2}}{2},$$
    and
        $$\hat{\mathbf{x}}^{k+1} = \mathbf{x}^{k+1} + \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}(\mathbf{x}^{k+1} - \mathbf{x}^k).$$
```

The theorem below gives the convergence rate of Algorithm 2 for both convex (i.e., $\mu = 0$) and strongly convex (i.e., $\mu > 0$) cases; see [2,32,33]. We will use the results to estimate iteration complexity of iALM.

Theorem 2 *Let $\{\mathbf{x}^k\}$ be the sequence generated from Algorithm 2. Assume $\mathbf{x}^*$ to be a minimizer of (2.8). The following results holds:*

1. *If $\mu = 0$ and $\alpha_0 = 1$, then*

$$\phi(\mathbf{x}^k) + \psi(\mathbf{x}^k) - \phi(\mathbf{x}^*) - \psi(\mathbf{x}^*) \leq \frac{2L_\phi\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{k^2}, \ \forall k \geq 1. \tag{2.9}$$

2. *If* $\mu > 0$ *and* $\alpha_0 = \sqrt{\frac{\mu}{L_\phi}}$, *then*

$$\phi(\mathbf{x}^k)+\psi(\mathbf{x}^k)-\phi(\mathbf{x}^*)-\psi(\mathbf{x}^*) \le \frac{(L_\phi+\mu)\|\mathbf{x}^0-\mathbf{x}^*\|^2}{2}\left(1-\sqrt{\frac{\mu}{L_\phi}}\right)^k, \ \forall k \ge 1. \tag{2.10}$$

3 Ergodic convergence rate and iteration complexity results for primal ε-solutions

In this section, we first establish an ergodic convergence rate result of Algorithm 1. From that result, we then specify algorithm parameters and estimate the total number of gradient evaluations in order to produce a primal ε-solution. Two different settings of the penalty parameters are studied: one with constant penalty and another with geometrically increasing penalty parameters. For each setting, the tolerance error parameter ε_k is chosen in an "optimal" way so that the total number of gradient evaluations is minimized.

Throughout this section, we make the following assumptions.

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

Assumption 2 For every k, there is $\mathbf{x}^{k+1}$ satisfying (1.9).

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

3.1 Convergence rate analysis of iALM

To show the convergence results of Algorithm 1, we first establish a few lemmas.

Lemma 3 *Let* $\{(\mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^K$ *be the sequence obtained from the updates* (1.10) *and* (1.11). *Then for any* $(\mathbf{y}, \mathbf{z})$ *and any* $0 \le k < K$, *it holds*

$$\frac{1}{2\rho_k}\left[\|\mathbf{y}^{k+1}-\mathbf{y}\|^2-\|\mathbf{y}^k-\mathbf{y}\|^2+\|\mathbf{y}^{k+1}-\mathbf{y}^k\|^2\right]-\langle \mathbf{y}^{k+1}-\mathbf{y}, \mathbf{r}^{k+1}\rangle = 0, \tag{3.1}$$

$$\frac{1}{2\rho_k}\left[\|\mathbf{z}^{k+1}-\mathbf{z}\|^2-\|\mathbf{z}^k-\mathbf{z}\|^2+\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2\right] - \sum_{i=1}^m (z_i^{k+1}-z_i)\cdot \max\left(-\frac{z_i^k}{\beta_k}, f_i(\mathbf{x}^{k+1})\right) = 0, \tag{3.2}$$

where $\mathbf{r}^k = \mathbf{Ax}^k - \mathbf{b}$.

Proof From (1.10), it follows that

$$\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \frac{1}{\rho_k}(\mathbf{y}^{k+1} - \mathbf{y}^k) - \mathbf{r}^{k+1} \right\rangle = 0.$$

Using the equality $2\mathbf{u}^\top \mathbf{v} = \|\mathbf{u}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2 + \|\mathbf{v}\|^2$, we have the result in (3.1). By similar arguments, one can show (3.2). □

Lemma 4 *For any* $\mathbf{z} \geq \mathbf{0}$, *we have*

$$\sum_{i=1}^m \left([z_i^k + \beta_k f_i(\mathbf{x}^{k+1})]_+ - z_i\right) f_i(\mathbf{x}^{k+1}) - \sum_{i=1}^m (z_i^{k+1} - z_i) \cdot \max\left(-\frac{z_i^k}{\beta_k}, f_i(\mathbf{x}^{k+1})\right)$$
$$\geq \frac{1}{\rho_k^2}(\beta_k - \rho_k)\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2. \tag{3.3}$$

Proof Denote

$$I_+^k = \{i \in [m] : z_i^k + \beta_k f_i(\mathbf{x}^{k+1}) \geq 0\}, \quad I_-^k = [m] \backslash I_+^k. \tag{3.4}$$

Then

$$\begin{aligned}
&\text{the left hand side of (3.3)}\\
&= \sum_{i \in I_+^k} \left[(z_i^k - z_i) f_i(\mathbf{x}^{k+1}) + \beta_k [f_i(\mathbf{x}^{k+1})]^2 - \left(z_i^k + \rho_k f_i(\mathbf{x}^{k+1}) - z_i\right) f_i(\mathbf{x}^{k+1})\right]\\
&\quad + \sum_{i \in I_-^k} \left[-z_i f_i(\mathbf{x}^{k+1}) - \left(z_i^k - \frac{\rho_k z_i^k}{\beta_k} - z_i\right)\left(-\frac{z_i^k}{\beta_k}\right)\right]\\
&= (\beta_k - \rho_k) \sum_{i \in I_+^k} [f_i(\mathbf{x}^{k+1})]^2 + \sum_{i \in I_-^k} \left[-z_i \left(f_i(\mathbf{x}^{k+1}) + \frac{z_i^k}{\beta_k}\right) + \frac{1}{\beta_k^2}(\beta_k - \rho_k)(z_i^k)^2\right]\\
&\geq (\beta_k - \rho_k) \sum_{i \in I_+^k} [f_i(\mathbf{x}^{k+1})]^2 + \frac{1}{\beta_k^2}(\beta_k - \rho_k) \sum_{i \in I_-^k} (z_i^k)^2\\
&= \frac{1}{\rho_k^2}(\beta_k - \rho_k)\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2,
\end{aligned}$$

where the inequality follows from $z_i \geq 0$ and $f_i(\mathbf{x}^{k+1}) + \frac{z_i^k}{\beta_k} \leq 0$, $\forall i \in I_-^k$, and the last equality holds due to the update (1.11). □

The next theorem is a fundamental result by running one iteration of Algorithm 1.

Theorem 3 (One-iteration progress of iALM) *Let $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}$ be the sequence generated from Algorithm 1. Then for any $\mathbf{x} \in \mathcal{X}$, any $\mathbf{y}$, and any $\mathbf{z} \geq \mathbf{0}$, it holds that*

$$\begin{aligned}
& f_0(\mathbf{x}^{k+1}) + \mathbf{y}^\top \mathbf{r}^{k+1} + \sum_{i=1}^m z_i f_i(\mathbf{x}^{k+1}) + \frac{\beta_k - \rho_k}{2} \|\mathbf{r}^{k+1}\|^2 \\
& \quad + \frac{\beta_k - \rho_k}{2\rho_k^2} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \frac{1}{2\rho_k} \|\mathbf{y}^{k+1} - \mathbf{y}\|^2 + \frac{1}{2\rho_k} \|\mathbf{z}^{k+1} - \mathbf{z}\|^2 \\
& \quad \leq \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \frac{1}{2\rho_k} \|\mathbf{y}^k - \mathbf{y}\|^2 + \frac{1}{2\rho_k} \|\mathbf{z}^k - \mathbf{z}\|^2 + \varepsilon_k. \qquad (3.5)
\end{aligned}$$

Proof From (1.9), it follows that for any $\mathbf{x} \in \mathcal{X}$,

$$f_0(\mathbf{x}^{k+1}) + \langle \mathbf{y}^k, \mathbf{r}^{k+1} \rangle + \frac{\beta_k}{2} \|\mathbf{r}^{k+1}\|^2 + \Psi_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) \leq \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \varepsilon_k. \quad (3.6)$$

Since $\langle \mathbf{y}^k, \mathbf{r}^{k+1} \rangle = \langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{r}^{k+1} \rangle + \langle \mathbf{y}, \mathbf{r}^{k+1} \rangle - \rho_k \|\mathbf{r}^{k+1}\|^2$, by adding (3.1) and (3.2) to the above inequality, we have

$$\begin{aligned}
& f_0(\mathbf{x}^{k+1}) + \mathbf{y}^\top \mathbf{r}^{k+1} + \sum_{i=1}^m z_i f_i(\mathbf{x}^{k+1}) + \sum_{i=1}^m \left([z_i^k + \beta_k f_i(\mathbf{x}^{k+1})]_+ - z_i \right) f_i(\mathbf{x}^{k+1}) \\
& \quad + \left(\frac{\beta_k}{2} - \rho_k \right) \|\mathbf{r}^{k+1}\|^2 + \Psi_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) - \sum_{i=1}^m [z_i^k + \beta_k f_i(\mathbf{x}^{k+1})]_+ f_i(\mathbf{x}^{k+1}) \\
& \quad + \frac{1}{2\rho_k} \left[\|\mathbf{y}^{k+1} - \mathbf{y}\|^2 - \|\mathbf{y}^k - \mathbf{y}\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 \right] \\
& \quad + \frac{1}{2\rho_k} \left[\|\mathbf{z}^{k+1} - \mathbf{z}\|^2 - \|\mathbf{z}^k - \mathbf{z}\|^2 + \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \right] \\
& \quad - \sum_{i=1}^m (z_i^{k+1} - z_i) \cdot \max\left(-\frac{z_i^k}{\beta_k}, f_i(\mathbf{x}^{k+1}) \right) \\
& \leq \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \varepsilon_k. \qquad (3.7)
\end{aligned}$$

Note that

$$\begin{aligned}
& \Psi_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) - \sum_{i=1}^m [z_i^k + \beta_k f_i(\mathbf{x}^{k+1})]_+ f_i(\mathbf{x}^{k+1}) \\
& \quad = \sum_{i \in I_+^k} \left[z_i^k f_i(\mathbf{x}^{k+1}) + \frac{\beta_k}{2} [f_i(\mathbf{x}^{k+1})]^2 - [z_i^k + \beta_k f_i(\mathbf{x}^{k+1})] f_i(\mathbf{x}^{k+1}) \right] \\
& \qquad + \sum_{i \in I_-^k} \left[-\frac{(z_i^k)^2}{2\beta_k} \right]
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i\in I_+^k} \frac{\beta_k}{2}[f_i(\mathbf{x}^{k+1})]^2 - \sum_{i\in I_-^k} \frac{(z_i^k)^2}{2\beta_k} \\
&= -\frac{\beta_k}{2\rho_k^2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2, \qquad (3.8)
\end{aligned}
$$

where the sets I_+^k and I_-^k are defined in (3.4). Hence, plugging (3.3) and (3.8) into (3.7) yields (3.5). □

By Lemma 2 and Theorem 3, we have the following convergence rate estimate of Algorithm 1.

Theorem 4 (Ergodic convergence rate of iALM) *Under Assumptions 1 and 2, let* $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^K$ *be the sequence generated from Algorithm 1 with* $\mathbf{y}^0 = \mathbf{0}, \mathbf{z}^0 = \mathbf{0}$ *and* $0 < \rho_k \le \beta_k$, $\forall k$. *Then*

$$
\left|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\right| \le \frac{1}{\sum_{t=0}^{K-1}\rho_t}\left(2\|\mathbf{y}^*\|^2 + 2\|\mathbf{z}^*\|^2 + \sum_{k=0}^{K-1}\rho_k\varepsilon_k\right), \qquad (3.9a)
$$

$$
\|\mathbf{A}\bar{\mathbf{x}}^K - \mathbf{b}\| + \left\|[\mathbf{f}(\bar{\mathbf{x}}^K)]_+\right\| \le \frac{1}{\sum_{t=0}^{K-1}\rho_t}\left(\frac{(1+\|\mathbf{y}^*\|)^2}{2} + \frac{(1+\|\mathbf{z}^*\|)^2}{2} + \sum_{k=0}^{K-1}\rho_k\varepsilon_k\right). \qquad (3.9b)
$$

In addition, if $\beta_k = \beta$ *and* $\rho_k = \rho, \forall k \ge 0$, *then*

$$
f_0(\mathbf{x}^*) - d_\beta(\bar{\mathbf{y}}^K, \bar{\mathbf{z}}^K) \le \frac{2}{K}\left(\frac{1}{\rho}\|\mathbf{y}^*\|^2 + \frac{1}{\rho}\|\mathbf{z}^*\|^2 + \sum_{k=0}^{K-1}\varepsilon_k\right). \qquad (3.10)
$$

In the above,

$$
\bar{\mathbf{x}}^K = \frac{\sum_{t=0}^{K-1}\rho_t\mathbf{x}^{t+1}}{\sum_{t=0}^{K-1}\rho_t},\ \bar{\mathbf{y}}^K = \frac{1}{K}\sum_{t=0}^{K-1}\mathbf{y}^t,\ \bar{\mathbf{z}}^K = \frac{1}{K}\sum_{t=0}^{K-1}\mathbf{z}^t. \qquad (3.11)
$$

Proof Since $\rho_k \le \beta_k$, the two terms $\frac{\beta_k-\rho_k}{2}\|\mathbf{r}^{k+1}\|^2$ and $\frac{\beta_k-\rho_k}{2\rho_k^2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2$ are nonnegative. Dropping them and multiplying ρ_k to both sides of (3.5) yields

$$
\begin{aligned}
&\rho_k\left[f_0(\mathbf{x}^{k+1}) + \mathbf{y}^\top\mathbf{r}^{k+1} + \sum_{i=1}^m z_i f_i(\mathbf{x}^{k+1})\right] + \frac{1}{2}\|\mathbf{y}^{k+1}-\mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{z}^{k+1}-\mathbf{z}\|^2 \\
&\le \rho_k\mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \frac{1}{2}\|\mathbf{y}^k-\mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{z}^k-\mathbf{z}\|^2 + \rho_k\varepsilon_k, \qquad (3.12)
\end{aligned}
$$

where $\mathbf{x} \in \mathcal{X}$, $\mathbf{y}$ is any vector, and $\mathbf{z} \geq \mathbf{0}$. Summing up (3.12) with $\mathbf{x} = \mathbf{x}^*$ and noting $\mathcal{L}_{\beta_k}(\mathbf{x}^*, \mathbf{y}^k, \mathbf{z}^k) \leq f_0(\mathbf{x}^*)$, we have

$$\begin{aligned}\sum_{k=0}^{K-1} \rho_k \left[f_0(\mathbf{x}^{k+1}) - f_0(\mathbf{x}^*) + \mathbf{y}^\top \mathbf{r}^{k+1} + \sum_{i=1}^m z_i f_i(\mathbf{x}^{k+1}) \right] \\ + \frac{1}{2}\|\mathbf{y}^K - \mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{z}^K - \mathbf{z}\|^2 \\ \leq \frac{1}{2}\|\mathbf{y}^0 - \mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{z}^0 - \mathbf{z}\|^2 + \sum_{k=0}^{K-1} \rho_k \varepsilon_k. \end{aligned} \tag{3.13}$$

By the convexity of f_i's and the nonnegativity of $\mathbf{z}$, we have

$$\begin{aligned} & f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*) + \mathbf{y}^\top (\mathbf{A}\bar{\mathbf{x}}^K - \mathbf{b}) + \sum_{i=1}^m z_i f_i(\bar{\mathbf{x}}^K) \\ & \leq \frac{1}{\sum_{t=0}^{K-1} \rho_t} \sum_{k=0}^{K-1} \rho_k \left[f_0(\mathbf{x}^{k+1}) - f_0(\mathbf{x}^*) + \mathbf{y}^\top \mathbf{r}^{k+1} + \sum_{i=1}^m z_i f_i(\mathbf{x}^{k+1}) \right], \end{aligned}$$

which together with (3.13) implies

$$\begin{aligned} & f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*) + \mathbf{y}^\top (\mathbf{A}\bar{\mathbf{x}}^K - \mathbf{b}) + \sum_{i=1}^m z_i f_i(\bar{\mathbf{x}}^K) \\ & \leq \frac{1}{\sum_{t=0}^{K-1} \rho_t} \left(\frac{1}{2}\|\mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{z}\|^2 + \sum_{k=0}^{K-1} \rho_k \varepsilon_k \right). \end{aligned}$$

The results in (3.9) thus follow from Lemma 2 with

$$\alpha = \frac{\sum_{k=0}^{K-1} \rho_k \varepsilon_k}{\sum_{k=0}^{K-1} \rho_k}, \quad c_1 = \frac{1}{2\sum_{k=0}^{K-1} \rho_k}, \quad c_2 = \frac{1}{2\sum_{k=0}^{K-1} \rho_k}.$$

When $\beta_k = \beta$ and $\rho_k = \rho$, $\forall k \geq 0$, letting $\mathbf{y} = \mathbf{0}$, $\mathbf{z} = \mathbf{0}$ in (3.12) and minimizing the right hand side about $\mathbf{x}$ give

$$f_0(\mathbf{x}^{k+1}) + \frac{1}{2\rho}\|\mathbf{y}^{k+1}\|^2 + \frac{1}{2\rho}\|\mathbf{z}^{k+1}\|^2 \leq d_\beta(\mathbf{y}^k, \mathbf{z}^k) + \frac{1}{2\rho}\|\mathbf{y}^k\|^2 + \frac{1}{2\rho}\|\mathbf{z}^k\|^2 + \varepsilon_k.$$

Summing the above inequality from $k = 0$ to $K - 1$, using the convexity of f_0 and concavity of d_β, and also noting $\mathbf{y}^0 = \mathbf{0}$, $\mathbf{z}^0 = \mathbf{0}$, we have

$$K f_0(\bar{\mathbf{x}}^K) \leq \sum_{k=0}^{K-1} f_0(\mathbf{x}^{k+1}) \leq \sum_{k=0}^{K-1} d_\beta(\mathbf{y}^k, \mathbf{z}^k) + \sum_{k=0}^{K-1} \varepsilon_k \leq K d_\beta(\bar{\mathbf{y}}^K, \bar{\mathbf{z}}^K) + \sum_{k=0}^{K-1} \varepsilon_k.$$

Now the result in (3.10) follows from (3.9a). □

Remark 1 Note that if $\rho_k \equiv \rho > 0, \forall k$ and $\sum_{k=0}^{\infty} \varepsilon_k < \infty$, then a sublinear convergence result follows from (3.9) and (3.10) in terms of both primal and dual variables. The work [38] has also analyzed the convergence of Algorithm 1 through the augmented dual function d_β. However, it requires $\sum_{k=0}^{\infty} \sqrt{\varepsilon_k} < \infty$, which is strictly stronger than the condition $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. The result in (3.10) seems also new. Without the $\mathbf{y}$-part, i.e., no linear constraint, [38, Equation (26)] shows that $\|\nabla_{\mathbf{z}} \mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{z}^k) - \nabla d_\beta(\mathbf{z}^k)\| \le \sqrt{\frac{2\varepsilon_k}{\beta}}$. Hence, to have $O(1/K)$ convergence rate about d_β, applying [40, Proposition 1] would require $\sum_{k=0}^{\infty} \sqrt{\varepsilon_k} < \infty$, and thus (3.10) is not implied.

3.2 Iteration complexity of iALM for primal ε-solutions

In this subsection, we apply Nesterov's optimal first-order method to each $\mathbf{x}$-subproblem (1.9) and estimate the total number of gradient evaluations to produce a primal ε-solution of (1.1). Note that the convergence rate results in Theorem 4 do not assume specific structures of (1.1) except convexity. If the problem (1.1) has richer structures than those in (1.3), more efficient methods can be applied to the subproblems in (1.9).

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

Proposition 1 *Assume* (1.3a)*,* (1.3b)*, and the boundedness of* $\text{dom}(h) \cap \mathcal{X}$*. Then there exist constants* $B_1, \ldots, B_m$ *such that*

$$\max\left(|f_i(\mathbf{x})|, \|\nabla f_i(\mathbf{x})\|\right) \le B_i, \ \forall \mathbf{x} \in \text{dom}(h) \cap \mathcal{X}, \forall i \in [m], \tag{3.14a}$$
$$|f_i(\hat{\mathbf{x}}) - f_i(\tilde{\mathbf{x}})| \le B_i \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|, \ \forall \hat{\mathbf{x}}, \tilde{\mathbf{x}} \in \text{dom}(h) \cap \mathcal{X}, \forall i \in [m]. \tag{3.14b}$$

Let the smooth part of $\mathcal{L}_\beta$ be denoted as

$$F_\beta(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \mathbf{z}) - h(\mathbf{x}).$$

Based on (3.14), we are able to show Lipschitz continuity of $\nabla_{\mathbf{x}} F_\beta(\mathbf{x}, \mathbf{y}, \mathbf{z})$ with respect to $\mathbf{x}$ for every $(\mathbf{y}, \mathbf{z})$.

Lemma 5 *Assume* (1.3a)*,* (1.3b)*, and the boundedness of* $\text{dom}(h) \cap \mathcal{X}$*. Let* B_i*'s be given in Proposition 1. Then* $\nabla_{\mathbf{x}} F_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k)$ *is Lipschitz continuous on* $\text{dom}(h) \cap \mathcal{X}$ *in terms of* $\mathbf{x}$ *with constant*

$$L(\mathbf{z}^k) = L_0 + \beta_k \|\mathbf{A}^\top \mathbf{A}\| + \sum_{i=1}^{m} \left(\beta_k B_i (B_i + L_i) + L_i |z_i^k|\right). \tag{3.15}$$

Proof For ease of description, we let $\beta = \beta_k$ and $(\mathbf{y}, \mathbf{z}) = (\mathbf{y}^k, \mathbf{z}^k)$ in the proof. First we notice that $\frac{\partial}{\partial u} \psi_\beta(u, v) = [\beta u + v]_+$, and thus for any v,

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

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

$$\begin{aligned}&\|\nabla_{\mathbf{x}}h_i(\hat{\mathbf{x}},z_i)-\nabla_{\mathbf{x}}h_i(\tilde{\mathbf{x}},z_i)\|\\
&\quad=\big\|\frac{\partial}{\partial u}\psi_\beta(f_i(\hat{\mathbf{x}}),z_i)\nabla f_i(\hat{\mathbf{x}})-\frac{\partial}{\partial u}\psi_\beta(f_i(\tilde{\mathbf{x}}),z_i)\nabla f_i(\tilde{\mathbf{x}})\big\|\\
&\quad\le\big\|\frac{\partial}{\partial u}\psi_\beta(f_i(\hat{\mathbf{x}}),z_i)\nabla f_i(\hat{\mathbf{x}})-\frac{\partial}{\partial u}\psi_\beta(f_i(\tilde{\mathbf{x}}),z_i)\nabla f_i(\hat{\mathbf{x}})\big\|\\
&\qquad+\big\|\frac{\partial}{\partial u}\psi_\beta(f_i(\tilde{\mathbf{x}}),z_i)\nabla f_i(\hat{\mathbf{x}})-\frac{\partial}{\partial u}\psi_\beta(f_i(\tilde{\mathbf{x}}),z_i)\nabla f_i(\tilde{\mathbf{x}})\big\|\\
&\quad\le\beta|f_i(\hat{\mathbf{x}})-f_i(\tilde{\mathbf{x}})|\cdot\|\nabla f_i(\hat{\mathbf{x}})\|+\big|\frac{\partial}{\partial u}\psi_\beta(f_i(\tilde{\mathbf{x}}),z_i)\big|\cdot\|\nabla f_i(\hat{\mathbf{x}})-\nabla f_i(\tilde{\mathbf{x}})\|\\
&\quad\le\beta B_i^2\|\hat{\mathbf{x}}-\tilde{\mathbf{x}}\|+L_i(\beta B_i+|z_i|)\|\hat{\mathbf{x}}-\tilde{\mathbf{x}}\|.\end{aligned}$$

Hence,

$$\begin{aligned}&\|\nabla_{\mathbf{x}}F_\beta(\hat{\mathbf{x}},\mathbf{y},\mathbf{z})-\nabla_{\mathbf{x}}F_\beta(\tilde{\mathbf{x}},\mathbf{y},\mathbf{z})\|\\
&\quad\le\|\nabla g(\hat{\mathbf{x}})-\nabla g(\tilde{\mathbf{x}})\|+\beta\|\mathbf{A}^\top\mathbf{A}(\hat{\mathbf{x}}-\tilde{\mathbf{x}})\|+\sum_{i=1}^m\|\nabla_{\mathbf{x}}h_i(\hat{\mathbf{x}},z_i)-\nabla_{\mathbf{x}}h_i(\tilde{\mathbf{x}},z_i)\|\\
&\quad\le\left(L_0+\beta\|\mathbf{A}^\top\mathbf{A}\|+\sum_{i=1}^m\left[\beta B_i^2+L_i(\beta B_i+|z_i|)\right]\right)\|\hat{\mathbf{x}}-\tilde{\mathbf{x}}\|,\end{aligned}$$

which completes the proof. □

Therefore, letting $\phi(\mathbf{x})=F_{\beta_k}(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k)$ and $\psi(\mathbf{x})=h(\mathbf{x})+\iota_{\mathcal{X}}(\mathbf{x})$, we can apply Nesterov's optimal first-order method in Algorithm 2 to find $\mathbf{x}^{k+1}$ in (1.9). From Theorem 2, we have the following results. Note that if the strong convexity constant $\mu=0$, the problem is just convex.

Lemma 6 *Assume that g is strongly convex with modulus $\mu\ge0$. Given $\varepsilon_k>0$, if we start from $\mathbf{x}^k$ and run Algorithm 2, then at most t_k iterations are needed to produce $\mathbf{x}^{k+1}$ such that (1.9) holds, where*

$$t_k=\begin{cases}\left\lceil\dfrac{\mathrm{dist}(\mathbf{x}^k,\mathcal{X}_k^*)\sqrt{2L(\mathbf{z}^k)}}{\sqrt{\varepsilon_k}}\right\rceil, & \text{if } \mu=0,\\[2ex] \left\lceil\dfrac{\log\left(\frac{L(\mathbf{z}^k)+\mu}{2\varepsilon_k}[\mathrm{dist}(\mathbf{x}^k,\mathcal{X}_k^*)]^2\right)}{\log 1/\left(1-\sqrt{\frac{\mu}{L(\mathbf{z}^k)}}\right)}\right\rceil, & \text{if } \mu>0,\end{cases}\tag{3.16}$$

and $\mathcal{X}_k^*$ denotes the set of optimal solutions to $\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}_{\beta_k}(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k)$.

Below we specify the sequences $\{\beta_k\}$, $\{\rho_k\}$ and $\{\varepsilon_k\}$ for a given $\varepsilon > 0$, and through combining Theorem 4 and Lemma 6, we give the iteration complexity results of iALM for producing a primal ε-solution. We study two cases. In the first case, a constant penalty parameter is used, and in the second case, we geometrically increase β_k and ρ_k.

Given $\varepsilon > 0$, we set $\{\beta_k\}$ and $\{\rho_k\}$ according to one of the follows:

Setting 1 (constant penalty) *Let K be a positive integer number and C_1 a positive real number. Set*

$$\rho_k = \beta_k = \beta = \frac{C_1}{K\varepsilon},\ \forall\, 0 \le k < K.$$

Setting 2 (geometrically increasing penalty) *Let K be a positive integer number, C_1 a positive real number, and $\sigma > 1$. Set*

$$\beta_0 = \frac{C_1}{\varepsilon}\frac{\sigma - 1}{\sigma^K - 1}, \tag{3.17}$$

and

$$\rho_k = \beta_k = \beta_0\sigma^k,\ \forall\, 0 \le k < K.$$

Note that if $K = 1$, the above two settings are the same, and in this case, Algorithm 1 simply reduces to the quadratic penalty method. For either of the above two settings, we have $\sum_{k=0}^{K-1} \rho_k = \frac{C_1}{\varepsilon}$, which is required in our analysis. To have this hold, we do not have to fix K first. Instead, we can keep $\rho_k = \beta_k, \forall k$, simply choose β_0 and C_1 first, and then run K outer iterations either with constant parameter β or geometrically increasing one such that $\sum_{k=0}^{K-1} \rho_k \ge \frac{C_1}{\varepsilon}$. The order of our complexity results will remain the same if β_0 is in the order of $\frac{1}{\varepsilon}$.

From (3.15), we see that the Lipschitz constant depends on $\mathbf{z}^k$. Hence, from (3.16), to solve the $\mathbf{x}$-subproblem to the accuracy ε_k, the number of gradient evaluations will depend on $\mathbf{z}^k$. Below we show that if ε_k is sufficiently small, $\mathbf{z}^k$ can be bounded and thus so is $L(\mathbf{z}^k)$.

Lemma 7 *Let $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^K$ be the sequence generated from Algorithm 1 with $\{\beta_k\}$ and $\{\rho_k\}$ set according to either Settings 1 or 2. If $\mathbf{y}^0 = \mathbf{0}, \mathbf{z}^0 = \mathbf{0}$, and ε_k's are chosen such that*

$$\sum_{k=0}^{K-1} \rho_k\varepsilon_k \le \frac{C_2}{2}, \tag{3.18}$$

for a certain constant $C_2 > 0$, then

$$L(\mathbf{z}^k) \le L_* + \beta_k H,\ \forall\, 0 \le k \le K, \tag{3.19}$$

where

$$H = \|\mathbf{A}^\top\mathbf{A}\| + \sum_{i=1}^m B_i(B_i + L_i),\ L_* = L_0 + \|\boldsymbol{\ell}\|\left(\|\mathbf{y}^*\| + 2\|\mathbf{z}^*\| + \sqrt{C_2}\right)$$

and $\boldsymbol{\ell}$ is given in (1.14).

Proof Letting $(\mathbf{x},\mathbf{y},\mathbf{z})=(\mathbf{x}^*,\mathbf{y}^*,\mathbf{z}^*)$ in (3.12), noting $\mathcal{L}_{\beta_k}(\mathbf{x}^*,\mathbf{y}^k,\mathbf{z}^k)\le f_0(\mathbf{x}^*)$, and using (2.2), we have

$$\frac{1}{2}\|\mathbf{y}^{k+1}-\mathbf{y}^*\|^2+\frac{1}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^*\|^2\le\frac{1}{2}\|\mathbf{y}^k-\mathbf{y}^*\|^2+\frac{1}{2}\|\mathbf{z}^k-\mathbf{z}^*\|^2+\rho_k\varepsilon_k.$$

Summing the above inequality yields

$$\begin{aligned}&\frac{1}{2}\|\mathbf{y}^k-\mathbf{y}^*\|^2+\frac{1}{2}\|\mathbf{z}^k-\mathbf{z}^*\|^2\\&\quad\le\frac{1}{2}\|\mathbf{y}^0-\mathbf{y}^*\|^2+\frac{1}{2}\|\mathbf{z}^0-\mathbf{z}^*\|^2+\sum_{t=0}^{k-1}\rho_t\varepsilon_t,\ \forall\,0\le k\le K,\end{aligned}$$

which implies

$$\|\mathbf{z}^k\|\le\|\mathbf{z}^*\|+\sqrt{\|\mathbf{y}^0-\mathbf{y}^*\|^2+\|\mathbf{z}^0-\mathbf{z}^*\|^2+2\sum_{t=0}^{k-1}\rho_t\varepsilon_t}.$$

Since $\|\mathbf{u}\|\le\|\mathbf{u}\|_1$ for any vector $\mathbf{u}$, we have from the above inequality that

$$\|\mathbf{z}^k\|\le\|\mathbf{z}^*\|+\|\mathbf{y}^0-\mathbf{y}^*\|+\|\mathbf{z}^0-\mathbf{z}^*\|+\sqrt{2\sum_{t=0}^{K-1}\rho_t\varepsilon_t},\ \forall\,0\le k\le K.\tag{3.20}$$

Hence, if $\mathbf{y}^0=\mathbf{0}$ and $\mathbf{z}^0=\mathbf{0}$, and (3.18) holds, it follows from the above inequality that

$$\|\mathbf{z}^k\|\le\|\mathbf{y}^*\|+2\|\mathbf{z}^*\|+\sqrt{C_2},\ \forall\,0\le k\le K,\tag{3.21}$$

By the Cauchy-Schwartz inequality, we have from (3.15) that for any $0\le k\le K$,

$$L(\mathbf{z}^k)\le L_0+\beta_k H+\|\mathbf{z}^k\|\cdot\|\boldsymbol{\ell}\|,$$

which together with (3.21) gives the result in (3.19). □

"Optimal" subproblem accuracy parameters If t_k gradient evaluations are required to produce $\mathbf{x}^{k+1}$, then the total number of gradient evaluations is $T_K=\sum_{k=0}^{K-1}t_k$ to generate $\{\mathbf{x}^k\}_{k=1}^K$. Given $\varepsilon>0$, and $\{\beta_k\}$,$\{\rho_k\}$ set according to either Settings 1 or 2, we can choose $\{\varepsilon_k\}$ to minimize T_K subject to the condition in (3.18).

When $\mu=0$, we solve the following problem:

$$\operatorname*{minimize}_{\boldsymbol{\varepsilon}>\mathbf{0}}\sum_{k=0}^{K-1}\frac{\operatorname{dist}(\mathbf{x}^k,\mathcal{X}_k^*)\sqrt{L(\mathbf{z}^k)}}{\sqrt{\varepsilon_k}},\ \text{ s.t. }\sum_{k=0}^{K-1}\beta_k\varepsilon_k\le\frac{C_2}{2},$$

where $\boldsymbol{\varepsilon} = [\varepsilon_0, \ldots, \varepsilon_{K-1}]$. Through formulating the KKT system of the above problem, one can easily find the optimal $\boldsymbol{\varepsilon}$ given by

$$\varepsilon_k = \frac{C_2}{2} \frac{[\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*)]^{\frac{2}{3}} [L(\mathbf{z}^k)]^{\frac{1}{3}}}{\beta_k^{\frac{2}{3}} \sum_{t=0}^{K-1} \beta_t^{\frac{1}{3}} [\mathrm{dist}(\mathbf{x}^t, \mathcal{X}_t^*)]^{\frac{2}{3}} [L(\mathbf{z}^t)]^{\frac{1}{3}}}, \ \forall\, 0 \le k < K. \tag{3.22}$$

When $\mu > 0$, we solve the problem below:

$$\underset{\boldsymbol{\varepsilon} > \mathbf{0}}{\text{minimize}} \sum_{k=0}^{K-1} \sqrt{\frac{L(\mathbf{z}^k)}{\mu}} \log\left(\frac{L(\mathbf{z}^k) + \mu}{2\varepsilon_k} [\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*)]^2\right), \ \text{s.t.} \sum_{k=0}^{K-1} \beta_k \varepsilon_k \le \frac{C_2}{2}, \tag{3.23}$$

whose optimal solution is given by

$$\varepsilon_k = \frac{C_2}{2} \frac{\sqrt{L(\mathbf{z}^k)}}{\beta_k \sum_{t=0}^{K-1} \sqrt{L(\mathbf{z}^t)}}, \ \forall\, 0 \le k < K. \tag{3.24}$$

Note that the summand in the objective of (3.23) is not exactly the same as that in the second inequality of (3.16). They are close if $\mu \ll L(\mathbf{z}^k)$ since $\log(1+a) = a + o(a)$.

The optimal ε_k given in (3.22) and (3.24) depends on $\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*)$ and the future points $\mathbf{z}^{k+1}, \ldots, \mathbf{z}^{K-1}$, which are unknown at iteration k. We do not assume these unknowns. Instead, we set ε_k in two different ways. One way is to simply set

$$\varepsilon_k = \frac{\varepsilon}{2} \frac{C_2}{C_1}, \ \forall\, 0 \le k < K, \tag{3.25}$$

for both cases of $\mu = 0$ and $\mu > 0$. Another way is to let

$$\varepsilon_k = \frac{C_2}{2} \frac{1}{\beta_k^{\frac{1}{3}} \sum_{t=0}^{K-1} \beta_t^{\frac{2}{3}}}, \ \forall\, 0 \le k < K, \tag{3.26}$$

for the case of $\mu = 0$, and

$$\varepsilon_k = \frac{C_2}{2} \frac{1}{\sqrt{\beta_k} \sum_{t=0}^{K-1} \sqrt{\beta_t}}, \ \forall\, 0 \le k < K, \tag{3.27}$$

for the case of $\mu > 0$. We see that if $\beta_k H$ dominates L_* and $\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*)$ is roughly the same for all k's, then $\{\varepsilon_k\}$ in (3.26) and (3.27) well approximate those in (3.22) and (3.24). If $\{\beta_k\}$ and $\{\rho_k\}$ are set according to Setting 1, i.e., constant parameters, then the $\{\varepsilon_k\}$ in both (3.26) and (3.27) is constant as in (3.25).

Plugging these parameters into (3.16), we have the following estimates on the total number of gradient evaluations.

Theorem 5 (Iteration complexity with constant penalty and constant error) *For any given $\varepsilon > 0$, let K be a positive integer number and C_1, C_2 two positive real numbers.*

Set $\{\beta_k\}$ and $\{\rho_k\}$ according to Setting 1 and $\{\varepsilon_k\}$ by (3.25). Let $(\bar{\mathbf{x}}^K, \bar{\mathbf{y}}^K, \bar{\mathbf{z}}^K)$ be given in (3.11). Then

$$\big|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\big| \le \frac{\varepsilon\big(2\|\mathbf{y}^*\|^2 + 2\|\mathbf{z}^*\|^2\big)}{C_1} + \frac{\varepsilon}{2}\frac{C_2}{C_1}, \tag{3.28a}$$

$$\|\mathbf{A}\bar{\mathbf{x}}^K - \mathbf{b}\| + \big\|[\mathbf{f}(\bar{\mathbf{x}}^K)]_+\big\| \le \frac{\varepsilon\big[(1+\|\mathbf{y}^*\|)^2 + (1+\|\mathbf{z}^*\|)^2\big]}{2C_1} + \frac{\varepsilon}{2}\frac{C_2}{C_1}, \tag{3.28b}$$

$$f_0(\mathbf{x}^*) - d_\beta(\bar{\mathbf{y}}^K, \bar{\mathbf{z}}^K) \le \frac{\varepsilon\big(2\|\mathbf{y}^*\|^2 + 2\|\mathbf{z}^*\|^2\big)}{C_1} + \frac{\varepsilon C_2}{C_1}. \tag{3.28c}$$

Assume $\mu \le \frac{L_0}{4}$. Then Algorithm 1 can produce $(\bar{\mathbf{x}}^K, \bar{\mathbf{y}}^K, \bar{\mathbf{z}}^K)$ by evaluating gradients of g, f_i, $i \in [m]$ in at most T_K times, where

$$T_K = \left\lceil 2DK\sqrt{\frac{C_1}{C_2}}\left(\sqrt{\frac{L_*}{\varepsilon}} + \frac{1}{\varepsilon}\sqrt{\frac{C_1 H}{K}}\right) + K\right\rceil, \text{ if } \mu = 0, \tag{3.29}$$

and

$$T_K = \left\lceil 2K\left(\sqrt{\frac{L_*}{\mu}} + \sqrt{\frac{C_1 H}{\mu K\varepsilon}}\right)\log\left(\frac{D^2 C_1}{C_2}\left(\frac{L_* + \mu}{\varepsilon} + \frac{C_1 H}{K\varepsilon^2}\right)\right) + K\right\rceil, \text{ if } \mu > 0. \tag{3.30}$$

Proof The results in (3.28) directly follows from Theorem 4 and the settings of $\{\beta_k\}$, $\{\rho_k\}$, and $\{\varepsilon_k\}$. For the total number of gradient evaluations, we use (3.16). First, for the case of $\mu = 0$, from the first equation of (3.16) and the parameter setting, it follows that the total number of gradient evaluations

$$T_K \le \sum_{k=0}^{K-1} \frac{\text{dist}(\mathbf{x}^k, \mathcal{X}_k^*)\sqrt{2(L_* + \frac{C_1 H}{K\varepsilon})}}{\sqrt{\varepsilon/2}\sqrt{C_2/C_1}} + K. \tag{3.31}$$

Since $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any two nonnegative numbers a, b, we have from the above inequality and by noting $\text{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \le D$ that

$$T_K \le 2D\sqrt{\frac{C_1}{C_2}}\sum_{k=0}^{K-1}\frac{\sqrt{L_*} + \sqrt{\frac{C_1 H}{K\varepsilon}}}{\sqrt{\varepsilon}} + K = 2DK\sqrt{\frac{C_1}{C_2}}\left(\sqrt{\frac{L_*}{\varepsilon}} + \frac{1}{\varepsilon}\sqrt{\frac{C_1 H}{K}}\right) + K,$$

which gives (3.29).

For the case of $\mu > 0$, we first note that for any $0 < a \le 1$, it holds $\log(1+a) \ge a - \frac{a^2}{2} \ge \frac{a}{2}$. Hence, if $\mu \le \frac{L_0}{4}$, we have $\mu \le \frac{L(\mathbf{z}^k)}{4}$ and $\frac{\sqrt{\mu/L(\mathbf{z}^k)}}{1-\sqrt{\mu/L(\mathbf{z}^k)}} \le 1$. Therefore,

$$\log\frac{1}{1-\sqrt{\mu/L(\mathbf{z}^k)}} = \log\left(1 + \frac{\sqrt{\mu/L(\mathbf{z}^k)}}{1-\sqrt{\mu/L(\mathbf{z}^k)}}\right) \ge \frac{1}{2}\frac{\sqrt{\mu/L(\mathbf{z}^k)}}{1-\sqrt{\mu/L(\mathbf{z}^k)}},$$

and thus

$$\frac{1}{\log \frac{1}{1-\sqrt{\mu/L(\mathbf{z}^k)}}} \leq 2\sqrt{\frac{L(\mathbf{z}^k)}{\mu}}\left(1-\sqrt{\mu/L(\mathbf{z}^k)}\right) \leq 2\sqrt{\frac{L(\mathbf{z}^k)}{\mu}}. \tag{3.32}$$

Using the above inequality and the second inequality of (3.16), we have that the total number of gradient evaluations

$$T_K \leq \sum_{k=0}^{K-1} 2\sqrt{\frac{L_* + \frac{C_1 H}{K\varepsilon}}{\mu}} \log\left(\frac{L_* + \frac{C_1 H}{K\varepsilon} + \mu}{\varepsilon C_2/C_1}[\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*)]^2\right) + K. \tag{3.33}$$

Since $\sqrt{L_* + \frac{C_1 H}{K\varepsilon}} \leq \sqrt{L_*} + \sqrt{\frac{C_1 H}{K\varepsilon}}$ and $\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \leq D$, the above inequality implies (3.30). This completes the proof. □

From Theorem 5, we can immediately obtain the following corollary about primal ε-solutions.

Corollary 1 (Iteration complexity for primal ε-solutions) *Let $\varepsilon > 0$ be given. To produce a primal ε-solution, Algorithm 1 needs to evaluate gradients of g, f_i, $i \in [m]$ in at most $O(\varepsilon^{-1})$ times for convex case of $\mu = 0$ and $O(\varepsilon^{-1/2}|\log \varepsilon|)$ times for strongly convex case of $\mu > 0$.*

Proof Let C_1 and C_2 be two constants such that

$$C_1 \geq \max\left(2\|\mathbf{y}^*\|^2 + 2\|\mathbf{z}^*\|^2, \frac{(1+\|\mathbf{y}^*\|)^2}{2} + \frac{(1+\|\mathbf{z}^*\|)^2}{2}\right) + \frac{C_2}{2}. \tag{3.34}$$

From the error bounds in (3.28a) and (3.28b), it follows that $\bar{\mathbf{x}}^K$ is a primal ε-solution. Set K independent of ε. Then the total number of gradient evaluations $T_K = O(\varepsilon^{-1})$ in (3.29) and $T_K = O(\varepsilon^{-1/2}|\log \varepsilon|)$ in (3.30). This completes the proof. □

We make two observations below about the results in Theorem 5 and Corollary 1.

Remark 2 The choices of C_1 and C_2 in (3.34) assume the knowledge of $\|\mathbf{y}^*\|$ and $\|\mathbf{z}^*\|$, which are often unknown. Practically, we can simply set C_1 and C_2 as certain constants, and the errors in (3.28) would be multiples of ε. In this case, Algorithm 1 will produce a primal $O(\varepsilon)$-solution.

If we represent ε by the total number t of gradient evaluations, we can obtain the convergence rate result in terms of t. For simplicity, let $C_1 = C_2$ and $K = 1$ in (3.29). Then the total number of gradient evaluations is about $t = 2D\left(\sqrt{\frac{L_*}{\varepsilon}} + \frac{1}{\varepsilon}\sqrt{C_1 H}\right)$. By quadratic formula, one can easily show that

$$\varepsilon = \frac{\left(D\sqrt{L_*} + \sqrt{L_* D^2 + 2Dt\sqrt{C_1 H}}\right)^2}{t^2} \leq \frac{4L_* D^2}{t^2} + \frac{4D\sqrt{C_1 H}}{t}.$$

Let $\hat{\mathbf{x}}^t = \bar{\mathbf{x}}^K$ to specify the dependence of the iterate on the number of gradient evaluations. Plugging the above ε into (3.28a) and (3.28b), we have

$$\left|f_0(\hat{\mathbf{x}}^t) - f_0(\mathbf{x}^*)\right| \le \left(\frac{2\|\mathbf{y}^*\|^2 + 2\|\mathbf{z}^*\|^2}{C_1} + \frac{1}{2}\right)\left(\frac{4L_* D^2}{t^2} + \frac{4D\sqrt{C_1 H}}{t}\right), \tag{3.35a}$$

$$\begin{aligned}&\|\mathbf{A}\hat{\mathbf{x}}^t - \mathbf{b}\| + \left\|[\mathbf{f}(\hat{\mathbf{x}}^t)]_+\right\| \\ &\quad\le \left(\frac{(1+\|\mathbf{y}^*\|)^2 + (1+\|\mathbf{z}^*\|)^2}{2C_1} + \frac{1}{2}\right)\left(\frac{4L_* D^2}{t^2} + \frac{4D\sqrt{C_1 H}}{t}\right).\end{aligned} \tag{3.35b}$$

If there are no equality or inequality constraints, then $H = 0$, $\mathbf{y}^* = \mathbf{0}$, $\mathbf{z}^* = \mathbf{0}$, and the rate of convergence in (3.35a) matches with the optimal one in (2.9); if the objective $f_0(\mathbf{x}) \equiv 0$ and there are no inequality constraints, then $H = \|\mathbf{A}^\top \mathbf{A}\|$, $\mathbf{y}^* = \mathbf{0}$, $\mathbf{z}^* = \mathbf{0}$, $L_* = 0$, and the rate of convergence with $C_1 = 2$ in (3.35b) roughly becomes

$$\|\mathbf{A}\hat{\mathbf{x}}^t - \mathbf{b}\|^2 \le \frac{8\sqrt{2}D^2\|\mathbf{A}^\top \mathbf{A}\|}{t^2},$$

whose order is also optimal. Therefore, the order of convergence rate in (3.35) is optimal, and so is the iteration complexity result in (3.29) to obtain a primal ε-solution.

For the strongly convex case, if there are no equality or inequality constraints, the iteration complexity result in (3.30) is optimal by comparing it to (2.10). With the existence of constraints and nonsmooth term in the objective, $O(\varepsilon^{-\frac{1}{2}})$ is a lower complexity bound for first-order methods to find a primal ε-solution [35]. Hence, our iteration complexity result is nearly optimal.

Remark 3 From both (3.29) and (3.30), we see that $T_1 \le T_K$, $\forall K \ge 1$, i.e., $K = 1$ is the best. Note that if $\mathbf{y}^0 = \mathbf{0}$, $\mathbf{z}^0 = \mathbf{0}$, and $K = 1$, Algorithm 1 reduces to the quadratic penalty method by solving a single penalty problem. However, practically $K > 1$ could be better since $\text{dist}(\mathbf{x}^k, \mathcal{X}_k^*)$ usually decreases as k increases. Hence, from (3.31) or (3.33), T_K can be smaller than T_1 if $K > 1$; see our numerical results in Sect. 6.

The rest part of this section assumes geometrically increasing penalty parameters. We do not have a fixed augmented dual function, and thus we only consider primal error in the iteration complexity results.

Theorem 6 (Iteration complexity with geometrically increasing penalty and constant error) *For any given $\varepsilon > 0$, let K be a positive integer number and C_1, C_2 two positive real numbers. Set $\{\beta_k\}$ and $\{\rho_k\}$ according to Setting 2 and $\{\varepsilon_k\}$ to (3.25). Assume $\mu \le \frac{L_0}{4}$. Let $\bar{\mathbf{x}}^K$ be given in (3.11). Then the inequalities in (3.28a) and (3.28b) hold, and Algorithm 1 can produce $\bar{\mathbf{x}}^K$ by evaluating gradients of g, f_i, $i \in [m]$ in at most T_K times, where*

$$T_K = \left\lceil 2D\sqrt{\frac{C_1}{C_2}}\left(K\sqrt{\frac{L_*}{\varepsilon}} + \frac{\sqrt{C_1 H(\sigma-1)}}{\varepsilon(\sqrt{\sigma}-1)}\right) + K\right\rceil, \text{ if } \mu = 0, \tag{3.36}$$

and

$$T_K = \left\lceil 2G_\varepsilon \left(K\sqrt{\frac{L_*}{\mu}} + \sqrt{\frac{H}{\mu}} \frac{\sqrt{C_1(\sigma-1)}}{\sqrt{\varepsilon}(\sqrt{\sigma}-1)} \right) + K \right\rceil, \text{ if } \mu > 0. \tag{3.37}$$

where

$$G_\varepsilon = \log \frac{C_1 D^2}{\varepsilon C_2} + \log \left(L_* + \mu + \frac{H\big(C_1(\sigma-1) + \beta_0 \varepsilon\big)}{\sigma \varepsilon} \right).$$

Proof When $\mu = 0$, we have from the first inequality in (3.16) that the total number of gradient evaluations satisfies

$$T_K \le \sum_{k=0}^{K-1} \frac{\operatorname{dist}(\mathbf{x}^k, \mathcal{X}_k^*)\sqrt{2(L_* + \beta_k H)}}{\sqrt{\varepsilon_k}} + K. \tag{3.38}$$

Plugging into (3.38) the ε_k given in (3.25) and noting $\operatorname{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \le D$ yields

$$T_K \le 2D\sqrt{\frac{C_1}{C_2}} \sum_{k=0}^{K-1} \frac{\sqrt{L_* + \beta_k H}}{\sqrt{\varepsilon}} + K. \tag{3.39}$$

Note that $\sum_{k=0}^{K-1} \sqrt{\beta_k} = \sqrt{\beta_0} \frac{\sigma^{\frac{K}{2}} - 1}{\sqrt{\sigma} - 1}$. From (3.17), it holds

$$\sigma^K = \frac{C_1(\sigma - 1)}{\beta_0 \varepsilon} + 1, \tag{3.40}$$

and thus $\sigma^{\frac{K}{2}} - 1 \le \sqrt{\frac{C_1(\sigma-1)}{\beta_0 \varepsilon}}$. Therefore,

$$\sum_{k=0}^{K-1} \sqrt{\beta_k} \le \frac{\sqrt{C_1(\sigma-1)}}{\sqrt{\varepsilon}(\sqrt{\sigma}-1)}, \tag{3.41}$$

and using $\sqrt{L_* + \beta_k H} \le \sqrt{L_*} + \sqrt{\beta_k H}$, we have

$$\sum_{k=0}^{K-1} \sqrt{L_* + \beta_k H} \le \sum_{k=0}^{K-1} \left(\sqrt{L_*} + \sqrt{\beta_k H} \right) \le K\sqrt{L_*} + \frac{\sqrt{C_1 H(\sigma-1)}}{\sqrt{\varepsilon}(\sqrt{\sigma}-1)}, \tag{3.42}$$

which together with (3.39) gives (3.36).

For the strongly convex case, we use (3.32) and the second inequality of (3.16) to have

$$T_K \leq 2\sum_{k=0}^{K-1}\sqrt{\frac{L_* + \beta_k H}{\mu}}\log\left(\frac{L_* + \beta_k H + \mu}{2\varepsilon_k}[\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*)]^2\right) + K. \tag{3.43}$$

Since $\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \leq D$ and ε_k's are set to those in (3.25), the above inequality indicates

$$T_K \leq 2\sum_{k=0}^{K-1}\sqrt{\frac{L_* + \beta_k H}{\mu}}\log\frac{C_1 D^2(L_* + \beta_k H + \mu)}{\varepsilon C_2} + K. \tag{3.44}$$

For $0 \leq k < K$,

$$\beta_k \leq \beta_{K-1} = \beta_0\sigma^{K-1} = \frac{\beta_0}{\sigma}\sigma^K \overset{(3.40)}{=} \frac{\beta_0}{\sigma}\left(\frac{C_1(\sigma-1)}{\beta_0\varepsilon} + 1\right) = \frac{C_1(\sigma-1) + \beta_0\varepsilon}{\sigma\varepsilon}. \tag{3.45}$$

Plugging into (3.44) the second inequality in (3.42) and the above bound on β_k, we have (3.37) and thus complete the proof. □

Remark 4 Comparing the iteration complexity results in Theorems 5 and 6, we see that if $K = 1$, the number T_K in either case of $\mu = 0$ or $\mu > 0$ is the same for both penalty parameter settings as $\sigma \to \infty$. That is because when $K = 1$, iALM with either of the two settings reduces to the penalty method. If $K > 1$, the number T_K for the setting of geometrically increasing penalty can be smaller than that for the constant parameter setting as σ is big; see numerical results in Sect. 6.

Theorem 7 (Iteration complexity with geometrically increasing penalty and adaptive error) *For any given $\varepsilon > 0$, let K be a positive integer number and C_1, C_2 two positive real numbers. Set $\{\beta_k\}$ and $\{\rho_k\}$ according to Setting 2. Assume $\mu \leq \frac{L_0}{4}$. If $\mu = 0$, set $\{\varepsilon_k\}$ as in (3.26), and if $\mu > 0$, set $\{\varepsilon_k\}$ as in (3.27). Let $\bar{\mathbf{x}}^K$ be given in (3.11). Then the inequalities in (3.28a) and (3.28b) hold, and Algorithm 1 can produce $\bar{\mathbf{x}}^K$ by evaluating gradients of g, f_i, $i \in [m]$ in at most T_K times, where*

$$T_K = \left\lceil 2D\sqrt{\frac{C_1}{C_2}}\left(\sqrt{\frac{L_*}{\varepsilon}}\frac{(\sigma-1)^{\frac{1}{2}}}{(\sigma^{\frac{1}{6}}-1)(\sigma^{\frac{2}{3}}-1)^{\frac{1}{2}}} + \frac{\sqrt{HC_1}(\sigma-1)}{\varepsilon(\sigma^{\frac{2}{3}}-1)^{\frac{3}{2}}}\right) + K\right\rceil, \text{ if } \mu = 0, \tag{3.46}$$

and

$$T_K = \left\lceil 2G_\varepsilon\left(K\sqrt{\frac{L_*}{\mu}} + \sqrt{\frac{H}{\mu}}\frac{\sqrt{C_1(\sigma-1)}}{\sqrt{\varepsilon}(\sqrt{\sigma}-1)}\right) + K\right\rceil, \text{ if } \mu > 0. \tag{3.47}$$

where

$$G_\varepsilon = \log \frac{C_1 D^2}{\varepsilon C_2} + \log \left(L_* + \mu + \frac{H\big(C_1(\sigma - 1) + \beta_0 \varepsilon\big)}{\sigma \varepsilon} \right) + \log \frac{\sqrt{(\sigma - 1)^2 + \beta_0 \varepsilon (\sigma - 1)/C_1}}{\sigma - \sqrt{\sigma}}.$$

Proof For the case of $\mu = 0$, we have (3.38), plugging into which the ε_k given in (3.26) yields

$$T_K \leq \frac{2}{\sqrt{C_2}} \sqrt{\sum_{t=0}^{K-1} \beta_t^{\frac{2}{3}}} \sum_{k=0}^{K-1} \operatorname{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \beta_k^{\frac{1}{6}} (L_* + \beta_k H)^{\frac{1}{2}} + K.$$

Since $\operatorname{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \leq D$, the above inequality implies

$$T_K \leq \frac{2D}{\sqrt{C_2}} \sqrt{\sum_{t=0}^{K-1} \beta_t^{\frac{2}{3}}} \sum_{k=0}^{K-1} \beta_k^{\frac{1}{6}} (L_* + \beta_k H)^{\frac{1}{2}} + K. \tag{3.48}$$

Note that

$$\sqrt{\sum_{t=0}^{K-1} \beta_t^{\frac{2}{3}}} = \sqrt{\sum_{t=0}^{K-1} \beta_0^{\frac{2}{3}} \sigma^{\frac{2t}{3}}} = \beta_0^{\frac{1}{3}} \sqrt{\frac{\sigma^{\frac{2K}{3}} - 1}{\sigma^{\frac{2}{3}} - 1}},$$

and

$$\begin{aligned} \sum_{k=0}^{K-1} \beta_k^{\frac{1}{6}} (L_* + \beta_k H)^{\frac{1}{2}} &\leq \sum_{k=0}^{K-1} \beta_k^{\frac{1}{6}} \left(\sqrt{L_*} + \sqrt{\beta_k H} \right) = \sqrt{L_*} \sum_{k=0}^{K-1} \beta_k^{\frac{1}{6}} + \sqrt{H} \sum_{k=0}^{K-1} \beta_k^{\frac{2}{3}} \\ &= \sqrt{L_*} \beta_0^{\frac{1}{6}} \frac{\sigma^{\frac{K}{6}} - 1}{\sigma^{\frac{1}{6}} - 1} + \sqrt{H} \beta_0^{\frac{2}{3}} \frac{\sigma^{\frac{2K}{3}} - 1}{\sigma^{\frac{2}{3}} - 1}. \end{aligned}$$

Hence, it follows from (3.48) that

$$T_K \leq \frac{2D}{\sqrt{C_2}} \beta_0^{\frac{1}{3}} \sqrt{\frac{\sigma^{\frac{2K}{3}} - 1}{\sigma^{\frac{2}{3}} - 1}} \left(\sqrt{L_*} \beta_0^{\frac{1}{6}} \frac{\sigma^{\frac{K}{6}} - 1}{\sigma^{\frac{1}{6}} - 1} + \sqrt{H} \beta_0^{\frac{2}{3}} \frac{\sigma^{\frac{2K}{3}} - 1}{\sigma^{\frac{2}{3}} - 1} \right) + K. \tag{3.49}$$

From (3.40) and the fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $\forall a, b \geq 0$, it follows that

$$\sigma^{\frac{2K}{3}} - 1 \leq \left(\frac{C_1(\sigma - 1)}{\beta_0 \varepsilon} \right)^{\frac{2}{3}}, \quad \sigma^{\frac{K}{6}} - 1 \leq \left(\frac{C_1(\sigma - 1)}{\beta_0 \varepsilon} \right)^{\frac{1}{6}}. \tag{3.50}$$

Therefore, plugging the two inequalities in (3.50) into (3.49) yields (3.46).

For the case of $\mu > 0$, we have (3.43). Since $\mathrm{dist}(\mathbf{x}^k, \mathcal{X}_k^*) \le D$ and ε_k's are set to those in (3.27), the inequality in (3.43) indicates

$$\begin{aligned} T_K &\le 2\sum_{k=0}^{K-1}\sqrt{\frac{L_* + \beta_k H}{\mu}}\log\left(\sqrt{\beta_k}\sum_{t=0}^{K-1}\sqrt{\beta_t}\frac{L_*+\beta_k H+\mu}{C_2}D^2\right) + K \\ &= 2\sum_{k=0}^{K-1}\sqrt{\frac{L_* + \beta_k H}{\mu}}\left(\log\frac{\sqrt{\beta_k}D^2\sum_{t=0}^{K-1}\sqrt{\beta_t}}{C_2} + \log\left(L_* + \beta_k H + \mu\right)\right) + K. \end{aligned} \tag{3.51}$$

Therefore, plugging into (3.51) the inequality in (3.41), the upper bounds of $\sum_{k=0}^{K-1}\sqrt{L_* + \beta_k H}$ and β_k in (3.42) and (3.45) respectively, we obtain (3.47) and complete the proof. □

Remark 5 Let us compare the iteration complexity results in Theorems 6 and 7. We see that for the case of $\mu = 0$, as $K > 1$ and σ is big, if $\sqrt{\frac{L_*}{\varepsilon}}$ dominates $\frac{\sqrt{HC_1}}{\varepsilon}$, the iteration complexity result in Theorem 7 is better than that in Theorem 6 (see the numerical results in Table 2), and if $\frac{\sqrt{HC_1}}{\varepsilon}$ dominates $\sqrt{\frac{L_*}{\varepsilon}}$, the two results are similar. For the case of $\mu > 0$, as $K > 1$, the iteration complexity result in Theorem 6 is better than that in Theorem 7.

4 Iteration complexity for primal-dual ε-solutions and nonergodic results

In this section, we show iteration complexity result for obtaining a primal-dual ε-solution by employing the relation between iALM and the inexact proximal point algorithm (iPPA). Also we establish a nonergodic convergence rate result of Algorithm 1 through existing bounds on the primal objective and feasibility errors. Throughout this section, we assume there is no affine equality constraint in (1.1), i.e., we consider the problem

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\ f_0(\mathbf{x}),\ \text{s.t. } f_i(\mathbf{x}) \le 0,\ \forall i \in [m], \tag{4.1}$$

where $f_i, i = 0, 1, \ldots, m$, satisfy the assumptions through (1.2)–(1.3b). We do not include affine equality constraints for the purpose of directly applying existing results in [37,39]. Although results similar to those in [37,39] can possibly be shown for the equality and inequality constrained problem (1.1), we do not extend our discussion but instead formulate any affine equality constraint $\mathbf{a}^\top\mathbf{x} = b$ by two affine inequality constraints $\mathbf{a}^\top\mathbf{x} - b \le 0$ and $-\mathbf{a}^\top\mathbf{x} + b \le 0$ if there is any. Without causing confusion, we will directly use the results established in the previous section by regarding $\mathbf{A}$ and $\mathbf{b}$ as a zero matrix and vector, and thus $\mathbf{y}^k = \mathbf{0}$, $\forall k \ge 0$ if $\mathbf{y}^0 = \mathbf{0}$.

4.1 Relation between iALM and iPPA

Let $\mathcal{L}_0(\mathbf{x}, \mathbf{z})$ be the Lagrangian function of (4.1), namely,

$$\mathcal{L}_0(\mathbf{x}, \mathbf{z}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} z_i f_i(\mathbf{x}),$$

and let $\mathcal{L}_\beta(\mathbf{x}, \mathbf{z})$ be the augmented Lagrangian function of (4.1), defined in the same way as that in (1.8). In addition, let $d_0(\mathbf{z})$ be the Lagrangian dual function, defined as

$$d_0(\mathbf{z}) = \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_0(\mathbf{x}, \mathbf{z}), \text{ for } \mathbf{z} \geq \mathbf{0},$$

and let $d_\beta(\mathbf{z}) \triangleq \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z})$ be the augmented dual function.

Applying Algorithm 1 with $\rho_k = \beta_k$ to (4.1), we have iterates $\{(\mathbf{x}^k, \mathbf{z}^k)\}$ that satisfy:

$$\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k) \leq d_{\beta_k}(\mathbf{z}^k) + \varepsilon_k, \tag{4.2a}$$

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \beta_k \nabla_{\mathbf{z}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k). \tag{4.2b}$$

The iPPA applied to the Lagrangian dual problem $\max_{\mathbf{z}\geq\mathbf{0}} d_0(\mathbf{z})$ iteratively performs the updates:

$$\mathbf{z}^{k+1} \approx \mathcal{M}_{\beta_k}(\mathbf{z}^k), \tag{4.3}$$

where the operator $\mathcal{M}_\beta$ is the proximal mapping of $-\beta d_0$, defined as

$$\mathcal{M}_\beta(\mathbf{z}) = \arg\max_{\mathbf{u}\geq\mathbf{0}} d_0(\mathbf{u}) - \frac{1}{2\beta}\|\mathbf{u} - \mathbf{z}\|^2.$$

In (4.3), the approximation could be measured by the objective error as in (4.2a) or by the gradient norm at the returned point $\mathbf{z}^{k+1}$; see [15] for example.

It was noted in [37] that

$$d_\beta(\mathbf{z}) = \max_{\mathbf{u}\geq\mathbf{0}} d_0(\mathbf{u}) - \frac{1}{2\beta}\|\mathbf{u} - \mathbf{z}\|^2, \tag{4.4}$$

and in addition, if $\hat{\mathbf{x}} \in \mathcal{X}$ satisfies $\mathcal{L}_\beta(\hat{\mathbf{x}}, \mathbf{z}) \leq d_\beta(\mathbf{z}) + \varepsilon$, then (c.f., [23])

$$\|\mathbf{z} + \beta \nabla_{\mathbf{z}} \mathcal{L}_\beta(\hat{\mathbf{x}}, \mathbf{z}) - \mathcal{M}(\mathbf{z})\| \leq \sqrt{2\beta\varepsilon}. \tag{4.5}$$

Therefore, iALM with updates in (4.2) reduces to iPPA in (4.3) with approximation error

$$\|\mathbf{z}^{k+1} - \mathcal{M}_{\beta_k}(\mathbf{z}^k)\| \leq \sqrt{2\beta_k\varepsilon_k}. \tag{4.6}$$

4.2 Iteration complexity for primal-dual ε-solutions

In this subsection, we start from a dual variable that is nearly optimal in terms of an augmented dual objective and obtain a nearly optimal dual variable in terms of a Lagrangian dual function by approximately solving one additional primal subproblem. We first establish the following result.

Lemma 8 *Given $\beta > 0$, assume $\bar{\mathbf{z}}$ to satisfy $d_\beta(\bar{\mathbf{z}}) \ge f_0^* - \delta_1$ for a certain $\delta_1 \ge 0$. Let $\bar{\mathbf{z}}^+ = \bar{\mathbf{z}} + \beta\nabla_{\mathbf{z}}\mathcal{L}_\beta(\hat{\mathbf{x}}, \bar{\mathbf{z}})$, where $\hat{\mathbf{x}} \in \mathcal{X}$ satisfies $\mathcal{L}_\beta(\hat{\mathbf{x}}, \bar{\mathbf{z}}) \le d_\beta(\bar{\mathbf{z}}) + \delta_2$ for some $\delta_2 \ge 0$. Then*

$$d_0(\bar{\mathbf{z}}^+) \ge f_0^* - \delta_1 - \bar{B}\sqrt{2\beta\delta_2}, \tag{4.7}$$

where $\bar{B} = \sqrt{\sum_{i=1}^m B_i^2}$ and B_i's are constants in (3.14).

Proof Denote $\tilde{\mathbf{z}} = \mathcal{M}_\beta(\bar{\mathbf{z}})$. From (4.4), it follows that $d_0(\tilde{\mathbf{z}}) = d_\beta(\bar{\mathbf{z}}) + \frac{1}{2\beta}\|\tilde{\mathbf{z}} - \bar{\mathbf{z}}\|^2$, and thus $d_0(\tilde{\mathbf{z}}) \ge f_0^* - \delta_1$. In addition, we have from (4.5) that $\|\bar{\mathbf{z}}^+ - \tilde{\mathbf{z}}\| \le \sqrt{2\beta\delta_2}$. Note that d_0 is Lipschitz continuous with constant $\bar{B}$; cf. [1, Theorem 6.3.7]. Hence,

$$d_0(\bar{\mathbf{z}}^+) \ge d_0(\tilde{\mathbf{z}}) - \bar{B}\|\bar{\mathbf{z}}^+ - \tilde{\mathbf{z}}\| \ge d_0(\tilde{\mathbf{z}}) - \bar{B}\sqrt{2\beta\delta_2} \ge f_0^* - \delta_1 - \bar{B}\sqrt{2\beta\delta_2},$$

and we complete the proof. □

Let δ_1 be the right hand side of (3.28c) and choose $\delta_2 = \frac{\varepsilon^2}{8\beta\bar{B}^2}$ in Lemma 8. Then from the result in (3.28c), we have the next lemma.

Lemma 9 *Let $\bar{\mathbf{z}}^K$ be the dual solution in Theorem 5 and set $\bar{\mathbf{z}}^{K+} = \bar{\mathbf{z}}^K + \beta\nabla_{\mathbf{z}}\mathcal{L}_\beta(\hat{\mathbf{x}}^K, \bar{\mathbf{z}}^K)$, where $\hat{\mathbf{x}}^K$ satisfies*

$$\mathcal{L}_\beta(\hat{\mathbf{x}}^K, \bar{\mathbf{z}}^K) \le d_\beta(\bar{\mathbf{z}}^K) + \frac{\varepsilon^2}{8\beta\bar{B}^2}. \tag{4.8}$$

Then

$$f_0^* - d_0(\bar{\mathbf{z}}^{K+}) \le \frac{2\varepsilon\|\mathbf{z}^*\|^2}{C_1} + \frac{\varepsilon C_2}{C_1} + \frac{\varepsilon}{2}. \tag{4.9}$$

From (3.16), we are able to find $\hat{\mathbf{x}}^K$ satisfying (4.8) by applying Algorithm 2 and running it to t_K iterations, where

$$t_K = \begin{cases} \left\lceil \dfrac{4\bar{B}D\sqrt{\beta L(\bar{\mathbf{z}}^K)}}{\varepsilon} \right\rceil, & \text{if } \mu = 0, \\ \left\lceil \dfrac{\log\left(\frac{L(\bar{\mathbf{z}}^K)+\mu}{\varepsilon^2/4}\beta\bar{B}^2D^2\right)}{\log 1/\left(1 - \sqrt{\frac{\mu}{L(\bar{\mathbf{z}}^K)}}\right)} \right\rceil, & \text{if } \mu > 0. \end{cases} \tag{4.10}$$

Below we estimate the iteration complexity of obtaining a primal-dual ε-solution.

Theorem 8 (Iteration complexity for primal-dual ε-solution) *Under the assumptions of Theorem 5, let $\bar{\mathbf{x}}^K$ and $\bar{\mathbf{z}}^{K+}$ be respectively given in (3.11) and Lemma 9. Then we have (4.9) and also*

$$\left|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\right| \le \frac{2\varepsilon\|\mathbf{z}^*\|^2}{C_1} + \frac{\varepsilon}{2}\frac{C_2}{C_1}, \tag{4.11a}$$

$$\left\|[\mathbf{f}(\bar{\mathbf{x}}^K)]_+\right\| \le \frac{\varepsilon(1 + \|\mathbf{z}^*\|)^2}{2C_1} + \frac{\varepsilon}{2}\frac{C_2}{C_1}. \tag{4.11b}$$

In addition, to produce $(\bar{\mathbf{x}}^K, \bar{\mathbf{z}}^{K+})$, *at most* $\hat{T}_K$ *gradient evaluations on* $g, f_i, i \in [m]$ *are required, where*

$$\hat{T}_K = T_K + \left\lceil \frac{4\bar{B}D\sqrt{\frac{C_1}{K\varepsilon}(L_* + \frac{C_1 H}{K\varepsilon})}}{\varepsilon} \right\rceil, \text{ if } \mu = 0, \tag{4.12}$$

and

$$\hat{T}_K = T_K + \left\lceil 2\left(\sqrt{\frac{L_*}{\mu}} + \sqrt{\frac{C_1 H}{\mu K \varepsilon}}\right) \log\left(\frac{4\bar{B}^2 D^2 C_1}{K\varepsilon^3}\left(L_* + \mu + \frac{C_1 H}{K\varepsilon}\right)\right) \right\rceil, \text{ if } \mu > 0. \tag{4.13}$$

In the above, T_K *is defined in* (3.29) *for* $\mu = 0$ *and in* (3.30) *for* $\mu > 0$.

Proof We only need to estimate $\hat{T}^K$. From (3.19) and (3.11), it follows that $L(\bar{\mathbf{z}}^K) \le L_* + \beta H$. For $\mu = 0$, plug $\beta = \frac{C_1}{K\varepsilon}$ into the first equation of (4.10) and also note $L(\bar{\mathbf{z}}^K) \le L_* + \frac{C_1 H}{K\varepsilon}$. Then we have $t_K \le \left\lceil \frac{4\bar{B}D\sqrt{\frac{C_1}{K\varepsilon}(L_* + \frac{C_1 H}{K\varepsilon})}}{\varepsilon} \right\rceil$ and obtain (4.12). For $\mu > 0$, we plug $\beta = \frac{C_1}{K\varepsilon}$ and the upper bound of $L(\bar{\mathbf{z}}^K)$ into the second equation of (4.10), and in addition, we use (3.32) to conclude that t_K is no greater than the second term in the right hand side of (4.13). Therefore, we complete the proof. □

Remark 6 Choose C_1 and C_2 such that $C_1 \ge \max\left(4\|\mathbf{z}^*\|^2 + 2C_2, \frac{(1+\|\mathbf{z}^*\|)^2}{2} + \frac{C_2}{2}\right)$. Then (4.9) and (4.11) imply that $(\bar{\mathbf{x}}^K, \bar{\mathbf{z}}^{K+})$ is a primal-dual ε-solution. In addition, for $\mu = 0$, we set $K = \lceil \varepsilon^{-\frac{2}{3}} \rceil$ and have $\hat{T}^K = O(\varepsilon^{-\frac{4}{3}})$; for $\mu > 0$, we set K independent of ε and have $\hat{T}^K = O(\varepsilon^{-\frac{1}{2}}|\log \varepsilon|)$.

4.3 Nonergodic convergence rate and iteration complexity of iALM for primal ε-solutions

For iALM with updates in (4.2) on solving (4.1), [39, Theorem 4] establishes the following bounds on the objective error and feasibility violation:

$$f_0(\mathbf{x}^{k+1}) - f_0(\mathbf{x}^*) \le \varepsilon_k + \frac{\|\mathbf{z}^k\|^2 - \|\mathbf{z}^{k+1}\|^2}{2\beta_k}, \tag{4.14a}$$

$$f_i(\mathbf{x}^{k+1}) \le \frac{|z_i^k - z_i^{k+1}|}{\beta_k}, \ \forall i \in [m]. \tag{4.14b}$$

If in (4.2a), $\varepsilon_k = 0$, $\forall k$, [14, Theorem 2.2] shows that

$$\frac{\|\mathbf{z}^k - \mathbf{z}^{k+1}\|}{\beta_k} \le \frac{\|\mathbf{z}^0 - \mathbf{z}^*\|}{\sum_{t=0}^k \beta_t}. \tag{4.15}$$

Therefore, combining the results in (4.14) with $\varepsilon_k = 0,\ \forall k$ and (4.15), and also noting the boundedness of $\mathbf{z}^k$ from (3.21), one can easily obtain a nonergodic convergence rate result of exact ALM on solving (4.1). However, if $\varepsilon_k > 0$, we do not notice any existing result on estimating $\frac{\|\mathbf{z}^k - \mathbf{z}^{k+1}\|}{\beta_k}$. By bounding $\{\mathbf{z}^k\}$, we can easily establish a bound on this quantity and thus show a nonergodic convergence rate result of iALM.

Theorem 9 (nonergodic convergence rate) *Given a positive integer K and a nonnegative number C_2, choose positive sequences $\{\beta_k\}$ and $\{\varepsilon_k\}$ such that $\sum_{k=0}^{K-1} \beta_k \varepsilon_k \le \frac{C_2}{2}$. Let $\{(\mathbf{x}^k, \mathbf{z}^k)\}_{k=0}^K$ be the sequence generated from the updates in (4.2) with $\mathbf{z}^0 = \mathbf{0}$ on solving (4.1). Then it holds that for any $0 \le k < K$,*

$$\left|f_0(\mathbf{x}^{k+1}) - f_0(\mathbf{x}^*)\right| \le \varepsilon_k + \frac{\left(2\|\mathbf{z}^*\| + \sqrt{C_2}\right)^2}{2\beta_k}, \tag{4.16a}$$

$$\left\|[\mathbf{f}(\mathbf{x}^{k+1})]_+\right\| \le \frac{1}{\beta_k}\left(4\|\mathbf{z}^*\| + 2\sqrt{C_2}\right). \tag{4.16b}$$

Proof Using (3.21) with $\mathbf{y}^* = \mathbf{0}$, we have $\|\mathbf{z}^k\| \le 2\|\mathbf{z}^*\| + \sqrt{C_2}$. By triangle inequality, it holds $\|\mathbf{z}^k - \mathbf{z}^{k+1}\| \le 4\|\mathbf{z}^*\| + 2\sqrt{C_2}$. Then the results in (4.16) directly follow from (4.14). □

Remark 7 From the results in (4.16), we see that to have $\{\mathbf{x}^k\}$ to be a minimizing sequence of (4.1), we need $\beta_k \to \infty$ and $\varepsilon_k \to 0$ as $k \to \infty$. Hence, setting $\{\beta_k\}$ to a constant sequence will not be a valid option.

Below we set parameters according to Setting 2 and estimate the iteration complexity of iALM on solving (4.1) by applying Nesterov's optimal first-order method to (4.2a). Again, note that the results in Theorem 9 do not need specific structure of (4.1) except convexity. Hence, if the problem has richer structures, one can apply more efficient methods to find $\mathbf{x}^{k+1}$ that satisfies (4.2a).

Theorem 10 (nonergodic iteration complexity) *Given a positive integer K and positive numbers C_1, C_2, choose positive sequences $\{\rho_k\}$ and $\{\beta_k\}$ according to Setting 2. In addition, choose $\{\varepsilon_k\}$ according to (3.25) for both cases of $\mu = 0$ and $\mu > 0$, or choose $\{\varepsilon_k\}$ according to (3.26) for the case of $\mu = 0$ and (3.27) for $\mu > 0$. Let $\{(\mathbf{x}^k, \mathbf{z}^k)\}_{k=0}^K$ be the sequence generated from Algorithm 1 with $\mathbf{y}^k = \mathbf{0},\ \forall k$, and $\mathbf{z}^0 = \mathbf{0}$ on solving (4.1). Then*

$$\left|f_0(\mathbf{x}^K) - f_0(\mathbf{x}^*)\right| \le \frac{\varepsilon}{2}\frac{C_2}{C_1} + \frac{\varepsilon\sigma}{2C_1(\sigma-1)}\left(2\|\mathbf{z}^*\| + \sqrt{C_2}\right)^2, \tag{4.17a}$$

$$\left\|[\mathbf{f}(\mathbf{x}^K)]_+\right\| \le \frac{\varepsilon\sigma}{C_1(\sigma-1)}\left(4\|\mathbf{z}^*\| + 2\sqrt{C_2}\right). \tag{4.17b}$$

If $\{\varepsilon_k\}$ is chosen according to (3.25) for both cases of $\mu = 0$ and $\mu > 0$, the total number T_K of gradient evaluations is given in (3.36) and (3.37) respectively; if $\{\varepsilon_k\}$ is set according to (3.26) for the case of $\mu = 0$ and (3.27) for $\mu > 0$, then T_K is given in (3.46) for $\mu = 0$ and (3.47) for $\mu > 0$.

Proof Note that β_k is increasing with respect to k. Hence, the ε_k given in both (3.26) and (3.27) is decreasing, and thus

$$\varepsilon_{K-1} \le \frac{\sum_{t=0}^{K-1} \beta_t \varepsilon_t}{\sum_{t=0}^{K-1} \beta_t} \le \frac{\varepsilon}{2} \frac{C_2}{C_1}.$$

If $\{\varepsilon_k\}$ is chosen according to (3.25) for both cases of $\mu = 0$ and $\mu > 0$, then the above bound on ε_{K-1} obviously holds. In addition, from (3.45), we have

$$\beta_{K-1} \ge \frac{C_1(\sigma - 1)}{\varepsilon\sigma}.$$

Therefore, plugging into (4.16) the bounds on ε_{K-1} and β_{K-1} gives the desired results in (4.17).

The bounds on the total number T_K of gradient evaluations follow from the same arguments as in the proofs of Theorems 6 and 7. Hence, we complete the proof. □

Remark 8 From the results in (4.17), we see that if

$$C_1 \ge \max\left(\frac{C_2}{2} + \frac{\sigma}{2(\sigma-1)}\left(2\|\mathbf{z}^*\| + \sqrt{C_2}\right)^2, \ \frac{\sigma}{(\sigma-1)}\left(4\|\mathbf{z}^*\| + 2\sqrt{C_2}\right)\right), \quad (4.18)$$

then $\mathbf{x}^K$ is a primal ε-solution to (4.1). If $\|\mathbf{z}^*\| \ge \frac{6}{5}$, $C_2 = \|\mathbf{z}^*\|^2$, and $\frac{\sigma}{\sigma-1} \approx 1$ (e.g., $\sigma = 10$ is often used), then the C_1 in (4.18) is roughly twice of that in (3.34) by assuming no affine constraint. For the iteration complexity, if $\sqrt{\frac{L_*}{\varepsilon}}$ dominates $\frac{\sqrt{H}\|\mathbf{z}^*\|}{\varepsilon}$, then the nonergodic result is roughly $\sqrt{2}$ times of the ergodic result for both convex and strongly convex cases. If $\frac{\sqrt{H}\|\mathbf{z}^*\|}{\varepsilon}$ dominates, then the former would be roughly twice of the latter for the convex case, but still roughly $\sqrt{2}$ times for the strongly convex case. However, in either case, both ergodic and nonergodic results have the same order of complexity.

5 Comparison with several existing results

In this section, we compare our iteration complexity results to several existing ones.

5.1 Affinely constrained convex problems

Let us compare our iteration complexity to those in Nedelcu et al. [29] and Liu et al. [25], both of which consider the affinely constrained convex problem (1.13) with possibly nonsmooth f_0. The former defines a primal-dual ε-solution in a way similar to ours. It shows that to reach a primal-dual ε-solution,[4] a nonaccelerated iDGM requires

[4] Nedelcu et al. [29] assumes every subproblem solved to the condition $\langle \tilde{\nabla}\mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k), \mathbf{x} - \mathbf{x}^{k+1}\rangle \ge -O(\varepsilon)$, $\forall \mathbf{x} \in \mathcal{X}$, which is implied by $\mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k) - \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^k) \le O(\varepsilon^2)$ if $\mathcal{L}_\beta$ is smooth with respect to $\mathbf{x}$.

$O(\varepsilon^{-1})$ outer iterations and every $\mathbf{x}$-subproblem solved to an accuracy $O(\varepsilon^2)$, and an accelerated iDGM requires $O(\varepsilon^{-\frac{1}{2}})$ outer iterations and every $\mathbf{x}$-subproblem solved to an accuracy $O(\varepsilon^3)$. Assume the composite structure of the objective, i.e., $f_0 = g + h$. Then by applying Nesterov's optimal first-order method to each subproblem, both the nonaccelerated iDGM and accelerated iDGM in Nedelcu et al. [29] would need $O(\varepsilon^{-2})$ gradient evaluations to produce a primal-dual ε-solution for convex problems. Hence, as mentioned in Remark 6, our result is better by an order of $\varepsilon^{-\frac{2}{3}}$. For strongly convex problems, the accelerated iDGM would need $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ gradient evaluations, which is in the same order as our result.

Assume that $f_0 = g + h$ in (1.13) and g is smooth. In [25], a point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is defined as a primal-dual ε-solution of (1.13) if

$$\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| \le \sqrt{\varepsilon}, \quad \left\langle \nabla g(\bar{\mathbf{x}}) + \mathbf{A}^\top \bar{\mathbf{y}}, \bar{\mathbf{x}} - \mathbf{x} \right\rangle + h(\bar{\mathbf{x}}) - h(\mathbf{x}) \le \varepsilon, \ \forall \mathbf{x} \in \mathcal{X}. \tag{5.1}$$

This definition is different from ours. In addition, [25] adopts a directly verifiable stopping condition. It is shown that $O(\varepsilon^{-2})$ gradient evaluations are required to produce a primal-dual ε-solution. In the appendix, we show that if $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ satisfies (5.1), it must be an $O(\sqrt{\varepsilon})$-solution in Definition 2. Hence directly applying the result in [25] gives the iteration complexity $O(\varepsilon^{-4})$ to produce an ε-solution in our definition. On the other hand, let $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ be a primal-dual ε-solution in Definition 2. We can obtain an $O(\sqrt{\varepsilon})$-solution of (5.1) by performing one additional proximal gradient update. Hence, directly applying our result in Theorem 8 gives the iteration complexity $O(\varepsilon^{-\frac{8}{3}})$ to produce an ε-solution in (5.1). Therefore, it is not clear whether the result in [25] or our result is better.

5.2 General convex problems

In this subsection, we compare our complexity result to those in [26], which was published online after our first submission. A more general convex cone program is considered in [26]. Specialized to the functional constrained problem (1.1), [26, Algorithm 4] also solves the ALM subproblem inexactly to update the primal iterate, the same as in (1.9). It requires $O(\varepsilon^{-\frac{7}{4}})$ gradient evaluations to produce an ε-KKT point $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$ that satisfies $\bar{\mathbf{z}} \ge \mathbf{0}$ and

$$\operatorname{dist}\left(\mathbf{0}, \partial f_0(\bar{\mathbf{x}}) + \mathcal{N}_{\mathcal{X}}(\bar{\mathbf{x}}) + \mathbf{A}^\top \bar{\mathbf{y}} + \sum_{j=1}^m \bar{z}_j \nabla f_j(\bar{\mathbf{x}})\right) \le \varepsilon, \tag{5.2a}$$

$$\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| + \left\|[\mathbf{f}(\bar{\mathbf{x}})]_+\right\| \le \varepsilon, \tag{5.2b}$$

$$\sqrt{\sum_{j:\bar{z}_j > 0} f_j(\bar{\mathbf{x}})^2} \le \varepsilon. \tag{5.2c}$$

A modified method, i.e., [26, Algorithm 5], is also given, and at each outer iteration, it inexactly solves a perturbed subproblem that is strongly convex. More specifically, its k-th subproblem is

$$\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + \frac{1}{2\beta_k}\|\mathbf{x} - \mathbf{x}^k\|^2, \tag{5.3}$$

which is solved by Nesterov's optimal first-order method until a point $\mathbf{x}^{k+1}$ is found such that

$$\operatorname{dist}\left(\mathbf{0}, \partial_{\mathbf{x}}\mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{y}^k, \mathbf{z}^k) + \mathcal{N}_{\mathcal{X}}(\mathbf{x}^{k+1}) + \frac{1}{\beta_k}(\mathbf{x}^{k+1} - \mathbf{x}^k)\right) \le \varepsilon_k, \tag{5.4}$$

for some $\varepsilon_k > 0$. The modified method can achieve a significantly better complexity result $O(\varepsilon^{-1}|\log\varepsilon|)$ to yield an ε-KKT point defined in (5.2). In addition, note that the stopping condition in (5.4) can be checked.

By the convexity of f_j's and the optimality condition (2.2), one can show that if $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$ satisfies all conditions in (5.2), then it must be a primal-dual $O(\varepsilon)$-solution in Definition 2. Hence, the complexity result of the modified method in [26] is better than ours. As shown in Sect. 6, however, its numerical performance can be significantly worse than the iALM under our setting. Similar to the discussion in Sect. 5.1, for the unmodified method in [26], it is not clear whether its complexity result or ours is better.

5.3 Iteration complexity from existing results on iPPA

Through relating iALM and iPPA, iteration complexity result can be obtained from existing results about iPPA to produce near-optimal dual solution. On solving problem $\min_{\mathbf{z}} \phi(\mathbf{z})$, [15] analyzes the iPPA with iterative update:

$$\mathbf{z}^{k+1} \approx \arg\min_{\mathbf{z}} \phi(\mathbf{z}) + \frac{1}{2\beta_k}\|\mathbf{z} - \hat{\mathbf{z}}^k\|^2.$$

If the above approximation error satisfies

$$\|\mathbf{z}^{k+1} - \mathbf{prox}_{\beta_k\phi}(\hat{\mathbf{z}}^k)\| = O(1/k^a), \tag{5.5}$$

for a certain number $a > \frac{1}{2}$, and the parameter β_k is increasing, then by choosing specifically designed $\hat{\mathbf{z}}^k$, [15] shows that

$$\phi(\mathbf{z}^k) - \phi(\mathbf{z}^*) = O(1/k^2) + O(1/k^{2a-1}).$$

From our discussion in Sect. 4.1, if $\varepsilon_k = O(\frac{1}{k^{2a}\beta_k})$ in (4.2a), then we have (5.5) holds with $\phi = -d_0$, and thus obtain the convergence rate in terms of dual function:

$$d_0(\mathbf{z}^*) - d_0(\mathbf{z}^k) = O(1/k^2) + O(1/k^{2a-1}).$$

Note that $\mathbf{z}^k$ is bounded from the summability of $\beta_k\varepsilon_k$ and the proof of Lemma 7. Hence, setting β_k to a constant for all k and applying Nesterov's optimal first-order method to each subproblem in (4.2a), we need $O(k^a)$ gradient evaluations.

Let $a = \frac{3}{2}$. Then $K = O(1/\sqrt{\varepsilon})$ iPPA iterations are required to obtain a dual ε-solution, i.e., $d_0(\mathbf{z}^K) \geq d_0(\mathbf{z}^*) - \varepsilon$, and the total number of gradient evaluations is

$$T_K = \sum_{k=1}^{K} O(k^{\frac{3}{2}}) = O(K^{\frac{5}{2}}) = O(\varepsilon^{-\frac{5}{4}}).$$

However, it is not clear how to measure the quality of the primal iterates.

6 Numerical results

In this section, we conduct numerical experiments on the quadratically constrained quadratic programming (QCQP):

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \ & \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x} + \mathbf{c}_0^\top \mathbf{x}, \\ \text{s.t.} \ & \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_j \mathbf{x} + \mathbf{c}_j^\top \mathbf{x} + d_j \leq 0, \ j = 1, \ldots, m, \\ & x_i \in [l_i, u_i], \ i = 1, \ldots, n. \end{aligned} \tag{6.1}$$

Clearly, (6.1) is one example of (1.1) with $\mathcal{X} = \times_{i=1}^n [l_i, u_i]$, $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x} + \mathbf{c}_0^\top \mathbf{x}$, $h \equiv 0$, $f_j(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_j \mathbf{x} + \mathbf{c}_j^\top \mathbf{x} + d_j$ for $j \in [m]$, $\mathbf{A} = \mathbf{0}$, and $\mathbf{b} = \mathbf{0}$.

We conduct two sets of tests. The first one is to verify the established theoretical results and compare the iALM with three different settings of parameters, and the second is to compare the iALM with our setting to a modified iALM in [26].

6.1 First set of tests

Three QCQP instances are made. The first two instances are convex, and the third one is strongly convex. For all three instances, we set $n = 100$, $m = 5$ and $l_i = -1$, $u_i = 1$, $\forall i$. The vectors $\mathbf{c}_j$, $j = 0, 1, \ldots, m$ are generated following Gaussian distribution, and the scalars d_j, $j = 1, \ldots, m$ are made negative. This way, all inequalities in (6.1) hold strictly at the origin $\mathbf{x} = \mathbf{0}$, and thus the KKT conditions are satisfied at the optimal solution. $\mathbf{Q}_j$, $j = 0, 1, \ldots, m$ are randomly generated and symmetric positive semidefinite. $\mathbf{Q}_0$ is rank-deficient for the first two instances and full-rank for the third one. The data in the first two instances are the same except $\mathbf{Q}_0$, which is 100 times in the second instance of that in the first instance.

For all instances, we set $\varepsilon = 10^{-3}$, $C_1 = 1$, $C_2 = \|\mathbf{u} - \mathbf{l}\|$, and $K = 10$, and the initial primal-dual point is set to zero vector. The algorithm parameters $\{(\beta_k, \rho_k, \varepsilon_k)\}_{k=0}^{K-1}$ are set in three different ways corresponding to Theorems 5, 6, and 7 respectively, where $\sigma = 10$ is used for the geometrically increasing penalty. On finding $\mathbf{x}^{k+1}$ by applying Algorithm 2 to $\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$, we terminate the algorithm if the iteration number exceeds 10^6 or

$$\text{dist}\left(-\nabla_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{z}^k), \mathcal{N}_{\mathcal{X}}(\mathbf{x}^{k+1})\right) \leq \frac{\varepsilon_k}{\|\mathbf{u} - \mathbf{l}\|}, \tag{6.2}$$

where $\mathcal{X} = \times_{i=1}^{n}[l_i, u_i]$. Since $\mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{z}^k)$ is convex about $\mathbf{x}$, and $\|\mathbf{u}-\mathbf{l}\|$ is the diameter of the feasible set $\mathcal{X}$, the condition in (6.2) guarantees that $\mathbf{x}^{k+1}$ satisfies (1.9).

We report the difference of primal objective value and optimal value, the feasibility violation at both actual iterate $\mathbf{x}^k$ and the weighted averaged point $\bar{\mathbf{x}}^k = \sum_{t=1}^{k} \mathbf{x}^t / \sum_{t=1}^{k} \beta_t$, and also the difference of dual objective value and optimal value at the actual dual iterate $\mathbf{z}^k$. Since a KKT point exists for the instances, the optimal dual objective value equals the optimal primal objective value. The optimal solution and dual objective values are computed by CVX [13]. In addition, to compare the iteration complexity, we also report the number of gradient evaluations and function evaluations for each outer iteration. The results are provided in Tables 1, 2, and 3 respectively for the three instances. We also report the results from quadratic penalty method, which corresponds to setting $K = 1$ (see the discussions in Remark 3).

From the results, we can clearly see that the quadratic penalty method is worse, namely, running a single iALM step with a big penalty parameter is significantly worse than running multiple steps with smaller penalty parameters. Also, we see that the iALM with three different settings yields the last actual iterate $\mathbf{x}^K$ and the averaged point $\bar{\mathbf{x}}^K$ of similar accuracy. For all three instances, to produce similarly accurate solutions, the iALM with constant penalty requires more gradient and function evaluations than that with geometrically increasing penalty. Furthermore, the iALM with geometrically increasing penalty and constant error requires fewest gradient and function evaluations on the first and third instances. However, the setting of geometrically increasing penalty and adaptive error is the best for iALM on the second instance. That is because the gradient Lipschitz constant of the objective in the second instance is significantly bigger than that in the first instance, in which case the bound on T_K in (3.46) is smaller than that in (3.36).

6.2 Second set of tests

We randomly generate 20 convex QCQP instances, in the same way as we generate the first instance in the previous subsection. Among them, 10 instances have size of $n = 100$ and $m = 5$ and another 10 of $n = 1,000$ and $m = 10$. The parameters $\{(\beta_k, \rho_k, \varepsilon_k)\}$ of the iALM are set according to Theorem 6, and all other settings are the same as in the first set of tests. We compare to the modified method [26, Algorithm 5], which inexactly solves the perturbed subproblem (5.3) at the k-th outer iteration until the stopping condition (5.4) holds. Its parameters are set to $\beta_k = 1.5\beta_0$ and $\varepsilon_k = 0.6\varepsilon_0$ with $\beta_0 = \varepsilon_0 = 0.1$. This setting appears to be the best for the modified method in this test. The iALM runs to 10 outer iterations, and the modified method is terminated once it produces a point satisfying all conditions in (5.2) with $\varepsilon = 10^{-3}$. We report the results in Table 4 for the size of $n = 100, m = 5$ and in Table 5 for the size of $n = 1,000, m = 10$. In the tables, `objErr` is computed as $|f(\bar{\mathbf{x}}) - f^*|$, where $\bar{\mathbf{x}}$ is the last iterate; `pres`, `dres`, and `compl` respectively stand for the primal residual, dual residual, and the violation of complementarity condition computed by the measures in (5.2). From the results, we see that for each tested instance, the iALM under our setting takes significantly shorter time and also achieves higher accuracy (by any measure among `objErr`, `pres`, `dres`, and `compl`) than the modified method in [26]. Although we cannot guarantee an ε-KKT point, the numerical results clearly show that it is achieved.

7 Concluding remarks

We have established ergodic and also nonergodic convergence rate results of iALM for general constrained convex programs. In addition, we have shown that to reach a primal ε-solution, it is sufficient to evaluate gradients of smooth part in the objective and the functions in the inequality constraints for $O(\varepsilon^{-1})$ times if the objective is convex and $O(\varepsilon^{-\frac{1}{2}}|\log \varepsilon|)$ times if the objective is strongly convex. For the convex case, the iteration complexity result is optimal, and for the strongly convex case, the result is nearly optimal. Furthermore, we have shown that to produce a primal-dual ε-solution, the result is $O(\varepsilon^{-\frac{4}{3}})$ for convex case and still $O(\varepsilon^{-\frac{1}{2}}|\log \varepsilon|)$ for strongly convex case.

A Relation of the primal-dual ε-solutions in Definition 2 and (5.1)

In this section, on linearly constrained problems in the form of (1.13) with $f_0 = g + h$, we compare the two different definitions of primal-dual ε-solutions given in Definition 2 and (5.1). The analysis in the second part follows from the proof of Theorem 2.1 in [25].

First, let $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ be a point satisfying (5.1). Then it follows from (2.2) that

$$f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) \geq -\langle \mathbf{y}^*, \mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\rangle \geq -\|\mathbf{y}^*\|\sqrt{\varepsilon}.$$

In addition, we have from the convexity of g and (5.1) that for any $\mathbf{x} \in \mathcal{X}$ and any constant $\beta > 0$,

$$\begin{aligned} & f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}) - \langle \bar{\mathbf{y}}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle \\ &\quad = f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}) - \langle \mathbf{A}^\top \bar{\mathbf{y}}, \mathbf{x} - \bar{\mathbf{x}}\rangle - \langle \bar{\mathbf{y}}, \mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\rangle \\ &\quad \leq \langle \nabla g(\bar{\mathbf{x}}) + \mathbf{A}^\top \bar{\mathbf{y}}, \bar{\mathbf{x}} - \mathbf{x}\rangle + h(\bar{\mathbf{x}}) - h(\mathbf{x}) - \langle \bar{\mathbf{y}}, \mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\rangle \\ &\quad \leq \varepsilon + \|\bar{\mathbf{y}}\|\sqrt{\varepsilon}. \end{aligned}$$

Letting $\mathbf{x} = \mathbf{x}^*$ in the above inequality gives $f_0(\bar{\mathbf{x}}) - f_0(\mathbf{x}^*) \leq \varepsilon + \|\bar{\mathbf{y}}\|\sqrt{\varepsilon}$, and minimizing the left hand side about $\mathbf{x} \in \mathcal{X}$ yields $f_0(\bar{\mathbf{x}}) - d_0(\bar{\mathbf{y}}) \leq \varepsilon + \|\bar{\mathbf{y}}\|\sqrt{\varepsilon}$. Hence, $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is an $O(\sqrt{\varepsilon})$-solution in Definition 2.

On the other hand, let $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ be a primal-dual ε-solution in Definition 2. Let

$$\mathcal{L}_0(\mathbf{x}, \mathbf{y}) = f_0(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle$$

and

$$\bar{\mathbf{x}}^+ = \operatorname*{arg\,min}_{\mathbf{x}\in\mathcal{X}} \left\langle \nabla g(\bar{\mathbf{x}}) + \mathbf{A}^\top \bar{\mathbf{y}}, \mathbf{x} \right\rangle + h(\mathbf{x}) + \frac{L_0}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|^2. \tag{A.1}$$

Table 1 Results by quadratic penalty method (i.e., iALM with $K = 1$) and iALM with three different settings on solving an instance of the QCQP problem (6.1)

#OutIter	#Grad	#Fun	$\lvert f_0(\mathbf{x}^k) - f_0^*\rvert$	$\lVert[\mathbf{f}(\mathbf{x}^k)]_+\rVert$	$\lvert f_0(\bar{\mathbf{x}}^k) - f_0^*\rvert$	$\lVert[\mathbf{f}(\bar{\mathbf{x}}^k)]_+\rVert$	$f_0^* - d_0(\mathbf{z}^k)$
Quadratic penalty method							
	1,000,000	2,709,547	3.4843e−05	8.6455e−05	3.4843e−05	8.6455e−05	9.3664e−06
Constant penalty and constant error							
0			1.9949e+01	0.0000e+00	1.9949e+01	0.0000e+00	4.8272e+00
1	561,406	1,521,166	7.4624e−05	8.6422e−04	7.4624e−05	8.6422e−04	7.2143e−09
2	18	48	6.3627e−08	1.1199e−08	3.7280e−05	4.3212e−04	7.9194e−09
3	1	2	6.4334e−08	7.5897e−09	2.4832e−05	2.8808e−04	6.4559e−09
4	1	2	6.4465e−08	4.4346e−09	1.8608e−05	2.1606e−04	9.0242e−09
5	12	33	6.4518e−08	3.6546e−09	1.4873e−05	1.7285e−04	7.4387e−09
6	1	2	6.4424e−08	9.6058e−10	1.2384e−05	1.4404e−04	8.0724e−09
7	1	2	6.4468e−08	6.9566e−09	1.0605e−05	1.2346e−04	5.4207e−09
8	7	19	6.4575e−08	5.4913e−09	9.2717e−06	1.0803e−04	7.9159e−09
9	5	14	6.4194e−08	1.0084e−08	8.2344e−06	9.6026e−05	6.0310e−09
10	25	68	6.4380e−08	3.6994e−09	7.4045e−06	8.6423e−05	7.0640e−09
Geometrically increasing penalty and constant error							
0			1.9949e+01	0.0000e+00	1.9949e+01	0.0000e+00	4.8272e+00
1	79	219	4.8272e+00	1.4965e+02	4.8272e+00	1.4965e+02	4.8070e+00
2	25	68	4.8244e+00	1.4620e+02	4.8249e+00	1.4651e+02	4.6124e+00
3	63	171	4.6110e+00	1.1372e+02	4.6449e+00	1.1676e+02	3.2827e+00
4	48	131	2.6482e+00	3.9786e+01	2.9365e+00	4.6130e+01	7.7828e−01
5	148	404	3.3960e−01	4.1060e+00	6.4690e−01	7.8394e+00	8.1817e−03
6	419	1141	3.6545e−03	4.4951e−02	6.9539e−02	8.0933e−01	1.1014e−06
7	68	191	4.5173e−06	6.0954e−05	6.9754e−03	8.0843e−02	2.3645e−08
8	28	81	8.1256e−08	9.5097e−08	6.9764e−04	8.0830e−03	2.4096e−08
9	4	17	8.3883e−08	7.1671e−09	6.9690e−05	8.0828e−04	2.3640e−08
10	3	15	8.3654e−08	4.6715e−10	6.8937e−06	8.0828e−05	2.3715e−08
Geometrically increasing penalty and adaptive error							
0			1.9949e+01	0.0000e+00	1.9949e+01	0.0000e+00	4.8272e+00
1	12	37	4.7787e+00	1.5604e+02	4.7787e+00	1.5604e+02	4.8062e+00
2	2	5	4.8180e+00	1.5082e+02	4.8163e+00	1.5128e+02	4.6055e+00
3	6	18	4.6268e+00	1.1583e+02	4.6614e+00	1.1912e+02	3.2565e+00
4	17	46	2.6514e+00	3.9850e+01	2.9424e+00	4.6322e+01	7.6990e−01
5	41	116	3.3800e−01	4.1017e+00	6.4579e−01	7.8268e+00	8.0789e−03
6	103	285	3.5713e−03	4.4256e−02	6.9331e−02	8.0726e−01	5.3699e−06
7	56	158	8.1959e−08	9.2275e−05	6.9512e−03	8.0661e−02	2.0134e−06
8	208	570	1.9113e−06	3.8904e−06	6.9371e−04	8.0654e−03	2.4591e−07
9	788	2141	5.3716e−07	1.0078e−07	6.8921e−05	8.0612e−04	6.1617e−08
10	2751	7460	1.2807e−07	4.2665e−09	6.7830e−06	8.0458e−05	8.2154e−09

In this instance, $\mathbf{Q}_j$ is symmetric positive semidefinite for each $j = 0, 1, \ldots, m$, and $\mathbf{Q}_0$ is singular. All $\mathbf{Q}_j$'s have similarly large spectral norm

Table 2 Results by iALM with three different settings on solving an instance of the QCQP problem (6.1)

#OutIter	#Grad	#Fun	$\lvert f_0(\mathbf{x}^k) - f_0^*\rvert$	$\lVert[\mathbf{f}(\mathbf{x}^k)]_+\rVert$	$\lvert f_0(\bar{\mathbf{x}}^k) - f_0^*\rvert$	$\lVert[\mathbf{f}(\bar{\mathbf{x}}^k)]_+\rVert$	$f_0^* - d_0(\mathbf{z}^k)$
Quadratic penalty method							
	600, 703	1, 627, 648	1.2262e−04	3.5018e−04	1.2262e−04	3.5018e−04	5.5078e−08
Constant penalty and constant error							
0			2.4292e+00	0.0000e+00	2.4292e+00	0.0000e+00	8.7897e+00
1	106, 555	288, 743	1.2258e−03	3.5007e−03	1.2258e−03	3.5007e−03	3.5986e−07
2	858	2325	3.4720e−07	1.0421e−06	6.1307e−04	1.7507e−03	3.4319e−08
3	14	38	1.2626e−08	3.4507e−09	4.0871e−04	1.1671e−03	3.3900e−08
4	4	11	1.3058e−08	1.4235e−10	3.0653e−04	8.7535e−04	3.3936e−08
5	1	2	1.2586e−08	3.7554e−08	2.4522e−04	7.0027e−04	3.4521e−08
6	3	9	1.3122e−08	4.5633e−10	2.0435e−04	5.8356e−04	3.4121e−08
7	1	2	1.3257e−08	5.8908e−10	1.7516e−04	5.0019e−04	3.3762e−08
8	3	10	1.2685e−08	2.8823e−08	1.5326e−04	4.3767e−04	3.4946e−08
9	2	8	1.3242e−08	1.6142e−09	1.3623e−04	3.8904e−04	3.3745e−08
10	3	11	1.2717e−08	2.3642e−08	1.2261e−04	3.5013e−04	3.4793e−08
Geometrically increasing penalty and constant error							
0			2.4292e+00	0.0000e+00	2.4292e+00	0.0000e+00	8.7897e+00
1	1006	2741	8.7897e+00	9.3636e+01	8.7897e+00	9.3636e+01	8.7818e+00
2	91	248	8.7896e+00	9.3581e+01	8.7896e+00	9.3586e+01	8.7029e+00
3	2400	6503	8.7021e+00	7.5041e+01	8.7109e+00	7.6215e+01	8.1272e+00
4	1885	5108	8.3403e+00	5.7872e+01	8.3897e+00	5.9057e+01	4.8927e+00
5	1397	3787	3.2123e+00	1.3502e+01	3.7378e+00	1.6466e+01	8.5509e−01
6	803	2178	4.9393e−01	1.5197e+00	8.2698e−01	2.6284e+00	3.0074e−02
7	501	1360	1.5245e−02	4.4612e−02	9.7567e−02	2.8394e−01	3.0183e−05
8	468	1273	5.2378e−05	1.5371e−04	9.8217e−03	2.8277e−02	6.8119e−09
9	38	109	6.4356e−08	2.2690e−07	9.8242e−04	2.8254e−03	2.3421e−08
10	4	16	1.0609e−08	8.8464e−10	9.8234e−05	2.8252e−04	2.3899e−08
Geometrically increasing penalty and adaptive error							
0			2.4292e+00	0.0000e+00	2.4292e+00	0.0000e+00	8.7897e+00
1	127	359	8.7040e+00	8.9660e+01	8.7040e+00	8.9660e+01	8.7821e+00
2	10	27	8.7100e+00	8.9176e+01	8.7097e+00	8.9220e+01	8.7072e+00
3	94	255	8.7087e+00	8.0403e+01	8.7093e+00	8.1163e+01	8.1040e+00
4	355	962	8.3368e+00	5.7827e+01	8.3873e+00	5.9319e+01	4.8760e+00
5	532	1443	3.1974e+00	1.3416e+01	3.7241e+00	1.6386e+01	8.6191e−01
6	221	600	4.9539e−01	1.5239e+00	8.2695e−01	2.6281e+00	3.0145e−02
7	216	588	1.5121e−02	4.4260e−02	9.7439e−02	2.8355e−01	4.8526e−05
8	287	783	5.8283e−05	1.7224e−04	9.8150e−03	2.8258e−02	6.1104e−07
9	244	667	2.5987e−07	9.9441e−07	9.8197e−04	2.8243e−03	1.5846e−07
10	927	2518	2.2079e−09	6.0770e−08	9.8206e−05	2.8247e−04	3.2921e−08

In this instance, $\mathbf{Q}_j$ is symmetric positive semidefinite for each $j = 0, 1, \ldots, m$. $\mathbf{Q}_0$ is singular, and its spectral norm is about 100 times of that of every other $\mathbf{Q}_j$

Table 3 Results by iALM with three different settings on solving a strongly convex instance of the QCQP problem (6.1)

#OutIter	#Grad	#Fun	$\lvert f_0(\mathbf{x}^k) - f_0^*\rvert$	$\Vert[\mathbf{f}(\mathbf{x}^k)]_+\Vert$	$\lvert f_0(\bar{\mathbf{x}}^k) - f_0^*\rvert$	$\Vert[\mathbf{f}(\bar{\mathbf{x}}^k)]_+\Vert$	$f_0^* - d_0(\mathbf{z}^k)$
Quadratic penalty method							
	11,407	30,943	1.6555e−06	4.1227e−05	1.6555e−06	4.1227e−05	9.8318e−11
Constant penalty and constant error							
0			1.3704e+01	0.0000e+00	1.3704e+01	0.0000e+00	7.7888e−01
1	4111	11,170	1.6951e−05	4.1227e−04	1.6951e−05	4.1227e−04	2.4024e−10
2	10	28	4.5144e−08	9.5417e−09	8.4530e−06	2.0614e−04	2.6835e−11
3	1	2	4.5496e−08	1.1679e−09	5.6202e−06	1.3742e−04	5.0360e−12
4	1	2	4.5470e−08	0.0000e+00	4.2038e−06	1.0307e−04	2.0108e−11
5	1	2	4.5410e−08	5.2874e−10	3.3539e−06	8.2455e−05	1.5403e−11
6	1	2	4.5423e−08	1.1857e−10	2.7874e−06	6.8712e−05	1.6200e−11
7	1	2	4.5417e−08	1.4938e−10	2.3827e−06	5.8896e−05	1.5124e−11
8	1	2	4.5420e−08	1.4935e−11	2.0792e−06	5.1534e−05	1.5547e−11
9	1	2	4.5417e−08	4.2566e−11	1.8431e−06	4.5808e−05	1.5184e−11
10	1	2	4.5417e−08	5.2216e−12	1.6543e−06	4.1227e−05	1.5563e−11
Geometrically increasing penalty and constant error							
0			1.3704e+01	0.0000e+00	1.3704e+01	0.0000e+00	7.7888e−01
1	16	47	7.7888e−01	4.0032e+01	7.7888e−01	4.0032e+01	7.7743e−01
2	6	18	7.7879e−01	3.9621e+01	7.7880e−01	3.9658e+01	7.6323e−01
3	9	26	7.7142e−01	3.5936e+01	7.7269e−01	3.6303e+01	6.4109e−01
4	11	31	5.6681e−01	1.8568e+01	6.0051e−01	2.0298e+01	1.8942e−01
5	26	75	8.2135e−02	2.0563e+00	1.4945e−01	3.8384e+00	2.6057e−03
6	56	158	1.1249e−03	2.7458e−02	1.6675e−02	4.0689e−01	4.4251e−07
7	31	91	1.5987e−06	4.0394e−05	1.6772e−03	4.0708e−02	5.9485e−09
8	2	12	4.9762e−08	0.0000e+00	1.6776e−04	4.0703e−03	2.6423e−09
9	2	11	3.6101e−08	3.0525e−08	1.6745e−05	4.0705e−04	2.9782e−09
10	2	10	3.6919e−08	1.0977e−09	1.6412e−06	4.0706e−05	5.3375e−09
Geometrically increasing penalty and adaptive error							
0			1.3704e+01	0.0000e+00	1.3704e+01	0.0000e+00	7.7888e−01
1	1	6	2.1524e+00	1.8206e+01	2.1524e+00	1.8206e+01	7.7827e−01
2	1	6	1.8807e−01	2.3036e+01	2.4599e−01	2.2315e+01	7.7005e−01
3	1	6	3.6907e−01	2.8225e+01	3.3200e−01	2.7560e+01	6.7280e−01
4	3	12	5.1655e−01	1.7712e+01	5.1962e−01	1.8616e+01	2.2134e−01
5	8	25	8.4506e−02	2.2023e+00	1.4307e−01	3.8034e+00	3.4037e−03
6	24	71	8.4618e−04	3.1440e−02	1.5993e−02	4.0623e−01	3.8825e−06
7	75	210	4.8570e−05	3.3014e−05	1.5905e−03	4.0504e−02	1.7042e−06
8	250	684	5.5023e−06	4.6695e−06	1.5741e−04	4.0367e−03	5.0884e−08
9	801	2178	5.7505e−07	1.2058e−07	1.5607e−05	4.0198e−04	3.6995e−09
10	2603	7062	5.8030e−08	0.0000e+00	1.5503e−06	4.0009e−05	1.0897e−10

Table 4 Comparison results by the iALM with settings in this paper and the modified iALM in [26] for the QCQP problem (6.1) with $n = 100$ and $m = 5$

#Test	iALM with settings in this paper							A modified iALM in [26]						
	Time	#Grad	#Fun	ObjErr	Pres	Dres	Compl	Time	#Grad	#Fun	ObjErr	Pres	Dres	Compl
1	2.56e−1	563	1577	6.41e−8	8.71e−10	4.96e−4	2.15e−09	7.64e−1	1590	4342	4.49e−7	3.17e−6	8.68e−4	3.38e−6
2	1.84e−1	630	1744	5.57e−8	3.04e−10	4.88e−4	3.59e−10	3.20e−1	1171	3193	2.51e−7	7.91e−7	7.30e−4	1.32e−6
3	1.46e−1	623	1727	1.12e−7	6.89e−10	4.92e−4	1.15e−09	2.98e−1	1564	4274	3.17e−7	4.32e−6	8.88e−4	4.88e−6
4	1.00e−1	496	1381	7.43e−8	1.16e−09	4.89e−4	1.16e−09	2.42e−1	1192	3259	2.36e−7	1.68e−6	7.65e−4	2.74e−6
5	1.10e−1	610	1698	9.29e−8	3.53e−10	4.89e−4	1.45e−09	2.71e−1	1463	4002	1.97e−7	1.96e−6	7.07e−4	1.98e−6
6	1.13e−1	572	1583	6.17e−8	9.60e−11	4.88e−4	1.43e−09	3.29e−1	1814	4956	3.35e−7	1.77e−6	6.90e−4	2.48e−6
7	1.04e−1	594	1659	9.75e−8	1.32e−09	4.81e−4	1.40e−09	3.30e−1	1810	4935	2.17e−7	2.34e−6	7.07e−4	2.47e−6
8	1.25e−1	729	2018	9.09e−8	2.94e−10	4.84e−4	3.96e−10	3.62e−1	1942	5301	2.82e−7	1.81e−6	6.68e−4	2.09e−6
9	9.93e−2	505	1406	6.59e−8	2.24e−09	4.75e−4	2.26e−09	2.82e−1	1272	3478	2.36e−7	2.18e−6	7.98e−4	2.24e−6
10	1.52e−1	686	1896	5.86e−8	2.64e−10	4.99e−4	3.49e−09	4.17e−1	1416	3880	3.33e−7	3.62e−6	8.98e−4	3.78e−6

Table 5 Comparison results by the iALM with settings in this paper and the modified iALM in [26] for the QCQP problem (6.1) with $n = 1000$ and $m = 10$

#Test	iALM with settings in this paper							A modified iALM in [26]						
	Time	#Grad	#Fun	ObjErr	Pres	Dres	Compl	Time	#Grad	#Fun	ObjErr	Pres	Dres	Compl
1	8.13e+0	739	2047	1.03e−7	3.24e−10	4.85e−4	7.40e−10	6.29e+1	5569	15,122	2.79e−7	6.29e−7	7.03e−4	6.29e−7
2	7.81e+0	711	1981	6.76e−8	3.38e−10	4.99e−4	3.59e−10	4.62e+1	4333	11,786	3.02e−7	7.63e−7	8.45e−4	7.66e−7
3	8.75e+0	733	2033	6.22e−8	6.68e−10	4.77e−4	1.04e−09	6.02e+1	4863	13,217	2.62e−7	6.25e−7	7.17e−4	7.98e−7
4	9.97e+0	789	2195	1.13e−7	9.97e−10	4.72e−4	1.02e−09	6.59e+1	5515	14,984	2.85e−7	8.29e−7	7.04e−4	8.68e−7
5	8.17e+0	713	1988	7.91e−8	7.22e−10	4.88e−4	7.36e−10	5.69e+1	5145	13,991	5.10e−7	1.08e−6	9.42e−4	1.12e−6
6	7.21e+0	656	1836	2.94e−8	2.68e−10	4.79e−4	9.25e−10	4.56e+1	4014	10,909	2.43e−7	7.76e−7	8.07e−4	7.91e−7
7	8.70e+0	715	1987	5.13e−8	7.60e−10	4.59e−4	9.05e−10	5.10e+1	4461	12,124	3.93e−7	9.23e−7	9.39e−4	1.09e−6
8	7.56e+0	694	1934	5.78e−8	1.80e−10	4.76e−4	5.11e−10	4.60e+1	4379	11,908	3.23e−7	9.43e−7	8.80e−4	9.43e−7
9	8.79e+0	802	2219	8.33e−8	5.59e−11	4.83e−4	2.65e−10	5.40e+1	5173	14,051	4.08e−7	9.61e−7	8.85e−4	9.62e−7
10	1.07e+1	770	2144	5.74e−8	5.45e−10	4.82e−4	1.03e−09	4.69e+1	4210	11,446	3.03e−7	9.79e−7	8.56e−4	9.80e−7

where L_0 is the Lipschitz constant of ∇g. Then we have (cf. [45, Lemma 2.1]) $\mathcal{L}_0(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - \mathcal{L}_0(\bar{\mathbf{x}}^+, \bar{\mathbf{y}}) \ge \frac{L_0}{2}\|\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}\|^2$. Since $\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| \le \varepsilon$ and $f_0(\bar{\mathbf{x}}) - d_0(\bar{\mathbf{y}}) \le 2\varepsilon$, we have $\mathcal{L}_0(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - d_0(\bar{\mathbf{y}}) \le \varepsilon\|\bar{\mathbf{y}}\| + 2\varepsilon$. Noting $d_0(\bar{\mathbf{y}}) \le \mathcal{L}_0(\bar{\mathbf{x}}^+, \bar{\mathbf{y}})$, we have $\frac{L_0}{2}\|\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}\|^2 \le \varepsilon\|\bar{\mathbf{y}}\| + 2\varepsilon$, and thus $\|\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}\| \le \sqrt{\frac{2\varepsilon(\|\bar{\mathbf{y}}\|+2)}{L_0}}$. By the triangle inequality, it holds that

$$\|\mathbf{A}\bar{\mathbf{x}}^+ - \mathbf{b}\| \le \|\mathbf{A}\| \cdot \|\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}\| + \|\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}\| \le \|\mathbf{A}\|\sqrt{\frac{2\varepsilon(\|\bar{\mathbf{y}}\| + 2)}{L_0}} + \varepsilon. \tag{A.2}$$

In addition, we have from (A.1) the optimality condition

$$\langle \nabla g(\bar{\mathbf{x}}) + \mathbf{A}^\top \bar{\mathbf{y}} + L_0(\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}}^+ \rangle + h(\mathbf{x}) - h(\bar{\mathbf{x}}^+) \ge 0,$$

and thus

$$\begin{aligned}
&\langle \nabla g(\bar{\mathbf{x}}^+) + \mathbf{A}^\top \bar{\mathbf{y}}, \bar{\mathbf{x}}^+ - \mathbf{x} \rangle + h(\bar{\mathbf{x}}^+) - h(\mathbf{x}) \\
&\quad = \langle \nabla g(\bar{\mathbf{x}}^+) - \nabla g(\bar{\mathbf{x}}), \bar{\mathbf{x}}^+ - \mathbf{x} \rangle + \langle \nabla g(\bar{\mathbf{x}}) + \mathbf{A}^\top \bar{\mathbf{y}}, \bar{\mathbf{x}}^+ - \mathbf{x} \rangle + h(\bar{\mathbf{x}}^+) - h(\mathbf{x}) \\
&\quad \le 2L_0\|\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}\| \cdot \|\bar{\mathbf{x}}^+ - \mathbf{x}\| \le 2DL_0\sqrt{\frac{2\varepsilon(\|\bar{\mathbf{y}}\| + 2)}{L_0}}.
\end{aligned}$$

Therefore, $(\bar{\mathbf{x}}^+, \bar{\mathbf{y}})$ is an $O(\sqrt{\varepsilon})$-solution in the sense of (5.1).

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms. Wiley, New York (2006)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
3. Ben-Tal, A., Zibulevsky, M.: Penalty/barrier multiplier methods for convex programming problems. SIAM J. Optim. **7**(2), 347–366 (1997)
4. Bertsekas, D.P.: Convergence rate of penalty and multiplier methods. In: 1973 IEEE Conference on Decision and Control Including the 12th Symposium on Adaptive Processes, vol. 12, pp. 260–264. IEEE (1973)
5. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont (1999)
6. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic press, London (2014)
7. Birgin, E.G., Castillo, R., Martínez, J.M.: Numerical comparison of augmented lagrangian algorithms for nonconvex problems. Comput. Optim. Appl. **31**(1), 31–55 (2005)
8. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach Learn. **3**(1), 1–122 (2011)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2016)
11. Gao, X., Xu, Y., Zhang, S.: Randomized primal-dual proximal block coordinate updates. J. Oper. Res. Soc. China **7**(2), 205–250 (2019)
12. Glowinski, R.: On alternating direction methods of multipliers: a historical perspective. In: Fitzgibbon, W., Kuznetsov, Y., Neittaanmäki, P., Pironneau, O. (eds.) Modeling, Simulation and Optimization for

Science and Technology. Computational Methods in Applied Sciences, vol. 34. Springer, Dordrecht (2014)
13. Grant, M., Boyd, S., Ye, Y.: CVX: Matlab Software for Disciplined Convex Programming (2008)
14. Güler, O.: On the convergence of the proximal point algorithm for convex minimization. SIAM J. Control Optim. **29**(2), 403–419 (1991)
15. Güler, O.: New proximal point algorithms for convex minimization. SIAM J. Optim. **2**(4), 649–664 (1992)
16. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm for general convex-concave saddle point problems. arXiv preprint arXiv:1803.01401 (2018)
17. He, B., Yuan, X.: On the acceleration of augmented Lagrangian method for linearly constrained optimization. Optimization Online (2010)
18. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the douglas-rachford alternating direction method. SIAM J. Numer. Anal. **50**(2), 700–709 (2012)
19. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969)
20. Kang, M., Kang, M., Jung, M.: Inexact accelerated augmented Lagrangian methods. Comput. Optim. Appl. **62**(2), 373–404 (2015)
21. Kang, M., Yun, S., Woo, H., Kang, M.: Accelerated bregman method for linearly constrained ℓ_1-ℓ_2 minimization. J. Sci. Comput. **56**(3), 515–534 (2013)
22. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented lagrangian methods for convex programming. Math. Program. **155**(1–2), 511–547 (2016)
23. Li, Z., Xu, Y.: First-order inexact augmented lagrangian methods for convex and nonconvex programs: nonergodic convergence and iteration complexity. Preprint (2019)
24. Lin, T., Ma, S., Zhang, S.: Iteration complexity analysis of multi-block admm for a family of convex minimization without strong convexity. J. Sci. Comput. **69**(1), 52–81 (2016)
25. Liu, Y.-F., Liu, X., Ma, S.: On the non-ergodic convergence rate of an inexact augmented lagrangian framework for composite convex programming. Math. Oper. Res. **44**(2), 632–650 (2019)
26. Lu, Z., Zhou, Z.: Iteration-complexity of first-order augmented lagrangian methods for convex conic programming. ArXiv preprint arXiv:1803.09941 (2018)
27. Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(2), 475–507 (2013)
28. Necoara, I., Nedelcu, V.: Rate analysis of inexact dual first-order methods application to dual decomposition. IEEE Trans. Autom. Control **59**(5), 1232–1243 (2014)
29. Nedelcu, V., Necoara, I., Tran-Dinh, Q.: Computational complexity of inexact gradient augmented lagrangian methods: application to constrained mpc. SIAM J. Control Optim. **52**(5), 3109–3134 (2014)
30. Nedić, A., Ozdaglar, A.: Approximate primal solutions and rate analysis for dual subgradient methods. SIAM J. Optim. **19**(4), 1757–1780 (2009)
31. Nedić, A., Ozdaglar, A.: Subgradient methods for saddle-point problems. J. Optim. Theory Appl. **142**(1), 205–228 (2009)
32. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publisher, Norwell (2004)
33. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
34. Ouyang, Y., Chen, Y., Lan, G., Pasiliao Jr., E.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imaging Sci. **8**(1), 644–681 (2015)
35. Ouyang, Y., Xu, Y.: Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. ArXiv preprint arXiv:1808.02901 (2018)
36. Powell, M.J.: A method for non-linear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization. Academic Press, New York (1969)
37. Rockafellar, R.T.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Program. **5**(1), 354–373 (1973)
38. Rockafellar, R.T.: The multiplier method of hestenes and powell applied to convex programming. J. Optim. Theory Appl. **12**(6), 555–562 (1973)
39. Rockafellar, R.T.: Augmented lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976)
40. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems, pp. 1458–1466 (2011)

41. Tseng, P., Bertsekas, D.P.: On the convergence of the exponential multiplier method for convex programming. Math. Program. **60**(1), 1–19 (1993)
42. Xu, Y.: Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. SIAM J. Optim. **27**(3), 1459–1484 (2017)
43. Xu, Y.: Primal-dual stochastic gradient method for convex programs with many functional constraints. ArXiv preprint arXiv:1802.02724 (2018)
44. Xu, Y.: Asynchronous parallel primal-dual block coordinate update methods for affinely constrained convex programs. Comput. Optim. Appl. **72**(1), 87–113 (2019)
45. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**(3), 1758–1789 (2013)
46. Xu, Y., Zhang, S.: Accelerated primal-dual proximal block coordinate updating methods for constrained convex optimization. Comput. Optim. Appl. **70**(1), 91–128 (2018)
47. Yu, H., Neely, M.J.: A primal-dual type algorithm with the $O(1/t)$ convergence rate for large scale constrained convex programs. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 1900–1905. IEEE (2016)
48. Yu, H., Neely, M.J.: A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs. SIAM J. Optim. **27**(2), 759–783 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Yangyang Xu[1]

✉ Yangyang Xu
xuy21@rpi.edu

[1] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA