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

Moreau Envelope Augmented Lagrangian Method for Nonconvex Nonsmooth Optimization with Linear Constraints

Jinshan Zeng · Wotao Yin · Ding-Xuan Zhou

Received: date / Accepted: date

Abstract The augmented Lagrangian method (ALM) is one of the most useful methods for constrained optimization. Its convergence has been well established under convexity assumptions or smoothness assumptions, or under both assumptions. ALM may experience oscillations and divergence facing nonconvexity and nonsmoothness simultaneously. In this paper, we consider the linearly constrained problem with a nonconvex (in particular, weakly convex) and nonsmooth objective. We modify ALM using a Moreau envelope of the augmented Lagrangian and establish its convergence under conditions that are weaker than those in the literature. We call it the *Moreau envelope augmented Lagrangian (MEAL)* method. We also show that the iteration complexity of MEAL is $o(\varepsilon^{-2})$ to yield an ε-accurate first-order stationary point. We establish its whole sequence convergence (regardless of the initial guess) and a rate when a Kurdyka-Łojasiewicz property is assumed. Moreover, when the subproblem of MEAL has no closed-form solution and is difficult to solve, we propose two practical variants of MEAL, an inexact version called *iMEAL* with an approximate proximal

We thank Kaizhao Sun for discussions that help us complete this paper, as well as presenting to us an additional approach to ensure boundedness. The work of J. Zeng is partly supported by National Natural Science Foundation of China (No. 61977038) and the Thousand Talents Plan of Jiangxi Province (No. jxsq2019201124). The work of D.-X. Zhou is partly supported by Research Grants Council of Hong Kong (No. CityU 11307319) and the Hong Kong Institute for Data Science.

J. Zeng
School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China.
Liu Bie Ju Centre for Mathematical Sciences, City University of Hong Kong, Hong Kong.
E-mail: jinshanzeng@jxnu.edu.cn

W. Yin
Department of Mathematics, University of California, Los Angeles, CA.
E-mail: wotaoyin@math.ucla.edu

D.X. Zhou
School of Data Science, Department of Mathematics, and Liu Bie Ju Centre for Mathematical Sciences, City University of Hong Kong, Hong Kong.
E-mail: mazhou@cityu.edu.hk

update, and a linearized version called *LiMEAL* for the constrained problem with a composite objective. Their convergence is also established.

Keywords Nonconvex nonsmooth optimization · augmented Lagrangian method · Moreau envelope · proximal augmented Lagrangian method · Kurdyka-Łojasiewicz inequality

1 Introduction

In this paper, we consider the following optimization problem with linear constraints

$$\begin{array}{ll} \text{minimize}_{x\in\mathbb{R}^n} & f(x) \\ \text{subject to} & Ax = b, \end{array} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a proper, lower-semicontinuous *weakly convex* function, which is possibly nonconvex and nonsmooth, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are some given matrix and vector, respectively. A function f is said to be *weakly convex* with a modulus $\rho > 0$ if $f(x) + \frac{\rho}{2}\|x\|^2$ is convex on $\mathbb{R}^n$, where $\|\cdot\|$ is the Euclidean norm. The class of weakly convex functions is broad [50], including all convex functions, smooth but nonconvex functions with Lipschitz continuous gradient, and their composite forms (say, $f(x) = h(x)+g(x)$ with both h and g being weakly convex, and $f(x) = g(h(x))$ with g being convex and Lipschitz continuous and h being a smooth mapping with Lipschitz Jacobian [31, Lemma 4.2]).

The augmented Lagrangian method (ALM) is a well-known algorithm for constrained optimization by Hestenes [37] and Powell [53]. ALM has been extensively studied and has a large body of literature ([54,9,25,24,14] just to name a few), yet *no ALM algorithm can solve the underlying problem (1) without at least one of the following assumptions*: convexity [52,54,9,10,33], or smoothness [2,3,5,4,27], or solving nonconvex subproblems to their global minima [14,16], or (when running ALM) the observation of a bounded nondecreasing penalty sequence [34,18]. Indeed, ALM may oscillate and even diverge unboundedly on simple quadratic programs [68, 59], where the objectives are weakly convex as shown in Sec. 7.1 later.

At a high level, we introduce a Moreau-envelope modification of the ALM for solving (1) and show the method can converge under weaker conditions. In particular, convexity is relaxed to weak convexity; nonsmooth functions are allowed; the subproblems can be solved inexactly; linearization can be applied to a Lipschitz-differential function; and, rank of A is not assumed. On the other hand, we introduce certain subgradient properties[1] as our main assumptions. By also assuming either a bounded energy sequence or bounded primal-dual sequence, we derive present subsequence rates of convergence. We introduce a novel way to establish it based on a feasible coercivity assumption and a local-stability assumption on the subproblem. Finally, with the additional assumption of Kurdyka-Łojasiewicz (KŁ) inequality, we establish global convergence.

[1] In this paper, we use either certain *implicit Lipschitz subgradient property* or *implicit bounded subgradient property* (see, Definition 1) to yield the convergence of proposed methods, where the *implicit Lipschitz subgradient property* is much weaker than the smoothness assumptions used in the literature.

1.1 Proposed Algorithms

To present our algorithm, define the augmented Lagrangian:

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

and the *Moreau envelope* of $\mathcal{L}_\beta(x,\lambda)$:

$$\phi_\beta(z,\lambda) = \min_x \left\{\mathcal{L}_\beta(x,\lambda) + \frac{1}{2\gamma}\|x-z\|^2\right\}, \tag{3}$$

where $\lambda \in \mathbb{R}^m$ is a multiplier vector, $\beta > 0$ is a penalty parameter, and $\gamma > 0$ is a proximal parameter. The Moreau envelope applies to the primal variable x for each fixed dual variable λ.

We introduce *Moreau Envelope Augmented Lagrangian method* (dubbed *MEAL*) as follows: given an initialization (z^0, λ^0), $\gamma > 0$, a sequence of penalty parameters $\{\beta_k\}$ and a step size $\eta \in (0,2)$, for $k = 0, 1, \ldots$, run

$$\text{(MEAL)} \quad \begin{cases} z^{k+1} = z^k - \eta\gamma\nabla_z \phi_{\beta_k}(z^k,\lambda^k), \\ \lambda^{k+1} = \lambda^k + \beta_k \nabla_\lambda \phi_{\beta_k}(z^k,\lambda^k). \end{cases} \tag{4}$$

The penalty parameter β_k can either vary or be fixed.

Introduce

$$x^{k+1} = \mathrm{Prox}_{\gamma,\mathcal{L}_{\beta_k}(\cdot,\lambda^k)}(z^k) := \operatorname*{argmin}_x \left\{\mathcal{L}_{\beta_k}(x,\lambda^k) + \frac{1}{2\gamma}\|x-z^k\|^2\right\},\ \forall k \in \mathbb{N},$$

which yields $\nabla_z\phi_{\beta_k}(z^k,\lambda^k) = \gamma^{-1}(z^k - x^{k+1})$ and $\nabla_\lambda \phi_{\beta_k}(z^k,\lambda^k) = Ax^{k+1} - b$. Then, MEAL (4) is equivalent to:

$$\text{(MEAL Reformulated)} \quad \begin{cases} x^{k+1} = \mathrm{Prox}_{\gamma,\mathcal{L}_{\beta_k}(\cdot,\lambda^k)}(z^k), \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b). \end{cases} \tag{5}$$

For $\mathrm{Prox}_{\gamma,\mathcal{L}_\beta}$ without closed-form solutions, we provide two practical variants of MEAL.

Inexact MEAL (iMEAL) We call x^{k+1} an ϵ_k-accurate stationary point of the x-subproblem in (5) if there exists

$$s^k \in \partial_x \mathcal{L}_{\beta_k}(x^{k+1},\lambda^k) + \gamma^{-1}(x^{k+1} - z^k) \quad \text{such that } \|s^k\| \le \epsilon_k. \tag{6}$$

iMEAL is described as follows: given an initialization (z^0,λ^0), $\gamma > 0$, $\eta \in (0,2)$, and two positive sequences $\{\epsilon_k\}$ and $\{\beta_k\}$, for $k = 0, 1, \ldots$, run

$$\text{(iMEAL)} \quad \begin{cases} \text{find an } x^{k+1} \text{ to satisfy (6)}, \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b). \end{cases} \tag{7}$$

Linearized MEAL (LiMEAL) When problem (1) has the following form

$$\begin{array}{ll} \text{minimize}_{x\in\mathbb{R}^n} & f(x) := h(x) + g(x) \\ \text{subject to} & Ax = b, \end{array} \tag{8}$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz-continuous differentiable and $g : \mathbb{R}^n \to \mathbb{R}$ is weakly convex and has an easy proximal operator (in particular, admitting a closed-form solution) [61,36,59,65], we shall use ∇h. Write $f^k(x) := h(x^k) + \langle \nabla h(x^k), x - x^k\rangle + g(x)$ and $\mathcal{L}_{\beta,f^k}(x,\lambda) := f^k(x) + \langle \lambda, Ax - b\rangle + \frac{\beta}{2}\|Ax - b\|^2$. We describe *LiMEAL* for (8) as: given (z^0, λ^0), $\gamma > 0$, $\eta \in (0, 2)$ and $\{\beta_k\}$, for $k = 0, 1, \ldots,$ run

$$\text{(LiMEAL)} \quad \begin{cases} x^{k+1} = \mathrm{Prox}_{\gamma,\mathcal{L}_{\beta_k,f^k}(\cdot,\lambda^k)}(z^k), \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b). \end{cases} \tag{9}$$

1.2 Relation to ALM and Proximal ALM

Like ALM, MEAL alternatively updates primal and dual variables; but unlike ALM, MEAL applies the update to the Moreau envelope of augmented Lagrangian. By [56], the Moreau envelope $\phi_{\beta_k}(z, \lambda^k)$ provides a smooth approximation of $\mathcal{L}_{\beta_k}(x, \lambda^k)$ from below and shares the same minima. The smoothness of Moreau envelope alleviates the possible oscillation that arises when ALM is applied to certain nonconvex optimization problems; see the example in [59, Proposition 1].

For the problems satisfying the conditions in this paper, ALM may require a sequence of possibly unbounded $\{\beta_k\}$. When β_k is large, the ALM subproblem is ill-conditioned. Therefore, bounding β_k is practically desirable [25,17]. MEAL and its practical variants can use a fixed penalty parameter under the *implicit Lipschitz subgradient* assumption introduced in Definition 1 later.

Proximal ALM was introduced in [55]. Its variants were recently studied in [38, 36,68,67]. These methods add a proximal term to the augmented Lagrangian. By the reformulation (5) of MEAL, proximal ALM [55] for problem (1) is a special case of MEAL with the step size $\eta = 1$. In [38], a proximal primal-dual algorithm called *Prox-PDA* was proposed for problem (1). Certain non-Euclidean matrix norms were adopted in Prox-PDA to guarantee the strong convexity of the ALM subproblem. A proximal linearized version of Prox-PDA for the composite optimization problem (8) was studied in [36]. These methods are closely related to MEAL, but their convergence conditions in the literature are stronger.

Recently, [68,67] modified proximal inexact ALM for the linearly constrained problems with an additional bounded box constraint set or polyhedral constraint set, denoted by C. Our method is partially motivated by their methods. Their problems are equivalent to the composite optimization problems (8) with $g(x) = \iota_C(x)$, where $\iota_C(x) = 0$ when $x \in C$ and $+\infty$ otherwise. In this setting, the methods in [68,67] can be regarded as prox-linear versions of LiMEAL (9), that is, yielding x^{k+1} via a prox-linear scheme [61] instead of the minimization scheme as used in LiMEAL (9), together with an additional dual step size and a sufficiently small primal step size in

[68,67]. Specifically, in the case of $g(x) = \iota_C(x)$, the updates of x^{k+1} in methods in [68,67] are yielded by

$$x^{k+1} = \mathrm{Proj}_C(x^k - s\nabla K(x^k, z^k, \lambda^k)),$$

where $K(x^k, z^k, \lambda^k)) = \mathcal{L}_{\beta^k, f}(x, \lambda^k) + \frac{1}{2\gamma}\|x - z^k\|^2$, and $\mathrm{Proj}_C(x)$ is the projection of x onto C. Besides, LiMEAL handles proximal functions beyond the indicator function and permits the wider choice $\eta \in (0, 2)$.

1.3 Other Related Literature

On convex and constrained problems, locally linear convergence[2] of ALM has been extensively studied in the literature [52,9,10,33,11,48,26], mainly under the second order sufficient condition (SOSC) and constraint conditions such as the linear independence constraint qualification (LICQ). Global convergence (i.e., convergence regardless of the initial guess) of ALM and its variants was studied in [58,54,25,24, 1,6,13,14,15], mainly under constraint qualifications and assumed boundedness of nondecreasing penalty parameters. On nonconvex and constrained problems, convergence of ALM was recently studied in [14,16,2,3,5,4,27], mainly under the following assumptions: solving nonconvex subproblems to their approximate global minima or stationary points [14,16], or boundedness of the nondecreasing penalty sequence [34, 18]. Most of them require *Lipschitz differentiability* of the objective.

Convergence of proximal ALM and its variants was established under the assumptions of either convexity in [55] or smoothness (in particular, Lipschitz differentiablity) in [38,36,39,68,67,60]. Besides proximal ALM, other related works for nonconvex and constrained problems include [12,35,49,51], which also assume smoothness of the objective, plus either gradient or Hessian information.

1.4 Contribution and Novelty

MEAL, iMEAL, and LiMEAL achieve the same order of iteration complexity $o(\varepsilon^{-2})$ to reach an ε-accurate first-order stationary point, slightly better than those in the ALM literature [38,36,60,68,67] but require weaker conditions. Our methods have convergence guarantees for a broader class of objective functions, for example, nonsmooth and nonconvex functions like $|x^2 - 1|$, $|xy - 1|$ (where $x, y \in \mathbb{R}$), the smoothly clipped absolute deviation (SCAD) regularization [32] and minimax concave penalty (MCP) regularization [66], which are underlying the applications of low-rank matrix/tensor factorization, phase retrieval, blind deconvolution, robust principal component analysis, statistical learning and beyond [28,30,59].

It should be pointed out that we only assume the feasibility of $Ax = b$, instead of these commonly used but stricter hypotheses in the literature, such as: the strict complementarity condition used in [68], certain rank assumption (such as $\mathrm{Im}(A) \subseteq$

[2] Locally linear convergence means exponentially fast convergence to a local minimum from a sufficiently close initial point.

$\text{Im}(B)$ when considering the two- (multi-)block case $Ax + By = 0$) used in [59], and the linear independence constrained qualification (LICQ) used in [11,48] (implying the full-rank assumption in the linear constraint case).

Our analysis is noticeably different from those in the literature [55,36,38,39,68, 67,60,59]. We base our analysis on new potential functions. The Moreau envelope in the potential functions is partially motivated by [28]. Our potential functions are tailored for MEAL, iMEAL, and LiMEAL and include the augmented Lagrangian with additional terms. The technique of analysis may have its own value for further generalizing and improving ALM-type methods.

1.5 Notation and Organization

We let $\mathbb{R}$ and $\mathbb{N}$ denote the sets of real and natural numbers, respectively. Given a matrix A, $\text{Im}(A)$ denotes its image, and $\tilde{\sigma}_{\min}(A^T A)$ denotes the smallest positive eigenvalue of $A^T A$. $\|\cdot\|$ is the Euclidean norm for a vector. Given any two nonnegative sequences $\{\xi_k\}$ and $\{\zeta_k\}$, we denote by $\xi_k = o(\zeta_k)$ if $\lim_{k\to\infty} \frac{\xi_k}{\zeta_k} = 0$, and $\xi_k = O(\zeta_k)$ if there exists a positive constant c such that $\xi_k \leq c\zeta_k$ for any sufficiently large k.

In the rest of this paper, Section 2 presents background and preliminary techniques. Section 3 states convergence results of MEAL and iMEAL. Section 4 presents the results of LiMEAL. Section 5 includes main proofs. Section 6 provides sufficient conditions for certain boundedness assumptions in above results along with comparisons with the related work. Section 7 provides some numerical experiments to demonstrate the effectiveness of proposed methods. We conclude this paper in Section 8.

2 Background and Preliminaries

This paper uses extended-real-valued functions, for example, $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Define the domain of h as $\text{dom}(h) := \{x \in \mathbb{R}^n : h(x) < +\infty\}$ and its range as $\text{ran}(h) := \{y : y = h(x), \forall x \in \text{dom}(h)\}$. For each $x \in \text{dom}(h)$, the *Fréchet subdifferential* of h at x, written as $\widehat{\partial} h(x)$, is the set of vectors $v \in \mathbb{R}^n$ satisfying

$$\liminf_{u \neq x, u \to x} \frac{h(u) - h(x) - \langle v, u - x\rangle}{\|x - u\|} \geq 0.$$

When $x \notin \text{dom}(h)$, we define $\widehat{\partial} h(x) = \emptyset$. The *limiting-subdifferential* (or simply *subdifferential*) of h [46] at $x \in \text{dom}(h)$ is defined as

$$\partial h(x) := \{v \in \mathbb{R}^n : \exists x^t \to x,\ h(x^t) \to h(x),\ \widehat{\partial} h(x^t) \ni v^t \to v\}. \tag{10}$$

A necessary (but not sufficient) condition for $x \in \mathbb{R}^n$ to be a minimizer of h is $0 \in \partial h(x)$. A point that satisfies this inclusion is called *limiting-critical* or simply *critical*. The distance between a point x and a subset $\mathcal{S}$ of $\mathbb{R}^n$ is defined as $\text{dist}(x, \mathcal{S}) = \inf_u\{\|x - u\| : u \in \mathcal{S}\}$.

2.1 Moreau Envelope

Given a function $h : \mathbb{R}^n \to \mathbb{R}$, we define its *Moreau envelope* [47,56]:

$$\mathcal{M}_{\gamma,h}(z) = \min_x \left\{ h(x) + \frac{1}{2\gamma}\|x - z\|^2 \right\}, \tag{11}$$

where $\gamma > 0$ is a parameter. We define its associated proximity operator

$$\mathrm{Prox}_{\gamma,h}(z) = \operatorname*{argmin}_x \left\{ h(x) + \frac{1}{2\gamma}\|x - z\|^2 \right\}. \tag{12}$$

If h is ρ-weakly convex and $\gamma \in (0, \rho^{-1})$, then $\mathrm{Prox}_{\gamma,h}$ is monotone, single-valued, and Lipschitz, and $\mathcal{M}_{\gamma,h}$ is differentiable with

$$\nabla \mathcal{M}_{\gamma,h}(z) = \gamma^{-1}\left(z - \mathrm{Prox}_{\gamma,h}(z)\right) \in \partial h(\mathrm{Prox}_{\gamma,h}(z)); \tag{13}$$

see [56, Proposition 13.37]. From [30,31], we also have

$$\begin{aligned}
&\mathcal{M}_{\gamma,h}(\mathrm{Prox}_{\gamma,h}(z)) \le h(z),\\
&\|\mathrm{Prox}_{\gamma,h}(z) - z\| = \gamma\|\nabla \mathcal{M}_{\gamma,h}(z)\|,\\
&\mathrm{dist}(0, \partial h(\mathrm{Prox}_{\gamma,h}(z))) \le \|\nabla \mathcal{M}_{\gamma,h}(z)\|.
\end{aligned}$$

The first relation above presents Moreau envelope as a smooth lower approximation of h. By the second and third relations, small $\|\nabla \mathcal{M}_{\gamma,h}(z)\|$ implies that z is *near* its proximal point $\mathrm{Prox}_{\gamma,h}(z)$ and z is *nearly stationary* for h [28]. Therefore, $\|\nabla \mathcal{M}_{\gamma,h}(z)\|$ can be used as a *continuous stationarity measure*. Hence, replacing the augmented Lagrangian with its Moreau envelope not only generates a strongly convex subproblem but also yields a stationarity measure.

2.2 Implicit Regularity Properties

Let h be a proper, lower semicontinuous, ρ-weakly convex function. Given a $\gamma \in (0, \rho^{-1})$, we define the *generalized inverse mapping* $\mathrm{Prox}_{\gamma,h}^{-1}$ of $\mathrm{Prox}_{\gamma,h}$ as

$$\mathrm{Prox}_{\gamma,h}^{-1}(x) := \{w : \mathrm{Prox}_{\gamma,h}(w) = x\}, \quad \forall x \in \mathrm{ran}(\mathrm{Prox}_{\gamma,h}). \tag{14}$$

Based on (14), we introduce the following definitions, which impose implicitly some regularity properties including Lipschitz continuity and boundedness on subgradient of a weakly convex function.

Definition 1 Let h be a proper, lower semicontinuous and ρ-weakly convex function.

(a) We say h satisfies the **implicit Lipschitz subgradient** property if for any $\gamma \in (0, \rho^{-1})$, there exists a constant $L > 0$ (possibly depending on γ) such that for any $u, v \in \mathrm{ran}(\mathrm{Prox}_{\gamma,h})$,

$$\|\nabla \mathcal{M}_{\gamma,h}(w) - \nabla \mathcal{M}_{\gamma,h}(w')\| \le L\|u - v\|, \ \forall w \in \mathrm{Prox}_{\gamma,h}^{-1}(u), w' \in \mathrm{Prox}_{\gamma,h}^{-1}(v);$$

(b) We say h satisfies the **implicit bounded subgradient** property if for any $\gamma \in (0, \rho^{-1})$, there exists a constant $\hat{L} > 0$ (possibly depending on γ) such that for any $u \in \mathrm{ran}(\mathrm{Prox}_{\gamma,h})$,

$$\|\nabla \mathcal{M}_{\gamma,h}(w)\| \leq \hat{L}, \ \forall w \in \mathrm{Prox}_{\gamma,h}^{-1}(u).$$

Since $\nabla \mathcal{M}_{\gamma,h}(x) \in \partial h(\mathrm{Prox}_{\gamma,h}(x))$ for any $x \in \mathbb{R}^n$, we have $\nabla \mathcal{M}_{\gamma,h}(w) \in \partial h(u), \forall u \in \mathrm{ran}(\mathrm{Prox}_{\gamma,h})$ and $w \in \mathrm{Prox}_{\gamma,h}^{-1}(u)$. Hence, the *implicit Lipschitz subgradient* and *implicit bounded subgradient* imply, respectively, the *Lipschitz continuity* and *boundedness* of only the components of ∂h that are Moreau envelope gradients, but not those of all components of ∂h. When h is differentiable, *implicit Lipschitz subgradient* implies *Lipschitz gradient*. Nonsmooth and nonconvex functions like $|x^2 - 1|$, $|xy - 1|$, $x, y \in \mathbb{R}$, SCAD regularization and MCP regularization which appear in phase retrieval, blind deconvolution, robust principal component analysis and statistical learning [28,30,59], have *implicit Lipschitz subgradients* but not gradients. They have not been covered in the ALM literature. Having *implicit bounded subgradients* is weaker than having bounded ∂h, which is commonly assumed in the analysis of nonconvex algorithms (cf. [65,28,36]).

2.3 Kurdyka-Łojasiewicz Inequality

The Kurdyka-Łojasiewicz (KŁ) inequality [44,45,41,20,21] is a property that leads to global convergence of nonconvex algorithms in the literature (see, [8,61,22,59,63, 64]). The following definition of Kurdyka-Łojasiewicz property is adopted from [20].

Definition 2 A function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to have the Kurdyka-Łojasiewicz property at $x^* \in \mathrm{dom}(\partial h)$ if there exist a neighborhood $\mathcal{U}$ of x^*, a constant $\nu > 0$, and a continuous concave function $\varphi(s) = cs^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$ such that the Kurdyka-Łojasiewicz inequality holds: for all $x \in \mathcal{U} \cap \mathrm{dom}(\partial h)$ and $h(x^*) < h(x) < h(x^*) + \nu$,

$$\varphi'(h(x) - h(x^*)) \cdot \mathrm{dist}(0, \partial h(x)) \geq 1, \tag{15}$$

(we use the conventions: $0^0 = 1, \infty/\infty = 0/0 = 0$), where θ is called the KŁ exponent of h at x^*. Proper lower semicontinuous functions satisfying the KŁ inequality at every point of $\mathrm{dom}(\partial h)$ are called KŁ functions.

This property was firstly introduced by [45] on real analytic functions [40] for $\theta \in \left[\frac{1}{2}, 1\right)$, was then extended to functions defined on the o-minimal structure in [41], and was later extended to nonsmooth subanalytic functions in [20]. KŁ functions include real analytic functions [40], semialgebraic functions [19], tame functions defined in some o-minimal structures [41], continuous subanalytic functions [20], definable functions [21], locally strongly convex functions [61], as well as many deep-learning training models [63,64].

3 Convergence of MEAL

In this section, we state our convergence results of MEAL and iMEAL, whose proofs are postponed in Section 5.

3.1 Assumptions and Stationarity Measure

Assumption 1 *The set* $\mathcal{X} := \{x : Ax = b\}$ *is nonempty.*

Assumption 2 *The objective* f *in problem* (1) *satisfies:*

(a) f *is proper lower semicontinuous and* ρ*-weakly convex; and for any* $\gamma \in (0, \rho^{-1})$, *either*
(b) f *satisfies the* ***implicit Lipschitz subgradient*** *property with a constant* $L_f > 0$ *(possibly depending on* γ*); or,*
(c) f *satisfies the* ***implicit bounded subgradient*** *property with a constant* $\hat{L}_f > 0$ *(possibly depending on* γ*).*

We emphasize that we are not assuming commonly used hypotheses in the literature, such as: the strict complementarity condition used in [68], any rank assumption (such as $\mathrm{Im}(A) \subseteq \mathrm{Im}(\mathrm{B})$ when considering the two- (multi-)block case $Ax + By = 0$) used in [59], the linear independence constrained qualification (LICQ) used in [11,48] (implying the full-rank assumption in the linear constraint case). Assumption 2 is mild as discussed in Section 2.2 and satisfied by some important nonconvex nonsmooth functions that have not been considered in the literature on ALM.

According to (3) and the update (4) of MEAL, we have

$$\nabla \phi_{\beta_k}(z^k, \lambda^k) = \begin{pmatrix} (\eta\gamma)^{-1}(z^k - z^{k+1}) \\ \beta_k^{-1}(\lambda^{k+1} - \lambda^k) \end{pmatrix} \in \begin{pmatrix} \partial f(x^{k+1}) + A^T \lambda^{k+1} \\ Ax^{k+1} - b \end{pmatrix}. \tag{16}$$

Let

$$\xi_{\mathrm{meal}}^k := \min_{0 \le t \le k} \|\nabla \phi_{\beta_t}(z^t, \lambda^t)\|, \ \forall k \in \mathbb{N}. \tag{17}$$

Then according to (16), for accuracy $\varepsilon > 0$, requiring $\xi_{\mathrm{meal}}^k \le \varepsilon$ implies

$$\min_{0 \le t \le k} \mathrm{dist}\left\{0, \begin{pmatrix} \partial f(x^{t+1}) + A^T \lambda^{t+1} \\ Ax^{t+1} - b \end{pmatrix}\right\} \le \xi_{\mathrm{meal}}^k \le \varepsilon,$$

which implies that MEAL can achieve an ε-accurate first-order stationary point of problem (1) within k iterations. Thus, the defined ξ_{meal}^k in (17) can be used as an effective stationarity measure of MEAL.

Based on (17), given an ε, the iteration complexity T_ε used in this paper is defined as follows:

$$T_\varepsilon = \inf\left\{t \ge 1 : \|\nabla \phi_{\beta_t}(z^t, \lambda^t)\| \le \varepsilon\right\}, \tag{18}$$

which is stronger than the commonly used measure of iteration complexity in the literature (e.g., [68, 60]), that is,[3]

$$\hat{T}_\varepsilon = \inf\left\{t \geq 1 : \mathrm{dist}(0, \partial f(x^t) + A^T\lambda^t) \leq \epsilon \text{ and } \|Ax^t - b\| \leq \varepsilon\right\}$$

in the sense that $T_\varepsilon \geq \hat{T}_\varepsilon$. Thus, the stationarity measure defined in (16) is slightly stronger than that used in the literature [68, 60].

3.2 Convergence Theorems of MEAL

We present the quantities used to state the convergence results of MEAL. Let

$$\mathcal{P}_\beta(x, z, \lambda) = \mathcal{L}_\beta(x, \lambda) + \frac{1}{2\gamma}\|x - z\|^2, \tag{19}$$

for some $\beta, \gamma > 0$. Then according to (5), MEAL can be interpreted as a primal-dual update with respect to $\mathcal{P}_{\beta_k}(x, z, \lambda)$ at the k-th iteration, that is, updating x^{k+1}, z^{k+1}, and λ^{k+1} by minimization, gradient descent, and gradient ascent respectively.

Based on (19), we introduce the following Lyapunov functions for MEAL:

$$\mathcal{E}^k_{\mathrm{meal}} := \mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) + 2\alpha_k\|z^k - z^{k-1}\|^2, \ \forall k \geq 1, \tag{20}$$

associated with the *implicit Lipschitz subgradient* assumption and

$$\tilde{\mathcal{E}}^k_{\mathrm{meal}} := \mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) + 3\alpha_k\|z^k - z^{k-1}\|^2, \ \forall k \geq 1, \tag{21}$$

associated with the *implicit bounded subgradient* assumption, where

$$\alpha_k := \frac{\beta_k + \beta_{k+1} + \gamma\eta(1 - \eta/2)}{2c_{\gamma,A}\beta_k^2}, \ \forall k \in \mathbb{N}, \tag{22}$$

and $c_{\gamma,A} := \gamma^2\tilde{\sigma}_{\min}(A^TA)$. When β is fixed, we also fix

$$\alpha := \frac{2\beta + \gamma\eta(1 - \eta/2)}{2c_{\gamma,A}\beta^2}. \tag{23}$$

Theorem 1 (Iteration Complexity of MEAL) *Suppose that Assumptions 1 and 2(a) hold. Pick $\gamma \in (0, \rho^{-1})$ and $\eta \in (0, 2)$. Let $\{(x^k, z^k, \lambda^k)\}$ be a sequence generated by MEAL (5). The following claims hold:*

(a) Set β sufficiently large such that in (23), $\alpha < \min\left\{\frac{1-\gamma\rho}{4\gamma(1+\gamma L_f)^2}, \frac{1}{8\gamma}(\frac{2}{\eta} - 1)\right\}$. Under Assumption 2(b), if $\{\mathcal{E}^k_{\mathrm{meal}}\}$ is lower bounded, then $\xi^k_{\mathrm{meal}} = o(1/\sqrt{k})$ for ξ^k_{meal} in (17), i.e., $\sqrt{k}\cdot\xi^k_{\mathrm{meal}} \to 0$ as $k \to \infty$.

(b) Pick any $K \geq 1$. Set $\{\beta_k\}$ so that in (22), $\alpha_k \equiv \frac{\alpha^*}{K}$ for $\alpha^* = \min\left\{\frac{1-\rho\gamma}{6\gamma}, \frac{1}{12\gamma}\left(\frac{2}{\eta} - 1\right)\right\}$. Under Assumption 2(c), if $\{\tilde{\mathcal{E}}^k_{\mathrm{meal}}\}$ is lower bounded, then $\xi^K_{\mathrm{meal}} \leq \tilde{c}_1/\sqrt{K}$ for some constant $\tilde{c}_1 > 0$.

[3] When f is differentiable, $\mathrm{dist}(0, \partial f(x^t) + A^T\lambda^t)$ reduces to $\|\nabla f(x^t) + A^T\lambda^t\|$.

In Section 6.1, we provide sufficient conditions for the lower-boundedness assumptions. Let us interpret the theorem. To achieve an ε-accurate stationary point, the iteration complexity of MEAL is $o(\varepsilon^{-2})$ assuming the implicit Lipschitz subgradient property and $O(\varepsilon^{-2})$ assuming the implicit bounded subgradient property. Both iteration complexities are consistent with the existing results of $O(\varepsilon^{-2})$ in [38,36,60,67]. The established results of MEAL also hold for proximal ALM by setting $\eta = 1$. We note that it is not our goal to pursue any better complexity (e.g., using momentum) in this paper.

Remark 1 Let $\bar{\alpha} := \min\left\{\frac{1-\gamma\rho}{4\gamma(1+\gamma L_f)^2}, \frac{1}{8\gamma}(\frac{2}{\eta}-1)\right\}$. By (23), the requirement $0 < \alpha < \bar{\alpha}$ in Theorem 1(a) is met by setting

$$\beta > \frac{1+\sqrt{1+\eta(2-\eta)\gamma c_{\gamma,A}\bar{\alpha}}}{2c_{\gamma,A}\bar{\alpha}}. \tag{24}$$

Similarly, the assumption $\alpha_k = \frac{\alpha^*}{K}$ in Theorem 1(b) is met by setting

$$\beta_k = \frac{K\left(1+\sqrt{1+\eta(2-\eta)\gamma c_{\gamma,A}\alpha^*/K}\right)}{2c_{\gamma,A}\alpha^*},\ k = 1,\ldots,K. \tag{25}$$

Next, we establish global convergence (whole sequence convergence regardless of initial points) and its rate for MEAL under the KŁ inequality (Definition 2). Let $\hat{z}^k := z^{k-1}$, $y^k := (x^k, z^k, \lambda^k, \hat{z}^k)$, $\forall k \geq 1$, $y := (x, z, \lambda, \hat{z}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$, and

$$\mathcal{P}_{\text{meal}}(y) := \mathcal{P}_\beta(x, z, \lambda) + 3\alpha\|z - \hat{z}\|^2 \tag{26}$$

where α is defined in (23).

Proposition 1 (Global convergence and rate of MEAL) *Suppose that the assumptions required for Theorem 1(a) hold and that $\{(x^k, z^k, \lambda^k)\}$ generated by MEAL* (5) *is bounded. If $\mathcal{P}_{\text{meal}}$ satisfies the KŁ property at some point $y^* := (x^*, x^*, \lambda^*, x^*)$ with an exponent of $\theta \in [0, 1)$, where (x^*, λ^*) is a limit point of $\{(x^k, \lambda^k)\}$, then*

(a) the whole sequence $\{\hat{y}^k := (x^k, z^k, \lambda^k)\}$ converges to $\hat{y}^* := (x^*, x^*, \lambda^*)$; and
(b) the following rate-of-convergence results hold: (1) if $\theta = 0$, then $\{\hat{y}^k\}$ converges within a finite number of iterations; (2) if $\theta \in (0, \frac{1}{2}]$, then $\|\hat{y}^k - \hat{y}^*\| \leq c\tau^k$ for all $k \geq k_0$, for certain $k_0 > 0, c > 0, \tau \in (0, 1)$; and (3) if $\theta \in (\frac{1}{2}, 1)$, then $\|\hat{y}^k - \hat{y}^*\| \leq ck^{-\frac{1-\theta}{2\theta-1}}$ for all $k \geq k_0$, for certain $k_0 > 0, c > 0$.

From Proposition 1, the KŁ property of $\mathcal{P}_{\text{meal}}$ defined in (26) plays a central role in the establishment of global convergence of MEAL, and its KŁ exponent determines the convergence speed of MEAL, particularly, the exponent $\theta = 1/2$ is desired due to it implies the linear rate of convergence. In the following, we provide some preliminary results on these, which can be yielded by the existing results ([57, page 43], [20, Theorem 3.1], [63, Lemma 5], [42, Theorem 3.6 and Corollary 5.2]).

Proposition 2 *The following claims hold:*

(a) *If f is subanalytic with a closed domain and continuous on its domain, then $\mathcal{P}_{\text{meal}}$ defined in* (26) *is a KŁ function;*
(b) *If $\mathcal{L}_\beta(x,\lambda)$ defined in* (2) *has the KŁ property at some point (x^*,λ^*) with exponent $\theta \in [1/2,1)$, then $\mathcal{P}_{\text{meal}}$ has the KŁ property at (x^*,x^*,λ^*,x^*) with exponent θ;*
(c) *If f has the following form (see, [42, Eq. (35)]):*

$$f(x) = \min_{1\le i\le r}\left\{\frac{1}{2}x^T M_i x + u_i^T x + c_i + P_i(x)\right\}, \tag{27}$$

where P_i are proper closed polyhedral functions, M_i are symmetric matrices of size n, $u_i \in \mathbb{R}^n$ and $c_i \in \mathbb{R}$ for $i=1,\dots,r$, then $\mathcal{L}_\beta$ is a KŁ function with an exponent of $1/2$.

Claim (a) can be argued as follows. The terms in $\mathcal{P}_{\text{meal}}$ besides f are polynomial functions, which are both real analytic and semialgebraic [19]. Since f is subanalytic with a closed domain and continuous on its domain, by [63, Lemma 5], $\mathcal{P}_{\text{meal}}$ is also subanalytic with a closed domain and continuous on its domain. By [20, Theorem 3.1], $\mathcal{P}_{\text{meal}}$ is a KŁ function. Claim (b) can be verified by applying [42, Theorem 3.6] to $\mathcal{P}_{\text{meal}}$. Claim (c) can be argued as follows. It can be shown that the defined class of functions f (27) are weakly convex with a modulus $\rho = 2\max_{1\le i\le r}\|M_i\|$. As pointed out in [42, Sec. 5.2], the class of functions defined in (27) covers many nonconvex functions such as the SCAD regularization [32] and MCP regularization [66] in statistical learning, and $-|x^2-1|$ and $-|xy-1|$ $(x,y\in\mathbb{R})$ in phase retrieval and blind deconvolution [28]. Notice that $\mathcal{L}_\beta(x,\lambda) = \frac{\beta}{2}\|Ax+\beta^{-1}\lambda-b\|^2 + (f(x) - \frac{1}{2\beta}\|\lambda\|^2)$, which falls into the form of regularized least square, then according to [42, Corollary 5.2], $\mathcal{L}_\beta$ is a KŁ function with an exponent of $1/2$. More results on the KŁ functions with exponent $1/2$ can be found in the recent literature [42,62] and references therein.

3.3 Convergence of iMEAL

When considering iMEAL, the Lyapunov functions need to be slightly modified into

$$\mathcal{E}^k_{\text{imeal}} := \mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) + 3\alpha_k\|z^k - z^{k-1}\|^2,\ \forall k\ge 1, \tag{28}$$

associated with the implicit Lipschitz subgradient assumption, and

$$\tilde{\mathcal{E}}^k_{\text{imeal}} := \mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) + 4\alpha_k\|z^k - z^{k-1}\|^2,\ \forall k\ge 1, \tag{29}$$

associated with the implicit bounded subgradient assumption, where α_k is defined in (22).

Theorem 2 **(Iteration Complexity of iMEAL)** *Let Assumptions 1 and 2(a) hold, $\gamma\in(0,\rho^{-1})$, and $\eta\in(0,2)$. Let $\{(x^k,z^k,\lambda^k)\}$ be a sequence generated by iMEAL* (7) *with $\sum_{k=0}^\infty \epsilon_k^2 < \infty$. The following claims hold:*

(a) *Set β sufficiently large such that in* (23), $\alpha < \min\left\{\frac{1-\gamma\rho}{6\gamma(1+\gamma L_f)^2}, \frac{1}{12\gamma}(\frac{2}{\eta}-1)\right\}$. *Under Assumption 2(b), if $\{\mathcal{E}^k_{\text{imeal}}\}$ is lower bounded, then $\xi^k_{\text{meal}} = o(1/\sqrt{k})$ (cf.* (17)*).*

(b) Pick $K \geq 1$. Set $\{\beta_k\}$ such that in (22), $\alpha_k \equiv \frac{\hat{\alpha}^*}{K}$ for $\hat{\alpha}^* := \min\left\{\frac{1-\rho\gamma}{8\gamma}, \frac{1}{16\gamma}(\frac{2}{\eta}-1)\right\}$. Under Assumption 2(c), if $\{\tilde{\mathcal{E}}^k_{\mathrm{imeal}}\}$ is lower bounded, then $\xi^K_{\mathrm{meal}} \leq \tilde{c}_2/\sqrt{K}$ for some constant $\tilde{c}_2 > 0$.

By Theorem 2, the iteration complexity of iMEAL is the same as that of MEAL and also consistent with that of inexact proximal ALM [60] (when the stationary accuracy ϵ_k is square summable). Moreover, if the condition on ϵ_k is strengthened to be $\sum_{k=0}^{\infty}\epsilon_k < +\infty$ as required in the literature [55,59], then following a proof similar for Proposition 1, global convergence and similar rates of MEAL also hold for iMEAL under the assumptions required for Theorem 2(a) and the KŁ property.

Remark 2 (Extension to multiblock case) We generalize MEAL to a class of linearly constrained optimization problems with multi-block variables of the following forms:

$$\begin{array}{ll} \mathrm{minimize}_{x_1,\ldots,x_p} & f(x_1,\ldots,x_p) := \sum_{i=1}^p r_i(x_i) \\ \text{subject to} & \sum_{i=1}^p A_i x_i = b, \end{array} \tag{30}$$

where $x_i \in \mathbb{R}^{n_i}$ with $n = \sum_{i=1}^p n_i$, $A_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$, and $b \in \mathbb{R}^m$, $r_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is weakly convex with a modulus $\rho > 0$ and its proximity operator is assumed to be easy to calculate (generally admitting a closed-form solution)[4]. Many kinds of problems can be formulated into the form (30) such as the regularized statistical learning [59, Sec. 5.A] and sparse regularized phase retrieval [43] shown later. Let $x := [x_1; x_2; \ldots; x_p]$, $x_{<i} := [x_1; \ldots; x_{i-1}]$ and $x_{>i} := [x_{i+1}; \ldots; x_p]$ for $i = 1, \ldots, p$ and for the notational convention, let $x_{<1} = \emptyset$ and $x_{>p} = \emptyset$. Similarly, we let $A_{<i} = [A_1, \ldots, A_{i-1}]$, $A_{>i} = [A_{i+1}, \ldots, A_p]$, $A_{<i}x_{<i} := \sum_{j=1}^{i-1} A_j x_j$ and $A_{>i}x_{>i} = \sum_{j=i+1}^{p} A_j x_j$ for $i = 1, \ldots, p$, and we can also define $x_{\leq i}, x_{\geq i}, A_{\leq i}, A_{\geq i}, A_{\leq i}x_{\leq i}$ and $A_{\geq i}x_{\geq i}$ similarly. Similar to MEAL (5), we adopt the alternating direction scheme (also called block coordinate descent scheme) [61] to deal with each block of variable. Specifically, at the k-th iteration, for $i = 1, \ldots, p$, let $x^k_{\setminus i} := [x^{k+1}_{<i}; x^k_{>i}]$, $A_{\setminus i}x^k_{\setminus i} := A_{<i}x^{k+1}_{<i} + A_{>i}x^k_{>i}$, $x^k_{[i]} := [x^{k+1}_{<i}; x^k_i; x^k_{>i}]$ and

$$\mathcal{L}_{\beta_k, r_i}(x_i, \lambda) := r_i(x_i) + \langle \lambda, A_i x_i + A_{\setminus i}x^k_{\setminus i} - b\rangle + \frac{\beta_k}{2}\|A_i x_i + A_{\setminus i}x^k_{\setminus i} - b\|^2,$$

then the iterate of *Moreau Envelope Alternating Direction* method (dubbed *MEAD*) for (30) can be described as follows: given an initialization (z^0, λ^0), $\gamma > 0$, $\eta > 0$ and $\{\beta_k\}$, for $k = 0, 1, \ldots,$

$$(\text{MEAD}) \quad \begin{cases} \text{for } i = 1, \ldots, p, \ x_i^{k+1} = \mathrm{Prox}_{\gamma, \mathcal{L}_{\beta_k, r_i}(\cdot, \lambda^k)}(z_i^k), \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b). \end{cases} \tag{31}$$

Convergence results of MEAL shall also hold for MEAD under similar assumptions.

[4] Without loss of generality, we use a uniform weakly convex modulus for all functions r_i's for the simplicity

4 Convergence of LiMEAL for Composite Objective

In this section, we present the convergence results of LiMEAL (9) for the constrained problem with a composite objective (8), of which proofs are presented in Section 5 later. Similar to Assumption 2, we make the following assumptions.

Assumption 3 *The objective $f(x) = h(x) + g(x)$ in problem* (8) *satisfies:*

(a) h is differentiable and ∇h is Lipschitz continuous with a constant $L_h > 0$;
(b) g is proper lower-semicontinuous and ρ_g-weakly convex; and ***either***
(c) g has the ***implicit Lipschitz subgradient*** *property with a constant $L_g > 0$;* ***or***
(d) g has the ***implicit bounded subgradient*** *property with a constant $\hat{L}_g > 0$.*

In (c) and (d), L_g and $\hat{L}_g$ may depend on γ.

By the update (9) of LiMEAL, some simple derivations show that

$$x^{k+1} = \mathrm{Prox}_{\gamma,g}(z^k - \gamma(\nabla h(x^k) + A^T \lambda^{k+1})) \tag{32}$$

and

$$g^k_{\text{limeal}} := \begin{pmatrix} \gamma^{-1}(z^k - x^{k+1}) + (\nabla h(x^{k+1}) - \nabla h(x^k)) \\ \beta_k^{-1}(\lambda^{k+1} - \lambda^k) \end{pmatrix} \in \begin{pmatrix} \partial f(x^{k+1}) + A^T \lambda^{k+1} \\ Ax^{k+1} - b \end{pmatrix}. \tag{33}$$

Actually, the term $\gamma^{-1}(z^k - x^{k+1})$ represents some *prox-gradient sequence* frequently used in the analysis of algorithms for the unconstrained composite optimization (e.g., [28]). Thus, let

$$\xi^k_{\text{limeal}} := \min_{0 \le t \le k} \|g^t_{\text{limeal}}\|, \ \forall k \in \mathbb{N}, \tag{34}$$

which can be taken as an effective stationarity measure of LiMEAL for problem (8).

In the following, we present the iteration complexity of LiMEAL for problem (8). Since the prox-linear scheme is adopted in the update of x^{k+1} in LiMEAL as described in (9), thus, the proximal term (i.e., $\|x^k - x^{k-1}\|^2$) should be generally included in the associated Lyapunov functions of LiMEAL, shown as follows:

$$\mathcal{E}^k_{\text{limeal}} := \mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) + 3\alpha_k(\gamma^2 L_h^2 \|x^k - x^{k-1}\|^2 + \|z^k - z^{k-1}\|^2) \tag{35}$$

associated with the *implicit Lipschitz gradient* assumption, and

$$\tilde{\mathcal{E}}^k_{\text{limeal}} := \mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) + 4\alpha_k(\gamma^2 L_h^2 \|x^k - x^{k-1}\|^2 + \|z^k - z^{k-1}\|^2), \tag{36}$$

associated with the *implicit bounded subgradient* assumption, where α_k is defined in (22).

The iteration complexity of MEAL can be similarly generalized to LiMEAL as follows.

Theorem 3 (Iteration Complexity of LiMEAL) *Take Assumptions 1 and 3(a)-(b). Pick $\eta \in (0,2)$ and $0 < \gamma < \frac{2}{(\rho_g + L_h)\left(1 + \sqrt{1 + \frac{2(2-\eta)\eta L_h^2}{(\rho_g+L_h)^2}}\right)}$. Let $\{(x^k, z^k, \lambda^k)\}$ be a sequence generated by LiMEAL* (9)*. The following claims hold:*

(a) *Set β sufficiently large such that* $\alpha < \min\left\{\frac{1}{12\gamma}(\frac{2}{\eta}-1), \frac{1-\gamma(\rho_g+L_h)-\eta(1-\eta/2)\gamma^2L_h^2}{6\gamma\left((1+\gamma L_g)^2+\gamma^2L_h^2\right)}\right\}$.
Under Assumption 3(c), if $\{\mathcal{E}_{\rm limeal}^k\}$ *is lower bounded, then* $\xi_{\rm limeal}^k = o(1/\sqrt{k})$.

(b) *Pick* $K \geq 1$. *Set* $\{\beta_k\}$ *such that* $\alpha_k \equiv \frac{\tilde{\alpha}^*}{K}$ *for* $\tilde{\alpha}^* = \min\left\{\frac{1-\gamma\left(\rho_g+L_h\right)-\eta(1-\eta/2)\gamma^2L_h^2)}{8\gamma(1+\gamma^2L_h^2)}, \frac{1}{16\gamma}\left(\frac{2}{\eta}-1\right)\right\}$.
Under Assumption 3(d), if $\{\tilde{\mathcal{E}}_{\rm limeal}^k\}$ *is lower bounded, then* $\xi_{\rm limeal}^K \leq \tilde{c}_3/\sqrt{K}$ *for some constant* $\tilde{c}_3 > 0$.

Similar to the discussions following Theorem 1, to yield an ε-accurate first-order stationary point, the iteration complexity of LiMEAL is $o(\varepsilon^{-2})$ under the *implicit Lipschitz subgradient* assumption and $O(\varepsilon^{-2})$ under the *implicit bounded subgradient* assumption, as demonstrated by Theorem 3. The conditions on β and β_k in these two cases can be derived similarly to (24) and (25), respectively.

In the following, we establish the global convergence and rate of LiMEAL under assumptions required for Theorem 3(a) and the KŁ property. Specifically, let $\hat{x}^k := x^{k-1}$, $\hat{z}^k := z^{k-1}$, $y^k := (x^k, z^k, \lambda^k, \hat{x}^k, \hat{z}^k)$, $\forall k \geq 1$, $y := (x, z, \lambda, \hat{x}, \hat{z}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n$, and

$$\mathcal{P}_{\rm limeal}(y) := \mathcal{P}_\beta(x,z,\lambda) + 4\alpha\left(\|z-\hat{z}\|^2 + \gamma^2L_h^2\|x-\hat{x}\|^2\right). \tag{37}$$

Proposition 3 (Global convergence and rate of LiMEAL) *Suppose that Assumptions 1 and 3(a)-(c) hold and that the sequence* $\{(x^k, z^k, \lambda^k)\}$ *generated by LiMEAL (9) is bounded. If* $\gamma \in (0, \frac{1}{\rho_g+L_h})$, $\eta \in (0,2)$, $0 < \alpha < \min\left\{\frac{1}{8\gamma}\left(\frac{2}{\eta}-1\right), \frac{1-\gamma(\rho_g+L_h)}{8\gamma\left((1+\gamma L_g)^2+\gamma^2L_h^2\right)}\right\}$, *and* $\mathcal{P}_{\rm limeal}$ *satisfies the KŁ property at some point* $y^* := (x^*, x^*, \lambda^*, x^*, x^*)$ *with an exponent of* $\theta \in [0,1)$, *where* (x^*, λ^*) *is a limit point of* $\{(x^k, \lambda^k)\}$, *then*

(a) *the whole sequence* $\{\hat{y}^k := (x^k, z^k, \lambda^k)\}$ *converges to* $\hat{y}^* := (x^*, x^*, \lambda^*)$; *and*
(b) *all the rate of convergence results in Proposition 1(b) also hold for LiMEAL.*

Remark 3 The established results in this section is more general than those in [68] and done under weaker assumptions on h and for more general class of g. Specifically, as discussed in Section 1.2, the algorithm studied in [68] is a prox-linear version of LiMEAL with g being an indicator function of a box constraint set. In [68], global convergence and a linear rate of proximal inexact ALM were proved for quadratic programming, where that the augmented Lagrangian satisfies the KŁ inequality with exponent $1/2$. Besides, the strict complementarity condition required in [68] is also removed in this paper for LiMEAL.

Remark 4 (Extension to multiblock composite objective) Similar to the analysis in Remark 2, we generalize LiMEAL to a class of linearly constrained optimization problems with multi-block compositions of the following forms:

$$\begin{array}{ll} \text{minimize}_{x_1,\ldots,x_p} & f(x_1,\ldots,x_p) := h(x_1,\ldots,x_p) + \sum_{i=1}^p r_i(x_i) \\ \text{subject to} & \sum_{i=1}^p A_i x_i = b, \end{array} \tag{38}$$

where $x_i \in \mathbb{R}^{n_i}$ with $n = \sum_{i=1}^p n_i$, $A_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$, and $b \in \mathbb{R}^m$, $r_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is weakly convex with a modulus $\rho > 0$ and its proximity operator is assumed to be easy to calculate, $h : \mathbb{R}^n \to \mathbb{R}$ is differentiable with respect to each block x_i and has L_h-Lipschitz continuous gradient in the sense that $\|\nabla_{x_i} h(x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_p) - \nabla_{x_i} h(x_1, \ldots, x_{i-1}, v, x_{i+1}, \ldots, x_p)\| \leq L_h \|u - v\|$, $\forall u, v \in \mathbb{R}^{n_i}$. Similar to LiMEAL (9), we adopt the block coordinate descent and prox-linear schemes [61] to deal with the smooth function h, while keep the nonsmooth functions r_i's. Specifically, at the k-th iteration, for $i = 1, \ldots, p$, let

$$f_i^k(x_i) := h(x_{[i]}^k) + \langle \nabla_{x_i} h(x_{[i]}^k), x_i - x_i^k \rangle + r_i(x_i) + \sum_{j<i} r_j(x_j^{k+1}) + \sum_{j>i} r_j(x_j^k) \quad (39)$$

and

$$\mathcal{L}_{\beta_k, f_i^k}(x_i, \lambda) := f_i^k(x_i) + \langle \lambda, A_i x_i + A_{\backslash i} x_{\backslash i}^k - b \rangle + \frac{\beta_k}{2} \|A_i x_i + A_{\backslash i} x_{\backslash i}^k - b\|^2,$$

(here we use the same notations as in Remark 2), then the iterate of *multi-block linearized MEAL* (mLiMEAL) for (38) can be described as follows: given an initialization (z^0, λ^0), $\gamma > 0$, $\eta > 0$ and $\{\beta_k\}$, for $k = 0, 1, \ldots,$

$$(\text{mLiMEAL}) \quad \begin{cases} \text{for } i = 1, \ldots, p, \ x_i^{k+1} = \mathrm{Prox}_{\gamma, \mathcal{L}_{\beta_k, f_i^k}(\cdot, \lambda^k)}(z_i^k), \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b). \end{cases} \quad (40)$$

Convergence results of LiMEAL shall also hold for mLiMEAL under similar assumptions.

5 Main Proofs

In this section, we first prove some lemmas and then present the proofs of our main convergence results.

5.1 Preliminary Lemmas

5.1.1 Lemmas on Iteration Complexity and Global Convergence

The first lemma concerns the convergence speed of a nonenegative sequence $\{\xi_k\}$ satisfying the following relation

$$\tilde{\eta} \xi_k^2 \leq (\mathcal{E}_k - \mathcal{E}_{k+1}) + \tilde{\epsilon}_k^2, \ \forall k \in \mathbb{N}, \quad (41)$$

where $\tilde{\eta} > 0$, $\{\mathcal{E}_k\}$ and $\{\tilde{\epsilon}_k\}$ are two nonnegative sequences, and $\sum_{k=1}^{\infty} \tilde{\epsilon}_k^2 < +\infty$.

Lemma 1 *For any sequence* $\{\xi_k\}$ *satisfying* (41), $\tilde{\xi}_k := \min_{1 \leq t \leq k} \xi_t = o(1/\sqrt{k})$.

Proof Summing (41) over k from 1 to K and letting $K \to +\infty$ yields

$$\sum_{k=1}^{\infty} \xi_k^2 \le \tilde{\eta}^{-1} \left(\mathcal{E}_1 + \sum_{k=1}^{\infty} \tilde{\epsilon}_k^2 \right) < +\infty,$$

which implies the desired convergence speed by $\frac{k}{2}\tilde{\xi}_k^2 \le \sum_{\frac{k}{2} \le j \le k} \xi_j^2 \to 0$ as $k \to \infty$, as proved in [29, Lemma 1.1].

Then we provide a lemma to show the convergence speed of a nonenegative sequence $\{\xi_k\}$ satisfying the following relation instead of (41)

$$\tilde{\eta}\xi_k^2 \le (\mathcal{E}_k - \mathcal{E}_{k+1}) + \tilde{\epsilon}_k^2 + \alpha_k \tilde{L}, \ \forall k \in \mathbb{N}, \tag{42}$$

where $\tilde{\eta} > 0$, $\tilde{L} > 0$, $\{\mathcal{E}_k\}$, $\{\alpha_k\}$ and $\{\tilde{\epsilon}_k\}$ are nonnegative sequences, and $\sum_{k=1}^{\infty} \tilde{\epsilon}_k^2 < +\infty$.

Lemma 2 *Pick* $K \ge 1$. *Let* $\{\xi_k\}$ *be a nonnegative sequence satisfying* (42). *Set* $\alpha_k \equiv \frac{\tilde{\alpha}}{K}$ *for some* $\tilde{\alpha} > 0$. *Then* $\tilde{\xi}_K := \min_{1\le k\le K} \xi_k \le \tilde{c}/\sqrt{K}$ *for some constant* $\tilde{c} > 0$.

Proof Summing (42) over k from 1 to K yields

$$\frac{1}{K}\sum_{k=1}^{K} \xi_k^2 \le \frac{\mathcal{E}_1 + \sum_{k=1}^{K} \tilde{\epsilon}_k^2 + \tilde{L}\sum_{k=1}^{K} \alpha_k}{K\tilde{\eta}}.$$

The result follows from the choice of $\tilde{\alpha}_k$ and $\sum_{k=1}^{\infty} \tilde{\epsilon}_k^2 < +\infty$.

In both Lemmas 1 and 2, the nonnegative assumption on the sequence $\{\mathcal{E}_k\}$ can be relaxed to its lower boundedness.

The following lemma presents the global convergence and rate of a sequence generated by some algorithm for the nonconvex optimization problem, based on the Kurdyka-Łojasiewicz inequality, where the global convergence result is from [8, Theorem 2.9] while the rate results are from [7, Theorem 5].

Lemma 3 (Existing global convergence and rate) *Let* $\mathcal{L}$ *be a proper, lower semicontinuous function, and* $\{u^k\}$ *be a sequence that satisfies the following three conditions:*

(P1) (Sufficient decrease condition) there exists a constant $a_1 > 0$ *such that* $\mathcal{L}(u^{k+1}) + a_1\|u^{k+1} - u^k\|^2 \le \mathcal{L}(u^k)$, $\forall k \in \mathbb{N}$;
(P2) (Bounded subgradient condition) for each $k \in \mathbb{N}$, *there exists* $v^{k+1} \in \partial\mathcal{L}(u^{k+1})$ *such that* $\|v^{k+1}\| \le a_2\|u^{k+1} - u^k\|$ *for some constant* $a_2 > 0$;
(P3) (Continuity condition) there exist a subsequence $\{u^{k_j}\}$ *and* $\tilde{u}$ *such that* $u^{k_j} \to \tilde{u}$ *and* $\mathcal{L}(u^{k_j}) \to \mathcal{L}(\tilde{u})$ *as* $j \to \infty$.

If $\mathcal{L}$ *satisfies the KŁ inequality at* $\tilde{u}$ *with an exponent of* θ, *then*

(1) $\{u^k\}$ *converges to* $\tilde{u}$; *and*
(2) depending on θ, *(i) if* $\theta = 0$, *then* $\{u^k\}$ *converges within a finite number of iterations; (ii) if* $\theta \in (0, \frac{1}{2}]$, *then* $\|u^k - \tilde{u}\| \le c\tau^k$ *for all* $k \ge k_0$, *for certain* $k_0 > 0, c > 0, \tau \in (0,1)$; *and (iii) if* $\theta \in (\frac{1}{2}, 1)$, *then* $\|u^k - \tilde{u}\| \le ck^{-\frac{1-\theta}{2\theta-1}}$ *for all* $k \ge k_0$, *for certain* $k_0 > 0, c > 0$.

5.1.2 Lemmas on controlling dual ascent by primal descent

In the following, we establish several lemmas to show that the dual ascent quantities of proposed algorithms can be controlled by the primal descent quantities.

Lemma 4 (MEAL: controlling dual by primal) *Let $\{(x^k, z^k, \lambda^k)\}$ be a sequence generated by MEAL* (5). *Take Assumptions 1 and 2(a) and $\gamma \in (0, \rho^{-1})$.*

(a) Under Assumption 2(b), for any $k \geq 1$,

$$\|A^T(\lambda^{k+1} - \lambda^k)\| \leq (L_f + \gamma^{-1})\|x^{k+1} - x^k\| + \gamma^{-1}\|z^k - z^{k-1}\|, \tag{43}$$

$$\|\lambda^{k+1} - \lambda^k\|^2 \leq 2c_{\gamma,A}^{-1}\left[(\gamma L_f + 1)^2\|x^{k+1} - x^k\|^2 + \|z^k - z^{k-1}\|^2\right], \tag{44}$$

where $c_{\gamma,A} = \gamma^2\tilde{\sigma}_{\min}(A^TA)$.

(b) Under Assumption 2(c), for any $k \geq 1$,

$$\|\lambda^{k+1} - \lambda^k\|^2 \leq 3c_{\gamma,A}^{-1}\left[4\gamma^2\hat{L}_f^2 + \|x^{k+1} - x^k\|^2 + \|z^k - z^{k-1}\|^2\right]. \tag{45}$$

Proof The update (5) of x^{k+1} implies

$$x^{k+1} = \operatorname*{argmin}_x \left\{ f(x) + \langle \lambda^k, Ax - b\rangle + \frac{\beta_k}{2}\|Ax - b\|^2 + \frac{1}{2\gamma}\|x - z^k\|^2 \right\}.$$

Its optimality condition and the update (5) of λ^{k+1} in MEAL together give us

$$0 \in \partial\left(f + \frac{1}{2\gamma}\|\cdot - (z^k - \gamma A^T\lambda^{k+1})\|^2\right)(x^{k+1}). \tag{46}$$

Let $w^{k+1} := z^k - \gamma A^T\lambda^{k+1}$, $\forall k \in \mathbb{N}$. The above inclusion implies

$$x^{k+1} = \operatorname{Prox}_{\gamma,f}(w^{k+1}), \tag{47}$$

and thus by (13),

$$A^T\lambda^{k+1} = -\nabla\mathcal{M}_{\gamma,f}(w^{k+1}) - \gamma^{-1}(x^{k+1} - z^k), \tag{48}$$

which further implies

$$\|A^T(\lambda^{k+1} - \lambda^k)\| = \|(\nabla\mathcal{M}_{\gamma,f}(w^{k+1}) - \nabla\mathcal{M}_{\gamma,f}(w^k)) + \gamma^{-1}(x^{k+1} - x^k) - \gamma^{-1}(z^k - z^{k-1})\|.$$

(a) With Assumption 2(b), the above equality yields

$$\|A^T(\lambda^{k+1} - \lambda^k)\| \leq (L_f + \gamma^{-1})\|x^{k+1} - x^k\| + \gamma^{-1}\|z^k - z^{k-1}\|,$$

which leads to (43). By Assumption 1 and the relation $\lambda^{k+1} - \lambda^k = \beta_k(Ax^{k+1} - b)$, $(\lambda^{k+1} - \lambda^k) \in \operatorname{Im}(A)$. Thus, from the above inequality, we deduce

$$\|\lambda^{k+1} - \lambda^k\| \leq \tilde{\sigma}_{\min}^{-1/2}(A^TA)\left[(L_f + \gamma^{-1})\|x^{k+1} - x^k\| + \gamma^{-1}\|z^k - z^{k-1}\|\right],$$

and, further by $(u + v)^2 \leq 2(u^2 + v^2)$ for any $u, v \in \mathbb{R}$,

$$\|\lambda^{k+1} - \lambda^k\|^2 \leq 2\tilde{\sigma}_{\min}^{-1}(A^TA)\left[(L_f + \gamma^{-1})^2\|x^{k+1} - x^k\|^2 + \gamma^{-2}\|z^k - z^{k-1}\|^2\right].$$

(b) From Assumption 2(c), we have

$$\|A^T(\lambda^{k+1}-\lambda^k)\| \le 2\hat{L}_f + \gamma^{-1}(\|x^{k+1}-x^k\| + \|z^k - z^{k-1}\|),$$

which implies

$$\|\lambda^{k+1}-\lambda^k\| \le \tilde{\sigma}_{\min}^{-1/2}(A^TA)\left[2\hat{L}_f + \gamma^{-1}(\|x^{k+1}-x^k\| + \|z^k - z^{k-1}\|)\right],$$

and further by $(a+c+d)^2 \le 3(a^2+c^2+d^2)$ for any $a,c,d \in \mathbb{R}$,

$$\|\lambda^{k+1}-\lambda^k\|^2 \le 3\tilde{\sigma}_{\min}^{-1}(A^TA)\left[4\hat{L}_f^2 + \gamma^{-2}(\|x^{k+1}-x^k\|^2 + \|z^k - z^{k-1}\|^2)\right].$$

The similar lemma also holds for iMEAL shown as follows.

Lemma 5 (iMEAL: dual controlled by primal) *Let* (x^k, z^k, λ^k) *be a sequence generated by iMEAL* (7)*. Take Assumptions 1 and 2(a) hold, and* $\gamma \in (0, \rho^{-1})$.

(a) Under Assumption 2(b), for any $k \ge 1$,

$$\|\lambda^{k+1}-\lambda^k\|^2 \le 3c_{\gamma,A}^{-1}\left[(\gamma L_f + 1)^2\|x^{k+1}-x^k\|^2 + \|z^k - z^{k-1}\|^2 + \gamma^2(\epsilon_k + \epsilon_{k-1})^2\right].$$

(b) Under Assumption 2(c), for any $k \ge 1$,

$$\|\lambda^{k+1}-\lambda^k\|^2 \le 4c_{\gamma,A}^{-1}\left[4\gamma^2\hat{L}_f^2 + \|x^{k+1}-x^k\|^2 + \|z^k - z^{k-1}\|^2 + \gamma^2(\epsilon_k + \epsilon_{k-1})^2\right].$$

Proof The proof is similar to that of Lemma 4, but with (46) being replaced by

$$0 \in \partial\left(f + \frac{1}{2\gamma}\left\|\cdot - \left(z^k - \gamma(A^T\lambda^{k+1} - s^k)\right)\right\|^2\right)(x^{k+1}),$$

and thus $w^{k+1} := z^k - \gamma(A^T\lambda^{k+1} - s^k)$.

Lemma 6 (LiMEAL: controlling dual by primal) *Let* $\{(x^k, z^k, \lambda^k)\}$ *is a sequence generated by LiMEAL* (9)*. Take Assumptions 1 and 3(a)-(b), and* $\gamma \in (0, \rho_g^{-1})$.

(a) Under Assumption 3(c), for any $k \ge 1$,

$$\begin{aligned} &\|A^T(\lambda^{k+1}-\lambda^k)\| &(49)\\ &\le (L_g + \gamma^{-1})\|x^{k+1}-x^k\| + L_h\|x^k - x^{k-1}\| + \gamma^{-1}\|z^k - z^{k-1}\|, \\ &\|\lambda^{k+1}-\lambda^k\|^2 &(50)\\ &\le 3c_{\gamma,A}^{-1}\left[(\gamma L_g + 1)^2\|x^{k+1}-x^k\|^2 + \gamma^2 L_h^2\|x^k - x^{k-1}\|^2 + \|z^k - z^{k-1}\|^2\right]. \end{aligned}$$

(b) Under Assumption 3(d), for any $k \ge 1$,

$$\begin{aligned} &\|\lambda^{k+1}-\lambda^k\|^2 &(51)\\ &\le 4c_{\gamma,A}^{-1}\left[4\gamma^2\hat{L}_g^2 + \|x^{k+1}-x^k\|^2 + \gamma^2 L_h^2\|x^k - x^{k-1}\|^2 + \|z^k - z^{k-1}\|^2\right]. \end{aligned}$$

Proof The proof is also similar to that of Lemma 4, but (46) needs to be modified to

$$0 \in \partial\left(g + \frac{1}{2\gamma}\| \cdot - \left(z^k - \gamma(A^T\lambda^{k+1} + \nabla h(x^k))\right)\|^2\right)(x^{k+1}),$$

and thus $w^{k+1} := z^k - \gamma(A^T\lambda^{k+1} + \nabla h(x^k)))$.

5.1.3 Lemmas on One-step Progress

Here, we provide several lemmas to characterize the progress achieved by a single iterate of the proposed algorithms.

Lemma 7 (MEAL: one-step progress) *Let* $\{(x^k, z^k, \lambda^k)\}$ *be a sequence generated by MEAL* (4). *Take Assumption 2(a),* $\gamma \in (0, \rho^{-1})$*, and* $\eta \in (0,2)$*. Then for any* $k \in \mathbb{N}$*,*

$$\mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1}, z^{k+1}, \lambda^{k+1}) \geq \frac{(1-\gamma\rho)}{2\gamma}\|x^{k+1} - x^k\|^2 \tag{52}$$
$$+ \frac{1}{4\gamma}(\frac{2}{\eta} - 1)\|z^{k+1} - z^k\|^2 + \frac{1}{4}\gamma\eta(2-\eta)\|\nabla\phi_{\beta_k}(z^k, \lambda^k)\|^2 - \alpha_k c_{\gamma,A}\|\lambda^{k+1} - \lambda^k\|^2,$$

where α_k *is presented in* (22) *and* $c_{\gamma,A} = \gamma^2\tilde{\sigma}_{\min}(A^T A)$.

Proof By the update (5) of x^{k+1} in MEAL, x^{k+1} is updated via minimizing a strongly convex function $\mathcal{P}_{\beta_k}(x, z^k, \lambda^k)$ with modulus at least $(\gamma^{-1} - \rho)$, we have

$$\mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) - \mathcal{P}_{\beta_k}(x^{k+1}, z^k, \lambda^k) \geq \frac{\gamma^{-1} - \rho}{2}\|x^{k+1} - x^k\|^2. \tag{53}$$

Next, recall in (5), $z^{k+1} = z^k + \eta(x^{k+1} - z^k)$ implies

$$2x^{k+1} - z^k - z^{k+1} = (2\eta^{-1} - 1)(z^{k+1} - z^k). \tag{54}$$

So we have

$$\mathcal{P}_{\beta_k}(x^{k+1}, z^k, \lambda^k) - \mathcal{P}_{\beta_k}(x^{k+1}, z^{k+1}, \lambda^k) = \frac{1}{2\gamma}(\|x^{k+1} - z^k\|^2 - \|x^{k+1} - z^{k+1}\|^2)$$
$$= \frac{1}{2\gamma}\langle z^{k+1} - z^k, 2x^{k+1} - z^k - z^{k+1}\rangle = \frac{1}{2\gamma}(\frac{2}{\eta} - 1)\|z^{k+1} - z^k\|^2.$$

Moreover, by the update $\lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b)$, we have

$$\mathcal{P}_{\beta_k}(x^{k+1}, z^{k+1}, \lambda^k) - \mathcal{P}_{\beta_k}(x^{k+1}, z^{k+1}, \lambda^{k+1}) = -\beta_k^{-1}\|\lambda^{k+1} - \lambda^k\|^2,$$

and

$$\mathcal{P}_{\beta_k}(x^{k+1}, z^{k+1}, \lambda^{k+1}) - \mathcal{P}_{\beta_{k+1}}(x^{k+1}, z^{k+1}, \lambda^{k+1}) = \frac{\beta_k - \beta_{k+1}}{2\beta_k^2}\|\lambda^{k+1} - \lambda^k\|^2.$$

Combining the above four terms of estimates yields

$$\mathcal{P}_{\beta_k}(x^k, z^k, \lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1}, z^{k+1}, \lambda^{k+1}) \tag{55}$$
$$\geq \frac{(1-\rho\gamma)}{2\gamma}\|x^{k+1} - x^k\|^2 + \frac{1}{2\gamma}(\frac{2}{\eta} - 1)\|z^{k+1} - z^k\|^2 - \frac{\beta_k + \beta_{k+1}}{2\beta_k^2}\|\lambda^{k+1} - \lambda^k\|^2.$$

Then, we establish (52) from (55). By the definition (16) of $\nabla\phi_{\beta_k}(z^k, \lambda^k)$, we have

$$\|\nabla\phi_{\beta_k}(z^k, \lambda^k)\|^2 = (\eta\gamma)^{-2}\|z^k - z^{k+1}\|^2 + \beta_k^{-2}\|\lambda^{k+1} - \lambda^k\|^2,$$

which implies

$$(\eta\gamma)^{-2}\|z^k - z^{k+1}\|^2 = \|\nabla\phi_{\beta_k}(z^k,\lambda^k)\|^2 - \beta_k^{-2}\|\lambda^{k+1}-\lambda^k\|^2.$$

Substituting this into the above inequality yields

$$\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1}) \geq \frac{(1-\gamma\rho)}{2\gamma}\|x^{k+1}-x^k\|^2$$
$$+\frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2 + \frac{1}{4}\gamma\eta(2-\eta)\|\nabla\phi_{\beta_k}(z^k,\lambda^k)\|^2 - \alpha_k c_{\gamma,A}\|\lambda^{k+1}-\lambda^k\|^2,$$

where $\alpha_k = \frac{\beta_k+\beta_{k+1}+\gamma\eta(1-\eta/2)}{2c_{\gamma,A}\beta_k^2}$. This finishes the proof.

Next, we provide a lemma for iMEAL (7).

Lemma 8 (iMEAL: one-step progress) *Let* $\{(x^k,z^k,\lambda^k)\}$ *be a sequence generated by iMEAL* (7). *Take Assumptions 2(a) and (b),* $\gamma\in(0,\rho^{-1})$, *and* $\eta\in(0,2)$. *It holds that*

$$\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1}) \tag{56}$$
$$\geq \frac{(1-\gamma\rho)}{2\gamma}\|x^{k+1}-x^k\|^2 + \langle s^k, x^k - x^{k+1}\rangle + \frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2$$
$$+\frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_{\beta_k}(z^k,\lambda^k)\|^2 - \alpha_k c_{\gamma,A}\|\lambda^{k+1}-\lambda^k\|^2,\ \forall k\in\mathbb{N}.$$

Proof The proof of this lemma is similar to that of Lemma 7 and uses the descent quantity along the update of x^{k+1}. By the update (7) of x^{k+1} in iMEAL and noticing that $\mathcal{L}_{\beta_k}(x,\lambda^k) + \frac{\|x-z^k\|}{2\gamma}$ is strongly convex with modulus at least $(\gamma^{-1}-\rho)$, we have

$$\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) \geq \mathcal{P}_{\beta_k}(x^{k+1},z^k,\lambda^k) + \langle s^k, x^k-x^{k+1}\rangle + \frac{\gamma^{-1}-\rho}{2}\|x^{k+1}-x^k\|^2.$$

By replacing (53) in the proof of Lemma 7 with the above inequality and following the rest part of its proof, we obtain the following inequality

$$\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1}) \geq \frac{1-\gamma\rho}{2\gamma}\|x^{k+1}-x^k\|^2 + \langle s^k, x^k-x^{k+1}\rangle$$
$$+\frac{1}{2\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2 - \frac{\beta_k+\beta_{k+1}}{2\beta_k^2}\|\lambda^{k+1}-\lambda^k\|^2.$$

We can establish (56) with a derivation similar to that in the proof of Lemma 7.

Also, we state a similar lemma for one-step progress of LiMEAL (9) as follows.

Lemma 9 (LiMEAL: one-step progress) *Let* $\{(x^k,z^k,\lambda^k)\}$ *be a sequence generated by LiMEAL* (9). *Take Assumptions 3(a) and (b),* $\gamma\in(0,\rho_g^{-1})$, *and* $\eta\in(0,2)$. *We have*

$$\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1}) \tag{57}$$
$$\geq \left(\frac{1-\gamma(\rho_g+L_h)}{2\gamma} - \frac{1}{4}\gamma(2-\eta)\eta L_h^2\right)\|x^{k+1}-x^k\|^2$$
$$+\frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2 + \frac{1}{4}\gamma(1-\eta_k/2)\eta\|g^k_{\mathrm{limeal}}\|^2 - \alpha_k c_{\gamma,A}\|\lambda^{k+1}-\lambda^k\|^2,\ \forall k\in\mathbb{N}.$$

Proof The proof of this lemma is similar to that of Lemma 7. By the update (9) of x^{k+1} in LiMEAL, x^{k+1} is updated via minimizing $(\gamma^{-1}-\rho_g)$-strongly convex $\mathcal{L}_{\beta_k,f^k}(x,\lambda^k)+\frac{\|x-z^k\|}{2\gamma}$, so

$$\mathcal{L}_{\beta_k,f^k}(x^k,\lambda^k)+\frac{\|x^k-z^k\|^2}{2\gamma}\geq \mathcal{L}_{\beta_k,f^k}(x^{k+1},\lambda^k)+\frac{\|x-z^k\|^2}{2\gamma}+\frac{\gamma^{-1}-\rho_g}{2}\|x^{k+1}-x^k\|^2.$$

By definition, $\mathcal{L}_{\beta_k,f^k}(x,\lambda)=h(x^k)+\langle\nabla h(x^k),x-x^k\rangle+g(x)+\langle\lambda,Ax-b\rangle+\frac{\beta}{2}\|Ax-b\|^2$ and $\mathcal{P}_{\beta_k}(x,z,\lambda)=h(x)+g(x)+\langle\lambda,Ax-b\rangle+\frac{\beta_k}{2}\|Ax-b\|^2+\frac{\|x-z\|^2}{2\gamma}$, so the above inequality implies

$$\begin{aligned}\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k)&\geq \mathcal{P}_{\beta_k}(x^{k+1},z^k,\lambda^k)+\frac{\gamma^{-1}-\rho_g}{2}\|x^{k+1}-x^k\|^2\\&\quad-(h(x^{k+1})-h(x^k)-\langle\nabla h(x^k),x^{k+1}-x^k\rangle)\\&\geq \mathcal{P}_{\beta_k}(x^{k+1},z^k,\lambda^k)+\frac{\gamma^{-1}-\rho_g-L_h}{2}\|x^{k+1}-x^k\|^2,\end{aligned}$$

where the second inequality is due to the L_h-Lipschitz continuity of ∇h. By replacing (53) in the proof of Lemma 7 with the above inequality and following the rest part of that proof, we obtain

$$\begin{aligned}&\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k)-\mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1})\\&\geq\frac{1-\gamma(\rho_g+L_h)}{2\gamma}\|x^{k+1}-x^k\|^2+\frac{1}{2\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2-\frac{\beta_k+\beta_{k+1}}{2\beta_k^2}\|\lambda^{k+1}-\lambda^k\|^2.\end{aligned}\tag{58}$$

Next, based on the above inequality, we establish (57). By the definition (33) of g^k_{limeal} and noticing that $z^k-x^{k+1}=-\eta^{-1}(z^{k+1}-z^k)$ by the update (9) of z^{k+1}, we have

$$\|g^k_{\text{limeal}}\|^2\leq 2L_h^2\|x^{k+1}-x^k\|^2+2(\gamma\eta)^{-2}\|z^{k+1}-z^k\|^2+\beta_k^{-2}\|\lambda^{k+1}-\lambda^k\|^2,$$

which implies

$$(\gamma\eta)^{-2}\|z^{k+1}-z^k\|^2\geq\frac{1}{2}\|g^k_{\text{limeal}}\|^2-\frac{1}{2}\beta_k^{-2}\|\lambda^{k+1}-\lambda^k\|^2-L_h^2\|x^{k+1}-x^k\|^2.$$

Substituting this inequality into (58) yields

$$\begin{aligned}&\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k)-\mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1})\\&\geq\left(\frac{1-\gamma(\rho_g+L_h)}{2\gamma}-\frac{1}{4}\gamma(2-\eta)\eta L_h^2\right)\|x^{k+1}-x^k\|^2\\&+\frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2+\frac{1}{4}\gamma(1-\eta/2)\eta\|g^k_{\text{limeal}}\|^2-\alpha_k c_{\gamma,A}\|\lambda^{k+1}-\lambda^k\|^2.\end{aligned}$$

This finishes the proof of this lemma.

5.2 Proofs for Convergence of MEAL

Based on the above lemmas, we give proofs of Theorem 1 and Proposition 1.

5.2.1 *Proof of Theorem 1*

Proof We first establish the $o(1/\sqrt{k})$ rate of convergence under the *implicit Lipschitz subgradient* assumption (Assumption 2(b)) and then the convergen rate result under the *implicit bounded subgradient* assumption (Assumption 2(c)).

(a) In the first case, $\beta_k = \beta$ and $\alpha_k = \alpha$. Substituting (44) into (52) yields

$$\mathcal{P}_\beta(x^k, z^k, \lambda^k) - \mathcal{P}_\beta(x^{k+1}, z^{k+1}, \lambda^{k+1}) \ge \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_\beta(z^k, \lambda^k)\|^2 + \left(\frac{(1-\gamma\rho)}{2\gamma} - 2\alpha(1+\gamma L_f)^2\right)\|x^{k+1} - x^k\|^2 + \frac{1}{4\gamma}(\frac{2}{\eta} - 1)\|z^{k+1} - z^k\|^2 - 2\alpha\|z^k - z^{k-1}\|^2.$$

By the definition (20) of $\mathcal{E}^k_{\text{meal}}$, the above inequality implies

$$\begin{aligned}\mathcal{E}^k_{\text{meal}} - \mathcal{E}^{k+1}_{\text{meal}} &\ge \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_\beta(z^k, \lambda^k)\|^2 + \left(\frac{1}{4\gamma}(\frac{2}{\eta} - 1) - 2\alpha\right)\|z^{k+1} - z^k\|^2 \\ &\quad + \left(\frac{1-\gamma\rho}{2\gamma} - 2\alpha(1+\gamma L_f)^2\right)\|x^{k+1} - x^k\|^2 \\ &\ge \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_\beta(z^k, \lambda^k)\|^2,\end{aligned} \tag{59}$$

where the second inequality holds due to the condition on α. Thus, claim (a) follows from the above inequality, Lemma 1 with $\tilde{\epsilon}_k = 0$ and the lower boundedness of $\{\mathcal{E}^k_{\text{meal}}\}$.

(b) Similarly, substituting (45) into (52) and using the definition (21) of $\tilde{\mathcal{E}}^k_{\text{meal}}$, we have

$$\tilde{\mathcal{E}}^k_{\text{meal}} - \tilde{\mathcal{E}}^{k+1}_{\text{meal}} \ge \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_{\beta_k}(z^k, \lambda^k)\|^2 - 12\alpha_k\gamma^2\hat{L}_f^2 + \left(\frac{1-\gamma\rho}{2\gamma} - 3\alpha_k\right)\|x^{k+1} - x^k\|^2 + \left(\frac{1}{4\gamma}(\frac{2}{\eta} - 1) - 3\alpha_{k+1}\right)\|z^{k+1} - z^k\|^2.$$

With $\alpha_k = \frac{\alpha^*}{K}$,

$$\tilde{\mathcal{E}}^k_{\text{meal}} - \tilde{\mathcal{E}}^{k+1}_{\text{meal}} \ge \frac{1}{2}\gamma(1-\eta/2)\eta\|\nabla\phi_{\beta_k}(z^k, \lambda^k)\|^2 - 12\alpha_k\gamma^2\hat{L}_f^2,$$

which yields claim (b) by Lemma 2 with $\tilde{\epsilon}_k = 0$ and the lower boundedness of $\{\tilde{\mathcal{E}}^k_{\text{meal}}\}$.

5.2.2 Proof of Proposition 1

Proof With Lemma 3, we only need to check conditions $(P1)$-$(P3)$ hold for MEAL.

(a) Establishing $(P1)$: With $a := \frac{\gamma\eta(2-\eta)}{4\beta}$, we have $\frac{1+a}{\beta c_{\gamma,A}} = \alpha$ for α in (23). Substituting (44) into (55) with fixed β_k yields

$$\mathcal{P}_\beta(x^k, z^k, \lambda^k) - \mathcal{P}_\beta(x^{k+1}, z^{k+1}, \lambda^{k+1}) \geq (\frac{1-\rho\gamma}{2\gamma} - 2\alpha(\gamma L_f + 1)^2)\|x^{k+1} - x^k\|^2$$
$$+ \frac{1}{2\gamma}(\frac{2}{\eta} - 1)\|z^{k+1} - z^k\|^2 - 2\alpha\|z^k - z^{k-1}\|^2 + a\beta^{-1}\|\lambda^{k+1} - \lambda^k\|^2.$$

For the definition (26) of $\mathcal{P}_{\text{meal}}$ and the assumption on α, we deduce from the above inequality:

$$\begin{aligned}&\mathcal{P}_{\text{meal}}(y^k) - \mathcal{P}_{\text{meal}}(y^{k+1}) \geq (\frac{1-\rho\gamma}{2\gamma} - 2\alpha(\gamma L_f + 1)^2)\|x^{k+1} - x^k\|^2 \\ &+ \left(\frac{1}{2\gamma}(\frac{2}{\eta} - 1) - 3\alpha\right)\|z^{k+1} - z^k\|^2 + \alpha\|z^k - z^{k-1}\|^2 + a\beta^{-1}\|\lambda^{k+1} - \lambda^k\|^2 \\ &\geq c_1\|y^{k+1} - y^k\|^2,\end{aligned} \tag{60}$$

where $c_1 := \min\left\{\frac{1-\rho\gamma}{2\gamma} - 2\alpha(\gamma L_f + 1)^2, \alpha, a\beta^{-1}\right\}$ by $\frac{1}{2\gamma}(\frac{2}{\eta} - 1) - 3\alpha \geq \alpha$. This yields $(P1)$ for MEAL.

(b) Establishing $(P2)$: Note that $\mathcal{P}_{\text{meal}}(y) = f(x) + \langle\lambda, Ax - b\rangle + \frac{\beta}{2}\|Ax - b\|^2 + \frac{1}{2\gamma}\|x - z\|^2 + 3\alpha\|z - \hat{z}\|^2$. The optimality condition from the update of x^{k+1} in (5) is

$$0 \in \partial f(x^{k+1}) + A^T\lambda^{k+1} + \gamma^{-1}(x^{k+1} - z^k),$$

which implies $\gamma^{-1}(z^k - z^{k+1}) + A^T(\lambda^{k+1} - \lambda^k) \in \partial_x\mathcal{P}_{\text{meal}}(y^{k+1})$. From the update of z^{k+1} in (5), $z^{k+1} - x^{k+1} = -(1-\eta)\eta^{-1}(z^{k+1} - z^k)$ and thus

$$\partial_z\mathcal{P}_{\text{meal}}(y^{k+1}) = \gamma^{-1}(z^{k+1} - x^{k+1}) + 6\alpha(z^{k+1} - z^k) = \left(6\alpha - \frac{1-\eta}{\eta\gamma}\right)(z^{k+1} - z^k).$$

The update of λ^{k+1} in (5) yields $\partial_\lambda\mathcal{P}_{\text{meal}}(y^{k+1}) = Ax^{k+1} - b = \beta^{-1}(\lambda^{k+1} - \lambda^k)$. Moreover, it is easy to show $\partial_{\hat{z}}\mathcal{P}_{\text{meal}}(y^{k+1}) = 6\alpha(z^k - z^{k+1})$. Thus, let

$$v^{k+1} := \begin{pmatrix} \gamma^{-1}(z^k - z^{k+1}) + A^T(\lambda^{k+1} - \lambda^k) \\ \left(6\alpha - \frac{1-\eta}{\eta\gamma}\right)(z^{k+1} - z^k) \\ \beta^{-1}(\lambda^{k+1} - \lambda^k) \\ 6\alpha(z^k - z^{k+1}) \end{pmatrix},$$

which obeys $v^{k+1} \in \partial\mathcal{P}_{\text{meal}}(y^{k+1})$ and

$$\begin{aligned}\|v^{k+1}\| &\leq \left(\gamma^{-1} + \left|6\alpha - \frac{1-\eta}{\eta\gamma}\right| + 6\alpha\right)\|z^{k+1} - z^k\| + \beta^{-1}\|\lambda^{k+1} - \lambda^k\| + \|A^T(\lambda^{k+1} - \lambda^k)\| \\ &\leq \left(\gamma^{-1} + \left|6\alpha - \frac{1-\eta}{\eta\gamma}\right| + 6\alpha\right)\|z^{k+1} - z^k\| + \beta^{-1}\|\lambda^{k+1} - \lambda^k\| \\ &+ (L_f + \gamma^{-1})\|x^{k+1} - x^k\| + \gamma^{-1}\|\hat{z}^{k+1} - \hat{z}^k\|,\end{aligned}$$

where the second inequality is due to (43). This yields $(P2)$ for MEAL.

(c) Establishing $(P3)$: $(P3)$ follows from the boundedness assumption of $\{y^k\}$, and the convergence of $\{\mathcal{P}_{\rm meal}(y^k)\}$ is implied by $(P1)$. This finishes the proof.

5.3 Proofs for Convergence of iMEAL

In this subsection, we present the proof of Theorem 2 for iMEAL (7).

Proof (of Theorem 2) We first show the $o(1/\sqrt{k})$ rate of convergence under Assumption 2(b) and then the convergence rate result under Assumption 2(c).

(a) In this case, we use a fixed $\beta_k = \beta$ and thus $\alpha_k = \alpha$. Substituting the inequality in Lemma 5(a) into (56) in Lemma 8 yields

$$\begin{aligned}&\mathcal{P}_\beta(x^k,z^k,\lambda^k) - \mathcal{P}_\beta(x^{k+1},z^{k+1},\lambda^{k+1}) \geq \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_\beta(z^k,\lambda^k)\|^2 \\ &+\left(\frac{1-\gamma\rho}{2\gamma} - 3\alpha(1+\gamma L_f)^2\right)\|x^{k+1}-x^k\|^2 + \langle s^k, x^k - x^{k+1}\rangle - 3\alpha\gamma^2(\epsilon_k+\epsilon_{k-1})^2 \\ &+\frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2 - 3\alpha\|z^k - z^{k-1}\|^2.\end{aligned} \tag{61}$$

Let $\delta := 2\left(\frac{(1-\gamma\rho)}{2\gamma} - 3\alpha(1+\gamma L_f)^2\right)$. By the assumption $0<\alpha<\min\left\{\frac{1-\gamma\rho}{6\gamma(1+\gamma L_f)^2}, \frac{1}{12\gamma}(\frac{2}{\eta}-1)\right\}$, we have $\delta > 0$ and further

$$\langle s^k, x^k - x^{k+1}\rangle \geq -\frac{\delta}{2}\|x^{k+1}-x^k\|^2 - \frac{1}{2\delta}\|s^k\|^2 \geq -\frac{\delta}{2}\|x^{k+1}-x^k\|^2 - \frac{1}{2\delta}(\epsilon_k+\epsilon_{k-1})^2.$$

Substituting this into (61) and noting the definition (28) of $\mathcal{E}^k_{\rm imeal}$, we have

$$\mathcal{E}^k_{\rm imeal} - \mathcal{E}^{k+1}_{\rm imeal} \geq \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_\beta(z^k,\lambda^k)\|^2 - (3\alpha\gamma^2 + \frac{1}{2\delta})(\epsilon_k+\epsilon_{k-1})^2,$$

which yields claim (a) by the assumption $\sum_{k=1}^\infty(\epsilon_k)^2 < +\infty$ and Lemma 1.

(b) Then we establish claim (b) under Assumption 2(c). Substituting the inequality in Lemma 5(b) into (56) in Lemma 8 yields

$$\begin{aligned}&\mathcal{P}_{\beta_k}(x^k,z^k,\lambda^k) - \mathcal{P}_{\beta_{k+1}}(x^{k+1},z^{k+1},\lambda^{k+1}) \geq \frac{1}{2}\gamma\eta(1-\eta/2)\|\nabla\phi_{\beta_k}(z^k,\lambda^k)\|^2 \\ &+\left(\frac{(1-\gamma\rho)}{2\gamma} - 4\alpha_k\right)\|x^{k+1}-x^k\|^2 + \langle s^k, x^k - x^{k+1}\rangle - 4\gamma^2\alpha_k(\epsilon_k+\epsilon_{k-1})^2 \\ &+\frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2 - 4\alpha_k\|z^k - z^{k-1}\|^2 - 16\alpha_k\gamma^2\hat{L}_f^2.\end{aligned} \tag{62}$$

Let $\hat{\alpha}^* := \min\left\{\frac{1-\rho\gamma}{8\gamma}, \frac{1}{16\gamma}(\frac{2}{\eta}-1)\right\}$ and $\tilde{\delta} := 2\left(\frac{(1-\gamma\rho)}{2\gamma} - 4\hat{\alpha}^*\right) > 0$. We have

$$\langle s^k, x^k - x^{k+1}\rangle \geq -\frac{\tilde{\delta}}{2}\|x^{k+1}-x^k\|^2 - \frac{1}{2\tilde{\delta}}\|s^k\|^2 \geq -\frac{\tilde{\delta}}{2}\|x^{k+1}-x^k\|^2 - \frac{1}{2\tilde{\delta}}(\epsilon_k+\epsilon_{k-1})^2.$$

Substituting this into (62), and by the definition (29) of $\tilde{\mathcal{E}}^k_{\rm imeal}$ and setting of α_k, we have

$$\tilde{\mathcal{E}}^k_{\rm imeal} - \tilde{\mathcal{E}}^{k+1}_{\rm imeal} \geq \frac{1}{2}\gamma(1-\eta/2)\eta\|\nabla\phi_{\beta_k}(z^k,\lambda^k)\|^2 - (4\alpha_k\gamma^2 + \frac{1}{2\tilde{\delta}})(\epsilon_k+\epsilon_{k-1})^2 - 16\alpha_k\gamma^2\hat{L}_f^2,$$

which yields claim (b) by the assumption $\sum_{k=1}^{\infty}(\epsilon_k)^2 < +\infty$ and Lemma 2.

5.4 Proofs for Convergence of LiMEAL

Now, we show proofs of main convergence theorems for LiMEAL (9).

5.4.1 Proof of Theorem 3

Proof We first establish claim (a) and then claim (b) under the associated assumptions.

(a) In this case, a fixed β_k is used. Substituting (50) into (57) yields

$$\begin{aligned} &\mathcal{P}_\beta(x^k,z^k,\lambda^k) - \mathcal{P}_\beta(x^{k+1},z^{k+1},\lambda^{k+1}) \geq \frac{1}{4}\gamma(1-\eta/2)\eta\|g^k_{\rm limeal}\|^2 \\ &+\left(\frac{1-\gamma(\rho_g+L_h)}{2\gamma} - \frac{1}{4}\gamma(2-\eta)\eta L_h^2 - 3(1+\gamma L_g)^2\alpha\right)\|x^{k+1}-x^k\|^2 \\ &+\frac{1}{4\gamma}(\frac{2}{\eta}-1)\|z^{k+1}-z^k\|^2 - 3\alpha(\gamma^2L_h^2\|x^k-x^{k-1}\|^2 + \|z^k-z^{k-1}\|^2). \end{aligned}$$

By the definition (35) of $\mathcal{E}^k_{\rm limeal}$, the above inequality implies

$$\begin{aligned} &\mathcal{E}^k_{\rm limeal} - \mathcal{E}^{k+1}_{\rm limeal} \geq \frac{1}{4}\gamma(1-\eta/2)\eta\|g^k_{\rm limeal}\|^2 + \left(\frac{1}{4\gamma}(\frac{2}{\eta}-1) - 3\alpha\right)\|z^{k+1}-z^k\|^2 \quad (63)\\ &+\left(\frac{1-\gamma(\rho_g+L_h)}{2\gamma} - \frac{1}{4}\gamma(2-\eta)\eta L_h^2 - 3\alpha\left((1+\gamma L_g)^2+\gamma^2L_h^2\right)\right)\|x^{k+1}-x^k\|^2 \\ &\geq \frac{1}{4}\gamma(1-\eta/2)\eta\|g^k_{\rm limeal}\|^2, \end{aligned}$$

where the second inequality holds under the conditions in Theorem 3(a). This shows the claim (a) by Lemma 1 and the lower boundedness of $\{\mathcal{E}^k_{\rm limeal}\}$.

(b) Similarly, substituting (51) into (57) and using the definitions of α_k in (22) and $\tilde{\mathcal{E}}^k_{\rm limeal}$ in (36), we obtain

$$\begin{aligned} &\tilde{\mathcal{E}}^k_{\rm limeal} - \tilde{\mathcal{E}}^{k+1}_{\rm limeal} \geq \frac{1}{4}\gamma(1-\eta/2)\eta_k\|g^k_{\rm limeal}\|^2 - 16\alpha_k\gamma^2\hat{L}_g^2 + \left(\frac{1}{4\gamma}(\frac{2}{\eta}-1) - 4\alpha_{k+1}\right)\|z^{k+1}-z^k\|^2 \\ &+\left(\frac{(1-\gamma(\rho_g+L_h))}{2\gamma} - \frac{1}{4}\gamma(2-\eta)\eta L_h^2 - 4\alpha_k - 4\gamma^2L_h^2\alpha_{k+1}\right)\|x^{k+1}-x^k\|^2 \\ &\geq \frac{1}{4}c\gamma\eta\|g^k_{\rm limeal}\|^2 - 16\alpha_k\gamma^2\hat{L}_g^2, \end{aligned}$$

where the second inequality is due to the settings of parameters presented in Theorem 3(b). This inequality shows claim (b) by Lemma 2 and the lower boundedness of $\{\tilde{\mathcal{E}}^k_{\rm limeal}\}$.

5.4.2 *Proof of Proposition 3*

Proof By Lemma 3, we only need to verify conditions $(P1)$-$(P3)$ hold for LiMEAL.

(a) Establishing $(P1)$: Similar to the proof of Theorem 1, let $a := \frac{\gamma\eta(2-\eta)}{4\beta}$. Then $\frac{1+a}{\beta c_{\gamma,A}} = \alpha$, where α is defined in (23). Substituting (50) into (58) with fixed β_k yields

$$\begin{aligned}
&\mathcal{P}_\beta(x^k, z^k, \lambda^k) - \mathcal{P}_\beta(x^{k+1}, z^{k+1}, \lambda^{k+1}) \\
&\geq \left(\frac{1-\gamma(\rho_g + L_h)}{2\gamma} - 3\alpha(1+\gamma L_g)^2\right)\|x^{k+1} - x^k\|^2 - 3\alpha\gamma^2 L_h^2 \|x^k - x^{k-1}\|^2 \\
&+ \frac{1}{2\gamma}(\frac{2}{\eta} - 1)\|z^{k+1} - z^k\|^2 - 3\alpha\|z^k - z^{k-1}\|^2 + a\beta^{-1}\|\lambda^{k+1} - \lambda^k\|^2.
\end{aligned}$$

By the definition (37) of $\mathcal{P}_{\text{limeal}}$, the above inequality implies

$$\begin{aligned}
\mathcal{P}_{\text{limeal}}(y^k) - \mathcal{P}_{\text{limeal}}(y^{k+1}) \geq &\left(\frac{1-\gamma(\rho_g + L_h)}{2\gamma} - 4\alpha\left((1+\gamma L_g)^2 + \gamma^2 L_h^2\right)\right)\|x^{k+1} - x^k\|^2 \\
&+ \left(\frac{1}{2\gamma}\left(\frac{2}{\eta} - 1\right) - 4\alpha\right)\|z^{k+1} - z^k\|^2 + a\beta^{-1}\|\lambda^{k+1} - \lambda^k\|^2 \\
&+ \alpha\left(\gamma^2 L_h^2 \|\hat{x}^{k+1} - \hat{x}^k\|^2 + \|\hat{z}^{k+1} - \hat{z}^k\|^2\right),
\end{aligned}$$

which, with the assumptions on the parameters, implies $(P1)$ for LiMEAL.

(b) Establishing $(P2)$: Note that $\mathcal{P}_{\text{limeal}}(y) = f(x) + \langle\lambda, Ax - b\rangle + \frac{\beta}{2}\|Ax - b\|^2 + \frac{1}{2\gamma}\|x - z\|^2 + 4\alpha\gamma^2 L_h^2\|x - \hat{x}\|^2 + 4\alpha\|z - \hat{z}\|^2$. The update of x^{k+1} in (9) has the optimality condition

$$0 \in \partial g(x^{k+1}) + \nabla h(x^k) + A^T\lambda^{k+1} + \gamma^{-1}(x^{k+1} - z^k),$$

which implies

$$\begin{aligned}
&(\nabla h(x^{k+1}) - \nabla h(x^k)) + 8\gamma^2 L_h^2 \alpha(x^{k+1} - x^k) \\
&+ \gamma^{-1}(z^k - z^{k+1}) + A^T(\lambda^{k+1} - \lambda^k) \in \partial_x \mathcal{P}_{\text{limeal}}(y^{k+1}).
\end{aligned}$$

The derivations for the other terms are straightforward and similar to those in the proof of Proposition 1. We directly show the final estimate: for some $v^{k+1} \in \partial\mathcal{P}_{\text{limeal}}(y^{k+1})$,

$$\begin{aligned}
\|v^{k+1}\| \leq &\left(L_h + L_g + \gamma^{-1} + 16\alpha\gamma^2 L_h^2\right)\|x^{k+1} - x^k\| + \left(\gamma^{-1} + \left|8\alpha - \frac{1-\eta}{\eta}\right| + 8\alpha\right)\|z^{k+1} - z^k\| \\
&+ \beta^{-1}\|\lambda^{k+1} - \lambda^k\| + L_h\|\hat{x}^{k+1} - \hat{x}^k\| + \gamma^{-1}\|\hat{z}^{k+1} - \hat{z}^k\|,
\end{aligned}$$

which yields $(P2)$ for LiMEAL.

(c) Establishing $(P3)$: $(P3)$ follows from the boundedness assumption of $\{y^k\}$ and the convergence of $\{\mathcal{P}_{\text{limeal}}(y^k)\}$ by $(P1)$. This finishes the proof.

6 Discussions on Boundedness and Related Work

In this section, we firstly discuss how to ensure the bounded sequences and then provide some discussions on related work.

6.1 Discussions on Boundedness of Sequence

Theorem 1 imposes the condition of lower boundedness of $\{\mathcal{E}_{\text{meal}}^k\}$ and Proposition 1 does with boundedness of the generated sequence $\{(x^k, z^k, \lambda^k)\}$. In this section, we provide some sufficient conditions to guarantee the former and then the latter boundedness conditions.

Besides the ρ-weak convexity of f (implying the curvature of f is lower bounded by ρ), we impose the coerciveness on the constrained problem (1) as follows.

Assumption 4 (Coercivity) *The minimal value $f^* := \inf_{x\in\mathcal{X}} f(x)$ is finite (recall $\mathcal{X} := \{x : Ax = b\}$), and f is coercive over the set $\mathcal{X}$, that is, $f(x) \to \infty$ if $x \in \mathcal{X}$ and $\|x\| \to \infty$.*

The coercive assumption is a common condition used to obtain the boundedness of the sequence, for example, used in [59, Assumption A1] for the nonconvex ADMM. Particularly, let (x^0, z^0, λ^0) be a finite initial guess of MEAL and

$$\mathcal{E}^0 := \mathcal{E}_{\text{meal}}^1 < +\infty. \tag{64}$$

By Assumption 4, if $x \in \mathcal{X}$ and $f(x) \le \mathcal{E}^0$, then there exists a positive constant $\mathcal{B}_0$ (possibly depending on $\mathcal{E}^0$) such that $\|x\| \le \mathcal{B}_0$. Define another positive constant as

$$\mathcal{B}_1 := \mathcal{B}_0 + \sqrt{2\rho^{-1} \cdot \max\{0, \mathcal{E}^0 - f^*\}}. \tag{65}$$

Given a $\gamma \in (0, 1/\rho)$ and $z \in \mathbb{R}^n$ with $\|z\| \le \mathcal{B}_1$ and $u \in \text{Im}(A)$, we define

$$x(u; z) := \operatorname*{argmin}_{\{x:Ax=u\}} \left\{ f(x) + \frac{1}{2\gamma}\|x - z\|^2 \right\}. \tag{66}$$

Since f is ρ-weakly convex by Assumption 2(a), then for any $\gamma \in (0, 1/\rho)$, the function $f(x) + \frac{1}{2\gamma}\|x - z\|^2$ is strongly convex with respect to x, and thus the above $x(u; z)$ is well-defined and unique for any given $z \in \mathbb{R}^n$ and $u \in \text{Im}(A)$. Motivated by [23, Ch 5.6.3], we impose some *local stability* on $x(u; z)$ defined in (66).

Assumption 5 (Local stability) *For any given $z \in \mathbb{R}^n$ with $\|z\| \le \mathcal{B}_1$, there exist a $\delta > 0$ and a finite positive constant $\bar{M}$ (possibly depending on A, $\mathcal{B}_1$ and δ) such that*

$$\|x(u; z) - x(b; z)\| \le \bar{M}\|u - b\|, \ \forall u \in \text{Im}(A) \cap \{v : \|v - b\| \le \delta\}.$$

The above *local stability* assumption is also related to the *Lipschitz sub-minimization path* assumption suggested in [59, Assumption A3]. As discussed in [59], the *Lipschitz sub-minimization path* assumption relaxes the more stringent *full-rank* assumption used in the literature (see the discussions in [59, Sections 2.2 and 4.1] and references therein). As $\{z \in \mathbb{R}^n : \|z\| \le \mathcal{B}_1\}$ is a compact set, $\bar{M}$ can be taken as the supremum of these stability constants over this compact set. Based on Assumption 5, we have the following lemma.

Lemma 10 *Let $\{(x^k, z^k, \lambda^k)\}$ be the sequence generated by MEAL (5) with fixed $\beta > 0$ and $\eta > 0$. If $\gamma \in (0, 1/\rho)$, $\|z^k\| \le \mathcal{B}_1$ and $\|Ax^{k+1} - b\| \le \delta$, there holds*

$$\|x^{k+1} - x(b; z^k)\| \le \bar{M}\|Ax^{k+1} - b\|, \ \forall k \in \mathbb{N}.$$

Proof Let $u^{k+1} = Ax^{k+1}$. By the update of x^{k+1} in (5), there holds

$$\mathcal{P}_\beta(x^{k+1}, z^k, \lambda^k) \le \mathcal{P}_\beta(x(u^{k+1}; z^k), z^k, \lambda^k),$$

Noting that $Ax(u^{k+1}; z^k) = Ax^{k+1}$ due to its definition in (66), the above inequality implies

$$f(x^{k+1}) + \frac{1}{2\gamma}\|x^{k+1} - z^k\|^2 \le f(x(u^{k+1}; z^k)) + \frac{1}{2\gamma}\|x(u^{k+1}; z^k) - z^k\|^2.$$

By the definition of $x(u^{k+1}; z^k)$ in (66) again and noting that $Ax^{k+1} = u^{k+1}$, we have

$$f(x^{k+1}) + \frac{1}{2\gamma}\|x^{k+1} - z^k\|^2 \ge f(x(u^{k+1}; z^k)) + \frac{1}{2\gamma}\|x(u^{k+1}; z^k) - z^k\|^2.$$

These two inequalities imply

$$f(x^{k+1}) + \frac{1}{2\gamma}\|x^{k+1} - z^k\|^2 = f(x(u^{k+1}; z^k)) + \frac{1}{2\gamma}\|x(u^{k+1}; z^k) - z^k\|^2,$$

which yields

$$x^{k+1} = x(u^{k+1}; z^k) = x(Ax^{k+1}; z^k)$$

by the strong convexity of function $f(x) + \frac{1}{2\gamma}\|x - z^k\|^2$ for any $\gamma \in (0, 1/\rho)$ and thus the uniqueness of $x(u^{k+1}; z^k)$. Then by Assumption 5, we yield the desired result.

Based on the above assumptions, we establish the lower boundedness of $\{\mathcal{E}^k_{\text{meal}}\}$ and the boundedness of $\{(x^k, z^k, \lambda^k)\}$ as follows.

Proposition 4 *Let $\{(x^k, z^k, \lambda^k)\}_{k\in\mathbb{N}}$ be a sequence generated by MEAL (5) with a finite initial guess (x^0, z^0, λ^0) such that $\|z^0\| \le \mathcal{B}_1$, where $\mathcal{B}_1$ is defined in* (65)*. Suppose that Assumptions 1, 2(a)-(b) and 4 hold and further Assumption 5 holds with some $0 < \bar{M} < \frac{2}{\sqrt{\sigma_{\min}(A^TA)}}$. If $\gamma \in (0, \rho^{-1})$, $\eta \in (0, 2)$ and $\beta >$ $\max\left\{\frac{1+\sqrt{1+\eta(2-\eta)\gamma c_{\gamma,A}\alpha_{\max}}}{2c_{\gamma,A}\alpha_{\max}}, \frac{a_2+\sqrt{a_2^2+4a_1a_3}}{2a_1}\right\}$, where $\alpha_{\max} := \min\left\{\frac{1-\gamma\rho}{4\gamma(1+\gamma L_f)^2}, \frac{1}{8\gamma}(\frac{2}{\eta} - 1)\right\}$, $c_{\gamma,A} = \gamma^2\sigma_{\min}(A^TA)$, $a_1 = 4 - \bar{M}^2\sigma_{\min}(A^TA)$, $a_2 = 4(\bar{L} + \gamma^{-1})\bar{M}^2 - \gamma\eta(2-\eta)$, $a_3 = (1 + \gamma\bar{L})\eta(2-\eta)\bar{M}^2$ and $\bar{L} = \rho + 2L_f$, then the following hold:*

(a) $\{\mathcal{E}^k_{\text{meal}}\}$ is lower bounded;
(b) $\{(x^k, z^k)\}$ is bounded; and
(c) if further $\lambda^0 \in \text{Null}(A^T)$ (the null space of A^T) and $\|\nabla\mathcal{M}_{\gamma,f}(w^1)\|$ is finite with $w^1 = z^0 - \gamma A^T\lambda^1$, then $\{\lambda^k\}$ is bounded.

Proof In order to prove this proposition, we firstly establish the following claim for sufficiently large k:

Claim A: *If $\|z^{k-1}\| \le \mathcal{B}_1$, $\|Ax^k - b\| \le \delta, \forall k \ge k_0$ for some sufficiently large k_0, then $\mathcal{E}^k_{\text{meal}} \ge f^*$, and $\|z^k\| \le \mathcal{B}_1$ and $\|x^k\| \le \mathcal{B}_2$.*

By Theorem 1(a), such k_0 does exist due to the lower boundedness of $\{\mathcal{E}^k_{meal}\}$ for all finite k and thus $\xi^k_{\text{meal}} \le \hat{c}/\sqrt{k}$ for some constant $\hat{c} > 0$ (implying $\|Ax^k - b\|$ is sufficiently small with a sufficiently large k).

In the next, we show **Claim A**. By the definition (20) of $\mathcal{E}^k_{\rm meal}$, we have

$$\begin{aligned}\mathcal{E}^k_{\rm meal} &= f(x^k) + \langle \lambda^k, Ax^k - b\rangle + \frac{\beta}{2}\|Ax^k - b\|^2 + \frac{1}{2\gamma}\|x^k - z^k\|^2 + 2\alpha\|z^k - z^{k-1}\|^2 \\ &= f(x^k) + \langle A^T\lambda^k, x^k - \bar{x}^k\rangle + \frac{\beta}{2}\|Ax^k - b\|^2 + \frac{1}{2\gamma}\|x^k - z^k\|^2 + 2\alpha\|z^k - z^{k-1}\|^2,\end{aligned}$$

where

$$\bar{x}^k := x(b; z^{k-1})$$

as defined in (66). Let $\bar{\lambda}^k$ be the associated optimal Lagrangian multiplier of $\bar{x}^k$ and $\bar{w}^k = z^{k-1} - \gamma A^T\bar{\lambda}^k$. Then we have

$$\bar{x}^k = \mathrm{Prox}_{\gamma,f}(\bar{w}^k),$$

and $\nabla\mathcal{M}_{\gamma,f}(\bar{w}^k) \in \partial f(\bar{x}^k)$. By (48) in the proof of Lemma 4, we have

$$A^T\lambda^k = -\nabla\mathcal{M}_{\gamma,f}(w^k) - \gamma^{-1}(x^k - z^{k-1}),$$

and $\nabla\mathcal{M}_{\gamma,f}(w^k) \in \partial f(x^k)$, where $w^k = z^{k-1} - \gamma A^T\lambda^k$. Substituting the above equation into the previous equality yields

$$\begin{aligned}\mathcal{E}^k_{\rm meal} = f(x^k) &+ \langle \nabla\mathcal{M}_{\gamma,f}(w^k), \bar{x}^k - x^k\rangle + \frac{\beta}{2}\|Ax^k - b\|^2 \\ &+ \gamma^{-1}\langle x^k - z^{k-1}, \bar{x}^k - x^k\rangle + \frac{1}{2\gamma}\|x^k - z^k\|^2 + 2\alpha\|z^k - z^{k-1}\|^2.\end{aligned} \tag{67}$$

Noting that $\nabla\mathcal{M}_{\gamma,f}(\bar{w}^k) \in \partial f(\bar{x}^k)$ and by the ρ-weak convexity of f, we have

$$f(x^k) \geq f(\bar{x}^k) + \langle \nabla\mathcal{M}_{\gamma,f}(\bar{w}^k), x^k - \bar{x}^k\rangle - \frac{\rho}{2}\|x^k - \bar{x}^k\|^2,$$

which implies

$$\begin{aligned}&f(x^k) + \langle \nabla\mathcal{M}_{\gamma,f}(w^k), \bar{x}^k - x^k\rangle \\ &\geq f(\bar{x}^k) - \frac{\rho}{2}\|\bar{x}^k - x^k\|^2 - \langle \nabla\mathcal{M}_{\gamma,f}(\bar{w}^k) - \nabla\mathcal{M}_{\gamma,f}(w^k), \bar{x}^k - x^k\rangle \\ &\geq f(\bar{x}^k) - \frac{\rho}{2}\|\bar{x}^k - x^k\|^2 - \|\nabla\mathcal{M}_{\gamma,f}(\bar{w}^k) - \nabla\mathcal{M}_{\gamma,f}(w^k)\| \cdot \|\bar{x}^k - x^k\|.\end{aligned}$$

By the implicit Lipschitz subgradient assumption (i.e., Assumption 2 (b)) and the definition of $\bar{L} := \rho + 2L_f$, the above inequality yields

$$f(x^k) + \langle \nabla\mathcal{M}_{\gamma,f}(w^k), \bar{x}^k - x^k\rangle \geq f(\bar{x}^k) - \frac{\bar{L}}{2}\|\bar{x}^k - x^k\|^2. \tag{68}$$

Moreover, it is easy to show that

$$\begin{aligned}&\gamma^{-1}\langle x^k - z^{k-1}, \bar{x}^k - x^k\rangle + \frac{1}{2\gamma}\|x^k - z^k\|^2 + 2\alpha\|z^k - z^{k-1}\|^2 \\ &= \frac{1}{2\gamma}\|\bar{x}^k - z^k\|^2 - \frac{1}{2\gamma}\|\bar{x}^k - x^k\|^2 + \gamma^{-1}\langle z^k - z^{k-1}, \bar{x}^k - x^k\rangle + 2\alpha\|z^k - z^{k-1}\|^2 \\ &= \frac{1}{2\gamma}\|\bar{x}^k - z^k\|^2 - \left(\frac{1}{2\gamma} + \frac{1}{8\alpha\gamma^2}\right)\|\bar{x}^k - x^k\|^2 + 2\alpha\left\|(z^k - z^{k-1}) + \frac{1}{4\alpha\gamma}(\bar{x}^k - x^k)\right\|^2.\end{aligned} \tag{69}$$

Substituting (68)-(69) into (67) and by Lemma 10, we have

$$
\begin{aligned}
\mathcal{E}_{\text{meal}}^k &\geq f(\bar{x}^k) + \frac{1}{2\gamma}\|\bar{x}^k - z^k\|^2 + 2\alpha \left\|(z^k - z^{k-1}) + \frac{1}{4\alpha\gamma}(\bar{x}^k - x^k)\right\|^2 \\
&\quad + \frac{1}{2}\left[\beta - \left(\frac{1}{4\alpha\gamma^2} + \bar{L} + \gamma^{-1}\right)\bar{M}^2\right]\|Ax^k - b\|^2 \\
&\geq f(\bar{x}^k) + \frac{1}{2\gamma}\|\bar{x}^k - z^k\|^2 + 2\alpha \left\|(z^k - z^{k-1}) + \frac{1}{4\alpha\gamma}(\bar{x}^k - x^k)\right\|^2 \qquad (70) \\
&\geq f^* + \frac{1}{2\gamma}\|\bar{x}^k - z^k\|^2 + 2\alpha \left\|(z^k - z^{k-1}) + \frac{1}{4\alpha\gamma}(\bar{x}^k - x^k)\right\|^2 \qquad (71) \\
&> -\infty, \qquad (72)
\end{aligned}
$$

where the second inequality follows from the definition of $\alpha = \frac{2\beta+\gamma\eta(1-\eta/2)}{2\gamma^2\sigma_{\min}(A^TA)\beta^2}$ and the condition on β, the third inequality holds for $\bar{x}^k := x(b; z^{k-1})$ and thus $A\bar{x}^k = b$ and $f(\bar{x}^k) \geq f^*$, and the final inequality is due to Assumption 4. The above inequality yields the lower boundedness of $\{\mathcal{E}_{\text{meal}}^k\}$ in **Claim A**. Thus, clam (a) in this proposition holds.

Then, we show the boundedness of $\{(x^k, z^k)\}$ in **Claim A**. By (70) and (59), we have

$$
f(\bar{x}^k) \leq \mathcal{E}^0 := \mathcal{E}_{\text{meal}}^1,
$$

which implies $\|\bar{x}^k\| \leq \mathcal{B}_0$ by Assumption 4. By (71) and the condition on $\gamma \in (0, \rho^{-1})$, we have $f^* + \frac{\rho}{2}\|\bar{x}^k - z^k\|^2 \leq f^* + \frac{1}{2\gamma}\|\bar{x}^k - z^k\|^2 \leq \mathcal{E}^0$, which implies

$$
\|z^k\| \leq \mathcal{B}_0 + \sqrt{2(\mathcal{E}^0 - f^*)/\rho} = \mathcal{B}_1.
$$

By (71) again, we have $\left\|(z^k - z^{k-1}) + \frac{1}{4\alpha\gamma}(\bar{x}^k - x^k)\right\|^2 \leq \frac{\mathcal{E}^0 - f^*}{2\alpha}$, which, together with these existing bounds $\|z^{k-1}\| \leq \mathcal{B}_1$, $\|z^k\| \leq \mathcal{B}_1$ and $\|\bar{x}^k\| \leq \mathcal{B}_0$, yields

$$
\|x^k\| \leq \mathcal{B}_0 + 4\alpha\gamma\left(2\mathcal{B}_1 + \sqrt{\frac{\mathcal{E}^0 - f^*}{2\alpha}}\right) =: \mathcal{B}_2. \qquad (73)
$$

Thus, we have shown **Claim A**. Recursively, we can show that $\{x^k\}$ and $\{z^k\}$ are respectively bounded by $\mathcal{B}_2$ and $\mathcal{B}_1$ for any $k \geq 1$, that is, claim (b) in this proposition holds.

In the following, we show claim (c) of this proposition. By the update of λ^{k+1} in (5), it is easy to show $\lambda^k = \lambda^0 + \hat{\lambda}^k$, where $\hat{\lambda}^k = \beta\sum_{t=1}^k(Ax^t - b) \in \text{Im}(A)$ by Assumption 1. Furthermore, by the assumption that $\lambda^0 \in \text{Null}(A^T)$, we have

$$
\langle \lambda^0, \hat{\lambda}^k \rangle = 0, \ \forall k \geq 1. \qquad (74)
$$

By (48), for any $k \geq 1$, we have

$$
A^T\lambda^k = -(\nabla\mathcal{M}_{\gamma,f}(w^k) - \nabla\mathcal{M}_{\gamma,f}(w^1)) - \nabla\mathcal{M}_{\gamma,f}(w^1) - \gamma^{-1}(x^k - z^{k-1}),
$$

Table 1 Comparisons on convergence results of existing algorithms for problem (1), where *imp-Lip* and *imp-bound* represent the *implicit Lipschitz subgradient* (see, Assumption 2(b)) and *implicit bounded subgradient* (see, Assumption 2(c)) assumptions, respectively. In [60], the general nonlinear equality constraints $c(x) = 0$ is considered, where ∇c is Lipschitz and bounded.

Algorithm	MEAL (our)	iMEAL (our)	Prox-PDA [38]	Prox-ALM [60]
Assumption	f : weakly convex, *imp-Lip* or *imp-bound*		∇f : Lipschitz	
Iteration complexity	*imp-Lip*: $o(\varepsilon^{-2})$ *imp-bound*: $O(\varepsilon^{-2})$	*imp-Lip*: $o(\varepsilon^{-2})$ *imp-bound*: $O(\varepsilon^{-2})$	$O(\varepsilon^{-2})$	$O(\varepsilon^{-2})$
Global Convergence	✓ (KŁ)	–	–	–

where $w^k = z^{k-1} - \gamma A^T \lambda^k$. By Assumption 2(b) and the boundedness of $\{(x^k, z^k)\}$ shown before, the above equation implies

$$\begin{aligned}\|A^T \lambda^k\| &\leq L_f \|x^k - x^1\| + \|\nabla \mathcal{M}_{\gamma,f}(w^1)\| + \gamma^{-1}\|x^k - z^{k-1}\| \\ &\leq \gamma^{-1}\mathcal{B}_1 + (2L_f + \gamma^{-1})\mathcal{B}_2 + \|\nabla \mathcal{M}_{\gamma,f}(w^1)\| < +\infty.\end{aligned}$$

By the relation $\lambda^k = \lambda^0 + \hat{\lambda}^k$ and (74), the above inequality implies

$$\|A^T \hat{\lambda}^k\| \leq \gamma^{-1}\mathcal{B}_1 + (2L_f + \gamma^{-1})\mathcal{B}_2 + \|\nabla \mathcal{M}_{\gamma,f}(w^1)\|.$$

Since $\hat{\lambda}^k \in \mathrm{Im}(A)$, the above inequality implies

$$\|\hat{\lambda}^k\| \leq \tilde{\sigma}_{\min}^{-1/2}(A^T A)\|A^T \hat{\lambda}^k\| \leq \tilde{\sigma}_{\min}^{-1/2}(A^T A)\left[\gamma^{-1}\mathcal{B}_1 + (2L_f + \gamma^{-1})\mathcal{B}_2 + \|\nabla \mathcal{M}_{\gamma,f}(w^1)\|\right],$$

which yields the boundedness of $\{\lambda^k\}$ by the triangle inequality. This finishes the proof.

The proof idea of claim (c) of this proposition is motivated by the proof of [68, Lemma 3.1]. Based on Proposition 4, we show the lower boundedness of the Lypunov function sequence and the boundedness of the sequence generated by MEAL. Following the similar analysis of this section, we can obtain the similar boundeness results for both iMEAL and LiMEAL.

6.2 Discussions on Related Work

When compared to these tightly related work [55,38,36,68,67,60,39], this paper provides some slightly stronger convergence results under weaker conditions. The detailed discussions and comparisons with these works are shown as follows and presented in Tables 1 and 2.

When reduced to the case of linear constraints, the proximal ALM suggested in [55] is a special case of MEAL with $\eta = 1$, and the Lipschitz continuity of certain fundamental mapping at the origin [55, pp. 100] generally implies the KŁ property of the proximal augmented Lagrangian with exponent $1/2$ at some stationary point, and thus, the linear convergence of proximal ALM can be directly yielded by Proposition

Table 2 Comparisons on convergence results of existing algorithms for the composite optimization problem (8).

Algorithm	LiMEAL (our)	PProx-PDA [36]	Prox-iALM [68]	S-prox-ALM [67]
Assumption	∇h: Lipschitz, g: weakly convex, *imp-Lip* or *imp-bound*	∇h: Lipschitz, g: convex, ∂g: bounded	∇h: Lipschitz, $g : \iota_C(x)$, C: box constraint	∇h: Lipschitz, $g : \iota_{\mathcal{P}}(x)$, $\mathcal{P}$: polyhedral set
Iteration complexity	*imp-Lip*: $o(\varepsilon^{-2})$ *imp-bound*: $O(\varepsilon^{-2})$	$O(\varepsilon^{-2})$	$O(\varepsilon^{-2})$	$O(\varepsilon^{-2})$
Global Convergence	✓ (KŁ)	–	✓ (quadratic programming)	–

1(b). Moreover, the proposed algorithms still work (in terms of convergence) for some constrained problems with nonconvex objectives and a fixed penalty parameter.

In [38], a proximal primal-dual algorithm (named *Prox-PDA*) was proposed for the linearly constrained problem (1) with $b = 0$. Prox-PDA is shown as follows:

$$(\mathrm{Prox-PDA}) \quad \begin{cases} x^{k+1} = \mathrm{argmin}_{x\in\mathbb{R}^n} \left\{ f(x) + \langle \lambda^k, Ax\rangle + \frac{\beta}{2}\|Ax\|^2 + \frac{\beta}{2}\|x - x^k\|^2_{B^TB} \right\}, \\ \lambda^{k+1} = \lambda^k + \beta A x^{k+1}, \end{cases}$$

where B is chosen such that $A^TA + B^TB \succeq \mathrm{I}_n$ (the identity matrix of size n). To achieve a $\sqrt{\varepsilon}$-accurate stationary point, the iteration complexity of Prox-PDA is $O(\varepsilon^{-1})$ under the Lipschitz differentiability of f (that is, f is differentiable and has Lipschitz gradient) and the assumption that there exists some $\underline{f} > -\infty$ and some $\delta > 0$ such that $f(x) + \frac{\delta}{2}\|Ax\|^2 \geq \underline{f}$ for any $x \in \mathbb{R}^n$. Such iteration complexity of Prox-PDA is consistent with the order of $O(\varepsilon^{-2})$ to achieve an ε-accurate stationary point. On one hand if we take $B = \mathrm{I}_n$ in Prox-PDA, then it reduces to MEAL with $\gamma = \beta^{-1}$ and $\eta = 1$. On the other hand, by our main Theorem 1(a), the iteration complexity of the order of $o(\varepsilon^{-2})$ is slightly better than that of Prox-PDA, under weaker conditions (see, Assumption 2(a)-(b)). Moreover, we established the global convergence and rate of MEAL under the KŁ inequality, while such global convergence result is missing (though obtainable) for Prox-PDA in [38].

A prox-linear variant of Prox-PDA (there dubbed *PProx-PDA*) was proposed in the recent paper [36] for the linearly constrained problem (8) with a composite objective. Besides Lipschitz differentiability of h, the nonsmooth function g is assumed to be convex with bounded subgradients. These assumptions used in [36] are stronger than ours in Assumption 3(a), (b) and (d), while the yielded iteration complexity of LiMEAL (Theorem 3(b)) is consistent with that of PProx-PDA in [36, Theorem 1]. Moreover, we establish the global convergence and rate of LiMEAL (Proposition 3), which is missing (though obtainable) for PProx-PDA.

In [60], an $O(\varepsilon^{-2})$-iteration complexity of proximal ALM was established for the constrained problem with nonlinear equality constraints, under assumptions that the objective is differentiable and its gradient is both Lipschitz continuous and bounded, and that the Jacobian of the constraints is also Lipschitz continuous and bounded and satisfies a *full-rank* property (see [60, Assumption 1]). If we reduce their setting to linear constraints, their iteration complexity is slightly worse than ours and their assumptions are stronger (of course, except for the part on nonlinear constraints).

In [68], a very related algorithm (called *Proximal Inexact Augmented Lagrangian Multiplier method*, dubbed Prox-iALM) was introduced for the following linearly constrained problem

$$\min_{x\in\mathbb{R}^n} h(x) \quad \text{subject to} \quad Ax = b,\ x \in C,$$

where C is a box constraint set. Subsequence convergence to a stationary point was established under the following assumptions: (a) the origin is in the relative interior of the set $\{Ax - b : x \in C\}$; (b) the strict complementarity condition [48] holds for the above constrained problem; (c) h is differentiable and has Lipschitz continuous gradient. Moreover, the global convergence and linear rate of this algorithm was established for the quadratic programming, in which case, the augmented Lagrangian satisfies the KŁ inequality with exponent $1/2$, by noticing the connection between Luo-Tseng error bound and KŁ inequality [42]. According to Theorem 3 and Proposition 3, the established convergence results in this paper is more general and stronger than that in [68] but under weaker assumptions. Particularly, besides the weaker assumption on h, the strict complementarity condition (b) is also removed in this paper for LiMEAL.

The algorithm studied in [68] has been recently generalized to handle the linearly constrained problem with the polyhedral set in [67] (dubbed *S-prox-ALM*). Under the Lipschitz differentiability of the objective, the iteration complexity of the order $O(\varepsilon^{-2})$ was established in [67] for the S-prox-ALM algorithm. Such iteration complexity is consistent with LiMEAL as shown in Theorem 3. Besides these major differences between this paper and [68,67], the step sizes η are more flexible for both MEAL and LiMEAL (only requiring $\eta \in (0,2)$), while the step sizes used in the algorithms in [68,67] should be sufficiently small to guarantee the convergence. Meanwhile, the Lyapunov function used in this paper is motivated by the Moreau envelope of the augmented Lagrangian, which is very different from the Lyapunov function used in [68,67]. Based on the defined Lyapunov function, our analysis is much simpler than that in [68,67].

In [39], a multiblock proximal ADMM algorithm (called *proximal ADMM-g*) was suggested for the constrained problem (38) with a multiblock composite objective, where $r_p(x) = 0, \forall x \in \mathbb{R}^n$ and $A_p = \mathrm{I}_{n_p}$ (the identity matrix of size n_p). The $O(\varepsilon^{-2})$-iteration complexity of the algorithm was established in [39] under the block-wise Lipschtz differentiability of h and the lower boundedness of the proximal functions r_i's. If we set the matrix norm used in *proximal ADMM-g* as the Euclidean norm, then the *proximal ADMM-g* algorithm proposed in [39] is a special case of mLiMEAL with $\eta = 1$ as described in (40). Following the similar analysis of LiMEAL, mLiMEAL shall permit the consistent iteration complexity of *proximal ADMM-g* in [39].

7 Numerical Experiments

In this section, we provide three experiments to show the effectiveness of the proposed algorithms. The first experiment is implemented to show the effectiveness of the propose LiMEAL for solving a specific quadratic programming problem, while it has been proved theoretically in [59, Proposition 1] that ALM with any bounded penalty parameter diverges when applied to this quadratic programming problem. In the second experiment, we consider a general quadratic programming problem with the same settings to that in [68, Sec. 6.2] and show the effectiveness of the proposed algorithm via comparing to the algorithm *Prox-iALM* recently suggested in [68]. In the third experiment, we consider a sparse regularized phase retrieval problem [43] that has not been considered in the ALM literature, to show the effectiveness of the multi-block version of MEAL, i.e., MEAD (31) via comparing to an ADMM method [59]. The codes are available at `https://github.com/JinshanZeng/MEAL`.

7.1 A Motivated Experiment

Motivated by [59, Proposition 1], we consider the following optimization problem:

$$\min_{x,y\in\mathbb{R}} x^2 - y^2, \quad \text{subject to } x = y,\ x \in [-1,1]. \tag{75}$$

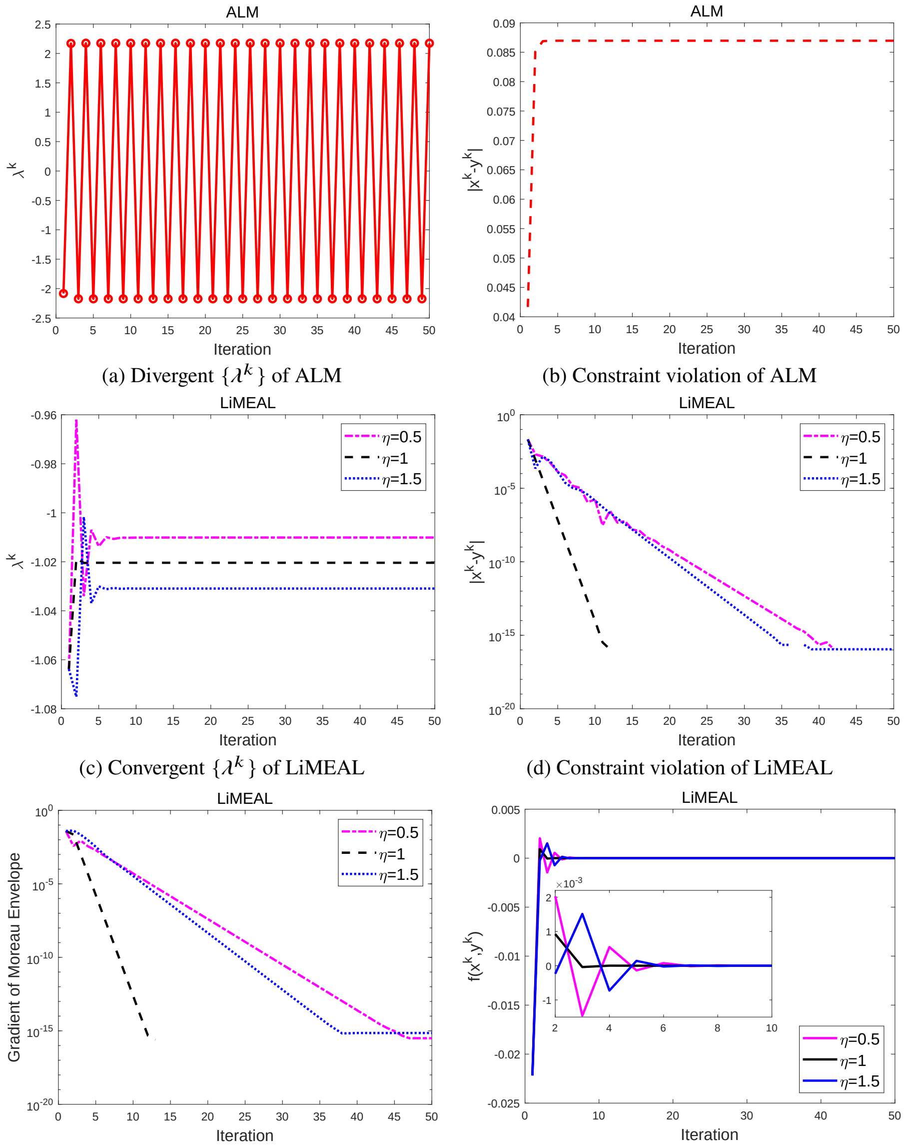
It has been shown in [59, Proposition 1] that ALM with any bounded penalty parameter β diverges when applied to solve the above problem, since the duality gap of this problem is non-zero, while according to Theorem 3 and Proposition 3, the proposed LiMEAL converges exponentially fast when applied to this problem since the augmented Lagrangian of this problem is a KŁ function with an exponent of $1/2$ (see, [68]). Specifically, for both ALM and LiMEAL, the penalty parameter β is empirically set to be 50. The proximal parameter γ used in LiMEAL is set to be $1/2$. We consider three different $\eta's$, that is, 0.5, 1, 1.5. The curves of objective $f(x^k,y^k) = (x^k)^2 - (y^k)^2$, constraint violation error $|x^k - y^k|$, multiplier sequences $\{\lambda^k\}$ and the norm of gradient of Moreau envelope defined as in (33) and used as the stationarity measure, are depicted in Fig. 1.

It can be observed from Fig. 1 that ALM diverges when applied to this problem, where the multiplier sequence $\{\lambda^k\}$ oscillates between two distinct values (Fig. 1 (a)) and the constraint violation converges to some positive value (Fig. 1 (b)), while the proposed LiMEAL converges exponentially fast (Fig. 1 (c)-(e)) and can achieve the optimal objective value 0 (Fig. 1 (f)) with different $\eta's$. This verifies the global convergence and rate results of LiMEAL established in Proposition 3. According to the detailed curves in Fig. 1(f), LiMEAL with all these three $\eta's$ can achieve the optimal objective value 0 within very few iterations, i.e., about 10 iterations. From Fig. 1(e), LiMEAL converges exponentially fast for all the concerned $\eta's$, and LiMEAL with $\eta = 1$ converges faster than the other two choices of $\eta's$ in this experiment.

7.2 Quadratic Programming

In the following, we consider the performance of proposed LiMEAL for the quadratic programming problem with box constraints, that is,

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^TQx + r^Tx \quad s.t. \quad Ax = b,\ \ell_i \le x_i \le u_i,\ i = 1,\dots,n, \tag{76}$$



(a) Divergent $\{\lambda^k\}$ of ALM (b) Constraint violation of ALM

(c) Convergent $\{\lambda^k\}$ of LiMEAL (d) Constraint violation of LiMEAL

(e) convergence rate of LiMEAL (f) objective sequence of LiMEAL

Fig. 1 Performance of ALM and LiMEAL for problem (75). It can be observed that ALM diverges while LiMEAL converges exponentially fast and can achieve the optimal objective value 0.

where $Q \in \mathbb{R}^{n\times n}$, $r \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\ell_i, u_i \in \mathbb{R}$, $i = 1, \ldots, n$. Let $\mathcal{C} := \{x : \ell_i \le x_i \le u_i, i = 1, \ldots, n\}$.

Following the update framework of LiMEAL (9) and with some easy computations, the specific updates of LiMEAL for problem (76) can be described as follows:

given (x^0, z^0, λ^0), $\gamma > 0$, $\eta \in (0,2)$ and $\beta > 0$, for $k = 0, 1, \ldots,$ run

$$\text{(LiMEAL)} \quad \begin{cases} \tilde{x}^k = (\beta A^T A + \gamma^{-1}\mathbf{I}_n)^{-1}(\gamma^{-1}z^k + \beta A^T b - r - Qx^k - A^T\lambda^k), \\ x^{k+1} = \mathrm{Proj}_C(\tilde{x}^k), \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b), \end{cases}$$

where $\mathrm{Proj}_C(x) := \mathrm{argmin}_{\bar{x}\in C} \|\bar{x} - x\|$ denotes the projection of x onto the box constraint set C with the i-th component of the projection $[\mathrm{Proj}_C(x)]_i$ being ℓ_i if $x_i \le \ell_i$, and u_i if $x_i \ge u_i$, and x_i itself otherwise, $\mathbf{I}_n$ represents an identity matrix of size n. We also take the *Prox-iALM* algorithm recently proposed in [68, Algorithm 2.2] as a competitor. Following [68, Algorithm 2.2], the update of Prox-iALM can be described as follows: given (x^0, z^0, λ^0), parameters $\beta, p, \alpha, s, \eta > 0$, for $k = 0, 1, \ldots,$ run

$$\text{(Prox}-\text{iALM)} \quad \begin{cases} \bar{x}^k = (\beta A^T A + p\mathbf{I}_n)x^k + Qx^k + A^T\lambda^k - pz^k - (\beta A^T b - r), \\ x^{k+1} = \mathrm{Proj}_C(x^k - s\bar{x}^k), \\ z^{k+1} = z^k - \eta(z^k - x^{k+1}), \\ \lambda^{k+1} = \lambda^k + \beta_k(Ax^{k+1} - b). \end{cases}$$

In particular, when $\eta = 1$, then Prox-iALM reduces to *Algorithm 2.1* in [68] (dubbed *iALM*).

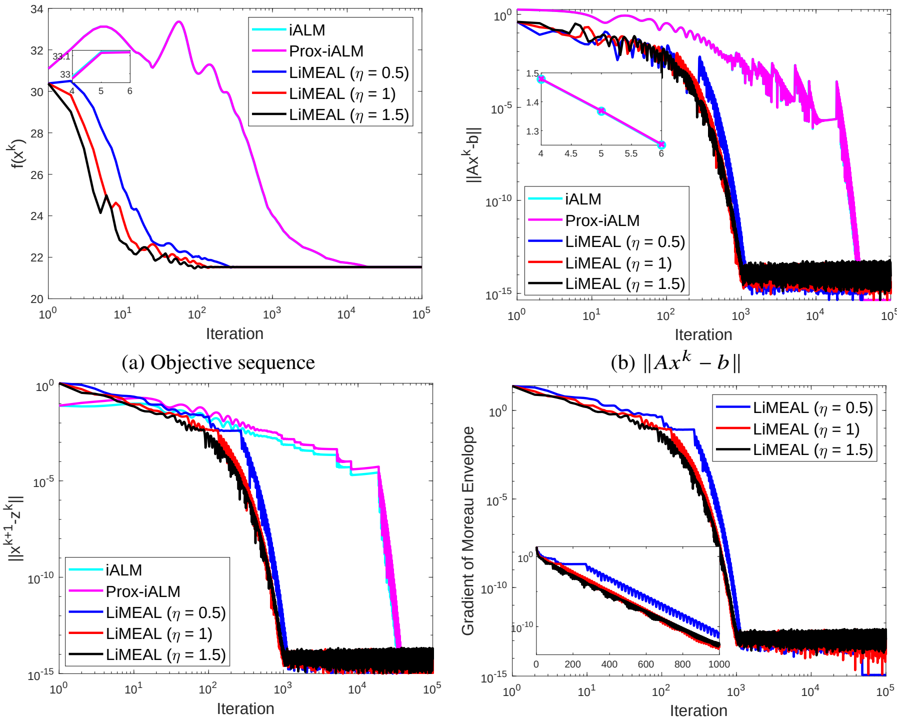
The experimental settings are similar to that in [68, Sec. 6.2]. Specifically, we set $m = 5, n = 20$. The entries of Q, A and b are generated according to the uniform distribution, and $b = A\tilde{x}$, where each entry of $\tilde{x}$ is generated according to the uniform distribution. For LiMEAL, we set $\beta = 50, \gamma = \frac{1}{2\|Q\|_2}$, and consider three $\eta's$, i.e., 0.5, 1, 1.5. For Prox-iALM, we use the similar parameter settings as in [68, Sec. 6.2], that is, $p = 2\|Q\|_2, \beta = 50, \alpha = \frac{\beta}{4}, s = \frac{1}{2(\|Q\|_2 + p + \beta\|A\|_2^2)}$. Moreover, we consider two $\eta's$, i.e., 1 and 0.5 for Prox-iALM. As pointed out before, Prox-iALM with $\eta = 1$ reduces to iALM. The curves of objective sequence, $\|Ax^k - b\|$, $\|x^{k+1} - z^k\|$ and norm of gradient of Moreau envelope are depicted in Fig. 2. From Fig. 2, the proposed LiMEAL converges faster than both iALM and Prox-iALM. Particularly, by Fig. 2(d), the proposed LiMEAL converges exponentially fast for all these three $\eta's$. This also verifies the developed theoretical results in Proposition 3(b) since the augmented Lagrangian of problem (76) is a KŁ function with an exponent 1/2 (see, [68]).

7.3 Sparse Regularized Phase Retrieval

We consider the following sparse regularized robust phase retrieval (e.g., [43]):

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} |\langle a_i, x\rangle^2 - b_i| + c\|x\|_1, \tag{77}$$

where the entries of $a_i \in \mathbb{R}^n$ $(i = 1, \ldots, m)$ are generated from the normal distribution, $b_i = \langle a_i, x^*\rangle^2 + \epsilon$ is the observation, $x^* \in \mathbb{R}^n$ is the sparse true signal with a sparsity level $s \ll n$, ϵ is some noise, $\|x\|_1 = \sum_{i=1}^n |x_i|$ is the ℓ_1-norm of x and $c > 0$ is a regularization parameter. By introducing m copies of x, i.e., $\mathbf{y} := [\mathbf{y}_1; \mathbf{y}_2, \ldots, \mathbf{y}_m]$



(a) Objective sequence (b) $\|Ax^k - b\|$

(c) $\|x^{k+1} - z^k\|$ (d) Convergence rates of LiMEAL

Fig. 2 Performance of LiMEAL and Prox-iALM for the quadratic programming problem (76).

(where $\mathbf{y}_i \in \mathbb{R}^n$), the above problem (77) can be reformulated to the following linearly constrained problem:

$$\min_{x \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^{mn}} \sum_{i=1}^{n} |\langle a_i, \mathbf{y}_i \rangle^2 - b_i| + c\|x\|_1 \quad \text{s.t.} \quad Ex + \mathbf{y} = 0, \tag{78}$$

where $E = -[-\mathbf{I}_n; \ldots; -\mathbf{I}_n] \in \mathbb{R}^{mn \times n}$ and $\mathbf{I}_n$ is the identity matrix of size n.

MEAD. Given the current iterate $(x^k, \mathbf{y}^k, u^k, \mathbf{v}^k, \lambda^k)$, the Moreau envelope alternating direction method (MEAD) for problem (78) can be described as follows.

1. (x-update) According to the update of MEAD (31), x^{k+1} is update via the following

$$x^{k+1} = \arg\min_x \left\{ c\|x\|_1 + \langle \lambda^k, Ex + \mathbf{y}^k \rangle + \frac{\beta}{2} \sum_{i=1}^{m} \|x - \mathbf{y}_i^k\|_2^2 + \frac{1}{2\gamma}\|x - u^k\|^2 \right\},$$

which has the following closed-form solution

$$x^{k+1} = \frac{1}{\beta m + \gamma^{-1}} \max\left\{0, |d^{k+1}| - c\right\} \cdot \operatorname{sign}(d^{k+1}),$$

where $d^{k+1} = \gamma^{-1} u^k + \sum_{i=1}^{m} (\lambda_i^k + \beta \mathbf{y}_i^k)$, and $\operatorname{sign}(d)$ represents the sign function of d, that is, $\operatorname{sign}(d) = 1$ if $d > 0$, and $\operatorname{sign}(d) = -1$ if $d < 0$, and $\operatorname{sign}(d) = 0$ otherwise.

2. ($\mathbf{y}$-update) Since the objective and linear constraints are separable and the same for each block of $\mathbf{y}_i$, thus we update them in a parallel way. Specifically, for $i = 1, \ldots, m$, $\mathbf{y}_i^{k+1}$ is updated according to the following

$$\mathbf{y}_i^{k+1} = \arg\min_{\mathbf{y}_i} \left\{ |\langle a_i, y_i \rangle^2 - b_i| + \langle \lambda_i^k, y_i \rangle + \frac{\beta}{2}\|x^{k+1} - y_i\|_2^2 + \frac{1}{2\gamma}\|\mathbf{v}_i^k - \mathbf{y}_i\|_2^2 \right\},$$

which has the following closed-form solution

$$\mathbf{y}_i^{k+1} = \mathrm{Prox}_{\frac{1}{\beta+\gamma^{-1}}, |\langle a_i, \cdot \rangle^2 - b_i|} \left(\frac{1}{\beta + \gamma^{-1}} (\beta x^{k+1} + \gamma^{-1}\mathbf{v}_i^k - \lambda_i^k) \right),$$

where $\mathrm{Prox}_{\delta, |\langle a, \cdot \rangle^2 - b|}(z) = \arg\min_x \left\{ |\langle a, x \rangle^2 - b| + \frac{1}{2\delta}\|x - z\|^2 \right\}$ for $\delta > 0$, $a \in \mathbb{R}^n$ and $b \geq 0$, which has the following closed form solution (see, [28, Sec. 5.1]):

$$\mathrm{Prox}_{\delta, |\langle a, \cdot \rangle^2 - b|}(z) = \arg\min_{x \in \mathcal{S}} \left\{ |\langle a, x \rangle^2 - b| + \frac{1}{2\delta}\|x - z\|^2 \right\},$$

where $\mathcal{S} := \left\{ z - \left(\frac{2\delta\langle a, z \rangle}{2\delta\|a\|^2 + 1} \right) a, z - \left(\frac{2\delta\langle a, z \rangle}{2\delta\|a\|^2 - 1} \right) a, z - \left(\frac{\langle a, z \rangle + \sqrt{b}}{\|a\|^2} \right) a, z - \left(\frac{\langle a, z \rangle - \sqrt{b}}{\|a\|^2} \right) a \right\}$.

3. ((u, v)-update) $u^{k+1} = u^k - \eta(u^k - x^{k+1})$, $\mathbf{v}^{k+1} = \mathbf{v}^k - \eta(\mathbf{v}^k - \mathbf{y}^{k+1})$.
4. (λ-update) $\lambda^{k+1} = \lambda^k + \beta(Ex^{k+1} + \mathbf{y}^{k+1})$.

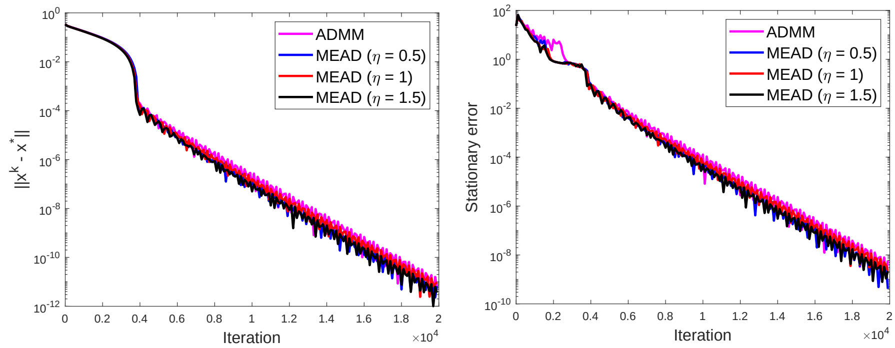
ADMM. By [59] and following the analysis similar to MEAD, the ADMM method for problem (78) can be described as follows: given the current update $(x^k, \mathbf{y}^k, \lambda^k)$,

1. (x-update) $x^{k+1} = (\beta m)^{-1} \max\{0, |\tilde{d}^{k+1} - c|\} \cdot \mathrm{sign}(\tilde{d}^{k+1})$, where $\tilde{d}^{k+1} = \sum_{i=1}^m (\lambda_i^k + \beta \mathbf{y}_i^k)$.
2. ($\mathbf{y}$-update) For $i = 1, \ldots, m$, $\mathbf{y}_i^{k+1} = \mathrm{Prox}_{\beta^{-1}, |\langle a_i, \cdot \rangle^2 - b_i|}(x^{k+1} - \beta^{-1}\lambda_i^k)$.
3. (λ-update) $\lambda^{k+1} = \lambda^k + \beta(Ex^{k+1} + \mathbf{y}^{k+1})$.

In this experiment, we set $n = 300$, $m = 100$, the sparsity level of true signal x^* is 10, where nonzero entries of x^* are generated from the normal distribution. For MEAD, $\gamma = 1/2$ and three different $\eta's$, i.e., 0.5, 1, 1.5 are considered. For both MEAD and ADMM, $\beta = 100$ and the regularization parameter c is tuned via a hand-optimal way. The curves of recovery error (i.e., $\|x^k - x^*\|$) and stationary error (measured by $\sqrt{\|\sum_{i=1}^m (\mathbf{y}_i^k - \mathbf{y}_i^{k+1})\|^2 + \|\lambda^{k+1} - \lambda^k\|^2}$) are presented in Fig. 3. From Fig. 3(a), both the suggested MEAD and ADMM can recover the sparse signal with a very high precision, while from Fig. 3(b), the rates of convergence of MEAD with three different η's are linear in terms of the defined stationary error. When compared to ADMM, the performance of the proposed MEAD is slightly beteer than ADMM.

8 Conclusion

This paper suggests a Moreau envelope augmented Lagrangian (MEAL) method for the linearly constrained weakly convex optimization problem. By leveraging the *implicit smoothing property* of Moreau envelope, the proposed MEAL generalizes the ALM and proximal ALM to the nonconvex and nonsmooth case. To yield an ε-accurate first-order stationary point, the iteration complexity of MEAL is $o(\varepsilon^{-2})$ under



(a) Recovery error (b) Stationary error

Fig. 3 Performance of MEAD and ADMM for the sparse regularized phase retrieval problem (77).

the *implicit Lipschitz subgradient* assumption and $O(\varepsilon^{-2})$ under the *implicit bounded subgradient* assumption. The global convergence and rate of MEAL is also established under the further Kurdyka-Łojasiewicz inequality. Moreover, an inexact variant (called *iMEAL*) and a prox-linear variant (called *LiMEAL*) for the composite objective case are suggested and analyzed for different practical settings. The convergence results established in this paper for MEAL and its variants are generally stronger than the existing ones, but under weaker assumptions.

One future direction of this paper is to get rid of the *implicit Lipschitz subgradient* and *implicit bounded subgradient* assumptions, which in some extent limit the applications of the suggested algorithms, though these two assumptions are respectively weaker than the *Lipschitz differentiable* and *bounded subgradient* assumptions commonly used in the literature. Another direction of this paper is to generalize this work to the constrained problem with nonlinear constraints. One possible application of our study is robustness and convergence of stochastic gradient descent in training parameters of structured deep neural networks such as deep convolutional neural networks [69], where linear constraints can be used to impose convolutional structures. We leave them in our future work.

References

1. Andreani, R., Birgin, E.G., Martinez, J.M., Schuverdt, M.L.: On augmented lagrangian methods with general lower-level constraints. SIAM J. Optim. **18**(4), 1286–1309 (2007)
2. Andreani, R., Birgin, E.G., Martinez, J.M., Schuverdt, M.L.: Augmented lagrangian methods under the constant positive linear dependence constraint qualification. Math. Program. **111**, 5–32 (2008)
3. Andreani, R., Birgin, E.G., Martinez, J.M., Schuverdt, M.L.: Second-order negative-curvature methods for box-constrained and general constrained optimization. Comput. Optim. Appl. **45**(2), 209–236 (2010)
4. Andreani, R., Fazzio, N., Schuverdt, M.L., Secchin, L.: A sequential optimality condition related to the quasi-normality constraint qualification and its algorithmic consequences. SIAM J. Optim. **29**(1), 743–766 (2019)
5. Andreani, R., Secchin, L., Silva, P.: Convergence properties of a second order augmented lagrangian method for mathematical programs with complementarity constraints. SIAM J. Optim. **28**(3), 2574–2600 (2018)

6. Armand, P., Omheni, R.: A globally and quadratically convergent primal-dual augmented lagrangian algorithm for equality constrained optimization. Optim. Methods Softw. **32**(1), 1–21 (2017)
7. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**, 5–16 (2009)
8. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi–algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. Math. Program. **137**, 91–219 (2013)
9. Bertsekas, D.P.: Convergence rate of penalty and multiplier methods. In: Proc. IEEE Conf. on Decision and Control), pp. 260–264. SanDiego, California (1973)
10. Bertsekas, D.P.: On penalty and multiplier methods for constrained minimization. SIAM J. Control Optim. **14**(2), 216–235 (1976)
11. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, London (1982)
12. Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. Math. Program. **149**(1), 301–327 (2005)
13. Birgin, E.G., Castillo, R., Martinez, J.M.: Numerical comparison of augmented lagrangian algorithms for nonconvex problems. Comput. Optim. Appl. **31**, 31–56 (2005)
14. Birgin, E.G., Floudas, C.A., Martinez, J.M.: Global minimization using an augmented lagrangian method with variable lower-level constraints. Math. Program. **125**, 139–162 (2010)
15. Birgin, E.G., Floudas, C.A., Martinez, J.M.: The boundedness of penalty parameters in an augmented lagrangian method with constrained subproblems. Optim. Methods Softw. **27**(6), 1001–1024 (2012)
16. Birgin, E.G., Haeser, G., Ramos, A.: Augmented lagrangians with constrained subproblems and convergence to second-order stationary points. Comput. Optim. Appl. **69**(1), 51–75 (2018)
17. Birgin, E.G., Martinez, J.M.: Practical Augmented Lagrangian Methods for Constrained Optimization, vol. vol. 10. SIAM, Philadelphia (2014)
18. Birgin, E.G., Martinez, J.M.: Complexity and performance of an augmented lagrangian algorithm. Optim. Methods Softw. (2020)
19. Bochnak, J., Coste, M., Roy, M.F.: Real algebraic geometry, vol. 36. Springer Science & Business Media, Berlin (1998)
20. Bolte, J., Daniilidis, A., Lewis, A.: The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
21. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clark subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
22. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1), 459–494 (2014)
23. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge, UK (2004)
24. Conn, A.R., Gould, N.I.M., Startenaer, A., Toint, P.L.: Convergence properties of an augmented lagrangian algorithm for optimization with a combination of general equality and linear constraints. SIAM J. Optim. **6**, 674–703 (1996)
25. Conn, A.R., Gould, N.I.M., Toint, P.L.: A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. SIAM J. Numer. Anal. **28**, 545–572 (1991)
26. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust-Region Methods. SIAM, Philadelphia (2000)
27. Curtis, F.E., Jiang, H., Robinson, D.P.: An adaptive augmented lagrangian method for large-scale constrained optimization. Math. Program. **152**(1), 201–245 (2015)
28. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
29. Deng, W., Lai, M.J., Peng, Z., Yin, W.: Parallel multi-block admm with $o(1/k)$ convergence. J. Sci. Comput. **71**, 712–736 (2017)
30. Drusvyatskiy, D.: The proximal point method revisited. SIAG/OPT Views and News **26**, 1–8 (2018)
31. Drusvyatskiy, D., Paquette, C.: Efficiency of minimizing compositions of convex functions and smooth maps. Math. Program. **178**, 503–558 (2019)
32. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc **96**, 1348–1360 (2001)
33. Fernadez, D., Solodov, M.V.: Local convergence of exact and inexact augmented lagrangian methods under the second-order sufficient optimality condition. SIAM J. Optim. **22**(2), 384–407 (2012)
34. Grapiglia, G.N., Yuan, Y.X.: On the complexity of an augmented Lagrangian method for nonconvex optimization. ArXiv e-prints (2019)

35. Haeser, G., Liu, H., Ye, Y.: Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. Math. Program. **178**, 263–299 (2019)
36. Hajinezhad, D., Hong, M.: Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. Math. Program. **176**, 207–245 (2019)
37. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
38. Hong, M., Hajinezhad, D., Zhao, M.M.: Prox-pda: The proximal primal-dual algorithm for fast dostributed nonconvex optimization and learning over networks. In: Proc. of the 34th International Conference on Machine Learning (ICML), pp. 1529–1538. Sydney, Australia (2017)
39. Jiang, B., Lin, T., Ma, S., Zhang, S.: Structured nonconvex and nonsmooth otpmization: algorithms and iteration complexity analysis. Comput. Optim. Appl. **72**(1), 115–157 (2019)
40. Krantz, S., Parks, H.R.: A Primer of Real Analytic Functions (2nd Edition). Birkhauser, Basel, Switzerland (2002)
41. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier **48**(3), 769–783 (1998)
42. Li, G., Pong, T.K.: Calculus of the exponent of kurdyka-łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. **18**, 1199–1232 (2018)
43. Li, Q., Huang, L., Liu, W., Sun, W., Zhang, P.: A fast gradient-based iterative algorithm for undersampled phase retrieval. IEEE Trans. Aerosp. Electron. Syst. **54**, 2086–2090 (2018)
44. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. In: Les Équations aux dérivées partielles. Éditions du centre National de la Recherche Scientifique, Paris pp. 87–89 (1963)
45. Łojasiewicz, S.: Sur la geometrie semi-et sous-analytique. Annales de l'institut Fourier **43**(5), 1575–1595 (1993)
46. Mordukhovich, B.S.: Variational analysis and generalized differentiation I: Basic Theory. Springer-Verlag, New York (2006)
47. Moreau, J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299 (1965)
48. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer-Verlag, New York (1999)
49. Nouiehed, M., Lee, J.D., Razaviyayn, M.: Convergence to second-order stationary for constrained non-convex optimization. ArXiv e-prints (2018)
50. Nurminskii, E.A.: The quasigradient method for the solving of the nonlinear programming problems. Cybernetics **9**, 145–150 (1973)
51. O'Neill, M., Wright, S.J.: A log-barrier newton-cg method for bound constrained optimization with complexity guarantees. IMA J. Numer. Anal. **00**, 1–38 (2020)
52. Polyak, B.T., Tretyakov, N.V.: The method of penalty bounds for constrained extremum problems. Zh. Vych Mat i Mat. Fiz, 13:34-46 = U.S.S.R. Computational Mathematics and Mathmatical Physics **13**, 42–58 (1973)
53. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. in Optimization, R. Fletcher, ed. Academic Press, London pp. 283–298 (1969)
54. Rockafellar, R.T.: The multiplier method of hestenes and powell applied to convex programming. J. Optim. Theory Appl. **12**, 555–562 (1973)
55. Rockafellar, R.T.: Augmented lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976)
56. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer-Verlag, New York (1997)
57. Shiota, M.: Geometry of Subanalytic and Semialgebraic Sets (Progress in Mathematics). Birkhauser, Basel, Switzerland (1997)
58. Tretykov, N.Y.: The method of penalty estimates of convex programming. Economics and Mathematical Methods (Russian) **9**, 525–540 (1973)
59. Wang, Y., Yin, W., Zeng, J.: Global convergence of admm in nonconvex nonsmooth optimization. J. Sci. Comput. **78**, 29–63 (2019)
60. Xie, Y., Wright, S.J.: Complexity of proximal augmented Lagrangian for nonconvex optimalization with nonlinear equality constraints. ArXiv e-prints (2019)
61. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimizaton with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**(3), 1758–1789 (2013)
62. Yu, P., Li, G., Pong, T.: Kurdyka-łojasiewicz exponent via inf-projection. Found. Comput. Math. (2021). DOI https://doi.org/10.1007/s10208-021-09528-6
63. Zeng, J., Lau, T.T.K., Lin, S.B., Yao, Y.: Global convergence of block coordinate descent in deep learning. In: Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, California, PMLR 97 (2019)

64. Zeng, J., Lin, S.B., Yao, Y., Zhou, D.X.: On ADMM in deep learning: Convergence and saturation-avoidance. ArXiv e-prints (2019)
65. Zeng, J., Yin, W.: On nonconvex descentralized gradient descent. IEEE Trans. Signal Process. **66**(11), 2834–2848 (2018)
66. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
67. Zhang, J., Luo, Z.Q.: A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. ArXiv e-prints (2020)
68. Zhang, J., Luo, Z.Q.: A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. SIAM J. Optim. **30**(3), 2272–2302 (2020)
69. Zhou, D.X.: Universality of deep convolutional neural networks. Appl. Comput. Harmonic Anal. **48**, 787–794 (2020)