Noname manuscript No.
(will be inserted by the editor)

On the non-ergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming

Ya-Feng Liu · Xin Liu · Shiqian Ma

Received: date / Accepted: date

Abstract In this paper, we consider the linearly constrained composite convex optimization problem, whose objective is a sum of a smooth function and a possibly nonsmooth function. We propose an inexact augmented Lagrangian (IAL) framework for solving the problem. The stopping criterion used in solving the augmented Lagrangian subproblem in the proposed IAL framework is weaker and potentially much easier to check than the one used in most of the existing IAL frameworks/methods. We analyze the global convergence and the non-ergodic convergence rate of the proposed IAL framework. Preliminary numerical results are presented to show the efficiency of the proposed IAL framework and the importance of the non-ergodic convergence and convergence rate analysis.

Keywords Inexact augmented Lagrangian framework · Non-ergodic convergence rate

Mathematics Subject Classification (2000) 90C25 · 65K05

Y.-F. Liu
State Key Laboratory of Scientific and Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, China
E-mail: yafliu@lsec.cc.ac.cn

X. Liu
State Key Laboratory of Scientific and Engineering Computing,
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences and University of Chinese Academy of Sciences, China
E-mail: liuxin@lsec.cc.ac.cn

S. Ma
Department of Mathematics, University of California, Davis, CA 95616, USA
E-mail: sqma@math.ucdavis.edu

1 Introduction

In this paper, we consider the linearly constrained composite convex optimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \ & F(x) := f(x) + g(x) \\ \text{s.t.} \ & Ax = b, \end{aligned} \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$; $f(x)$ is a convex smooth function with L_f-Lipschitz continuous gradient; and $g(x)$ is a closed convex (not necessarily smooth) function. An important example of problem (1.1) is that $g(x)$ is an indicator function of a closed convex set $\mathcal{X}$, that is,

$$g(x) = \mathrm{Ind}_{\mathcal{X}}(x) := \begin{cases} 0, \ \text{if } x \in \mathcal{X}, \\ +\infty, \ \text{otherwise.} \end{cases}$$

In this case, problem (1.1) can be rewritten as

$$\min_{x\in\mathcal{X}} \ f(x), \ \text{s.t.} \quad Ax = b. \tag{1.2}$$

One efficient approach to solving problem (1.1) is the augmented Lagrangian (AL) method [11,30,32]. The AL function of problem (1.1) is

$$\mathcal{L}_\beta(x;\lambda) := \hat{f}_\beta(x;\lambda) + g(x), \tag{1.3}$$

where

$$\hat{f}_\beta(x;\lambda) := f(x) + \langle \lambda, Ax - b\rangle + \frac{\beta}{2}\|Ax - b\|^2, \tag{1.4}$$

$\lambda \in \mathbb{R}^m$ is the Lagrange multiplier associated with the linear constraint, and $\beta > 0$ is the penalty parameter. The augmented Lagrangian dual of problem (1.1) is

$$\max_{\lambda\in\mathbb{R}^m} \ d(\lambda) \tag{1.5}$$

with

$$d(\lambda) := \min_{x\in\mathbb{R}^n} \mathcal{L}_\beta(x;\lambda). \tag{1.6}$$

It is well-known that the dual function $d(\lambda)$ in (1.6) is differentiable and its gradient is given by $\nabla d(\lambda) = Ax(\lambda) - b$, where $x(\lambda)$ is the solution of problem (1.6) (see [3]).

Given λ^k, the AL method for solving problem (1.1) updates the primal and dual variables via

$$x(\lambda^k) = \arg\min_{x\in\mathbb{R}^n} \mathcal{L}_\beta(x;\lambda^k) \tag{1.7}$$

and

$$\lambda^{k+1} = \lambda^k + \beta\left(Ax(\lambda^k) - b\right),$$

respectively. The AL method for solving problem (1.1) is essentially a dual gradient ascent method for solving the dual problem (1.5), which updates the dual variable by performing a dual gradient ascent step

$$\lambda^{k+1} = \lambda^k + \beta\nabla d(\lambda^k).$$

The AL method can also be interpreted as a proximal point algorithm applied to solve the classical Lagrangian dual problem [32]

$$\max_{\lambda} d'(\lambda), \tag{1.8}$$

where

$$d'(\lambda) := \min_{x\in\mathbb{R}^n} \{f(x) + \langle \lambda, Ax - b\rangle + g(x)\}.$$

In this paper, we always refer to the augmented Lagrangian dual (problem) whenever we talk about the Lagrangian dual (problem) or the dual (problem), unless otherwise specified. More generally, the AL method can be derived through the Bregman regularization approach [9, 29, 37] and it enjoys the so-called error-forgetting property [36] when applied to solve problem (1.1) where $F(x)$ is a piece-wise linear function. For various variants of the AL method with nonquadratic penalty terms and other multiplier update formulas, please see [2, Chapter 5], [5, 12, 34, 38].

When the problem dimension n is large, finding an exact solution of AL subproblem (1.7) can be computationally expensive and thus the exact gradient $\nabla d(\lambda^k)$ is often unavailable. As a result, many works focused on inexact versions of (dual) gradient methods; see [4,6,7,16,18–20,22–24,31–33,35] and references therein. For instance, Necoara and Patrascu [22] analyzed dual first-order methods for solving a class of *strongly* convex conic programs and provided a detailed (ergodic and non-ergodic) convergence rate analysis of the methods. The methods in [22] are the exact gradient methods applied to solve the dual problem (1.5) where the penalty parameter β in (1.4) is set to be zero (i.e., problem (1.8)) and thus is different from the AL method where the penalty parameter β in (1.4) is positive. For the inexact augmented Lagrangian (IAL) framework, Rockafellar [32] proposed an IAL framework, where the AL subproblem is solved until a point x^{k+1} is found such that

$$\mathcal{L}_\beta\left(x^{k+1};\lambda^k\right) - \mathcal{L}_\beta\left(x(\lambda^k);\lambda^k\right) \le \eta_k^2, \tag{1.9}$$

and showed that the proposed IAL framework converges if the nonnegative tolerance sequence $\{\eta_k\}$ is summable. Very recently, Devolder, Glineur, and Nesterov [6] proposed a general inexact gradient framework and analyzed the ergodic convergence rate of their framework when it is applied to solve dual problem (1.5). In [24], Nedelcu, Necoara, and Tran-Dinh proposed an IAL method, where the AL subproblem was approximately solved by Nesterov's gradient method [25–27] such that (1.9) is satisfied and showed again the ergodic convergence rate of the proposed IAL method. The non-ergodic convergence rate result for the IAL framework/method has been missing in the literature for a long time until in a very recent work by Lan and Monterio [16], where they proposed an IAL method (where the AL subproblems are approximately solved by Nesterov's gradient method) and analyzed the non-ergodic convergence rate for the proposed method.

We make the following assumptions throughout this paper.

A1 there exists a Lagrange multiplier λ^* such that the optimal value of problem (1.1) is equal to $d(\lambda^*)$;
A2 the function $g(x)$ has a bounded domain.

Assumption A1 is the strong duality assumption and assumption A2 is made in the paper mainly for the ease of presentation. In fact, for problem (1.1) arising

from many applications of interest such as machine learning, statistics, and signal processing, we often can easily find a bounded set $\mathcal{X}$ such that the solution of problem (1.1) lies in $\mathcal{X}$. Therefore, we can restrict the definition of $g(x)$ over this bounded set. Let us take the following basis pursuit problem in compressed sensing as an example:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \ & \|x\|_1 \\ \text{s.t.} \ & Ax = b, \end{aligned} \tag{1.10}$$

which is a special case of problem (1.1) with $f(x) = 0$ and $g(x) = \|x\|_1$. We can restrict the definition of $\|x\|_1$ over the bounded domain

$$\{x \,|\, \|x\|_1 \le \|\hat{x}\|_1\},$$

where $\hat{x}$ is any point satisfying $A\hat{x} = b$. It is worth remarking that problem (1.2) has been considered in the existing papers such as [16,23,24] and they all assumed that the set $\mathcal{X}$ is convex and compact.

The contribution of this paper is twofold. First, we propose a new IAL framework (see Algorithm 1) for solving problem (1.1), where the AL subproblem is approximately solved until a point x^{k+1} is found such that

$$\max_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(x^{k+1};\lambda^k),\, x^{k+1} - x \right\rangle + g(x^{k+1}) - g(x) \right\} \le \eta_k. \tag{1.11}$$

Here $\nabla \hat{f}_\beta(x;\lambda)$ is the gradient of $\hat{f}_\beta(x;\lambda)$ with respect to x. The termination condition (1.11) in our proposed IAL framework is weaker and (potentially) easier to check than (1.9) in most of the existing IAL frameworks/methods. More specifically, to check whether x^{k+1} satisfies (1.11) or not, we only need to solve the convex optimization problem on the left-hand side of (1.11), which can be solved exactly or to a high precision in time (essentially) linear to the size of the input for many $g(x)$ such as the ℓ_1-norm and the nuclear norm; see more examples in [13]. In contrast, it is generally hard to check whether x^{k+1} satisfies (1.9) or not (because $x(\lambda^k)$ is unknown). Second, we establish the global convergence of the proposed IAL framework under the assumption that the sequence $\{\eta_k\}$ in (1.11) is summable; see Theorem 3.4. We also show, in Theorems 4.1 and 4.3, the non-ergodic convergence rate (under weaker conditions than that in [16]) for the proposed IAL framework, which reveals how the error in solving the AL subproblem affects the convergence rate.

It is worth highlighting here that the non-ergodic analysis focuses on the iterates generated by the algorithm while the ergodic analysis focuses on some (weighted) average of the iterates generated by the algorithm. In practice, the non-ergodic iterates tend to share structural properties of the solution of the problem such as sparsity in ℓ_1 minimization problem (1.10), while the ergodic iterates tend to "average out" these properties. Therefore, the non-ergodic solution is more preferable in practical applications. In fact, our simulation results on the basis pursuit problem in Section 6 show that the last iterate indeed is much better than the average of all iterates in terms of the sparsity. From the perspective of theoretical analysis, the non-ergodic convergence implies and thus is stronger than the ergodic convergence. This paper will focus on the non-ergodic convergence analysis.

2 The IAL framework

In this section, we present the IAL framework for solving problem (1.1). The proposed IAL framework is given in Algorithm 1. At the k-th iteration, the IAL framework first solves AL subproblem (2.1) with fixed dual variable λ^k in an inexact manner until a point x^{k+1} satisfying (1.11) is found; then updates the dual variable by performing an inexact gradient ascent step (2.2).

Algorithm 1: The IAL framework for solving problem (1.1)

1 Initialize $x^1 \in \mathcal{X}$, $\lambda^1 \in \mathbb{R}^m$, and the nonnegative sequence $\{\eta_k\}$.
2 **for** $k \geq 1$: **do**
3 Find an approximate solution x^{k+1} of the AL subproblem

$$\min_{x\in\mathbb{R}^n} \left\{ \mathcal{L}_\beta(x;\lambda^k) := \hat{f}_\beta(x;\lambda^k) + g(x) \right\} \tag{2.1}$$

such that (1.11) is satisfied;
4 Update the dual variable via

$$\lambda^{k+1} = \lambda^k + \beta\left(Ax^{k+1} - b\right). \tag{2.2}$$

Three remarks on the proposed IAL framework are in order. First, the termination condition (1.11) in our proposed IAL framework is (potentially) easier to check than (1.9) in most of the existing IAL frameworks/methods. Let us take problem (1.10) as an example again. In this case, to check whether x^{k+1} satisfies (1.11) or not, we only need to solve the following convex optimization problem

$$\max_{\|x\|_1 \leq \|\hat{x}\|_1} \left\{ \left\langle \nabla \hat{f}_\beta(x^{k+1};\lambda^k),\, x^{k+1} - x \right\rangle + \left\| x^{k+1} \right\|_1 - \|x\|_1 \right\},$$

where $\nabla \hat{f}_\beta(x^{k+1};\lambda^k) = A^T\left(\lambda^k + \beta\left(Ax^{k+1} - b\right)\right)$. Let $\bar{i}_{k+1}$ be the index of the largest entry of $\nabla \hat{f}_\beta(x^{k+1};\lambda^k)$ in magnitude, then the solution to the above optimization problem is

$$\bar{x}^{k+1} = \begin{cases} -\|\hat{x}\|_1 \mathrm{sign}\left(\left[\nabla \hat{f}_\beta(x^{k+1};\lambda^k)\right]_{\bar{i}_{k+1}}\right) e_{\bar{i}_{k+1}}, \text{ if } \left|\left[\nabla \hat{f}_\beta(x^{k+1};\lambda^k)\right]_{\bar{i}_{k+1}}\right| \geq 1; \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise,} \end{cases}$$

where $e_{\bar{i}_{k+1}}$ is the n-dimensional vector with the $\bar{i}_{k+1}$-th entry being 1 and all other entries being 0 and $\mathrm{sign}(\cdot)$ is the sign function.

Second, the smaller the tolerance η_k is, the more computational cost is needed in Algorithm 1 to find the point x^{k+1} satisfying (1.11). On the other hand, the larger the tolerance η_k is, the larger the approximation error between the approximate gradient $Ax^{k+1} - b$ and the true gradient $\nabla d(\lambda^k)$ is (see Lemma 3.2 further ahead), which might lead to slow convergence or even divergence of the proposed Algorithm 1. Therefore, the choice of $\{\eta_k\}$ is important in balancing the computational cost (of finding the point x^{k+1} satisfying (1.11)) and the global convergence

and convergence rate (of the framework). We will discuss the possible choices of $\{\eta_k\}$ in more details in Section 4.

Third, AL subproblem (2.1) can be efficiently solved in an inexact manner by various (first-order) methods such as the accelerated proximal gradient methods in [1,25–27] and the Frank-Wolfe (a.k.a. conditional gradient) methods in [8,10,14, 15,28]. Next, we discuss the (inner) iteration complexity of finding the point x^{k+1} satisfying (1.11) when the accelerated proximal gradient methods and the Frank-Wolfe methods are applied to solve problem (2.1). Although any of the accelerated proximal gradient methods in [1,25–27] and also any of the Frank-Wolfe methods in [8, 10, 14, 15, 28] can be used in solving AL subproblem (2.1), we choose the algorithms in [1] and [28] in our following analysis.

Let us fist define

$$L_{\hat{f}} = L_f + \beta\|A\|^2, \tag{2.3}$$

where $\|A\|$ denotes the largest singular value of the matrix A. It is simple to see that $\nabla_x \hat{f}_\beta(x;\lambda)$ (with respect to x) is Lipschitz continuous with Lipschitz constant $L_{\hat{f}}$. Note that $L_{\hat{f}}$ does not depend on the Lagrange multiplier λ. Moreover, let $\mathcal{X}$ denote the bounded domain of the function $g(x)$ and let

$$D = \max_{x,y\in\mathcal{X}} \|x-y\| < +\infty \tag{2.4}$$

denote the diameter of the set $\mathcal{X}$.

Algorithm 2: The fast iterative shrinkage-thresholding algorithm (FISTA) for solving AL subproblem (2.1) [1]

Initialize $y^{k,1} = x^{k,0} \in \mathcal{X}$ and $t^1 = 1$.
for $\ell \geq 1$: **do**
  Set $t^{\ell+1} = \dfrac{1+\sqrt{1+4(t^\ell)^2}}{2}$;
  Compute

$$x^{k,\ell} = \arg\min_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(y^{k,\ell};\lambda^k),\, x - y^{k,\ell} \right\rangle + \frac{L_{\hat{f}}}{2}\left\| x - y^{k,\ell} \right\|^2 + g(x) \right\}$$

  and

$$y^{k,\ell+1} = x^{k,\ell} + \left(\frac{t^\ell - 1}{t^{\ell+1}}\right)\left(x^{k,\ell} - x^{k,\ell-1}\right);$$

Theorem 2.1 *Let $\left\{x^{k,\ell}\right\}_{\ell\geq 1}$ be the sequence generated by FISTA (i.e., Algorithm 2) when applied to solve AL subproblem* (2.1)*, where ℓ is the index of the inner iteration. Suppose that $\hat{x}^{k,\ell}$ is the point such that*

$$\hat{x}^{k,\ell} = \arg\min_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(x^{k,\ell};\lambda^k),\, x - x^{k,\ell} \right\rangle + \frac{L_{\hat{f}}}{2}\left\| x - x^{k,\ell} \right\|^2 + g(x) \right\}. \tag{2.5}$$

Then,

$$\max_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(\hat{x}^{k,\ell};\lambda^k),\, \hat{x}^{k,\ell} - x \right\rangle + g(\hat{x}^{k,\ell}) - g(x) \right\} \leq \frac{4L_{\hat{f}}D^2}{\ell+1},\ \forall\ \ell \geq 1. \tag{2.6}$$

In other words, it takes at most

$$\left\lceil \frac{4L_{\hat{f}}D^2}{\eta_k} \right\rceil - 1 \tag{2.7}$$

FISTA iterations and one proximal gradient iteration (equivalent to solving problem (2.5)*) to find the point* x^{k+1} *satisfying* (1.11)*.*

Proof For simplicity, denote $\hat{f}_\beta(x;\lambda)$ by $\hat{f}(x)$, $\mathcal{L}_\beta(x;\lambda)$ by $\mathcal{L}(x)$, and $x^{k,\ell}$ and $\hat{x}^{k,\ell}$ by x^ℓ and $\hat{x}^\ell$ (respectively) for all $\ell \geq 1$ in the proof. First, it follows from [1, Theorem 4.4] and the definitions of $L_{\hat{f}}$ in (2.3) and D in (2.4) that

$$\mathcal{L}(x^\ell) - \mathcal{L}(x(\lambda^k)) \leq \frac{2L_{\hat{f}}D^2}{(\ell+1)^2},\ \forall\ \ell \geq 1. \tag{2.8}$$

Note that $\hat{x}^\ell$ in (2.5) is obtained after performing a proximal gradient step from x^ℓ. Then, from [1, Lemma 2.3] and [27, Theorem 2], we get

$$\mathcal{L}(x^\ell) - \mathcal{L}(\hat{x}^\ell) \geq \frac{L_{\hat{f}}}{2}\left\|\hat{x}^\ell - x^\ell\right\|^2. \tag{2.9}$$

Combining (2.8) and (2.9) yields

$$\left\|\hat{x}^\ell - x^\ell\right\| \leq \frac{2D}{\ell+1},\ \forall\ \ell \geq 1. \tag{2.10}$$

By the optimality of $\hat{x}^\ell$, we have

$$\left\langle \nabla\hat{f}(x^\ell) + L_{\hat{f}}\left(\hat{x}^\ell - x^\ell\right),\ x - \hat{x}^\ell \right\rangle + g(x) - g(\hat{x}^\ell) \geq 0,\ \forall\ x \in \mathcal{X}, \tag{2.11}$$

where $\mathcal{X}$ is the domain of $g(x)$. Therefore, for any $x \in \mathcal{X}$,

$$\begin{aligned}
&\left\langle \nabla\hat{f}(\hat{x}^\ell),\ \hat{x}^\ell - x \right\rangle + g(\hat{x}^\ell) - g(x)\\
=\ &\left\langle \nabla\hat{f}(\hat{x}^\ell) - \nabla\hat{f}(x^\ell),\ \hat{x}^\ell - x \right\rangle + \left\langle \nabla\hat{f}(x^\ell),\ \hat{x}^\ell - x \right\rangle + g(\hat{x}^\ell) - g(x)\\
\leq\ &\left\langle \nabla\hat{f}(\hat{x}^\ell) - \nabla\hat{f}(x^\ell),\ \hat{x}^\ell - x \right\rangle + \left\langle L_{\hat{f}}\left(\hat{x}^\ell - x^\ell\right),\ x - \hat{x}^\ell \right\rangle\\
\leq\ &2L_{\hat{f}}D\left\|\hat{x}^\ell - x^\ell\right\|\\
\leq\ &4L_{\hat{f}}D^2/(\ell+1),
\end{aligned}$$

where the first inequality comes from (2.11); the second inequality is due to the Cauchy-Schwarz inequality, the definition of D in (2.4), and the fact that $\nabla\hat{f}(x)$ (with respect to x) is $L_{\hat{f}}-$Lipschitz continuous; the third inequality follows from (2.10). Taking the maximum over $x \in \mathcal{X}$ in the above inequality leads to the desired result (2.6). □

Algorithm 3: The Frank-Wolfe algorithm for solving AL subproblem (2.1) [28]

Initialize $x^{k,0} \in \mathcal{X}$.
for $\ell \geq 0$: **do**
  Set $\gamma^\ell = 2/(\ell+2)$;
  Compute

$$\upsilon^\ell = \arg\min_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(x^{k,\ell}; \lambda^k),\, x - x^{k,\ell} \right\rangle + g(x) \right\}; \tag{2.12}$$

  Update $x^{k,\ell+1} = x^{k,\ell} + \gamma^\ell(\upsilon^\ell - x^{k,\ell})$;

Theorem 2.2 *Let $\{x^{k,\ell}\}_{\ell\geq 1}$ be the sequence generated by the Frank-Wolfe method (i.e., Algorithm 3) when applied to solve AL subproblem* (2.1)*, where ℓ is the index of the inner iteration. Then there exists $1 \leq \hat{\ell} \leq \ell$ such that*

$$\max_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(x^{k,\hat{\ell}}; \lambda^k),\, x^{k,\hat{\ell}} - x \right\rangle + g(x^{k,\hat{\ell}}) - g(x) \right\} \leq \frac{6L_{\hat{f}}D^2}{\ell+2},\ \forall\ \ell \geq 1. \tag{2.13}$$

In other words, it takes at most

$$\left\lceil \frac{6L_{\hat{f}}D^2}{\eta_k} \right\rceil - 2 \tag{2.14}$$

Frank-Wolfe iterations to find the point x^{k+1} satisfying (1.11)*.*

Proof For simplicity, denote $\hat{f}_\beta(x;\lambda)$ by $\hat{f}(x)$, $\mathcal{L}_\beta(x;\lambda)$ by $\mathcal{L}(x)$, and $x^{k,\ell}$ by x^ℓ for all $\ell \geq 0$ in the proof. Define

$$\Delta^\ell = \mathcal{L}(x^\ell) - \mathcal{L}(x(\lambda^k))$$

and

$$V^\ell = \left\langle \nabla \hat{f}(x^\ell),\, x^\ell - \upsilon^\ell \right\rangle + g(x^\ell) - g(\upsilon^\ell).$$

From the convexity of $\hat{f}(x)$ and the definition of υ^ℓ in (2.12), we get

$$\Delta^\ell \leq \left\langle \nabla \hat{f}(x^\ell),\, x^\ell - x(\lambda^k) \right\rangle + g(x^\ell) - g(x(\lambda^k)) \leq V^\ell. \tag{2.15}$$

Recall $x^{\ell+1} = x^\ell + \gamma^\ell(\upsilon^\ell - x^\ell)$. Since $\nabla\hat{f}(x)$ is $L_{\hat{f}}$-Lipschitz continuous, it follows that

$$\hat{f}(x^{\ell+1}) \leq \hat{f}(x^\ell) + \gamma^\ell \left\langle \nabla \hat{f}(x^\ell), \upsilon^\ell - x^\ell \right\rangle + \frac{L_{\hat{f}}\left(\gamma^\ell\right)^2}{2} \left\| \upsilon^\ell - x^\ell \right\|^2. \tag{2.16}$$

By the convexity of $g(x)$, we get

$$g(x^{\ell+1}) \leq \gamma^\ell g(\upsilon^\ell) + \left(1 - \gamma^\ell\right) g(x^\ell). \tag{2.17}$$

Combining (2.16) and (2.17), we have, for $\ell = 0, 1, ...,$

$$\mathcal{L}(x^{\ell+1}) \leq \mathcal{L}(x^\ell) - \gamma^\ell V^\ell + \frac{L_{\hat{f}}\left(\gamma^\ell\right)^2}{2} \left\| \upsilon^\ell - x^\ell \right\|^2, \tag{2.18}$$

which, together with (2.15), implies

$$\Delta^{\ell+1} \leq \left(1-\gamma^{\ell}\right)\Delta^{\ell} + \frac{L_{\hat{f}}\left(\gamma^{\ell}\right)^{2}}{2}\left\|v^{\ell}-x^{\ell}\right\|^{2}$$

and thus

$$\Delta^{\ell+1} \leq \frac{\ell}{\ell+2}\Delta^{\ell} + \left(\frac{2}{\ell+2}\right)^{2}\frac{L_{\hat{f}}D^{2}}{2}.$$

By mathematical induction, it can be verified that

$$\Delta^{\ell} \leq \frac{2L_{\hat{f}}D^{2}}{\ell+2},\ \ell \geq 1. \tag{2.19}$$

Based on (2.19), we can use the same argument (by contradiction) as in [14,21] to show the desired result (2.13). The idea is to show that $\{V^{m}\}_{m=1}^{\ell}$ cannot stay large over many consecutive iterations. For completeness, we give the details below. Denote

$$\hat{C} = 2L_{\hat{f}}D^{2},\ \ell_{1} = \lceil\frac{\ell}{2}\rceil,\ \text{and}\ \nu = \frac{\ell_{1}+1}{\ell+2}.$$

Suppose on the contrary that

$$V^{m} > \frac{3\hat{C}}{\ell+2}, \forall\ m = \ell_{1}, \ell_{1}+1, \ldots, \ell. \tag{2.20}$$

From (2.18), we have,

$$\Delta^{\ell+1} \leq \Delta^{\ell} - \frac{2V^{\ell}}{\ell+2} + \frac{\hat{C}}{(\ell+2)^{2}},\ \ell \geq 0.$$

Summing this inequality for indices from ℓ_1 to ℓ yields

$$\begin{aligned}
\Delta^{\ell+1} &\leq \Delta^{\ell_1} - \sum_{m=\ell_1}^{\ell}\frac{2V^{m}}{m+2} + \sum_{m=\ell_1}^{\ell}\frac{\hat{C}}{(m+2)^{2}} \\
&< \Delta^{\ell_1} - \frac{6\hat{C}}{\ell+2}\sum_{m=\ell_1+2}^{\ell+2}\frac{1}{m} + \sum_{m=\ell_1+2}^{\ell+2}\frac{\hat{C}}{m^{2}} \\
&\leq \frac{\hat{C}}{\nu(\ell+2)} - \frac{6\hat{C}}{\ell+2}\frac{\ell-\ell_1+1}{\ell+2} + \frac{\hat{C}(\ell-\ell_1+1)}{(\ell+2)(\ell_1+1)} \\
&= \frac{\hat{C}}{\nu(\ell+2)}\left(2-6\nu(1-\nu)-\nu\right),
\end{aligned} \tag{2.21}$$

where the second inequality is due to (2.20), and the third inequality is due to (2.19) and the fact $\sum_{m=a}^{b}\frac{1}{k^{2}} \leq \frac{b-a+1}{b(a-1)}$ for any $b \geq a > 1$. Define $\phi(x) = 2 - 6x(1-x) - x$. Since $\nu \in [\frac{1}{2}, \frac{2}{3}]$, it follows from (2.21) that $\Delta^{\ell+1} < \frac{\hat{C}}{\nu(\ell+2)}\phi(\nu) \leq 0$, which is a contradiction. The proof is completed. □

The convergence rate result (2.19) is not new when $g(x) = \mathrm{Ind}_{\mathcal{X}}(x)$; see [14,17, 21] and references therein. For the general composite minimization case, Nesterov proved

$$\mathcal{L}(x^{\ell}) - \mathcal{L}(x(\lambda^k)) \leq \frac{4L_{\hat{f}}D^2}{\ell+1},\ \ell \geq 1;$$

see Eq. (2.16) in [28]. Roughly speaking, our result (2.19) improves the above bound by a factor of two.

Compared to accelerated proximal gradient methods when applied to solve AL subproblem (2.1), the Frank-Wolfe methods generally need more iterations to find the point x^{k+1} satisfying (1.11), while the computational cost per iteration in the Frank-Wolfe methods is generally cheaper.

3 Global convergence

In this section, we present the global convergence result of our IAL framework (Algorithm 1), which is independent of the methods used to find the point x^{k+1} satisfying (1.11). Theorem 3.4 shows global convergence of the IAL framework under the assumption that the nonnegative sequence $\{\eta_k\}$ is summable.

It is worth remarking that (3.4), (3.5), and (3.6), which build a bridge between the exact dual function value and dual gradient and the approximate ones, are crucial for establishing global convergence and non-ergodic convergence rate results in this paper. We shall show that condition (1.11) implies (3.4), (3.5), and (3.6); see the proofs of Lemma 3.1 and Lemma 3.2. Clearly, condition (1.11) can be replaced with some other conditions (e.g., condition (1.9)) in Algorithm 1 and global convergence and non-ergodic convergence results of Algorithm 1 will still follow as long as the new conditions imply (3.4), (3.5), and (3.6).

For the ease of presentation, we define

$$x(\lambda^k) := \arg\min_{x\in\mathbb{R}^n} \mathcal{L}_\beta(x;\lambda^k),\ k=1,2,\ldots, \tag{3.1}$$

$$\nabla d(\lambda^k) := Ax(\lambda^k) - b,\ k=1,2,\ldots,$$

$$\bar{d}(\lambda^k) := \mathcal{L}_\beta(x^{k+1};\lambda^k),\ k=1,2,\ldots, \tag{3.2}$$

$$\nabla \bar{d}(\lambda^k) := Ax^{k+1} - b,\ k=1,2,\ldots, \tag{3.3}$$

where x^{k+1} is generated by Algorithm 1 and satisfies (1.11).

We first prove the following two lemmas (Lemma 3.1 and Lemma 3.2), which have been proved for smooth function $F(x)$ in [6, 16, 24]. We now extend them to composite nonsmooth function $F(x)$. In particular, Lemma 3.1 shows that $d(\lambda^{k+1})$ can be bounded from both above and below and Lemma 3.2 shows that $\left\|\nabla d(\lambda^k) - \nabla \bar{d}(\lambda^k)\right\|$ is bounded by $\sqrt{\eta_k/\beta}$.

Lemma 3.1 *The following two inequalities hold:*

$$d(\lambda) \leq \bar{d}(\mu) + \left\langle \nabla \bar{d}(\mu), \lambda - \mu \right\rangle,\ \forall\ \lambda,\mu, \tag{3.4}$$

and

$$d(\lambda^{k+1}) \geq \bar{d}(\lambda^k) + \frac{\beta}{2}\left\|\nabla \bar{d}(\lambda^k)\right\|^2 - \eta_k. \tag{3.5}$$

Proof We first show (3.4). By the definitions of $d(\lambda)$ and $\mathcal{L}_\beta(x;\lambda)$, we have

$$d(\lambda) = \min_{x\in\mathbb{R}^n}\{\mathcal{L}_\beta(x;\lambda)\} \le \mathcal{L}_\beta(x_\mu;\lambda) = \mathcal{L}_\beta(x_\mu;\mu) + \langle \lambda-\mu, Ax_\mu - b\rangle,$$

where x_μ satisfies $\bar{d}(\mu) = \mathcal{L}_\beta(x_\mu;\mu)$. This, together with the definition of $\nabla\bar{d}(\mu)$ (cf. (3.2)), yields (3.4).

We now prove (3.5). By the convexity of $f(x)$, the definition of $\nabla\bar{d}(\lambda^k)$, and (2.2), we get

$$\begin{aligned}\mathcal{L}_\beta(x;\lambda^{k+1}) &\ge f(x^{k+1}) + \left\langle \nabla f(x^{k+1}), x-x^{k+1}\right\rangle + g(x) + \left\langle \lambda^{k+1}, Ax-b\right\rangle + \frac{\beta}{2}\|Ax-b\|^2\\ &= \mathcal{L}_\beta(x^{k+1};\lambda^k) + \beta\left(\left\langle \nabla\bar{d}(\lambda^k), Ax-b\right\rangle + \frac{1}{2}\left\|(Ax-b)-\nabla\bar{d}(\lambda^k)\right\|^2\right)\\ &\quad + \left\langle \nabla\hat{f}_\beta(x^{k+1};\lambda^k), x-x^{k+1}\right\rangle + g(x) - g(x^{k+1}).\end{aligned}$$

Taking the minimum over $x\in\mathbb{R}^n$ on both sides of the above inequality, we have

$$\begin{aligned}d(\lambda^{k+1}) \ge \mathcal{L}_\beta(x^{k+1};\lambda^k) &+ \beta\min_{x\in\mathbb{R}^n}\left\{\left\langle \nabla\bar{d}(\lambda^k), Ax-b\right\rangle + \frac{1}{2}\left\|(Ax-b)-\nabla\bar{d}(\lambda^k)\right\|^2\right\}\\ &+ \min_{x\in\mathbb{R}^n}\left\{\left\langle \nabla\hat{f}_\beta(x^{k+1};\lambda^k), x-x^{k+1}\right\rangle + g(x) - g(x^{k+1})\right\}.\end{aligned}$$

By using the definition of $\bar{d}(\lambda)$ and (1.11), we immediately get the desired result (3.5). □

Lemma 3.2 *The following inequality holds:*

$$\left\|\nabla d(\lambda^k) - \nabla\bar{d}(\lambda^k)\right\|^2 \le \frac{\eta_k}{\beta}. \tag{3.6}$$

Proof It follows from the optimality of $x(\lambda^k)$ (cf. (3.1)) that

$$\left\langle \nabla\hat{f}_\beta(x(\lambda^k);\lambda^k), x(\lambda^k)-x^{k+1}\right\rangle + g(x(\lambda^k)) - g(x^{k+1}) \le 0.$$

By setting $x = x(\lambda^k)$ in (1.11), we get

$$\left\langle \nabla\hat{f}_\beta(x^{k+1};\lambda^k), x^{k+1}-x(\lambda^k)\right\rangle - g(x(\lambda^k)) + g(x^{k+1}) \le \eta_k.$$

Adding the above two inequalities yields

$$\begin{aligned}\eta_k &\ge \left\langle \nabla\hat{f}_\beta(x^{k+1};\lambda^k) - \nabla\hat{f}_\beta(x(\lambda^k);\lambda^k), x^{k+1}-x(\lambda^k)\right\rangle\\ &= \left\langle \nabla f(x^{k+1}) + \beta A^T(Ax^{k+1}-b) - \nabla f(x(\lambda^k)) - \beta A^T(Ax(\lambda^k)-b), x^{k+1}-x(\lambda^k)\right\rangle\\ &\ge \left\langle \beta A^T(Ax^{k+1}-b) - \beta A^T(Ax(\lambda^k)-b), x^{k+1}-x(\lambda^k)\right\rangle\\ &= \beta\left\|\nabla d(\lambda^k) - \nabla\bar{d}(\lambda^k)\right\|^2,\end{aligned}$$

where the second inequality is due to the convexity of $f(x)$ and the second equality is due to the definitions of $\nabla d(\lambda^k)$ and $\nabla\bar{d}(\lambda^k)$. □

The following Lemma 3.3 shows that the sequence $\{\lambda^k\}$ generated by Algorithm 1 is bounded.

Lemma 3.3 *Let $\{\lambda^k\}$ be generated by Algorithm 1. Suppose the sequence $\{\eta_k\}$ satisfies* (3.10)*, then*

$$\left\|\lambda^k - \lambda^*\right\| \leq B,\ k = 1, 2, \ldots, \tag{3.7}$$

where

$$B := \sqrt{\|\lambda^1 - \lambda^*\|^2 + 2\beta \sum_{k=1}^{+\infty} \eta_k} < +\infty. \tag{3.8}$$

Proof We have

$$\begin{aligned}
\left\|\lambda^{k+1} - \lambda^*\right\|^2 &= \left\|\lambda^k - \lambda^* + \beta \nabla \bar{d}(\lambda^k)\right\|^2 \\
&= \left\|\lambda^k - \lambda^*\right\|^2 + 2\beta \left\langle \nabla \bar{d}(\lambda^k), \lambda^k - \lambda^* \right\rangle + \beta^2 \left\|\nabla \bar{d}(\lambda^k)\right\|^2 \\
&\leq \left\|\lambda^k - \lambda^*\right\|^2 + 2\beta \left(\bar{d}(\lambda^k) - d(\lambda^*)\right) + \beta^2 \left\|\nabla \bar{d}(\lambda^k)\right\|^2 \\
&= \left\|\lambda^k - \lambda^*\right\|^2 + 2\beta \left(d(\lambda^{k+1}) - d(\lambda^*)\right) + 2\beta \left(\bar{d}(\lambda^k) - d(\lambda^{k+1})\right) + \beta^2 \left\|\nabla \bar{d}(\lambda^k)\right\|^2 \\
&\leq \left\|\lambda^k - \lambda^*\right\|^2 + 2\beta \left(d(\lambda^{k+1}) - d(\lambda^*)\right) + 2\beta\eta_k \\
&\leq \left\|\lambda^k - \lambda^*\right\|^2 + 2\beta\eta_k,
\end{aligned}$$

where the first inequality is due to (3.4) (with λ and μ replaced by λ^* and λ^k respectively), the second inequality is due to (3.5), and the last inequality is due to the fact that $d(\lambda^{k+1}) \leq d(\lambda^*)$ for all $k \geq 1$. Summing the above inequality, we obtain

$$\left\|\lambda^k - \lambda^*\right\|^2 \leq \left\|\lambda^1 - \lambda^*\right\|^2 + 2\beta \sum_{i=1}^{k-1} \eta_i,\ k = 1, 2, \ldots,$$

which, together with (3.8), completes the proof. □

Before presenting the main result of this section, i.e., the global convergence result of our IAL framework, we define

$$\theta := \frac{\beta}{4B^2}, \tag{3.9}$$

where B is given in (3.8).

Theorem 3.4 *Let $\{x^k\}$ and $\{\lambda^k\}$ be generated by Algorithm 1. Suppose the non-negative sequence $\{\eta_k\}$ satisfies*

$$\sum_{k=1}^{+\infty} \eta_k < +\infty. \tag{3.10}$$

Then,

$$\delta_k := d\left(\lambda^*\right) - d(\lambda^k) \to 0 \text{ and } \left\|Ax^{k+1} - b\right\| \to 0,$$

where λ^* is an optimal solution to problem (1.5) *and $d(\lambda)$ is defined in* (1.6)*.*

Proof It suffices to show

$$\sum_k \delta_k^2 < +\infty, \tag{3.11}$$

and

$$\sum_k \left\| Ax^{k+1} - b \right\|^2 < +\infty. \tag{3.12}$$

By (3.5) and the definition of $\bar{d}(\lambda^k)$, we obtain

$$d(\lambda^{k+1}) \geq d(\lambda^k) + \frac{\beta}{2} \left\| \nabla \bar{d}(\lambda^k) \right\|^2 - \eta_k, \tag{3.13}$$

which, together with (3.6) and the inequality $a^2 \geq b^2/2 - (a-b)^2$, implies

$$d(\lambda^{k+1}) \geq d(\lambda^k) + \frac{\beta}{4} \left\| \nabla d(\lambda^k) \right\|^2 - \frac{3}{2}\eta_k. \tag{3.14}$$

Moreover, it follows from (3.7) and the concavity of $d(\lambda)$ that

$$d(\lambda^*) - d(\lambda^k) \leq \left\langle \nabla d(\lambda^k), \lambda^* - \lambda^k \right\rangle \leq \left\| \lambda^k - \lambda^* \right\| \left\| \nabla d(\lambda^k) \right\| \leq B \left\| \nabla d(\lambda^k) \right\|.$$

Combining the above, (3.9), and (3.14), we immediately obtain

$$\delta_{k+1} \leq \delta_k - \theta \delta_k^2 + \frac{3}{2}\eta_k, \ k = 1, 2, \ldots, \tag{3.15}$$

which further implies

$$\delta_k \leq \delta_1 - \theta \sum_{i=1}^{k-1} \delta_i^2 + \frac{3}{2} \sum_{i=1}^{k-1} \eta_i, \ k = 1, 2, \ldots \tag{3.16}$$

From the definition of δ_k, we know $\delta_k \geq 0$ for all $k \geq 1$. From this, (3.10), and (3.16), we obtain (3.11).

Next, we prove (3.12). It follows from (3.3) and (3.13) that

$$\left\| Ax^{k+1} - b \right\|^2 \leq \frac{2}{\beta} \left(d(\lambda^{k+1}) - d(\lambda^k) + \eta_k \right), \ k = 1, 2, \ldots. \tag{3.17}$$

Summing (3.17) from $i = 1$ to k yields

$$\begin{aligned} \sum_{i=1}^{k} \left\| Ax^{i+1} - b \right\|^2 &\leq \frac{2}{\beta} \left(d(\lambda^{k+1}) - d(\lambda^1) + \sum_{i=1}^{k} \eta_i \right) \\ &\leq \frac{2}{\beta} \left(d(\lambda^*) - d(\lambda^1) + \sum_{i=1}^{k} \eta_i \right) \\ &\leq \frac{2}{\beta} \left(\frac{1}{2\beta} \left\| \lambda^1 - \lambda^* \right\|^2 + \sum_{i=1}^{k} \eta_i \right) \\ &\leq \frac{B^2}{\beta^2}, \end{aligned}$$

where the third inequality is due to the facts that $\nabla d(\lambda^*) = 0$ and $\nabla d(\lambda)$ is $\frac{1}{\beta}$-Lipschitz continuous [3] and the last inequality is due to (3.8). The proof of Theorem 3.4 is completed. □

Theorem 3.4 shows the global convergence of Algorithm 1 under conditions (1.11) and (3.10). Classical conditions in [32] that guarantee the global convergence of the IAL framework are (1.9) and (3.10). Since (1.9) implies (1.11) (by Theorem 2.1), our conditions (1.11) and (3.10) are weaker than conditions (1.9) and (3.10) in [32].

4 Non-ergodic convergence rate

In this section, we present the non-ergodic convergence rate result of our IAL framework (Algorithm 1). Theorems 4.1 and 4.3 show the non-ergodic convergence rate of the IAL framework.

Since $\delta_k \to 0$ (cf. Theorem 3.4) and $\eta_k \to 0$, there exists $k_0 \geq 4$ such that

$$\max_{k\geq k_0}\{\delta_k\} \leq \frac{1}{2\theta} \text{ and } \max_{k\geq k_0}\{\eta_k\} \leq \frac{1}{24\theta}, \tag{4.1}$$

where θ is given in (3.9). Define

$$\tau_1 := \frac{k_0}{4\theta} \text{ and } \tau_2 := \frac{1}{4\theta\sqrt{\eta_{k_0}}}. \tag{4.2}$$

It is easy to verify

$$\tau_1 \geq \frac{1}{\theta} \text{ and } \tau_2 \geq \sqrt{\frac{3}{2\theta}}. \tag{4.3}$$

Theorem 4.1 *Let $\{\lambda^k\}$ be generated by Algorithm 1. Suppose that the positive sequence $\{\eta_k\}$ is nonincreasing and satisfies* (3.10) *and*

$$\sqrt{\frac{\eta_{k+1}}{\eta_k}} \geq \frac{k-2}{k},\ k = k_0, k_0+1, \ldots, \tag{4.4}$$

where k_0 satisfies (4.1)*. Then,*

$$\delta_k \leq \frac{\tau_1}{k} + \tau_2\sqrt{\eta_k},\ k = k_0, k_0+1, \ldots, \tag{4.5}$$

where τ_1 and τ_2 are defined in (4.2)*.*

Proof We prove the theorem by induction. From (4.1) and (4.2), we know

$$\delta_{k_0} \leq \frac{1}{2\theta} = \frac{\tau_1}{k_0} + \tau_2\sqrt{\eta_{k_0}}.$$

Therefore, the inequality (4.5) holds for $k = k_0$. Next, we assume that (4.5) holds for some $k \geq k_0$, and we consider the case $k+1$. We have

$$\begin{aligned}
\delta_{k+1} &\leq \delta_k - \theta\delta_k^2 + \frac{3}{2}\eta_k \\
&\leq \frac{\tau_1}{k} + \tau_2\sqrt{\eta_k} - \theta\left(\frac{\tau_1}{k} + \tau_2\sqrt{\eta_k}\right)^2 + \frac{3}{2}\eta_k \\
&= \frac{\tau_1}{k+1}\frac{(k+1)(k-\theta\tau_1)}{k^2} + \tau_2\sqrt{\eta_{k+1}}\frac{\sqrt{\eta_k}\left(1-\frac{2\theta\tau_1}{k}\right)}{\sqrt{\eta_{k+1}}} + \left(\frac{3}{2} - \theta\tau_2^2\right)\eta_k \\
&\leq \frac{\tau_1}{k+1} + \tau_2\sqrt{\eta_{k+1}},
\end{aligned}$$

where the first inequality is due to (3.15), the second inequality is due to the facts that the function $x - \theta x^2$ is an increasing function in $x \in (-\infty, \frac{1}{2\theta}]$ and $\frac{\tau_1}{k} + \tau_2\sqrt{\eta_k} \leq \frac{\tau_1}{k_0} + \tau_2\sqrt{\eta_{k_0}} = \frac{1}{2\theta}$ for all $k \geq k_0$ (because $\{\eta_k\}$ is nonincreasing), and the last inequality is due to (4.3) and (4.4). The proof of Theorem 4.1 is completed. □

As shown in (4.5), the rate that $\{\delta_k\}$ converges to zero depends on two terms, i.e., $\frac{\tau_1}{k}$ and $\tau_2\sqrt{\eta_k}$, and the rate is determined by the slower one of them: if $k\sqrt{\eta_k} \to 0$, then $\delta_k \to 0$ with a rate of $\mathcal{O}(1/k)$; otherwise, $\delta_k \to 0$ with a rate of $\mathcal{O}\left(\sqrt{\eta_k}\right)$. In particular, if $\eta_k = 0$ for all $k \geq 1$, then Algorithm 1 reduces to the exact dual gradient ascent method, and achieves the $\mathcal{O}(1/k)$ convergence rate.

These facts indicate that the sequence $\{\eta_k\}$ in Algorithm 1 should not be chosen such that $\left\{\sqrt{\eta_k}\right\}$ converges faster than $\{1/k\}$ to zero. This is because such a choice would increase the computational cost of solving the AL subproblem, but theoretically cannot improve the convergence rate of $\{\delta_k\}$, which is $\mathcal{O}(1/k)$ in this case. One possible choice of the sequence $\{\eta_k\}$ is

$$\eta_k = \frac{\sigma}{k^{2\alpha}}, \ k = 1, 2, \ldots \tag{4.6}$$

with some constant $\sigma > 0$ and $\alpha \in (\frac{1}{2}, 1]$. It is easy to check that (4.6) satisfies all conditions required in Theorem 4.1.

The following Theorem 4.3 gives the non-ergodic convergence rate of Algorithm 1 when η_k is chosen as in (4.6). We first present a lemma, which is useful in proving Theorem 4.3.

Lemma 4.2 *Suppose the nonnegative sequence $\{\delta_k\}$ satisfies*

$$\frac{E}{2}\delta_{k+1}^2 + \delta_{k+1} \leq \delta_k, \ k = 1, 2, \ldots, \tag{4.7}$$

where $E > 0$ is a constant. Then, we have

$$\delta_k \leq \frac{\max\left\{\delta_1, \frac{4}{E}\right\}}{k}, \ k = 1, 2, \ldots. \tag{4.8}$$

Proof Again we prove this by induction. Clearly, the inequality (4.8) is true for $k = 1$. Next, assuming (4.8) is true for some $k \geq 1$, we show it is also true for $k + 1$. In fact, we have

$$\begin{aligned}
\delta_{k+1} \leq \frac{-1 + \sqrt{1 + 2E\delta_k}}{E} &\leq \frac{-1 + \sqrt{1 + 2E\frac{\max\left\{\delta_1, \frac{4}{E}\right\}}{k}}}{E} \\
&= \frac{2\max\left\{\delta_1, \frac{4}{E}\right\}}{k + \sqrt{k^2 + 2E\max\left\{\delta_1, \frac{4}{E}\right\}k}} \\
&\leq \frac{\max\left\{\delta_1, \frac{4}{E}\right\}}{k+1},
\end{aligned}$$

where the first inequality is due to the inequality (4.7), and the second inequality is due to the assumption that (4.8) holds for k. □

Theorem 4.3 *Let $\{x^k\}$ and $\{\lambda^k\}$ be generated by Algorithm 1. Suppose that the positive sequence $\{\eta_k\}$ is chosen as in* (4.6)*. Then,*

$$\delta_k \leq \frac{C}{k^\alpha},\ k = 1, 2, \ldots, \tag{4.9}$$

where

$$C = 4\sqrt{\frac{3\left(\frac{3}{2}\theta\sigma + 1\right)\sigma}{\theta}} + \frac{4}{3}\max\left\{\delta_1, \frac{4}{\theta}\right\}, \tag{4.10}$$

and θ is given in (3.9)*;*

$$\left\|Ax^{k+1} - b\right\|^2 \leq \psi_k := \frac{2}{\beta}\left(\frac{C+\sigma}{k^\alpha}\right),\ k = 1, 2, \ldots; \tag{4.11}$$

$$-\left\|\lambda^*\right\|\sqrt{\psi_k} - \frac{\beta}{2}\psi_k \leq F(x^{k+1}) - F(x^*) \leq \left(\left\|\lambda^*\right\| + B\right)\sqrt{\psi_k} + \eta_k,\ k = 1, 2, \ldots, \tag{4.12}$$

where λ^* is an optimal solution to problem (1.5) *and B is given in* (3.9)*; and*

$$\max_{x\in\mathbb{R}^n}\left\{\left\langle \nabla f(x^{k+1}) + A^T\lambda^{k+1},\, x^{k+1} - x\right\rangle + g(x^{k+1}) - g(x)\right\} \leq \eta_k,\ k = 1, 2, \ldots \tag{4.13}$$

Proof We show (4.9), (4.11), (4.12), and (4.13) separately.

We first show (4.9) by induction. Clearly, the inequality (4.9) holds for $k = 1$. Next, we assume that (4.9) holds for some $k \geq 1$, and show it is also true for $k+1$. We use the contrapositive argument. Assume that (4.9) does not hold for $k+1$, i.e.,

$$\delta_{k+1} > \frac{C}{(k+1)^\alpha}, \tag{4.14}$$

where C is given in (4.10). Let $z^* = \frac{C}{4} - \frac{1}{\theta} > 0$. If

$$\frac{\theta}{2}\left(\delta_{k+1} - \frac{z^*}{(k+1)^\alpha}\right)^2 + \left(\delta_{k+1} - \frac{z^*}{(k+1)^\alpha}\right) \leq \left(\delta_k - \frac{z^*}{k^\alpha}\right) \tag{4.15}$$

holds, then it follows from Lemma 4.2 that

$$\delta_{k+1} - \frac{z^*}{(k+1)^\alpha} \leq \frac{\max\left\{\delta_1, \frac{4}{\theta}\right\}}{k+1},$$

and

$$\delta_{k+1} \leq \frac{z^* + \max\left\{\delta_1, \frac{4}{\theta}\right\}}{(k+1)^\alpha} < \frac{C}{(k+1)^\alpha}. \tag{4.16}$$

Clearly, (4.16) contradicts (4.14), which implies that (4.9) is true. Next, we prove (4.15), which is equivalent to

$$P(z^*) := \frac{\theta}{2}\left(\delta_{k+1} - \frac{z^*}{(k+1)^\alpha}\right)^2 + \left(\delta_{k+1} - \frac{z^*}{(k+1)^\alpha}\right) - \left(\delta_k - \frac{z^*}{k^\alpha}\right) \leq 0.$$

We consider the following quadratic function $Q(z)$ with respect to z :

$$Q(z) = \frac{\theta}{2}z^2 - \left(\frac{\theta C}{4} - 1\right) z + \frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\sigma.$$

It can be verified that the minimizer of $Q(z)$ is z^*. Since the discriminant of $Q(z)$ is nonnegative, it follows that the minimum value

$$Q(z^*) \leq 0. \tag{4.17}$$

Moreover, for any $k \geq 1$, we have

$$(3.15) \Longrightarrow \delta_{k+1} \leq \delta_k + \frac{3}{2}\eta_k \Longrightarrow \frac{1}{2}\delta_{k+1}^2 \leq \delta_k^2 + \frac{9}{4}\eta_k^2 \Longrightarrow -\delta_k^2 \leq -\frac{1}{2}\delta_{k+1}^2 + \frac{9}{4}\sigma\eta_k.$$

Combining the last inequality in the above with (3.15) yields

$$\frac{\theta}{2}\delta_{k+1}^2 + \delta_{k+1} \leq \delta_k + \frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\eta_k, \tag{4.18}$$

which further implies

$$\begin{aligned} P(z^*) &= \frac{\theta}{2}\delta_{k+1}^2 + \delta_{k+1} - \delta_k - \frac{\theta\delta_{k+1}z^*}{(k+1)^\alpha} + \frac{\theta(z^*)^2}{2(k+1)^{2\alpha}} - \frac{z^*}{(k+1)^\alpha} + \frac{z^*}{k^\alpha} \\ &\leq \frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\eta_k - \frac{\theta\delta_{k+1}z^*}{(k+1)^\alpha} + \frac{\theta(z^*)^2}{2(k+1)^{2\alpha}} - \frac{z^*}{(k+1)^\alpha} + \frac{z^*}{k^\alpha}. \end{aligned} \tag{4.19}$$

By $\eta_k \leq \frac{\sigma}{k^{2\alpha}}$, the assumption $\delta_{k+1} > \frac{C}{(k+1)^\alpha}$ (cf. (4.14)), the facts $(k+1)^{2\alpha} \leq 4k^{2\alpha}$ and $(k+1)^\alpha - k^\alpha \leq 1$ for all $k \geq 1$ and $\alpha \in (0, 1]$, we get

$$\begin{aligned} &\frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\eta_k - \frac{\theta\delta_{k+1}z^*}{(k+1)^\alpha} + \frac{\theta(z^*)^2}{2(k+1)^{2\alpha}} - \frac{z^*}{(k+1)^\alpha} + \frac{z^*}{k^\alpha} \\ \leq\ &\frac{\frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\sigma}{k^{2\alpha}} - \frac{\theta C}{4k^{2\alpha}}z^* + \frac{\theta}{2k^{2\alpha}}(z^*)^2 + \frac{1}{k^{2\alpha}}z^* \\ =\ &\frac{Q(z^*)}{k^{2\alpha}}, \end{aligned}$$

which, together with (4.17) and (4.19), yields $P(z^*) \leq 0$.

We now show (4.11). From (3.17), we obtain

$$\begin{aligned} \left\|Ax^{k+1} - b\right\|^2 \leq \frac{2}{\beta}\left(d(\lambda^{k+1}) - d(\lambda^k) + \eta_k\right) &\leq \frac{2}{\beta}\left(d(\lambda^*) - d(\lambda^k) + \eta_k\right) \\ &= \frac{2}{\beta}\left(\delta_k + \eta_k\right), \end{aligned}$$

which, together with (4.6) and (4.9), yields (4.11).

Next, we show (4.12). From the strong duality and the definition of $\mathcal{L}_\beta(x;\lambda)$ (cf. (1.3)), we obtain

$$\begin{aligned} F(x^*) \leq \mathcal{L}_\beta(x^{k+1};\lambda^*) &= F(x^{k+1}) + \left\langle \lambda^*, Ax^{k+1} - b\right\rangle + \frac{\beta}{2}\left\|Ax^{k+1} - b\right\|^2 \\ &\leq F(x^{k+1}) + \left\|\lambda^*\right\|\left\|Ax^{k+1} - b\right\| + \frac{\beta}{2}\left\|Ax^{k+1} - b\right\|^2. \end{aligned}$$

This, together with (4.11), implies

$$F(x^{k+1}) - F(x^*) \geq -\|\lambda^*\|\sqrt{\psi_k} - \frac{\beta}{2}\psi_k. \tag{4.20}$$

On the other hand, we have

$$\begin{aligned}\mathcal{L}_\beta(x^{k+1};\lambda^k) &\leq \hat{f}_\beta(x(\lambda^k);\lambda_k) + \left\langle \nabla \hat{f}_\beta(x^{k+1};\lambda^k), x^{k+1} - x(\lambda^k)\right\rangle + g(x^{k+1}) \\ &= d(\lambda^k) + \left\langle \nabla \hat{f}_\beta(x^{k+1};\lambda^k), x^{k+1} - x(\lambda^k)\right\rangle + g(x^{k+1}) - g(x(\lambda^k)) \\ &\leq d(\lambda^k) + \eta_k \\ &\leq F(x^*) + \eta_k,\end{aligned}$$

where the first inequality is due to the convexity of $\hat{f}_\beta(x;\lambda)$ with respect to x, the first equality is due to the definition of $d(\lambda^k)$, the second inequality is due to (1.11), and the last inequality is due to the fact $d(\lambda^k) \leq F(x^*)$. Recall the definition of $\mathcal{L}_\beta(x;\lambda)$, we get

$$F(x^{k+1}) + \left\langle \lambda^k, Ax^{k+1} - b\right\rangle + \frac{\beta}{2}\left\|Ax^{k+1} - b\right\|^2 \leq F(x^*) + \eta_k,$$

which, together with (3.7), immediately implies

$$F(x^{k+1}) - F(x^*) \leq \left(\|\lambda^*\| + B\right)\sqrt{\psi_k} + \eta_k. \tag{4.21}$$

Combining (4.20) and (4.21) yields (4.12).

Finally, we show (4.13). It follows from (2.2) and the definition of $\hat{f}_\beta(x;\lambda)$ in (1.4) that

$$\nabla f(x^{k+1}) + A^T\lambda^{k+1} = \nabla f(x^{k+1}) + A^T\left(\lambda^k + \beta\left(Ax^{k+1} - b\right)\right) = \nabla \hat{f}_\beta(x^{k+1};\lambda^k).$$

The above, together with (1.11), immediately implies (4.13). The proof of Theorem 4.3 is completed. □

As a direct consequence of Theorem 4.3, we obtain the following result.

Corollary 4.4 *Let $\{x^k\}$ and $\{\lambda^k\}$ be generated by Algorithm 1 with $\eta_k = \frac{\sigma}{k^2}$. Then,*

$$\delta_k = \mathcal{O}\left(\frac{1}{k}\right), \quad \left\|Ax^{k+1} - b\right\| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right), \quad \left|F(x^{k+1}) - F(x^*)\right| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right) \tag{4.22}$$

and

$$\max_{x\in\mathbb{R}^n}\left\{\left\langle \nabla f(x^{k+1}) + A^T\lambda^{k+1},\, x^{k+1} - x\right\rangle + g(x^{k+1}) - g(x)\right\} = \mathcal{O}\left(\frac{1}{k^2}\right).$$

Next, we present some iteration complexity results of Algorithm 1 to return an ϵ-optimal solution of problem (1.1). Our definition of the ϵ-optimal solution is given as follows, which is a perturbation of the KKT optimality conditions.

Definition 4.5 (ϵ-optimal solution) For any given $\epsilon > 0$, $(x_\epsilon, \lambda_\epsilon)$ is called an ϵ-optimal solution pair if they satisfy

$$\|Ax_\epsilon - b\| \leq \sqrt{\epsilon} \tag{4.23}$$

and

$$\max_{x \in \mathbb{R}^n} \left\{ \left\langle \nabla f(x_\epsilon) + A^T \lambda_\epsilon,\, x_\epsilon - x \right\rangle + g(x_\epsilon) - g(x) \right\} \leq \epsilon. \tag{4.24}$$

Theorem 4.6 (Iteration Complexity) *For any $\epsilon > 0$, set*

$$\alpha = \sigma = 1 \tag{4.25}$$

in (4.6) *and the penalty parameter*

$$\beta = \frac{2(C+1)}{\sqrt{\epsilon}}, \tag{4.26}$$

where C is defined in (4.10)*. Then, the total number of iterations for Algorithm 1, where AL subproblem* (2.1) *is approximately solved by Algorithms 2 or 3 until a point x^{k+1} satisfying* (1.11) *is found, to return an ϵ-optimal solution of problem* (1.1) *satisfying* (4.23) *and* (4.24) *is at most*

$$T_1 := \left\lceil 4 \left(L_f + \frac{2\,(C+1)\,\|A\|^2}{\sqrt{\epsilon}} \right) \frac{D^2}{\epsilon^{3/2}} \right\rceil \tag{4.27}$$

and

$$T_2 := \left\lceil 6 \left(L_f + \frac{2\,(C+1)\,\|A\|^2}{\sqrt{\epsilon}} \right) \frac{D^2}{\epsilon^{3/2}} \right\rceil, \tag{4.28}$$

respectively, where L_f is the Lipschitz constant of $\nabla f(x)$ and D is defined in (2.4)*.*

Proof Let $K = \lceil 1/\sqrt{\epsilon} \rceil$. Substituting $k = K$, α and σ in (4.25), and β in (4.26) into (4.11) and (4.13), we immediately see that the pair (x^{k+1}, λ^{k+1}) satisfies (4.23) and (4.24). Next, we compute the total iteration complexity of Algorithm 1 with Algorithm 2 being used to solve the AL subproblem. By invoking Theorem 2.1, we know that the total number of iterations is

$$\sum_{k=1}^{K} \left\{ \left\lceil \frac{4L_{\hat{f}} D^2}{\eta_k} \right\rceil - 1 \right\} \leq \frac{4\left(L_f + \beta \|A\|^2\right) D^2}{\sigma} \sum_{k=1}^{K} k^{2\alpha} \leq T_1.$$

Using the same argument, we can show that the iteration complexity of Algorithm 1 with Algorithm 3 being used to solve the AL subproblem is upper bounded by T_2. We omit the details for succinctness. □

5 Comparisons with existing works [16, 23]

In this section, we make some remarks on the comparison of our proposed IAL framework (Algorithm 1) and two closely related methods in [16, 23] for solving the linearly constrained convex programming problems.

We first compare our proposed IAL method with the one in [16]. The method in [16] is designed for solving problem (1.1) with $g(x) = \mathrm{Ind}_{\mathcal{X}}(x)$. It applies Nesterov's optimal first-order method to solve AL subproblem (2.1) until a point x^{k+1} satisfying (1.9) is found. Our IAL framework can be used to solve more general problem (1.1) (with a general composite function $g(x)$). Our framework requires approximately solving subproblem (2.1) until a point x^{k+1} satisfying (1.11) (which is easier to check than (1.9)) is found.

The work [16] shows the same non-ergodic convergence rate results as ours in (4.22) in Corollary 4.4, but under a much stronger condition that the sequence $\{\eta_k\}$ in (1.9) satisfies

$$\sum_{i=1}^{k} \eta_i^2 = \mathcal{O}\left(\frac{1}{k}\right).$$

To make it more clearly, consider the special case where we are interested in finding an exact solution of problem (1.1), which requires $k \to +\infty$ in Corollary 4.4. In this case, the method in [16] needs to solve each AL subproblem exactly (i.e., η_i in (1.9) needs to be zero for all $i = 1, 2, \ldots$), while our IAL framework only needs to solve each subproblem approximately (i.e., η_i in (1.11) only needs to be in the order of $\mathcal{O}(1/i^2)$ for $i = 1, 2, \ldots$).

Next, we compare our proposed IAL method with the one in [23]. The closest related method in [23] to our IAL method, called inexact gradient augmented Lagrangian, is designed for solving a class of convex conic problems

$$\min_{u \in \mathcal{U}} \ f(u), \ \text{s.t.} \quad Gu + g \in \mathcal{K}, \tag{5.1}$$

where $\mathcal{U} \subseteq \mathbb{R}^n$ is a convex compact set, $\mathcal{K} \subseteq \mathbb{R}^m$ is a convex cone, $G \in \mathbb{R}^{m \times n}$, and $g \in \mathbb{R}^m$. It is easy to show that problem (5.1) is a special case of problem (1.1). At the k-th iteration, the method in [23] applies Nesterov's optimal first-order method to solve AL subproblem (2.1) until a point x^{k+1} satisfying

$$\mathcal{L}_\beta\left(x^{k+1}; \lambda^k\right) - \mathcal{L}_\beta\left(x(\lambda^k); \lambda^k\right) \leq \delta \tag{5.2}$$

is found, where $\delta > 0$ is the given accuracy; then the method updates the dual variable by

$$\lambda^{k+1} = \lambda^k + \frac{\beta}{2}\left(Ax^{k+1} - b\right). \tag{5.3}$$

Again, our framework requires approximately solving subproblem (2.1) until a point x^{k+1} satisfying (1.11) is found and updates the dual variable via (2.2). Notice that the dual variable update formula (5.3) in [23] is different from (2.2) in the classical AL method.

The work [23, Corollary 3.3 and Theorem 3.4] shows the ergodic convergence rate results in terms of the dual function values, the primal infeasibility, and the primal function values. Moreover, the work [23, Theorem 3.5] shows that it takes

the algorithm (with an optimal choice of the penalty parameter β and the solution tolerance δ) a total number of $\mathcal{O}(1/\epsilon)$ iterations to return an ϵ-optimal solution u_ϵ defined as follows:

$$|f(u_\epsilon) - f^*| \leq \epsilon \text{ and } \text{dist}_{\mathcal{K}}(Gu_\epsilon + g) \leq \epsilon, \tag{5.4}$$

where f^* is the optimal value of problem (5.1). As mentioned in [23], the algorithm behind the above complexity result reduces to a quadratic penalty method (without any update of the dual variable) and therefore there is no convergence guarantee for the dual variable.

In summary, the results in our paper significantly differ from the ones in [23] in terms of global convergence results, convergence rate results, and the algorithms.

- **Global convergence**. Our IAL framework enjoys the global convergence (under the assumption that the error sequence $\{\eta_k\}$ is summable), but the inexact gradient augmented Lagrangian method in [23] does not have global convergence guarantee, due to the existence of the positive accuracy constant δ.
- **Convergence rate**. The convergence results for our IAL framework in terms of the dual objective values, the primal infeasibility, and the primal objective values are all for the *non-ergodic* solution, but the results for the inexact gradient augmented Lagrangian method in [23] are for the *ergodic* solution.
- **Algorithms and dual optimality guarantee**. The dual variable update formula in our IAL framework and the one in [23] are different. The algorithms behind the iteration complexity results (see Theorem 3.8 in [23] and Theorem 4.6 in our paper) are also sharply different from each other. The algorithm behind Theorem 3.8 in [23] is essentially a penalty method (without any update of the dual variable) but the algorithm behind Theorem 4.6 in our paper indeed is an IAL method. Therefore, there is no convergence guarantee for the dual variable in [23], while it is guaranteed in our paper. In particular, from Theorem 4.3 and with the choice of the parameters in Theorem 4.6, we get

  $$\max_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla f(x^{k+1}) + A^T\lambda^{k+1},\, x^{k+1} - x \right\rangle + g(x^{k+1}) - g(x) \right\} \leq \epsilon$$

  and

  $$d(\lambda^*) - d(\lambda^{k+1}) = \mathcal{O}\left(\sqrt{\epsilon}\right).$$

- **Definition of ϵ-optimal solution.** Our definition of the ϵ-optimal solution is a natural perturbation of the KKT optimality conditions of problem (1.1), which involves the dual variable. The definition of the ϵ-optimal solution in [23], i.e., (5.4), does not involve the dual variable.

6 Numerical results

In this section, we present some preliminary numerical results for the purpose of comparing the following two things: (i) the difference of the ergodic and nonergidoc solutions; (ii) the difference of the "exact" augmented Lagrangian (EAL) method and our IAL method. Our codes were written in MATLAB and the results were obtained on a standard PC.

The numerical experiments were conducted on basis pursuit problem (1.10). In Table 6.1 we report the results for some small problem with $m = 60$, $n = 100$

and sparsity (number of nonzero entries) $s = 15$. The 10 instances were randomly generated in the following manner: the entries of A were generated randomly following the standard Gaussian distribution $\mathcal{N}(0, 1)$; the positions of the nonzero entries in x^* were uniformly randomly chosen and their values were generated following the uniform distribution in $(0, 1)$; and finally b is set to $b = Ax^*$. To construct the bounded set containing the true solution, we set $\hat{x} = A_m^{-1}b$, where A_m is the square matrix formed by the first m columns of A. We use $\bar{x}$ to denote the solution returned by EAL or IAL.

We ran both of IAL and EAL for $K = 200$ (dual) iterations. We set both of the initial (primal) point x^1 and the initial (dual) Lagrange multiplier λ^1 to 0. We applied the proximal gradient method to solve the AL subproblem until (1.11) is satisfied with $\eta_k = 1/k^2$ for IAL and $\eta_k \equiv 10^{-4}$ for EAL. We reported the comparison results in Table 6.1. In particular, we reported the cpu time (in seconds), the relative error of the solution (denoted by relerr $= \|\bar{x} - x^*\|/\|x^*\|$), the residual of the linear constraint (denoted by resi $= \|A\bar{x} - b\|$), and the objective value error (denoted by objerr $= |\|\bar{x}\|_1 - \|x^*\|_1|$). Moreover, we also reported the sparsity (the number of the nonzero entries) of the returned non-ergodic solution $\bar{x}$ (denoted by s_n) and the sparsity of the ergodic solution $x_e := \sum_{k=1}^{K} x^k/K$ (denoted by s_e).

We see from Table 6.1 that IAL and EAL are comparable in terms of the solution quality measured by relerr, resi, and objerr, and there is no evidence showing that one is better than the other. However, IAL is much faster than EAL in terms of the cpu time. This is expected because the AL subproblems are solved much less accurately in IAL in the first 100 iterations (compared to EAL). It is worth mentioning that $\eta_k < 10^{-4}$ for $k > 100$ and $\eta_k = 0.25 * 10^{-4}$ for $k = 200$ in IAL, that is, the last 100 AL subproblems in IAL are solved slightly more accurately than EAL, but IAL is still much faster.

We also observe from Table 6.1 that for both IAL and EAL, the non-ergodic solution $\bar{x}$ is significantly more sparse than the ergodic solution x_e. In fact, the sparsity pattern of the non-ergodic solution always perfectly matches that of the true solution x^*. This well justifies the importance of our global convergence and convergence rate analysis on the *non-ergodic* solution in this paper.

We also report numerial results for some larger problem instances in Table 6.2. The problem instances are randomly generated in the same manner as before. From Table 6.2 we have the same observations as the ones from Table 6.1: IAL is much faster than EAL in terms of the cpu time and the non-ergodic solutions are significantly more sparse than the ergodic solutions.

Acknowledgements We are grateful to the editors and three anonymous referees for their insightful comments and suggestions that helped us improve the quality of the paper greatly. We would like also to thank Guanghui Lan, Zhaosong Lu, Zaiwen Wen, and Wotao Yin for their insightful comments, which helped us in improving the results in this paper. We thank Xiangfeng Wang for the useful discussion on an earlier version of this paper.

The work of Y.-F. Liu was supported in part by NSFC grants 11631013, 11331012, 11671419, and 11571221, and Beijing NSF grant L172020. The work of X. Liu was supported in part by NSFC grants 11622112, 11471325, 91530204, and 11688101, the National Center for Mathematics and Interdisciplinary Sciences, CAS, and Key Research Program of Frontier Sciences (QYZDJ-SSW-SYS010), CAS. The work of S. Ma was supported in part by a startup package in Department of Mathematics at UC Davis.

Table 6.1 Numerical results of IAL and EAL for solving basis pursuit problem (1.10) with $m = 60$, $n = 100$, $s = 15$, where ID denotes the problem instance index.

	IAL						EAL					
ID	s_e	s_n	relerr	resi	objerr	cpu	s_e	s_n	relerr	resi	objerr	cpu
1	40	15	2.0e-08	1.9e-07	2.8e-08	2.2	40	15	1.4e-08	1.3e-07	5.7e-08	3.9
2	27	15	1.0e-08	9.6e-08	1.2e-08	0.6	27	15	8.0e-09	7.0e-08	1.6e-08	1.3
3	16	15	3.3e-08	3.5e-07	1.2e-08	0.2	16	15	3.9e-08	4.0e-07	1.4e-08	0.3
4	23	15	7.2e-09	8.1e-08	2.7e-08	0.3	23	15	9.0e-09	9.8e-08	3.0e-08	0.5
5	25	15	1.7e-08	1.7e-07	9.1e-09	0.4	25	15	6.4e-09	5.7e-08	3.1e-09	0.8
6	26	15	3.0e-09	2.2e-08	8.5e-10	1.1	26	15	2.8e-09	2.1e-08	1.8e-09	2.0
7	26	15	5.3e-08	4.7e-07	1.7e-07	0.7	26	15	3.5e-08	3.0e-07	1.3e-07	1.7
8	23	15	2.9e-08	2.7e-07	2.5e-08	0.3	22	15	1.1e-08	1.0e-07	4.6e-08	0.7
9	25	18	6.0e-04	6.7e-03	1.3e-03	0.5	25	18	6.0e-04	6.7e-03	1.3e-03	1.0
10	17	15	6.4e-08	6.8e-07	7.6e-08	0.2	17	15	4.7e-08	4.6e-07	3.0e-08	0.3

References

1. Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
2. Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont, MA, USA, 1996.
3. Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, USA, 1999.
4. Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. Optim.*, 10(3):627–642, 2000.
5. J. D. Buys. *Dual algorithms for constrained optimization problems*. Ph. D. Thesis, University of Leiden, The Netherlands, 1972.
6. Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2):37–75, 2014.
7. Jonathan Eckstein and Paulo J. S. Silva. A practical relative error criterion for augmented Lagrangians. *Math. Program.*, 141(1-2):319–348, 2013.
8. Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.
9. Tom Goldstein and Stanley Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2(2):323–343, 2009.
10. Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152(1-2):75–112, 2015.
11. Magnus R. Hestenes. Multiplier and gradient methods. *J. Optim. Theory Appl.*, 4(5):303–330, 1969.
12. Alfredo N. Iusem and Marc Teboulle. Convergence rate analysis of nonquadratic proximal methods for convex and linear programming. *Math. Oper. Res.*, 20(3):657–677, 1995.
13. Martin Jaggi. *Sparse Convex Optimization Methods for Machine Learning*. PhD thesis, ETH Zurich, 2011.
14. Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceeding of ICML, Atlanta*, 2013.

Table 6.2 Numerical results of IAL and EAL for solving larger basis pursuit problem (1.10), where ID denotes the problem instance index.

	IAL						EAL					
	$m = 600$, $n = 1000$, $s = 150$											
ID	s_e	s_n	relerr	resi	objerr	cpu	s_e	s_n	relerr	resi	objerr	cpu
1	241	150	5.3e-12	5.2e-10	4.1e-11	588.5	241	150	3.9e-12	3.9e-10	2.0e-11	871.5
2	188	150	7.4e-11	7.1e-09	3.7e-10	139.6	188	150	6.8e-11	6.7e-09	3.4e-10	241.8
3	233	150	5.4e-11	5.1e-09	5.2e-10	239.0	233	150	2.5e-11	2.4e-09	1.4e-10	360.6
4	222	150	6.0e-11	5.5e-09	3.2e-10	130.5	222	150	3.1e-11	3.0e-09	3.1e-10	222.0
5	225	150	1.2e-11	1.2e-09	2.8e-11	452.9	225	150	6.4e-12	6.0e-10	5.3e-11	698.7
	$m = 1800$, $n = 3000$, $s = 450$											
1	1609	450	7.4e-12	2.2e-09	1.2e-10	2999.1	1609	450	6.8e-12	2.0e-09	7.8e-12	4424.2
2	1668	450	3.6e-12	9.9e-10	3.0e-11	3823.4	1668	451	4.2e-12	1.2e-09	2.2e-11	8075.7
3	1639	450	1.3e-11	3.5e-09	5.3e-11	5176.9	1639	450	8.6e-12	2.5e-09	5.7e-11	5921.8
4	1669	450	5.8e-12	1.6e-09	2.1e-11	7136.7	1669	451	7.9e-12	2.4e-09	1.1e-10	7858.3
5	1692	450	2.9e-13	7.9e-11	5.1e-13	2464.4	1692	450	3.1e-13	8.7e-11	6.8e-12	3146.6

15. Guanghui Lan. Gradient sliding for composite optimization. *Math. Program.*, 159(1-2):201–235, 2016.
16. Guanghui Lan and Renato D. C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.*, 155(1-2):511–547, 2016.
17. Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM J. Optim.*, 26(2):1379–1409, 2016.
18. Chengbo Li, Wotao Yin, Hong Jiang, and Yin Zhang. An efficient augmented Lagrangian method with applications to total variation minimization. *Comput. Optim. Appl.*, 56(3):507–530, 2013.
19. Xudong Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM J. Optim.*, 28(1):433–458, 2018.
20. Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.*, 46(1):157–178, 1993.
21. Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank–Wolfe meets proximal methods. *SIAM J. Sci. Comput.*, 38(5):3291–3317, 2016.
22. Ion Necoara and Andrei Patrascu. Iteration complexity analysis of dual first-order methods for conic convex programming. *Optim. Methods Softw.*, 31(3):645–678, 2016.
23. Ion Necoara, Andrei Patrascu, and Francois Glineur. Complexity of first order inexact Lagrangian and penalty methods for conic convex programming. *https://arxiv.org/abs/1506.05320*, 2017.
24. Valentin Nedelcu, Ion Necoara, and Quoc Tran-Dinh. Computational complexity of inexact gradient augmented Lagrangian methods: Application to constrained MPC. *SIAM J. Control Optim.*, 52(5):3109–3134, 2014.
25. Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math. Dokl.*, 27(2):372–376, 1983.

26. Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
27. Yurii Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1):125–161, 2013.
28. Yurri Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *to appear Math. Program.*, 2017.
29. Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Model. Simul.*, 4(2):460–489, 2005.
30. M. J. D. Powell. A method for nonlinear constraints in minimization problems. *In: R. Fletcher (ed.) Optimization*, pages 283–298, 1969.
31. Ralph Tyrrell Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *J. Optim. Theory Appl.*, 12(6):555–562, 1973.
32. Ralph Tyrrell Rockafellar. Augmented Lagrangian and applications of the proximal point algorithm in convex progremming. *Math. Oper. Res.*, 1(2):97–116, 1976.
33. Anthony Man-Cho So and Zirui Zhou. Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. *Optim. Methods Softw.*, 32(4):963–992, 2017.
34. R. A. Tapia. Newton's method for problems with equality constraints. *SIAM J. Numer. Anal.*, 11(5):874–886, 1974.
35. Liuqin Yang, Defeng Sun, and Kim-Chuan Toh. SDPNAL+: A majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints. *Math. Prog. Comp.*, 7(3):331–366, 2015.
36. Wotao Yin and Stanley Osher. Error forgetting of Bregman iteration. *J. Sci. Comput.*, 54(2-3):684–695, 2013.
37. Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143–168, 2008.
38. Yaxiang Yuan. Analysis on a superlinearly convergent augmented Lagrangian method. *Acta Math. Sinica*, 30(1):1–10, 2014.