SIAM J. Numer. Anal.
Vol. 55, No. 1, pp. 168–193

© 2017 Society for Industrial and Applied Mathematics

AN AUGMENTED LAGRANGIAN METHOD FOR NON-LIPSCHITZ NONCONVEX PROGRAMMING*

XIAOJUN CHEN[†], LEI GUO[‡], ZHAOSONG LU[§], AND JANE J. YE[¶]

Abstract. We consider a class of constrained optimization problems where the objective function is a sum of a smooth function and a nonconvex non-Lipschitz function. Many problems in sparse portfolio selection, edge preserving image restoration, and signal processing can be modelled in this form. First, we propose the concept of the Karush–Kuhn–Tucker (KKT) stationary condition for the non-Lipschitz problem and show that it is necessary for optimality under a constraint qualification called the relaxed constant positive linear dependence (RCPLD) condition, which is weaker than the Mangasarian–Fromovitz constraint qualification and holds automatically if all the constraint functions are affine. Then we propose an augmented Lagrangian (AL) method in which the augmented Lagrangian subproblems are solved by a nonmonotone proximal gradient method. Under the assumption that a feasible point is known, we show that any accumulation point of the sequence generated by our method must be a feasible point. Moreover, if RCPLD holds at such an accumulation point, then it is a KKT point of the original problem. Finally, we conduct numerical experiments to compare the performance of our AL method and the interior point (IP) method for solving two sparse portfolio selection models. The numerical results demonstrate that our method is not only comparable to the IP method in terms of solution quality, but also substantially faster than the IP method.

Key words. non-Lipschitz programming, sparse optimization, augmented Lagrangian method

AMS subject classifications. 90C26, 90C30, 90C59

DOI. 10.1137/15M1052834

1. Introduction.

In this paper, we consider the following non-Lipschitz nonlinear programming problem:

$$\begin{aligned} &\min \quad f(x)+\Phi(x) \\ &\text{s.t.} \quad c(x)=0,\ d(x)\le 0, \end{aligned} \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$, and $d : \mathbb{R}^n \to \mathbb{R}^p$ are continuously differentiable functions, and $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a lower semicontinuous function (possibly non-Lipschitz).

Problem (1.1) has an intriguing property in that its local minimizers are often sparse for a certain class of non-Lipschitz functions Φ with $\Phi(0) = 0$ (see, e.g., [15]). In fact, it has been shown in [13] that the magnitude of all nonzero entries of local minimizers of problem (1.1) is greater than a certain positive number under some

---

*Received by the editors December 14, 2015; accepted for publication (in revised form) November 15, 2016; published electronically January 24, 2017.
http://www.siam.org/journals/sinum/55-1/M105283.html

Funding: The work of the first author was supported in part by Hong Kong Research Grants Council PolyU5002/13p. The work of the second author was supported by NSFC grant (11401379) and the China Postdoctoral Science Foundation (2015T80428). The work of the third and fourth authors was supported in part by NSERC.

[†]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maxjchen@polyu.edu.hk).

[‡]Sino-US Global Logistics Institute, Shanghai Jiao Tong University, Shanghai 200030, China (guolayne@sjtu.edu.cn).

[§]Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (zhaosong@sfu.ca).

[¶]Department of Mathematics and Statistics, University of Victoria, Victoria, BC, V8W 2Y2, Canada (janeye@uvic.ca).

suitable conditions on the objective and constraint functions. Due to this property, problem (1.1) has been widely used to find a sparse solution in contexts such as sparse regression [32], sparse feature selection in machine learning [29], edge preserving image restoration [30], compressed sensing in signal processing [9, 22], and joint power and admission control [24]. It has also found applications in sparse portfolio selection [8, 12].

The augmented Lagrangian (AL) method and its various variants are a well-known class of optimization methods for solving constrained optimization problems that have been studied for more than four decades (see, e.g., [3]). Very recently, Curtis, Jiang, and Robinson [17] proposed an adaptive AL method for solving problem (1.1), where $\Phi(x) \equiv 0$, $d(x) \leq 0$ is a box constraint and all functions are smooth. The adaptive AL method has a novel adaptive updating scheme for the penalty parameter which greatly improves the overall performance of the algorithm. However, as observed from its convergence analysis, this method may converge to an infeasible point. This pathological behavior also exists in most of the existing AL methods in the literature.

To remedy the aforementioned pathological behavior of AL methods, Lu and Zhang [27] proposed an AL method for solving nonlinear programming problems where the objective function is a sum of a smooth term and a nonsmooth *convex* term, and established global convergence under Robinson's constraint qualification which reduces to the Mangasarian–Fromovitz constraint qualification (MFCQ) in the setting of problem (1.1). Their method differs from the classical AL methods in that the values of the AL function along the solution sequence generated by the method are bounded from above, and moreover, the magnitude of penalty parameters outgrows that of Lagrange multipliers. It is noteworthy that the results of [27] cannot be directly applied to problem (1.1) since the regularization function Φ is assumed to be convex and hence locally Lipschitz continuous in their problem setting.

In this paper, we propose an AL method for solving problem (1.1) that is considerably more general than the problems studied in [17, 27]. The non-Lipschitzness of the regularization term Φ creates some challenges and requires some special treatment when studying the necessary optimality conditions of problem (1.1) and the convergence of the proposed AL method. This is reflected in the horizon subdifferential of the non-Lipschitz term in the statement of the Fritz John type necessary optimality condition for non-Lipschitz optimization problems, which was first shown in Mordukhovich [28, Theorem 1(b)] and was reproved by Borwein, Treiman, and Zhu in [7, Corollary 2.6]. For smooth nonlinear programs, the Fritz John condition is equivalent to a Karush–Kuhn–Tucker (KKT) condition under MFCQ. However, MFCQ may be restrictive for some applications. For example, when all of the constraint functions are affine, MFCQ may not necessarily hold. Recently, Andreani et al. [1] proposed a constraint qualification that is called the relaxed constant positive linear dependence (RCPLD) condition, which is weaker than MFCQ and holds automatically when all constraints are affine.

The main purposes of this paper are to derive necessary optimality conditions of problem (1.1), propose an AL method for problem (1.1), and establish its convergence under RCPLD and a suitable condition called the basic qualification (BQ), which is used to handle the non-Lipschitzness of the objective function. We summarize our main contributions as follows:

- We derive a KKT necessary optimality condition (Theorem 2.1) for problem (1.1) under RCPLD and BQ. We also derive a KKT necessary optimality condition (Theorem 2.2) for some special cases of problem (1.1) under RCPLD only. Such special cases are extensions of the linearly constrained problems in

[4, 12, 25] to nonlinearly constrained problems. To the best of our knowledge, no such optimality conditions are available in the literature for non-Lipschitz nonconvex programming problems.

- We propose a new AL method for problem (1.1). We show in Theorem 3.1 that any accumulation point of the sequence generated by the AL method is a KKT point of problem (1.1) under RCPLD and BQ. Moreover, we establish in Theorem 3.2 the convergence result for some special cases of problem (1.1) under RCPLD only. These convergence results are new even for the case where the regularization function Φ is locally Lipschitz continuous.
- We conduct numerical experiments to compare the performance of our AL method and an interior point (IP) method for solving sparse portfolio selection models. The numerical results demonstrate that our method is not only comparable to the IP method in terms of solution quality, but also much faster than the IP method.

The rest of this paper is organized as follows. In section 2 we derive necessary optimality conditions for problem (1.1). In section 3 we propose an AL method for solving problem (1.1) and establish its global convergence. In section 4 we present numerical results of the AL method for solving three sparse portfolio selection models.

1.1. Notation and terminology. For any vector x, let x_i denote the ith entry of x, let $x_+ := \max\{x, 0\}$ be the nonnegative part of x, let $\mathrm{Diag}(x)$ be the diagonal matrix whose ith diagonal entry is x_i, and let $\|x\|_q := \left(\sum_{i=1}^n |x_i|^q\right)^{1/q}$ for any $q > 0$. Let $\|\cdot\|$ denote the Euclidean norm of a vector or the induced norm of a matrix. Let e_i be the ith standard unit vector, whose dimension shall be clear from the context. Given a point $x \in \mathbb{R}^n$ and $\delta > 0$, $\mathcal{B}_\delta(x)$ denotes a closed ball centered at x with radius δ. Given an index set $\mathcal{I} \subseteq \{1, \ldots, n\}$, $A_\mathcal{I}$ denotes the submatrix of A formed by its columns indexed by $\mathcal{I}$, and $x_\mathcal{I}$ denotes the subvector of x indexed by $\mathcal{I}$. We denote by A^T the transpose of matrix A. Let $\mathrm{dist}(x, \mathcal{D})$ denote the Euclidean distance from a point x to a closed set $\mathcal{D}$. Given a mapping $\psi : \mathbb{R}^n \to \mathbb{R}^m$ and $x \in \mathbb{R}^n$, $\nabla\psi(x) \in \mathbb{R}^{n\times m}$ stands for the transposed Jacobian of ψ at x.

We recall from [31, Definition 8.3] that for a lower semicontinuous function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ with $\overline{\mathbb{R}} := [-\infty, \infty]$ and a point $x \in \mathbb{R}^n$ where $\phi(x)$ is finite, the limiting and horizon subdifferentials are defined, respectively, as

$$\partial\phi(x) := \left\{ v \;\middle|\; \exists x^k \xrightarrow{\phi} x,\ v^k \to v \ \text{ with } \liminf_{z\to x^k} \frac{\phi(z) - \phi(x^k) - \langle v^k, z - x^k\rangle}{\|z - x^k\|} \geq 0\ \forall k \right\},$$

$$\partial^\infty\phi(x) := \left\{ v \;\middle|\; \exists x^k \xrightarrow{\phi} x, t_k v^k \to v, t_k \downarrow 0 \ \text{ with } \liminf_{z\to x^k} \frac{\phi(z) - \phi(x^k) - \langle v^k, z - x^k\rangle}{\|z - x^k\|} \geq 0\ \forall k \right\},$$

where $t_k \downarrow 0$ means $t_k > 0$ and $t_k \to 0$, and $x^k \xrightarrow{\phi} x$ means that $x^k \to x$ and $\phi(x^k) \to \phi(x)$. Moreover, if ϕ is convex, the limiting subdifferential coincides with the classical subdifferential in convex analysis [31, Proposition 8.12]. Furthermore, for a continuously differentiable function ϕ, we simply have $\partial\phi(x) = \{\nabla\phi(x)\}$ [31, Exercise 8.8(b)]. In addition, it follows from [31, Theorem 9.13] that ϕ is Lipschitz continuous at x if and only if $\partial^\infty\phi(x) = \{0\}$. Recall from [31, Definition 5.4] that a set-valued mapping $\Psi : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be outer semicontinuous at $x \in \mathbb{R}^n$ if

$$\left\{ v \mid \exists x^k \to x,\ v^k \to v\ , v^k \in \Psi(x^k) \right\} \subseteq \Psi(x).$$

The indicator function of a set $\mathcal{D}$ is denoted by $\delta_\mathcal{D}$, and $\mathcal{N}_\mathcal{D}(x) := \partial\delta_\mathcal{D}(x)$ is the limiting normal cone at $x \in \mathcal{D}$. If $\mathcal{D}$ is a convex set, then $\mathcal{N}_\mathcal{D}(x)$ coincides with the classical normal cone in the convex analysis. It is well known that the limiting normal cone mapping $\mathcal{N}_{\mathcal{D}}$, the limiting subdifferential mapping $\partial\phi$, and the horizon subdifferential mapping $\partial^\infty\phi$ are all outer semicontinuous everywhere (see, e.g., [31, Propositions 6.6 and 8.7]).

2. Necessary optimality conditions.

In this section, we derive constraint qualifications under which a local minimizer of problem (1.1) satisfies the KKT necessary optimality conditions, defined as follows.

Definition 2.1. *Let $\mathcal{F}$ be the feasible region of problem* (1.1). *We say that a point $x^* \in \mathbb{R}^n$ is a KKT point of problem* (1.1) *provided that $x^* \in \mathcal{F}$, and there exist $\mu \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^p_+$ such that*

$$0 \in \nabla f(x^*) + \partial\Phi(x^*) + \nabla c(x^*)\mu + \nabla d(x^*)\nu, \quad \nu_i d_i(x^*) = 0, \quad i = 1, \ldots, p.$$

Notice that Φ is possibly non-Lipschitz and $\mathcal{F}$ is possibly nonconvex. To ensure that a local minimizer x^* of problem (1.1) is a KKT point, one generally needs to consider not only the constraint qualification at x^*, but also the relation between the horizon subdifferential of the objective function and the limiting normal cone of $\mathcal{F}$ at x^*.

A standard constraint qualification for smooth nonlinear programs is MFCQ. Recall that MFCQ holds at $x^* \in \mathcal{F}$ if the gradients $\{\nabla c_1(x^*), \ldots, \nabla c_m(x^*)\}$ are linearly independent, and there exists a vector $v \in \mathbb{R}^n$ such that

$$\nabla c_i(x^*)^T v = 0, \quad i = 1, \ldots, m, \quad \nabla d_i(x^*)^T v < 0, \quad i \in \mathcal{I}_d(x^*),$$

where $\mathcal{I}_d(x^*) := \{i \mid d_i(x^*) = 0\}$ denotes the set of active inequality constraints at x^*. It is well known that MFCQ is equivalent to the positive linear independence of the gradient vectors, i.e., there is no $\mu \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^p_+$ not all zero such that

$$\nabla c(x^*)\mu + \nabla d(x^*)\nu = 0, \quad \nu_i d_i(x^*) = 0, \quad i = 1, \ldots, p.$$

MFCQ may be relatively restrictive for some applications. For example, it does not necessarily hold when the constraints are all affine, even though it is known that having affine constraints is itself a constraint qualification. Recently, Andreani et al. [1] introduced a weaker constraint qualification called the relaxed constant positive linear dependence (RCPLD) condition.

Definition 2.2 (see [1, Definition 4]). *Let $x^* \in \mathcal{F}$, and let $\mathcal{I} \subseteq \{1, \ldots, m\}$ be such that $\{\nabla c_i(x^*) \mid i \in \mathcal{I}\}$ is a basis for* span$\{\nabla c_i(x^*) \mid i = 1, \ldots, m\}$. *We say that RCPLD holds for the system $c(x) = 0, d(x) \le 0$ at x^* if there exists $\delta > 0$ such that*

- *$\{\nabla c_i(x) \mid i = 1, \ldots, m\}$ has the same rank for each $x \in \mathcal{B}_\delta(x^*)$;*
- *for each $\mathcal{J} \subseteq \mathcal{I}_d(x^*)$, if $(\{\nabla c_i(x^*) | i \in \mathcal{I}\}, \{\nabla d_j(x^*) | j \in \mathcal{J}\})$ are positively linearly dependent, i.e., there exist $\{\alpha_i \in \mathbb{R} \mid i \in \mathcal{I}\}$ and $\{\beta_j \ge 0 \mid j \in \mathcal{J}\}$ not all zero such that $\sum_{i\in\mathcal{I}} \alpha_i \nabla c_i(x^*) + \sum_{j\in\mathcal{J}} \beta_j \nabla d_j(x^*) = 0$, then $\{\nabla c_i(x), \nabla d_j(x) | i \in \mathcal{I}, j \in \mathcal{J}\}$ is linearly dependent for each $x \in \mathcal{B}_\delta(x^*)$.*

Note that the definition of RCPLD is independent of the choice of $\mathcal{I}$. In addition, RCPLD is stable in the sense that if it holds at a given feasible point, then it must hold at every feasible point in some neighborhood of that point. For ease of reference, we now state this property below, which will be used in the convergence analysis of our method proposed in the next section.

Proposition 2.1 (see [1, Theorem 4]). *If RCPLD holds at $x^* \in \mathcal{F}$, then there exists $\delta > 0$ such that RCPLD holds at any $x \in \mathcal{B}_\delta(x^*) \cap \mathcal{F}$.*

The following result shows that under the RCPLD condition, the normal cone can be represented as a finitely generated cone.

Proposition 2.2. *If RCPLD holds for the system* $c(x) = 0, d(x) \leq 0$ *at* $x^* \in \mathcal{F}$, *then*

$$\mathcal{N}_{\mathcal{F}}(x^*) = \left\{\nabla c(x^*)\mu + \sum_{i\in\mathcal{I}^*} \nu_i \nabla d_i(x^*) \mid \mu \in \mathbb{R}^m, \nu_i \geq 0 \ i \in \mathcal{I}^*\right\},$$

where $\mathcal{I}^* = \mathcal{I}_d(x^*)$.

Proof. On the one hand, by [23, Theorem 3.2], under RCPLD we have

$$\mathcal{N}_{\mathcal{F}}(x^*) \subseteq \left\{\nabla c(x^*)\mu + \sum_{i\in\mathcal{I}^*} \nu_i \nabla d_i(x^*) \mid \mu \in \mathbb{R}^m, \nu_i \geq 0 \ i \in \mathcal{I}^*\right\}.$$

On the other hand, by [31, Theorem 6.14], since all constraint functions are continuously differentiable, we have

$$\mathcal{N}_{\mathcal{F}}(x^*) \supseteq \left\{\nabla c(x^*)\mu + \sum_{i\in\mathcal{I}^*} \nu_i \nabla d_i(x^*) \mid \mu \in \mathbb{R}^m, \nu_i \geq 0 \ i \in \mathcal{I}^*\right\}.$$

The desired result follows immediately. The proof is now complete. ☐

Since the objective function of problem (1.1) is not locally Lipschitz continuous, the Fritz John necessary optimality condition involves the horizon subdifferential of the non-Lipschitz term (see [7, Example 2.8]). Hence a local minimizer of problem (1.1) may not be a KKT point under RCPLD only. To ensure a local minimizer is a KKT point, we also need the following basic qualification.

Definition 2.3. *We say that the BQ holds at* $x^* \in \mathcal{F}$ *if*

$$-\partial^\infty \Phi(x^*) \cap \mathcal{N}_{\mathcal{F}}(x^*) = \{0\}. \tag{2.1}$$

Obviously, BQ (2.1) holds automatically if Φ is locally Lipschitz continuous at x^*, or Φ is non-Lipschitz continuous at an interior point x^* of $\mathcal{F}$, since $\partial^\infty \Phi(x^*) = \{0\}$ in the first case and $\mathcal{N}_{\mathcal{F}}(x^*) = \{0\}$ in the second case. In the following, we show that the KKT condition is necessary for optimality under BQ and RCPLD.

Theorem 2.1. *Let* x^* *be a local minimizer of problem* (1.1). *If BQ* (2.1) *and RCPLD hold at* x^*, *then* x^* *is a KKT point of problem* (1.1).

Proof. Using the indicator function $\delta_{\mathcal{F}}$, problem (1.1) can be equivalently rewritten as

$$\min_x f(x) + \Phi(x) + \delta_{\mathcal{F}}(x).$$

It follows from Fermat's rule (see, e.g., [31, Theorem 10.1]) that

$$0 \in \nabla f(x^*) + \partial(\Phi + \delta_{\mathcal{F}})(x^*). \tag{2.2}$$

Notice that $\partial\delta_{\mathcal{F}}(x^*) = \partial^\infty\delta_{\mathcal{F}}(x^*) = \mathcal{N}_{\mathcal{F}}(x^*)$. By this and the assumption that BQ (2.1) holds at x^*, one has $-\partial^\infty\Phi(x^*) \cap \partial^\infty\delta_{\mathcal{F}}(x^*) = \{0\}$. Using this relation, $\partial\delta_{\mathcal{F}}(x^*) = \mathcal{N}_{\mathcal{F}}(x^*)$, and the sum rule of the limiting subdifferential (see, e.g., [31, Corollary 10.9]), we have

$$\partial(\Phi + \delta_{\mathcal{F}})(x^*) \subseteq \partial\Phi(x^*) + \partial\delta_{\mathcal{F}}(x^*) = \partial\Phi(x^*) + \mathcal{N}_{\mathcal{F}}(x^*),$$

which together with (2.2) yields $0 \in \nabla f(x^*) + \partial\Phi(x^*) + \mathcal{N}_{\mathcal{F}}(x^*)$. Since RCPLD holds at x^*, it follows from Proposition 2.2 that

$$\mathcal{N}_{\mathcal{F}}(x^*) = \left\{ \nabla c(x^*)\mu + \sum_{i\in\mathcal{I}^*} \nu_i \nabla d_i(x^*) \mid \mu \in \mathbb{R}^m, \nu_i \geq 0 \ i \in \mathcal{I}^* \right\}, \quad \mathcal{I}^* = \mathcal{I}_d(x^*).$$

Combining the last two relations implies that x^* is a KKT point. ∎

Now we show that if BQ (2.1) holds at a given point, then it must hold at any point in some neighborhood of that point. This result will be useful for proving the convergence of our proposed method in the next section.

PROPOSITION 2.3. *For any lower semicontinuous function* $\Psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *and closed set* $\mathcal{D}$*, if* $-\partial^\infty\Psi(x^*) \cap \mathcal{N}_{\mathcal{D}}(x^*) = \{0\}$*, then there exists* $\delta > 0$ *such that* $-\partial^\infty\Psi(x) \cap \mathcal{N}_{\mathcal{D}}(x) = \{0\}$ *for any* $x \in \mathcal{B}_\delta(x^*) \cap \mathcal{D}$*.*

Proof. Suppose for the sake of contradiction that the conclusion does not hold. Then there exists a sequence $\{x^k\} \subseteq \mathcal{D}$ converging to x^* and $v_k \neq 0$ such that $v^k \in -\partial^\infty\Psi(x^k) \cap \mathcal{N}_{\mathcal{D}}(x^k)$. Since $-\partial^\infty\Psi(x^k) \cap \mathcal{N}_{\mathcal{D}}(x^k)$ is a cone, it then follows that

$$\omega^k := \frac{v^k}{\|v^k\|} \in -\partial^\infty\Psi(x^k) \cap \mathcal{N}_{\mathcal{D}}(x^k). \tag{2.3}$$

Without loss of generality, we assume that $\omega^k \to \omega^*$ with $\|\omega^*\| = 1$ as $k \to \infty$. In addition, taking limits on both sides of (2.3) as $k \to \infty$, it follows from the outer semicontinuity of the horizon subdifferential and the limiting normal cone that $\omega^* \in -\partial^\infty\Psi(x^*) \cap \mathcal{N}_{\mathcal{D}}(x^*)$. This contradicts the assumption that $-\partial^\infty\Psi(x^*) \cap \mathcal{N}_{\mathcal{D}}(x^*) = \{0\}$. The proof is now complete. ∎

Recently, necessary optimality conditions of optimization problems with special non-Lipschitz objective functions and linear constraints have been derived in [4, 12, 25] without imposing BQ (2.1), but with the aid of a relatively restricted problem. In the following, we apply such an approach to these non-Lipschitz objective functions in [4, 12, 25] with nonlinear constraints.

THEOREM 2.2. *Let* x^* *be a local minimizer of problem* (1.1). *Suppose* $\Phi(x) = \sum_{i=1}^n \phi_i(x_i)$ *for some lower semicontinuous functions* $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$. *Let* $\mathcal{I} = \{i \mid \partial^\infty\phi_i(x_i^*) = \{0\}\}$*, and let* $\mathcal{I}^c$ *be the complement of* $\mathcal{I}$ *with respect to* $\{1, \ldots, n\}$*. Assume further that for any* $i \in \mathcal{I}^c$, $\partial\phi_i(x_i^*) = \mathbb{R}$*, and RCPLD holds at* $x_{\mathcal{I}}^*$ *for the system* $c(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) = 0,\ d(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) \leq 0$*. Then* x^* *is a KKT point of problem* (1.1).

Proof. Observe that $x_{\mathcal{I}}^*$ is a local minimizer of the restricted problem

$$\begin{array}{ll} \min\limits_{x_{\mathcal{I}}} & f(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) + \Phi(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) \\ \text{s.t.} & c(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) = 0,\ d(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) \leq 0, \end{array} \tag{2.4}$$

and RCPLD holds at $x_{\mathcal{I}}^*$ for the system $c(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) = 0,\ d(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) \leq 0$. In addition, by the assumption that $\partial^\infty\phi_i(x_i^*) = \{0\}$ for all $i \in I$, one can see that BQ holds at $x_{\mathcal{I}}^*$ for problem (2.4). It thus follows from Theorem 2.1 that $x_{\mathcal{I}}^*$ is a KKT point of problem (2.4). That is, there exist $\mu \in \mathbb{R}^m$ and $\nu \in \mathbb{R}_+^p$ such that

$$0 \in \nabla_{x_{\mathcal{I}}} f(x^*) + \partial_{x_{\mathcal{I}}}\Phi(x^*) + \nabla_{x_{\mathcal{I}}} c(x^*)\mu + \nabla_{x_{\mathcal{I}}} d(x^*)\nu, \quad \nu_i d_i(x^*) = 0, \quad i = 1, \ldots, p,$$

where $\nabla_{x_{\mathcal{I}}}$ and $\partial_{x_{\mathcal{I}}}$ denote, respectively, the gradient and the limiting subdifferential with respect to $x_{\mathcal{I}}$. Since the regularization term $\Phi(x)$ is assumed to be separable, it follows from [31, Proposition 10.5] that the limiting subdifferential is separable in the sense that

$$\partial\Phi(x^*) = \partial\phi_1(x_1^*) \times \partial\phi_2(x_2^*) \times \cdots \times \partial\phi_n(x_n^*),$$

and hence $\partial\Phi(x^*) = \partial_{x_{\mathcal{I}}}\Phi(x^*) \times \partial_{x_{\mathcal{I}^c}}\Phi(x^*)$. Since $\partial\phi_i(x_i^*)$ is the whole real line for each $i \in \mathcal{I}^c$, the inclusion

$$0 \in \nabla_{x_{\mathcal{I}^c}} f(x^*) + \partial_{x_{\mathcal{I}^c}}\Phi(x^*) + \nabla_{x_{\mathcal{I}^c}} c(x^*)\mu + \nabla_{x_{\mathcal{I}^c}} d(x^*)\nu$$

always holds. Therefore, x^* is a KKT point of problem (1.1). ∎

COROLLARY 2.1. *Let* x^* *be a local minimizer of problem* (1.1)*. Suppose* $\Phi(x) := \sum_{i=1}^n |x_i|^q$ *with* $q \in (0,1)$*. Let* $\mathcal{I} := \{i \mid x_i^* \neq 0\}$*, and let* $\mathcal{I}^c := \{i \mid x_i^* = 0\}$*. If RCPLD holds at* $x_{\mathcal{I}}^*$ *for the system* $c(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) = 0,\ d(x_{\mathcal{I}}, x_{\mathcal{I}^c}^*) \leq 0$*, then* x^* *is a KKT point.*

Proof. Since $\phi_i(x_i) = |x_i|^q$ for all i, it follows from [14, Lemma 2.5] that $\partial^\infty\phi(x_i^*) = \{0\}$ for every $i \in \mathcal{I}$ and $\partial\phi(x_i^*) = \mathbb{R}$ for any $i \in \mathcal{I}^c$. The conclusion of this corollary follows from Theorem 2.2 immediately. ∎

A special case of problem (1.1) was recently considered in [25], where $\Phi(x) = \sum_{i=1}^l (b_i - a_i^T x)_+^q$, and the mappings c and d are affine. We next study a more general problem with possibly nonlinear constraints. In particular, we establish a KKT necessary optimality condition for this problem.

COROLLARY 2.2. *Suppose* $\Phi(x) = \sum_{i=1}^l (b_i - a_i^T x)_+^q$*, where* $b_i \in \mathbb{R}, a_i \in \mathbb{R}^n$*, and* $q \in (0,1)$*. Let* x^* *be a local minimizer of problem* (1.1) *and*

$$\mathcal{J} := \{i \mid b_i - a_i^T x^* = 0\}, \quad \mathcal{K} := \{i \mid b_i - a_i^T x^* > 0\}.$$

If RCPLD holds at x^* *for the system*

$$c(x) = 0, \quad d(x) \leq 0, \quad b_i - a_i^T x \leq 0, \quad i \in \mathcal{J}, \tag{2.5}$$

then there exist $u \in \mathbb{R}^m$*,* $v \in \mathbb{R}_+^p$*, and* $\mu \in \mathbb{R}_+^{|\mathcal{J}|}$ *such that*

$$\begin{aligned} &\nabla f(x^*) + \nabla c(x^*)u + \nabla d(x^*)v - \sum_{i\in\mathcal{K}} q(b_i - a_i^T x^*)^{q-1} a_i - \sum_{i\in\mathcal{J}} \mu_i a_i = 0, \\ &d_i(x^*)v_i = 0, \quad i = 1, \ldots, p. \end{aligned}$$

Proof. Let $\mathcal{J}$ and $\mathcal{K}$ be defined above, let $\mathcal{I} := \{i \mid b_i - a_i^T x^* < 0\}$, and let $y_i^* := (b_i - a_i^T x^*)_+$ for every $i = 1, \ldots, l$. Clearly, $y_i^* = 0$ for every $i \in \mathcal{I} \cup \mathcal{J}$ and $y_i^* > 0$ for any $i \in \mathcal{K}$. Since x^* is a local minimizer of problem (1.1), it is easy to see that (x^*, y^*) is a local minimizer of the problem

$$\begin{aligned} &\min \quad f(x) + \sum_{i=1}^l |y_i|^q \\ &\text{s.t.} \quad c(x) = 0, \quad d(x) \leq 0, \quad b_i - a_i^T x \leq y_i, \quad y_i \geq 0, \quad i = 1, \ldots, l. \end{aligned} \tag{2.6}$$

Let $\phi(t) := |t|^q$. In view of the fact that $y_i^* = 0$ for all $i \in \mathcal{I} \cup \mathcal{J}$ and $y_i^* > 0$ for all $i \in \mathcal{K}$, one can see that $\partial\phi(y_i^*) = \mathbb{R}$ for $i \in \mathcal{I} \cup \mathcal{J}$ and $\partial^\infty\phi(y_i^*) = \{0\}$ for $i \in \mathcal{K}$. By the assumption that RCPLD holds at x^* for the system (2.5), it is not hard to verify that RCPLD also holds at $(x^*, y_{\mathcal{K}}^*)$ for the system

$$c(x) = 0; \quad d(x) \leq 0; \quad b_i - a_i^T x \leq y_i,\ i \in \mathcal{K}; \quad b_i - a_i^T x \leq 0,\ i \in \mathcal{J}.$$

Observe that the constraints

$$b_i - a_i^T x \le 0,\ i \in \mathcal{I}; \quad y_i \ge 0,\ i \in \mathcal{K}$$

are inactive at $(x^*, y^*_{\mathcal{K}})$. These imply that RCPLD holds at $(x^*, y^*_{\mathcal{K}})$ for the system

$$c(x) = 0; \quad d(x) \le 0; \quad b_i - a_i^T x \le y_i,\ i \in \mathcal{K}; \quad b_i - a_i^T x \le 0,\ i \in \mathcal{I} \cup \mathcal{J}; \quad y_i \ge 0,\ i \in \mathcal{K}.$$

In view of this and the fact that $\partial\phi(y_i^*) = \mathbb{R}$ for $i \in \mathcal{I} \cup \mathcal{J}$ and $\partial^\infty\phi(y_i^*) = \{0\}$ for $i \in \mathcal{K}$, one can see that the assumptions of Theorem 2.2 hold at (x^*, y^*) for problem (2.6). By applying Theorem 2.2 to problem (2.6), we conclude that (x^*, y^*) is a KKT point of (2.6). Hence, there exist $u \in \mathbb{R}^m$, $v \in \mathbb{R}^p_+$, $\mu \in \mathbb{R}^l_+$, and $\nu \in \mathbb{R}^l_+$ such that

$$\nabla f(x^*) + \nabla c(x^*)u + \nabla d(x^*)v - \sum_{i=1}^{l} \mu_i a_i = 0, \tag{2.7}$$

$$0 \in \partial\phi(y_i^*) - \mu_i - \nu_i, \quad i = 1, \ldots, l, \tag{2.8}$$

$$d_i(x^*)v_i = 0, \quad i = 1, \ldots, p, \tag{2.9}$$

$$(b_i - a_i^T x^* - y_i^*)\mu_i = 0, \quad y_i^*\nu_i = 0, \quad i = 1, \ldots, l. \tag{2.10}$$

By (2.10), we have $\mu_i = 0, i \in \mathcal{I}$ and $\nu_i = 0, i \in \mathcal{K}$. Then it follows from (2.8) that $q{y_i^*}^{q-1} - \mu_i = 0, i \in \mathcal{K}$. The last two relations together with (2.7) indicate that

$$\nabla f(x^*) + \nabla c(x^*)u + \nabla d(x^*)v - \sum_{i\in\mathcal{K}} q{y_i^*}^{q-1} a_i - \sum_{i\in\mathcal{J}} \mu_i a_i = 0. \tag{2.11}$$

The desired result follows from (2.9)–(2.11) and the definition of y^* immediately. The proof is now complete. □

3. An augmented Lagrangian method for problem (1.1).

In this section we present an AL method for solving problem (1.1). In practice, problem (1.1) often involves two types of constraints: easy and hard. For example, the lower and upper bound constraints can be viewed as easy constraints. For practical efficiency, we handle the easy constraints directly while penalizing the hard constraints into the AL function. To this end, we divide the mappings c and d into two parts: $c(x) = (c^{\mathbf{h}}(x), c^{\mathbf{e}}(x)), d(x) = (d^{\mathbf{h}}(x), d^{\mathbf{e}}(x))$, where $c^{\mathbf{h}} : \mathbb{R}^n \to \mathbb{R}^{m_1}, c^{\mathbf{e}} : \mathbb{R}^n \to \mathbb{R}^{m_2}$ and $d^{\mathbf{h}} : \mathbb{R}^n \to \mathbb{R}^{p_1}, d^{\mathbf{e}} : \mathbb{R}^n \to \mathbb{R}^{p_2}$ with $m_1 + m_2 = m$ and $p_1 + p_2 = p$. Here, the superscripts $\mathbf{h}$ and $\mathbf{e}$ stand for "hard" and "easy," respectively. For convenience of presentation, let

$$\mathcal{X} := \{x \mid c^{\mathbf{e}}(x) = 0, d^{\mathbf{e}}(x) \le 0\}$$

represent the set of easy constraints that can be handled in the AL method directly and efficiently. Therefore, problem (1.1) can be rewritten as

$$\begin{array}{rl} \min\limits_{x\in\mathcal{X}} & f(x) + \Phi(x) \\ \text{s.t.} & c^{\mathbf{h}}(x) = 0, \quad d^{\mathbf{h}}(x) \le 0. \end{array} \tag{3.1}$$

For any given penalty parameter $\rho > 0$ and Lagrange multipliers μ, ν, the AL function for problem (3.1) is defined as

$$\mathcal{L}(x, \mu, \nu, \rho) := f(x) + \Phi(x) + \frac{1}{2\rho}(\|\mu + \rho c^{\mathbf{h}}(x)\|^2 - \|\mu\|^2) + \frac{1}{2\rho}(\|[\nu + \rho d^{\mathbf{h}}(x)]_+\|^2 - \|\nu\|^2). \tag{3.2}$$

Inspired by the classical AL method, at each outer iteration we approximate problem (3.1) by an AL subproblem in the form of

$$(P^{\rho}_{\mu,\nu}) \qquad \min_{x\in\mathcal{X}} \quad \mathcal{L}(x,\mu,\nu,\rho). \tag{3.3}$$

For convenience, we write the AL function as a sum of a smooth term and the regularization term: $\mathcal{L}(x,\mu,\nu,\rho) = \varphi(x,\mu,\nu,\rho) + \Phi(x)$, where

$$\varphi(x,\mu,\nu,\rho) := f(x) + \frac{1}{2\rho}(\|\mu+\rho c^{\mathbf{h}}(x)\|^2 - \|\mu\|^2) + \frac{1}{2\rho}(\|[\nu+\rho d^{\mathbf{h}}(x)]_+\|^2 - \|\nu\|^2). \tag{3.4}$$

It is well known that the classical AL method may converge to an infeasible point. To remedy this pathological behavior, we adopt the strategies proposed by Lu and Zhang [27] so that the values of the AL function along the solution sequence generated by the method are bounded from above, and also the magnitude of penalty parameters outgrows that of Lagrange multipliers.

Throughout this section, assume that at least one feasible solution of problem (3.1) is known, which is denoted by $x^{\mathbf{feas}}$. It is generally not easy to find such $x^{\mathbf{feas}}$. However, for many practical problems, $x^{\mathbf{feas}}$ can be simply found without computation or can be easily computed, such as the problems with polyhedral constraints or Stiefel manifold. We are now ready to present the AL method for solving problem (3.1) (or, equivalently, (1.1)).

ALGORITHM 3.1. Choose $\mu^0 \in \mathbb{R}^{m_1}, \nu^0 \in \mathbb{R}^{p_1}_+$, $x^0_{\text{init}} \in \mathcal{X}$, $\rho_0 > 0$, $\gamma \in (1,\infty)$, $\tau,\eta \in (0,1)$, a nonnegative sequence $\{\epsilon_k\}$, and a constant

$$\Upsilon \geq \max\{f(x^{\mathbf{feas}}) + \Phi(x^{\mathbf{feas}}), \mathcal{L}(x^0_{\text{init}},\mu^0,\nu^0,\rho_0)\}.$$

Set $k = 0$.

(1) Solve problem (3.3) with $\mu = \mu^k$, $\nu = \nu^k$, and $\rho = \rho_k$ to find an approximate stationary point $x^k \in \mathcal{X}$ such that

$$\operatorname{dist}\left(0, \nabla_x\varphi(x^k,\mu^k,\nu^k,\rho_k) + \partial(\Phi+\delta_{\mathcal{X}})(x^k)\right) \leq \epsilon_k,\ \mathcal{L}(x^k,\mu^k,\nu^k,\rho_k) \leq \Upsilon. \tag{3.5}$$

(2) Set

$$\mu^{k+1} = \mu^k + \rho_k c^{\mathbf{h}}(x^k), \quad \nu^{k+1} = [\nu^k + \rho_k d^{\mathbf{h}}(x^k)]_+, \tag{3.6}$$

$$\zeta^{k+1} = \min\{\nu^{k+1}/\rho_k, -d^{\mathbf{h}}(x^k)\}. \tag{3.7}$$

(3) If $k > 0$ and

$$\max\left\{\|c^{\mathbf{h}}(x^k)\|, \|\zeta^{k+1}\|\right\} \leq \eta \max\left\{\|c^{\mathbf{h}}(x^{k-1})\|, \|\zeta^k\|\right\}, \tag{3.8}$$

then set $\rho_{k+1} = \rho_k$. Otherwise, set

$$\rho_{k+1} = \max\{\gamma\rho_k, \|\mu^{k+1}\|^{1+\tau}, \|\nu^{k+1}\|^{1+\tau}\}. \tag{3.9}$$

(4) Set $k \leftarrow k+1$ and go to step (1).

Remark. As will be discussed in subsection 3.1, x^k satisfying (3.5) can be found by a nonmonotone proximal gradient (NPG) method if a Lipschitz constant of $\nabla\varphi$ is known. Hence Algorithm 3.1 is well defined. In addition, x^0_{init} is used as an initial point for the NPG method when solving (3.3) with $\mu = \mu^0$, $\nu = \nu^0$, and $\rho = \rho_0$.

We next establish the convergence of Algorithm 3.1. Before proceeding, we state a result that may be viewed as a corollary of Carathéodory's lemma.

LEMMA 3.1 (see [1, Lemma 1]). *If* $x = \sum_{i=1}^{m+p} \alpha_i v_i$ *with* $\alpha_i \neq 0, i = 1, \dots, m$ *and* $\{v_i \mid i = 1, \dots, m\}$ *being linearly independent, then there exist* $\mathcal{J} \subseteq \{m+1, \dots, m+p\}$ *and* $\{\bar{\alpha}_i \mid i \in \{1, \dots, m\} \cup \mathcal{J}\}$ *such that*

- $x = \sum_{i \in \{1,\dots,m\} \cup \mathcal{J}} \bar{\alpha}_i v_i$ *with* $\alpha_i \bar{\alpha}_i > 0$ *for every* $i \in \mathcal{J}$, *and*
- $\{v_i \mid i \in \{1, \dots, m\} \cup \mathcal{J}\}$ *is linearly independent.*

The following result will also be useful in proving the convergence of the AL method.

PROPOSITION 3.1. *For any* $x \in \mathcal{F}$, *if* $-\partial^\infty \Phi(x) \cap \mathcal{N}_\mathcal{F}(x) = \{0\}$, *then* $-\partial^\infty \Phi(x) \cap \mathcal{N}_\mathcal{X}(x) = \{0\}$.

Proof. Since $\mathcal{F} \subseteq \mathcal{X}$, by the definition of the limiting normal cone, we have $\mathcal{N}_\mathcal{X}(x) \subseteq \mathcal{N}_\mathcal{F}(x)$ for any $x \in \mathcal{X}$. It thus follows that $-\partial^\infty \Phi(x) \cap \mathcal{N}_\mathcal{X}(x) \subseteq -\partial^\infty \Phi(x) \cap \mathcal{N}_\mathcal{F}(x)$. In addition, it is clear to observe that $0 \in -\partial^\infty \Phi(x) \cup \mathcal{N}_\mathcal{X}(x)$. The conclusion then immediately follows from these relations. ☐

We are now ready to establish the convergence of Algorithm 3.1.

THEOREM 3.1. *Suppose that* $\lim_{k\to\infty} \epsilon_k = 0$ *for Algorithm* 3.1. *Let* $\{x^k\}$ *be the sequence generated by Algorithm* 3.1, *and let* x^* *be an accumulation point of* $\{x^k\}$. *Assume that the function* $f + \Phi$ *is bounded below in* $\mathcal{X}$. *Then the following statements hold:*

(i) $c^{\mathbf{h}}(x^k) \to 0$ *and* $[d^{\mathbf{h}}(x^k)]_+ \to 0$ *as* $k \to \infty$.

(ii) x^* *is a feasible point of problem* (1.1).

(iii) *If BQ* (2.1) *holds at* x^* *and RCPLD holds at* x^*, *respectively, for the system* $c(x) = 0, d(x) \leq 0$ *and the system* $c^{\mathbf{e}}(x) = 0, d^{\mathbf{e}}(x) \leq 0$, *then* x^* *is a KKT point of problem* (1.1).

Proof. Notice that $\{x^k\} \subseteq \mathcal{X}$. This and the assumption that $f + \Phi$ is bounded below in $\mathcal{X}$ imply that $\{f(x^k) + \Phi(x^k)\}$ is bounded below. In addition, by the closedness of $\mathcal{X}$ and the fact that x^* is an accumulation point of $\{x^k\}$, one can easily see that $x^* \in \mathcal{X}$.

(i) We now prove statement (i) by considering two separate cases as follows.

Case (a): $\{\rho_k\}$ is bounded. In view of this and its updating scheme, one can see that $\{\rho_k\}$ is updated by (3.9) only for finite times. It thus implies that (3.8) holds for all $k \geq k_0$ for some k_0, that is,

$$\max\left\{\|c^{\mathbf{h}}(x^k)\|, \|\zeta^{k+1}\|\right\} \leq \eta \max\left\{\|c^{\mathbf{h}}(x^{k-1})\|, \|\zeta^k\|\right\} \quad \forall k \geq k_0.$$

It then follows that

$$\lim_{k\to\infty} \max\left\{\|c^{\mathbf{h}}(x^k)\|, \|\zeta^{k+1}\|\right\} = 0. \tag{3.10}$$

By the definition of ζ^{k+1}, one can observe that $d^{\mathbf{h}}(x^k) \leq -\zeta^{k+1}$ and hence $[d^{\mathbf{h}}(x^k)]_+ \leq [-\zeta^{k+1}]_+$, which implies $\|[d^{\mathbf{h}}(x^k)]_+\| \leq \|[-\zeta^{k+1}]_+\| \leq \|\zeta^{k+1}\|$. This relation along with (3.10) yields

$$\lim_{k\to\infty} \max\left\{\|c^{\mathbf{h}}(x^k)\|, \|[d^{\mathbf{h}}(x^k)]_+\|\right\} = 0,$$

which clearly implies that statement (i) holds.

Case (b): $\{\rho_k\}$ is unbounded. By its updating scheme, one can observe that $\{\rho_k\}$ must be updated by (3.9) for infinite times. Let $\{\rho_{j_1}, \rho_{j_2}, \dots\}$ denote all elements in $\{\rho_k\}$ that are updated by (3.9) with $\mathcal{J} = \{j_1, j_2, \dots\}$ arranged in increasing order. It then follows that $\{\rho_{j_\ell}\} \to \infty$ as $\ell \to \infty$ and

$$\rho_i = \rho_{j_\ell}, \quad j_\ell \le i < j_{\ell+1}, \quad \ell \ge 1, \tag{3.11}$$

$$\rho_{j_\ell} = \max\{\gamma \rho_{j_\ell - 1}, \|\mu^{j_\ell}\|^{1+\tau}, \|\nu^{j_\ell}\|^{1+\tau}\}, \quad \ell \ge 1. \tag{3.12}$$

Let $\underline{j}(k) := \max\{j \in \mathcal{J} \mid k \ge j\}$ for every $k \ge j_1$.

Claim that for every $k \ge j_1$,

$$\frac{\|\nu^k\|}{\rho_k} \le \frac{\|\nu^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \big\|[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+\big\| + \sum_{i=1}^{k-\underline{j}(k)-1} \big\|[d^{\mathbf{h}}(x^{\underline{j}(k)+i})]_+\big\|, \tag{3.13}$$

$$\frac{\|\mu^k\|}{\rho_k} \le \frac{\|\mu^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \|c^{\mathbf{h}}(x^{\underline{j}(k)})\| + \sum_{i=1}^{k-\underline{j}(k)-1} \|c^{\mathbf{h}}(x^{\underline{j}(k)+i})\|. \tag{3.14}$$

To this end, let $k \ge j_1$ be arbitrarily chosen. Clearly, (3.13) and (3.14) hold when $k = \underline{j}(k)$. We now suppose $k > \underline{j}(k)$. It then follows from the definition of $\underline{j}(k)$ that $\rho_{\underline{j}(k)+i} = \rho_{\underline{j}(k)}$ for $0 < i \le k - \underline{j}(k)$. In view of this and the second relation in (3.6), one has

$$\begin{aligned} \frac{\|\nu^{\underline{j}(k)+i}\|}{\rho_{\underline{j}(k)+i}} &= \frac{\|\nu^{\underline{j}(k)+i}\|}{\rho_{\underline{j}(k)+i-1}} = \left\| \left[\frac{\nu^{\underline{j}(k)+i-1}}{\rho_{\underline{j}(k)+i-1}} + d^{\mathbf{h}}(x^{\underline{j}(k)+i-1}) \right]_+ \right\| \\ &\le \frac{\|\nu^{\underline{j}(k)+i-1}\|}{\rho_{\underline{j}(k)+i-1}} + \left\| \left[d^{\mathbf{h}}(x^{\underline{j}(k)+i-1}) \right]_+ \right\|, \quad 0 < i \le k - \underline{j}(k), \end{aligned}$$

where the last inequality is due to $\nu^{\underline{j}(k)+i-1} \ge 0$. Summing up the above inequalities for $i = 1, \dots, k - \underline{j}(k)$ yields (3.13). The inequality (3.14) can be proved by using a similar argument and the first relation in (3.6).

We next show that for every $k \ge j_1$,

$$\frac{\|\nu^k\|}{\rho_k} \le \frac{\|\nu^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \big\|[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+\big\| + \frac{\eta}{1-\eta} \max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)})\|, \|\zeta^{\underline{j}(k)+1}\|\}, \tag{3.15}$$

$$\frac{\|\mu^k\|}{\rho_k} \le \frac{\|\mu^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \|c^{\mathbf{h}}(x^{\underline{j}(k)})\| + \frac{\eta}{1-\eta} \max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)})\|, \|\zeta^{\underline{j}(k)+1}\|\}. \tag{3.16}$$

Indeed, let $k \ge j_1$ be arbitrarily chosen. Clearly, it follows from (3.13) and (3.14) that (3.15) and (3.16) hold when $k = \underline{j}(k)$ or $\underline{j}(k) + 1$. We now suppose $k > \underline{j}(k) + 1$. It follows from the definition of $\underline{j}(k)$ and the updating scheme of $\{\rho_\ell\}$ that

$$\begin{aligned} &\max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)+i})\|, \|\zeta^{\underline{j}(k)+i+1}\|\} \\ &\qquad \le \eta \max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)+i-1})\|, \|\zeta^{\underline{j}(k)+i}\|\}, \quad 0 < i < k - \underline{j}(k), \end{aligned}$$

which leads to

$$\max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)+i})\|, \|\zeta^{\underline{j}(k)+i+1}\|\} \le \eta^i \max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)})\|, \|\zeta^{\underline{j}(k)+1}\|\}, \quad 0 < i < k - \underline{j}(k). \tag{3.17}$$

In addition, from the proof of case (a), we know that $\|[d^{\mathbf{h}}(x^{\underline{j}(k)+i})]_+\| \le \|\zeta^{\underline{j}(k)+i+1}\|$

for every i, which along with (3.13) and (3.17) implies

$$\begin{aligned}\frac{\|\nu^k\|}{\rho_k} &\le \frac{\|\nu^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \left\|[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+\right\| + \sum_{i=1}^{k-\underline{j}(k)-1} \left\|\zeta^{\underline{j}(k)+i+1}\right\| \\ &\le \frac{\|\nu^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \left\|[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+\right\| + \left(\sum_{i=1}^{k-\underline{j}(k)-1} \eta^i\right) \max\{\|c^{\mathbf{h}}(x^{\underline{j}(k)})\|, \|\zeta^{\underline{j}(k)+1}\|\}.\end{aligned}$$

The relation (3.15) follows from this and the fact that $\sum_{\ell=1}^{\infty} \eta^\ell \le \eta/(1-\eta)$. By (3.14), (3.17), and a similar argument, one can also see that (3.16) holds.

Next, we show that

$$\lim_{k\to\infty} \|\mu^{\underline{j}(k)}\|/\rho_{\underline{j}(k)} = 0, \qquad \lim_{k\to\infty} \|\nu^{\underline{j}(k)}\|/\rho_{\underline{j}(k)} = 0, \tag{3.18}$$

$$\lim_{k\to\infty} \|c^{\mathbf{h}}(x^{\underline{j}(k)})\| = 0, \quad \lim_{k\to\infty} \|[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+\| = 0, \quad \lim_{k\to\infty} \|\zeta^{\underline{j}(k)+1}\| = 0. \tag{3.19}$$

Indeed, recall that $\{\rho_{j_\ell}\} \to \infty$ as $\ell \to \infty$. It thus follows that $\rho_{\underline{j}(k)} \to \infty$ as $k \to \infty$. By (3.12) and the definition of $\underline{j}(k)$, one has

$$\rho_{\underline{j}(k)} = \max\left\{\gamma \rho_{\underline{j}(k)-1}, \|\mu^{\underline{j}(k)}\|^{1+\tau}, \|\nu^{\underline{j}(k)}\|^{1+\tau}\right\}, \qquad k \ge j_1.$$

This yields

$$\|\mu^{\underline{j}(k)}\|^{1+\tau} \le \rho_{\underline{j}(k)}, \quad \|\nu^{\underline{j}(k)}\|^{1+\tau} \le \rho_{\underline{j}(k)}, \qquad k \ge j_1,$$

which implies that

$$\|\mu^{\underline{j}(k)}\|/\rho_{\underline{j}(k)} \le (\rho_{\underline{j}(k)})^{-\frac{\tau}{1+\tau}}, \quad \|\nu^{\underline{j}(k)}\|/\rho_{\underline{j}(k)} \le (\rho_{\underline{j}(k)})^{-\frac{\tau}{1+\tau}}, \qquad k \ge j_1.$$

The relations (3.18) then follow from this and $\{\rho_{\underline{j}(k)}\} \to \infty$ as $k \to \infty$. Further, by the second relation in (3.5) and the definition of the AL function (3.2), one has

$$\begin{aligned}&f(x^{\underline{j}(k)}) + \Phi(x^{\underline{j}(k)}) \\ &+\frac{\|\mu^k + \rho_{\underline{j}(k)} c^{\mathbf{h}}(x^{\underline{j}(k)})\|^2 - \|\mu^{\underline{j}(k)}\|^2}{2\rho_{\underline{j}(k)}} + \frac{\|[\nu^{\underline{j}(k)} + \rho_{\underline{j}(k)} d^{\mathbf{h}}(x^{\underline{j}(k)})]_+\|^2 - \|\nu^{\underline{j}(k)}\|^2}{2\rho_{\underline{j}(k)}} \le \Upsilon,\end{aligned}$$

which leads to

$$\begin{aligned}&\left\|c^{\mathbf{h}}(x^{\underline{j}(k)}) + \frac{\mu^{\underline{j}(k)}}{\rho_{\underline{j}(k)}}\right\|^2 + \left\|\left[d^{\mathbf{h}}(x^{\underline{j}(k)}) + \frac{\nu^{\underline{j}(k)}}{\rho_{\underline{j}(k)}}\right]_+\right\|^2 \\ &\le \frac{2}{\rho_{\underline{j}(k)}}[\Upsilon - f(x^{\underline{j}(k)}) - \Phi(x^{\underline{j}(k)})] + \frac{\|\mu^{\underline{j}(k)}\|^2 + \|\nu^{\underline{j}(k)}\|^2}{\rho^2_{\underline{j}(k)}}.\end{aligned} \tag{3.20}$$

Using this, (3.18), the lower boundedness of $\{f(x^{\underline{j}(k)}) + \Phi(x^{\underline{j}(k)})\}$, and $\rho_{\underline{j}(k)} \to \infty$, one can see that

$$\lim_{k\to\infty} \left\|c^{\mathbf{h}}(x^{\underline{j}(k)}) + \frac{\mu^{\underline{j}(k)}}{\rho_{\underline{j}(k)}}\right\| = 0, \qquad \lim_{k\to\infty} \left\|\left[d^{\mathbf{h}}(x^{\underline{j}(k)}) + \frac{\nu^{\underline{j}(k)}}{\rho_{\underline{j}(k)}}\right]_+\right\| = 0, \tag{3.21}$$

which together with (3.18) implies that the first two relations in (3.19) hold. In addition, by (3.7) and the second relation in (3.6), one has

$$\zeta^{\underline{j}(k)+1} = \min\left\{\frac{\nu^{\underline{j}(k)+1}}{\rho_{\underline{j}(k)}}, -d^{\mathbf{h}}(x^{\underline{j}(k)})\right\} = \min\left\{\left[d^{\mathbf{h}}(x^{\underline{j}(k)}) + \frac{\nu^{\underline{j}(k)}}{\rho_{\underline{j}(k)}}\right]_+, -d^{\mathbf{h}}(x^{\underline{j}(k)})\right\},$$

which implies that

$$-[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+ \le \zeta^{\underline{j}(k)+1} \le \left[d^{\mathbf{h}}(x^{\underline{j}(k)}) + \frac{\nu^{\underline{j}(k)}}{\rho_{\underline{j}(k)}}\right]_+.$$

It then follows from this, (3.21), and $[d^{\mathbf{h}}(x^{\underline{j}(k)})]_+ \to 0$ that $\zeta^{\underline{j}(k)+1} \to 0$. Hence the last relation in (3.19) holds.

In view of (3.15), (3.16), (3.18), and (3.19), we conclude that

$$\lim_{k\to\infty} \|\mu^k\|/\rho_k = 0, \qquad \lim_{k\to\infty} \|\nu^k\|/\rho_k = 0. \tag{3.22}$$

By the same argument used for proving (3.20), one can show that

$$\left\|c^{\mathbf{h}}(x^k) + \frac{\mu^k}{\rho_k}\right\|^2 + \left\|\left[d^{\mathbf{h}}(x^k) + \frac{\nu^k}{\rho_k}\right]_+\right\|^2 \le \frac{2}{\rho_k}[\Upsilon - f(x^k) - \Phi(x^k)] + \frac{\|\mu^k\|^2 + \|\nu^k\|^2}{\rho_k^2},$$

which together with (3.22) and the lower boundedness of $\{f(x^k)+\Phi(x^k)\}$ implies that $c^{\mathbf{h}}(x^k) \to 0$ and $[d^{\mathbf{h}}(x^k)]_+ \to 0$ as $k \to \infty$. This completes the proof of statement (i).

(ii) Statement (ii) is a direct consequence of statement (i). Indeed, since x^* is an accumulation point of $\{x^k\}$, it follows from the continuity of $c^{\mathbf{h}}$ and $d^{\mathbf{h}}$ and statement (i) that $\max\{\|c^{\mathbf{h}}(x^*)\|, \|[d^{\mathbf{h}}(x^*)]_+\|\} = 0$, which clearly yields $c^{\mathbf{h}}(x^*) = 0$ and $d^{\mathbf{h}}(x^*) \le 0$. Recall that $x^* \in \mathcal{X}$. Hence, x^* is a feasible point of problem (1.1).

(iii) Since x^* is an accumulation point of $\{x^k\}$, there exists a subsequence $\mathcal{K}$ such that $\{x^k\}_{\mathcal{K}} \to x^*$. Claim that

$$\{\nu_i^{k+1}\}_{\mathcal{K}} \to 0 \quad \forall i \notin \mathcal{I}_{d^{\mathbf{h}}}(x^*) := \{i \mid d_i^{\mathbf{h}}(x^*) = 0\}. \tag{3.23}$$

Indeed, let $i \notin \mathcal{I}_{d^{\mathbf{h}}}(x^*)$ be arbitrarily chosen. We next prove (3.23) by considering two separate cases as follows.

Case (a): $\{\rho_k\}$ is bounded. By (3.10), one has that $\zeta^{k+1} \to 0$ as $k \to \infty$. Since $i \notin \mathcal{I}_{d^{\mathbf{h}}}(x^*)$ and $\{x^k\}_{\mathcal{K}} \to x^*$, we have $d_i^{\mathbf{h}}(x^k) < d_i^{\mathbf{h}}(x^*)/2 < 0$ for sufficiently large $k \in \mathcal{K}$. It follows from this and (3.7) that $\{\nu_i^{k+1}/\rho_k\}_{\mathcal{K}} \to 0$, which together with the boundedness of $\{\rho_k\}$ yields (3.23).

Case (b): $\{\rho_k\}$ is unbounded. Recall from the proof of statement (i) that $\|\nu^k\|/\rho_k \to 0$. From above, we know that $d_i^{\mathbf{h}}(x^k) < d_i^{\mathbf{h}}(x^*)/2 < 0$ for sufficiently large $k \in \mathcal{K}$. It follows from these and the second relation of (3.6) that for sufficiently large $k \in \mathcal{K}$,

$$\nu_i^{k+1} = \rho_k[\nu_i^k/\rho_k + d_i^{\mathbf{h}}(x^k)]_+ = 0,$$

and hence (3.23) holds.

For convenience, let

$$\mathcal{I}_{d^{\mathbf{e}}}(x) := \{i \mid d_i^{\mathbf{e}}(x) = 0\}, \quad \mathcal{I}_{d^{\mathbf{h}}}(x) := \{i \mid d_i^{\mathbf{h}}(x) = 0\} \quad \forall x \in \mathbb{R}^n.$$

Since BQ (2.1) holds at x^*, it follows from Proposition 3.1 that $-\partial^\infty\Phi(x^*)\cap\mathcal{N}_{\mathcal{X}}(x^*) = \{0\}$. Hence by Proposition 2.3 there exists $\delta_0 > 0$ such that for any $x \in \mathcal{B}_{\delta_0}(x^*) \cap \mathcal{X}$,

$$-\partial^\infty\Phi(x) \cap \mathcal{N}_{\mathcal{X}}(x) = \{0\}.$$

Moreover, by assumption, RCPLD holds at x^* for the system $c^e(x) = 0, d^e(x) \leq 0$. Then it follows from the sum rule of the limiting subdifferential [31, Corollary 10.9] and Propositions 2.1 and 2.2 that there exists $\delta \in (0, \delta_0)$ such that for any $x \in \mathcal{B}_\delta(x^*)$,

$$\begin{aligned} &\partial(\Phi + \delta_{\mathcal{X}})(x) \subseteq \partial\Phi(x) + \mathcal{N}_{\mathcal{X}}(x) \\ &= \partial\Phi(x) + \left\{ \nabla c^{\mathbf{e}}(x)\alpha + \sum_{i\in\mathcal{I}_{d^{\mathbf{e}}}(x)} \beta_i \nabla d_i^{\mathbf{e}}(x) \mid \alpha \in \mathbb{R}^{m_2}, \beta_i \geq 0\ i \in \mathcal{I}_{d^{\mathbf{e}}}(x) \right\}. \end{aligned} \tag{3.24}$$

Let $\mathcal{I}_e^k := \mathcal{I}_{d^{\mathbf{e}}}(x^k)$ and $\mathcal{I}_h^* := \mathcal{I}_{d^{\mathbf{h}}}(x^*)$. By (3.4), (3.24), the first relation in (3.5), and step (2) in Algorithm 3.1, there exist $\alpha^{k+1} \in \mathbb{R}^{m_2}$, $\beta^{k+1} \in \mathbb{R}_+^{|\mathcal{I}_e^k|}$, and $\xi^k \in \mathbb{R}^n$ such that

$$\begin{aligned} &\xi^k \in \nabla f(x^k) + \partial\Phi(x^k) + \sum_{i=1}^{m_1} \mu_i^{k+1}\nabla c_i^{\mathbf{h}}(x^k) \\ &\quad + \sum_{i=1}^{m_2} \alpha_i^{k+1}\nabla c_i^{\mathbf{e}}(x^k) + \sum_{i=1}^{p_1} \nu_i^{k+1}\nabla d_i^{\mathbf{h}}(x^k) + \sum_{i\in\mathcal{I}_e^k} \beta_i^{k+1}\nabla d_i^{\mathbf{e}}(x^k) \end{aligned} \tag{3.25}$$

and $\|\xi^k\| \leq \epsilon_k$ for all k. It then follows from $\epsilon_k \to 0$ that $\xi^k \to 0$.

Recall that $c(x) = (c^{\mathbf{h}}(x), c^{\mathbf{e}}(x))$. Let $\mathcal{I}$ be such that $\{\nabla c_i(x^*) \mid i \in \mathcal{I}\}$ is a basis for span$\{\nabla c_i(x^*) \mid i = 1, \ldots, m\}$. This together with $\{x^k\}_{\mathcal{K}} \to x^*$ implies that $\{\nabla c_i(x^k) \mid i \in \mathcal{I}\}$ is linearly independent for any sufficiently large $k \in \mathcal{K}$. Then it follows from RCPLD at x^* that $\{\nabla c_i(x^k) \mid i \in \mathcal{I}\}$ is a basis for span$\{\nabla c_i(x^k) \mid i = 1, \ldots, m\}$ for any sufficiently large $k \in \mathcal{K}$. Thus it follows from (3.25) that for every sufficiently large $k \in \mathcal{K}$, there exists $\{\hat{\mu}_i^{k+1} \mid i \in \mathcal{I}\}$ such that

$$\tilde{\xi}^k \in \nabla f(x^k) + \partial\Phi(x^k) + \sum_{i\in\mathcal{I}} \hat{\mu}_i^{k+1}\nabla c_i(x^k) + \sum_{i\in\mathcal{I}_h^*} \nu_i^{k+1}\nabla d_i^{\mathbf{h}}(x^k) + \sum_{i\in\mathcal{I}_e^k} \beta_i^{k+1}\nabla d_i^{\mathbf{e}}(x^k),$$

where $\tilde{\xi}^k := \xi^k - \sum_{i\notin\mathcal{I}_h^*} \nu_i^{k+1}\nabla d_i^{\mathbf{h}}(x^k)$. Then by Lemma 3.1, for every sufficiently large $k \in \mathcal{K}$, there exist

$$\{\bar{\mu}_i^{k+1} \mid i \in \mathcal{I}\}, \quad \{\bar{\nu}_i^{k+1} \geq 0 \mid i \in \mathcal{J}_1^k\}, \quad \{\bar{\beta}_i^{k+1} \geq 0 \mid i \in \mathcal{J}_2^k\}$$

with $\mathcal{J}_1^k \subseteq \mathcal{I}_h^*$ and $\mathcal{J}_2^k \subseteq \mathcal{I}_e^k$ such that

$$\begin{aligned} &\tilde{\xi}^k \in \nabla f(x^k) + \partial\Phi(x^k) \\ &\quad + \sum_{i\in\mathcal{I}} \bar{\mu}_i^{k+1}\nabla c_i(x^k) + \sum_{i\in\mathcal{J}_1^k} \bar{\nu}_i^{k+1}\nabla d_i^{\mathbf{h}}(x^k) + \sum_{i\in\mathcal{J}_2^k} \bar{\beta}_i^{k+1}\nabla d_i^{\mathbf{e}}(x^k) \end{aligned} \tag{3.26}$$

and

$$\{\nabla c_i(x^k), \nabla d_j^{\mathbf{h}}(x^k), \nabla d_\ell^{\mathbf{e}}(x^k) \mid i \in \mathcal{I}, j \in \mathcal{J}_1^k, \ell \in \mathcal{J}_2^k\} \text{ is linearly independent.} \tag{3.27}$$

Since the number of possible sets $\mathcal{J}_1^k$ and $\mathcal{J}_2^k$ is finite, we can find a subsequence $\mathcal{K}_1$ in $\mathcal{K}$ such that $\mathcal{J}_1^k \equiv \mathcal{J}_1$ and $\mathcal{J}_2^k \equiv \mathcal{J}_2$ for every $k \in \mathcal{K}_1$. Moreover, we may assume that

(3.26) and (3.27) hold for all $k \in \mathcal{K}_1$. Note also that $\mathcal{J}_1 \subseteq \mathcal{I}_h^*$ and $\mathcal{J}_2 \subseteq \mathcal{I}_e^* := \mathcal{I}_{d^{\mathbf{e}}}(x^*)$. In addition, by $\{x^k\}_{\mathcal{K}} \to x^*$, $\xi^k \to 0$, (3.23), and the definition of $\tilde{\xi}^k$, one can observe that $\{\tilde{\xi}^k\}_{\mathcal{K}_1} \to 0$.

If $\mathcal{I} \cup \mathcal{J}_1 \cup \mathcal{J}_2 = \emptyset$, then it follows from (3.26), $\{\tilde{\xi}^k\}_{\mathcal{K}_1} \to 0$, and the outer semicontinuity of the limiting subdifferential that $0 \in \nabla f(x^*) + \partial\Phi(x^*)$, which implies that x^* is a KKT point of problem (1.1). We now suppose $\mathcal{I} \cup \mathcal{J}_1 \cup \mathcal{J}_2 \neq \emptyset$. Let

$$t_k := \max\big\{|\bar{\mu}_i^{k+1}|, |\bar{\nu}_j^{k+1}|, |\bar{\beta}_\ell^{k+1}| \,|\, i \in \mathcal{I}, j \in \mathcal{J}_1, \ell \in \mathcal{J}_2\big\}.$$

We claim that $\{t_k\}_{\mathcal{K}_1}$ is bounded. Suppose, on the contrary, that $\{t_k\}_{\mathcal{K}_1}$ is unbounded. Without loss of generality, we assume that as $\mathcal{K}_1 \ni k \to \infty$, we have $t_k \to \infty$, and

$$\frac{\bar{\mu}_i^{k+1}}{t_k} \to \bar{\mu}_i^*, \quad i \in \mathcal{I}; \quad \frac{\bar{\nu}_j^{k+1}}{t_k} \to \bar{\nu}_j^*, \quad j \in \mathcal{J}_1; \quad \frac{\bar{\beta}_\ell^{k+1}}{t_k} \to \bar{\beta}_\ell^*, \quad \ell \in \mathcal{J}_2.$$

It is clear to see that

$$\max\big\{|\bar{\mu}_i^*|, |\bar{\nu}_j^*|, |\bar{\beta}_\ell^*| \,|\, i \in \mathcal{I}, j \in \mathcal{J}_1, \ell \in \mathcal{J}_2\big\} = 1. \tag{3.28}$$

Moreover, since $\bar{\nu}_j^{k+1} \geq 0, \bar{\beta}_\ell^{k+1} \geq 0$ for all $j \in \mathcal{J}_1$, $\ell \in \mathcal{J}_2$, and $k \in \mathcal{K}_1$, we have

$$\bar{\nu}_j^* \geq 0, \quad j \in \mathcal{J}_1; \quad \bar{\beta}_\ell^* \geq 0, \quad \ell \in \mathcal{J}_2. \tag{3.29}$$

Dividing (3.26) by t_k and taking limits on both sides as $\mathcal{K}_1 \ni k \to \infty$, it follows from $\{\tilde{\xi}^k\}_{\mathcal{K}_1} \to 0$ and the definition of the horizon subdifferential that

$$0 \in \partial^\infty \Phi(x^*) + \sum_{i\in\mathcal{I}} \bar{\mu}_i^* \nabla c_i(x^*) + \sum_{j\in\mathcal{J}_1} \bar{\nu}_j^* \nabla d_j^{\mathbf{h}}(x^*) + \sum_{\ell\in\mathcal{J}_2} \bar{\beta}_\ell^* \nabla d_\ell^{\mathbf{e}}(x^*).$$

Moreover, it follows from Proposition 2.2 that

$$\sum_{i\in\mathcal{I}} \bar{\mu}_i^* \nabla c_i(x^*) + \sum_{j\in\mathcal{J}_1} \bar{\nu}_j^* \nabla d_j^{\mathbf{h}}(x^*) + \sum_{\ell\in\mathcal{J}_2} \bar{\beta}_\ell^* \nabla d_\ell^{\mathbf{e}}(x^*) \in \mathcal{N}_{\mathcal{F}}(x^*).$$

Then it follows from BQ (2.1) and the last two relations that

$$\sum_{i\in\mathcal{I}} \bar{\mu}_i^* \nabla c_i(x^*) + \sum_{j\in\mathcal{J}_1} \bar{\nu}_j^* \nabla d_j^{\mathbf{h}}(x^*) + \sum_{\ell\in\mathcal{J}_2} \bar{\beta}_\ell^* \nabla d_\ell^{\mathbf{e}}(x^*) = 0.$$

This together with (3.28), (3.29), and RCPLD implies that for any x sufficiently close to x^*,

$$\{\nabla c_i(x), \nabla d_j^{\mathbf{h}}(x), \nabla d_\ell^{\mathbf{e}}(x) \,|\, i \in \mathcal{I}, j \in \mathcal{J}_1, \ell \in \mathcal{J}_2\}$$

is linearly dependent. This contradicts (3.27). Thus $\{t_k\}_{\mathcal{K}_1}$ is bounded as desired. It then follows from the definition of t_k that $\{\bar{\mu}_i^{k+1}\}_{\mathcal{K}_1}$, $\{\bar{\nu}_j^{k+1}\}_{\mathcal{K}_1}$, and $\{\bar{\beta}_\ell^{k+1}\}_{\mathcal{K}_1}$ are bounded for all $i \in \mathcal{I}$, $j \in \mathcal{J}_1$, and $\ell \in \mathcal{J}_2$. Without loss of generality, we assume that as $\mathcal{K}_1 \ni k \to \infty$,

$$\bar{\mu}_i^{k+1} \to \mu_i^*, \quad i \in \mathcal{I}; \quad \bar{\nu}_j^{k+1} \to \nu_j^*, \quad j \in \mathcal{J}_1; \quad \bar{\beta}_\ell^{k+1} \to \beta_\ell^*, \quad \ell \in \mathcal{J}_2.$$

Moreover, by a similar argument as in the proof of (3.29), one can have

$$\nu_j^* \geq 0, \quad j \in \mathcal{J}_1; \qquad \beta_\ell^* \geq 0, \quad \ell \in \mathcal{J}_2. \tag{3.30}$$

Taking limits on both sides of (3.26) as $\mathcal{K}_1 \ni k \to \infty$, we then have

$$0 \in \nabla f(x^*) + \partial\Phi(x^*) + \sum_{i\in\mathcal{I}} \mu_i^* \nabla c_i(x^*) + \sum_{j\in\mathcal{J}_1} \nu_j^* \nabla d_j^{\mathbf{h}}(x^*) + \sum_{\ell\in\mathcal{J}_2} \beta_j^* \nabla d_\ell^{\mathbf{e}}(x^*). \tag{3.31}$$

Since $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{I}_h^* \cup \mathcal{I}_e^*$, it follows from (3.30) and (3.31) that x^* is a KKT point of problem (1.1). The proof is now complete. ◻

If the set $\mathcal{X}$ is bounded, then the sequence $\{x^k\}$ generated by Algorithm 3.1 is bounded, and thus an accumulation point exists. Note that when the function Φ is Lipschitz continuous at x^*, $\partial^\infty\Phi(x^*) = \{0\}$, and so BQ (2.1) required in Theorem 3.1 is superfluous. We point out that even in the case where Φ is convex, Theorem 3.1 improves [27, Theorem 3.3] since RCPLD is weaker than Robinson's constraint qualification that is required in [27, Theorem 3.3]. Another case where BQ (2.1) is superfluous is considered in the following theorem.

THEOREM 3.2. *Assume that* $\Phi(x) = \sum_{i=1}^n \phi_i(x_i)$ *with lower semicontinuous functions* $\phi_i : \mathbb{R} \to \mathbb{R}, i = 1, \ldots, n$, *and* $\mathcal{X}$ *is a box in* $\mathbb{R}^n$. *Suppose that* $\lim_{k\to\infty} \epsilon_k = 0$ *for Algorithm* 3.1. *Let* $\{x^k\}$ *be a sequence generated by Algorithm* 3.1, *and let* x^* *be an arbitrary accumulation point of* $\{x^k\}$. *Let* $\mathcal{I} := \{i \mid \partial^\infty\phi_i(x_i^*) = \{0\}\}$, *and let* $\mathcal{I}^c$ *be the complement of* $\mathcal{I}$ *with respect to* $\{1, \ldots, n\}$. *Assume further that for any* $i \in \mathcal{I}^c$, $\partial\phi_i(x_i^*) = \mathbb{R}$, *and RCPLD holds for the system* $c(x_\mathcal{I}, x_{\mathcal{I}^c}^*) = 0, d(x_\mathcal{I}, x_{\mathcal{I}^c}^*) \le 0$ *at* $x_\mathcal{I}^*$. *Then* x^* *is a KKT point of problem* (1.1).

Proof. Similarly as in the proof of Theorem 2.2, by the separability of the function $\Phi(x)$ and the fact that $\partial\phi_i(x_i^*) = \mathbb{R}$ for any $i \in \mathcal{I}^c$, we need only focus on the components associated with the index set $\mathcal{I}$.

It was shown that x^* is a feasible point of problem (1.1) in Theorem 3.1. Now we show that x^* is a KKT point. Since the regularization term $\Phi(x)$ is assumed to be separable and the set $\mathcal{X}$ is a box, the function $\Phi + \delta_\mathcal{X}$ is also separable. It follows from [31, Proposition 10.5] that the partial limiting subdifferential is separable and hence

$$\partial(\Phi + \delta_\mathcal{X})(x) = \partial_{x_\mathcal{I}}(\Phi + \delta_\mathcal{X})(x) \times \partial_{x_{\mathcal{I}^c}}(\Phi + \delta_\mathcal{X})(x) \quad \forall x \in \mathcal{X}.$$

Thus it follows from the first relation in (3.5) that

$$\begin{aligned} &\operatorname{dist}\left(0, \nabla_{x_\mathcal{I}}\varphi(x^k, \mu^k, \nu^k, \rho_k) + \partial_{x_\mathcal{I}}(\Phi + \delta_\mathcal{X})(x^k)\right) \\ &\le \operatorname{dist}\left(0, \nabla_x\varphi(x^k, \mu^k, \nu^k, \rho_k) + \partial(\Phi + \delta_\mathcal{X})(x^k)\right) \le \epsilon_k. \end{aligned} \tag{3.32}$$

By the assumption $\partial_{x_\mathcal{I}}^\infty\Phi(x^*) = \{0\}$, one has $\partial_{x_\mathcal{I}}^\infty\Phi(x^*) \cap [\mathcal{N}_\mathcal{X}(x^*)]_\mathcal{I} = \{0\}$. This, along with Proposition 2.3 and the separability of Φ and $\mathcal{X}$, implies that there exists $\delta > 0$ such that $\partial_{x_\mathcal{I}}^\infty\Phi(x) \cap [\mathcal{N}_\mathcal{X}(x)]_\mathcal{I} = \{0\}$ for any $x \in \mathcal{B}_\delta(x^*)$. It then follows from this, the separability of Φ and $\mathcal{X}$, and the sum rule of the limiting subdifferential [31, Corollary 10.10] that $\partial_{x_\mathcal{I}}(\Phi + \delta_\mathcal{X})(x) \subseteq \Pi_{i\in\mathcal{I}}\partial\phi_i(x_i) + \mathcal{N}_{\mathcal{X}_\mathcal{I}}(x_\mathcal{I})$ for every $x \in \mathcal{B}_\delta(x^*)$, where $\mathcal{X}_\mathcal{I} := \{x_\mathcal{I} \mid c_i^e(x_i) = 0, d_i^e(x_i) \le 0, i \in \mathcal{I}\}$. Notice that $\mathcal{X}_\mathcal{I}$ is a linear system. Hence, RCPLD holds at every point in $\mathcal{X}_\mathcal{I}$. By these facts and Proposition 2.2, we have that for every $x \in \mathcal{B}_\delta(x^*)$,

$$\begin{aligned} &\partial_{x_\mathcal{I}}(\Phi + \delta_\mathcal{X})(x) \subseteq \Pi_{i\in\mathcal{I}}\partial\phi_i(x_i) + \mathcal{N}_{\mathcal{X}_\mathcal{I}}(x_\mathcal{I}) \subseteq \Pi_{i\in\mathcal{I}}\partial\phi_i(x_i) \\ &\quad + \left\{ \nabla_{x_\mathcal{I}} c^{\mathbf{e}}(x)\alpha + \sum_{i\in\mathcal{I}_{d^{\mathbf{e}}}(x)} \beta_i \nabla_{x_\mathcal{I}} d_i^{\mathbf{e}}(x) \mid \alpha \in \mathbb{R}^{m_2}, \beta_i \ge 0 \ i \in \mathcal{I}_{d^{\mathbf{e}}}(x) \right\}. \end{aligned} \tag{3.33}$$

Using (3.32) instead of the first relation in (3.5), and (3.33) instead of (3.24), the rest of the proof is in line with the proof process of Theorem 3.1 by focusing only on the components associated with the index set $\mathcal{I}$. The proof is complete. ☐

3.1. Nonmonotone proximal gradient method for subproblems (3.3). In this subsection, we discuss how to find an approximate stationary point x^k of the kth AL subproblem (3.3) satisfying (3.5) as required in step (1) of Algorithm 3.1.

In particular, we apply a nonmonotone proximal gradient (NPG) method to solve subproblem (3.3). As shown subsequently in Theorem 3.4, the point x^k obtained by the NPG method satisfies the first relation of (3.5) when the associated tolerance parameter is suitably chosen. Moreover, such x^k also satisfies the second relation of (3.5), provided that the initial point of the NPG method is properly chosen. Indeed, let $x^k_{\text{init}} \in \mathcal{X}$ be the initial point of the NPG method when applied to the kth subproblem (3.3), which is specified as follows:

$$x^k_{\text{init}} = \begin{cases} x^{\textbf{feas}} & \text{if } \mathcal{L}(x^{k-1}, \mu^k, \nu^k, \rho_k) > \Upsilon, \\ x^{k-1} & \text{otherwise}, \end{cases} \qquad k \geq 1, \tag{3.34}$$

where x^{k-1} is an approximate stationary point of the $(k-1)$th AL subproblem (3.3) obtained by the NPG method. Since $c^{\mathbf{h}}(x^{\textbf{feas}}) = 0$ and $d^{\mathbf{h}}(x^{\textbf{feas}}) \leq 0$, it follows from (3.2) and the definition of Υ that

$$\mathcal{L}(x^{\textbf{feas}}, \mu^k, \nu^k, \rho_k) \leq f(x^{\textbf{feas}}) + \Phi(x^{\textbf{feas}}) \leq \Upsilon.$$

This inequality and the above choice of x^k_{init} imply that $\mathcal{L}(x^k_{\text{init}}, \mu^k, \nu^k, \rho_k) \leq \Upsilon$. In addition, as observed below, a feature of the NPG method is that the objective function values at all iterates generated are bounded above by that at the initial point. Thus we have

$$\mathcal{L}(x^k, \mu^k, \nu^k, \rho_k) \leq \mathcal{L}(x^k_{\text{init}}, \mu^k, \nu^k, \rho_k) \leq \Upsilon,$$

and thus the second relation of (3.5) is satisfied at x^k.

We now discuss an NPG method for solving the AL subproblem (3.3). In particular, Wright, Robert and Figueiredo [33] recently proposed an NPG method for solving a class of problems in the form of

$$\min_x \ h(x) + \Psi(x),$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable with a globally Lipschitz continuous gradient and $\Psi : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous. More recently, it was shown in [14] that if the objective function of this problem has bounded level sets, then the global Lipschitz continuity of ∇h can be weakened to the local Lipschitz continuity of ∇h. We now adopt the NPG method [33] to solve the AL subproblem (3.3).

Algorithm 3.2. Let $\sigma \in (0, 1)$, $\theta > 1$, $L_{\max} > L_{\min} > 0$, and a nonnegative integer M be given. Choose an arbitrary point $z^0 \in \mathbb{R}^n$ and set $\jmath = 0$.

(1) Choose $L^0_\jmath \in [L_{\min}, L_{\max}]$ arbitrarily. Set $L_\jmath = L^0_\jmath$.

(1a) Solve the following subproblem to obtain $z^{\jmath+1}$:

$$\min_{x \in \mathcal{X}} \ \varphi(z^\jmath, \mu, \nu, \rho) + \nabla_x \varphi(z^\jmath, \mu, \nu, \rho)^T (x - z^\jmath) + \frac{L_\jmath}{2} \|x - z^\jmath\|^2 + \Phi(x). \tag{3.35}$$

(1b) If

$$\mathcal{L}(z^{\jmath+1}, \mu, \nu, \rho) \leq \max_{[\jmath - M]_+ \leq i \leq \jmath} \mathcal{L}(z^i, \mu, \nu, \rho) - \frac{\sigma}{2} \|z^{\jmath+1} - z^\jmath\|^2 \tag{3.36}$$

is satisfied, then go to step (2). Otherwise, set $L_\jmath \leftarrow \theta L_\jmath$ and go to step (1a).

(2) Set $\jmath \leftarrow \jmath + 1$ and go to step (1).

We now make some remarks regarding Algorithm 3.2. Observe that step (1a) in Algorithm 3.2 is a proximal gradient method since the first three terms in the objective function of (3.35) can be viewed as a quadratic separable approximation to φ near $z^\jmath$. The constant σ in (3.36) is usually chosen to be close to zero. Step (1b) means that the candidate will be accepted as the new iterate if its objective function value $\mathcal{L}(z^{\jmath+1}, \mu, \nu, \rho)$ is slightly smaller than the largest value of the objective function over the past $M + 1$ iterations. This method is nonmonotone since the objective function value is not always decreasing at each step. In addition, in our implementation we set $L_0^0 = 1$ and update $L_\jmath^0$ by the same strategy as in [2], that is,

$$L_\jmath^0 = \max\left\{L_{\min}, \min\left\{L_{\max}, \frac{(\Delta g)^T \Delta x}{\|\Delta x\|^2}\right\}\right\},$$

where $\Delta x = z^{\jmath+1} - z^\jmath$ and $\Delta g = \nabla_x \varphi(z^{\jmath+1}, \mu, \nu, \rho) - \nabla_x \varphi(z^\jmath, \mu, \nu, \rho)$.

It is not hard to observe that problem (3.35) is equivalent to the problem

$$\min_{x \in \mathcal{X}} \frac{1}{2}\|x - w^\jmath\|^2 + \frac{1}{L_\jmath}\Phi(x), \tag{3.37}$$

where $w^\jmath = z^\jmath - \nabla_x \varphi(z^\jmath, \mu, \nu, \rho)/L_\jmath$. If Φ and $\mathcal{X}$ have some special structure, one can solve problem (3.37) and hence problem (3.35) efficiently. For example, in practical applications, we often have $\Phi(x) = \sum_{i=1}^n \phi_i(x_i)$ for some lower semicontinuous function $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$, and $\mathcal{X} = [l, u]$ with $l < u$. Due to the separability of such Φ and $\mathcal{X}$, problem (3.37) can be solved as n number of one-dimensional problems:

$$\begin{array}{ll} \min & \frac{1}{2}(x_i - w_i^\jmath)^2 + \frac{1}{L_\jmath}\phi_i(x_i) \\ \text{s.t.} & l_i \le x_i \le u_i \end{array}$$

for $i = 1, \ldots, n$. These problems either have a closed-form solution or can be converted to a univariate root-finding problem for various ϕ, such as bridge penalty [20], capped-l_1 penalty [34], fraction penalty [30], logistic penalty [22], l_0-quasi-norm [6, 26], l_1-norm [32, 11], and smoothly clipped absolute deviation (SCAD) penalty [19].

For ease of presentation, let $\Omega(x^0) := \{x \in \mathcal{X} \mid \mathcal{L}(x, \mu, \nu, \rho) \le \mathcal{L}(x^0, \mu, \nu, \rho)$, where x^0 is the initial point of Algorithm 3.2. Throughout this subsection, we make the following assumption regarding φ, $\mathcal{L}$, and $\mathcal{X}$.

Assumption 3.1. $\nabla\varphi(\cdot, \mu, \nu, \rho)$ is globally Lipschitz continuous over $\mathcal{X}$ and $\mathcal{L}(\cdot, \mu, \nu, \rho)$ is uniformly continuous on $\mathcal{X}$; or $\nabla\varphi(\cdot, \mu, \nu, \rho)$ is locally Lipschitz continuous in $\mathcal{X}$ and $\Omega(x^0)$ is bounded.[1]

The following lemma shows that condition (3.36) is satisfied in finite iterations, whose proof is similar to [33, Lemma 3] and [14, Proposition A.1].

LEMMA 3.2. *There exists* $L_0 > 0$ *such that* $z^{\jmath+1}$ *satisfies condition* (3.36) *whenever* $L_\jmath \ge L_0$.

The following lemma follows from [33, Lemma 4] immediately.

---

[1] When $\Omega(x^0)$ is compact, Assumption 3.1 implies that $\nabla\varphi(\cdot, \mu, \nu, \rho)$ is globally Lipschitz on $\Omega(x^0)$.

LEMMA 3.3. *Let* $\{z^{\jmath}\}$ *be a sequence generated by Algorithm* 3.2*. Then* $z^{\jmath+1}-z^{\jmath} \to 0$ *as* $\jmath \to \infty$*.*

We are now ready to establish the convergence result for Algorithm 3.2.

THEOREM 3.3. *Let* $\{z^{\jmath}\}$ *be a sequence generated by Algorithm* 3.2*. Then any accumulation point* z^* *of* $\{z^{\jmath}\}$ *satisfies* $0 \in \nabla_x \varphi(z^*, \mu, \nu, \rho) + \partial(\Phi + \delta_{\mathcal{X}})(z^*)$*.*

Proof. Let $\mathcal{J}$ be a subsequence such that $z^{\jmath} \to z^*$ as $\mathcal{J} \ni \jmath \to \infty$. By Lemma 3.3, one can see that $z^{\jmath+1} \to z^*$ as $\mathcal{J} \ni \jmath \to \infty$. Since $z^{\jmath+1}$ is a minimizer of problem (3.35), it follows from the first-order optimality condition that

$$0 \in \nabla_x \varphi(z^{\jmath}, \mu, \nu, \rho) + L_{\jmath}(z^{\jmath+1} - z^{\jmath}) + \partial(\Phi + \delta_{\mathcal{X}})(z^{\jmath+1}). \tag{3.38}$$

By Lemma 3.2, we know that $\{L_{\jmath}\}$ is bounded. Using this fact, Lemma 3.3, and the outer semicontinuity of the limiting subdifferential, taking limits on both sides of (3.38) as $\mathcal{J} \ni \jmath \to \infty$ yields $0 \in \nabla_x \varphi(z^*, \mu, \nu, \rho) + \partial(\Phi + \delta_{\mathcal{X}})(z^*)$. The proof is complete. ◻

DEFINITION 3.1. *Let* $z \in \mathcal{X}$*. Given any* $\epsilon > 0$*, we say that* z *is an* ϵ*-approximate stationary point of problem* (3.3) *if* $\operatorname{dist}(0, \nabla_x \varphi(z, \mu, \nu, \rho) + \partial(\Phi + \delta_{\mathcal{X}})(z)) \le \epsilon$*.*

As a consequence of Theorem 3.3, we next show that when $\jmath$ is sufficiently large, $z^{\jmath+1}$ generated by Algorithm 3.2 is an ϵ-approximate stationary point of (3.3).

THEOREM 3.4. *Let* $\{z^{\jmath}\}$ *be generated by Algorithm* 3.2*, and let* $\epsilon > 0$ *be given. Then when* $\jmath$ *is sufficiently large,*

$$\|z^{\jmath+1} - z^{\jmath}\| \le \frac{\epsilon}{L_\varphi + \bar{L}} \tag{3.39}$$

holds, and as a consequence, $z^{\jmath+1}$ *is an* ϵ*-approximate stationary point of* (3.3)*, where* L_φ *is the Lipschitz constant of* $\nabla_x \varphi(\cdot, \mu, \nu, \rho)$ *in* $\Omega(x^0)$ *and* $\bar{L} := \max\{L_{\max}, \theta(L_\varphi + \eta)\}$*.*

Proof. Let $\bar{L}$ be defined above. By a similar argument as in [14, Proposition A.1], one can show that $0 < L_{\jmath} \le \bar{L}$ for all $\jmath$. It then follows from Lemma 3.3 that (3.39) holds when $\jmath$ is sufficiently large. Using this relation, the Lipschitz continuity of $\nabla_x \varphi$, and the fact that $0 < L_{\jmath} \le \bar{L}$ for all $\jmath$, we have

$$\begin{aligned} \|\nabla_x \varphi(z^{\jmath+1}, \mu, \nu, \rho) - \nabla_x \varphi(z^{\jmath}, \mu, \nu, \rho)\| + \|L_{\jmath}(z^{\jmath+1} - z^{\jmath})\| &\le (L_\varphi + L_{\jmath})\|z^{\jmath+1} - z^{\jmath}\| \\ &\le \epsilon. \end{aligned} \tag{3.40}$$

Notice from (3.38) that

$$\begin{aligned} &\nabla_x \varphi(z^{\jmath+1}, \mu, \nu, \rho) - \nabla_x \varphi(z^{\jmath}, \mu, \nu, \rho) - L_{\jmath}(z^{\jmath+1} - z^{\jmath}) \\ &\qquad\qquad \in \nabla_x \varphi(z^{\jmath+1}, \mu, \nu, \rho) + \partial(\Phi + \delta_{\mathcal{X}})(z^{\jmath+1}). \end{aligned}$$

This, together with (3.40), implies that when $\jmath$ is sufficiently large,

$$\begin{aligned} &\operatorname{dist}\left(0, \nabla_x \varphi(z^{\jmath+1}, \mu, \nu, \rho) + \partial(\Phi + \delta_{\mathcal{X}})(z^{\jmath+1})\right) \\ &\le \|\nabla_x \varphi(z^{\jmath+1}, \mu, \nu, \rho) - \nabla_x \varphi(z^{\jmath}, \mu, \nu, \rho) - L_{\jmath}(z^{\jmath+1} - z^{\jmath})\| \ \le \ \epsilon. \end{aligned}$$

The conclusion follows from this inequality and Definition 3.1. ◻

Remark. By Theorem 3.4 and the specific choice of x^k_{init} given in (3.34), one can see that x^k satisfying (3.5) can be found by the NPG method. However, for a given point x, $\partial(\Phi + \delta_{\mathcal{X}})(x)$ may not be easily evaluated. It thus can be hard to verify explicitly the first condition in (3.5). Fortunately, Theorem 3.4 provides a practical approach to verify it. Indeed, let $\{z^\jmath\}$ be the sequence generated by the NPG applied to the AL subproblem (3.3) with $\mu = \mu^k$, $\nu = \nu^k$, and $\rho = \rho_k$. By Theorem 3.4, one knows that there must exist some $\jmath$ such that (3.39) holds with $\epsilon = \epsilon_k$, which implies that $x^k = z^{\jmath+1}$ satisfies the first condition in (3.5). Therefore, to verify such a condition, it suffices to verify (3.39).

4. Numerical experiments.

In this section we conduct numerical experiments to test the performance of the proposed AL method (Algorithm 3.1). In particular, we apply it to solve some ℓ_q-regularized portfolio selection models with $q \in (0, 1)$, and also compare it with an interior point (IP) method proposed in [12].

Given a set of assets $\mathcal{S} = \{s_1, \dots, s_N\}$, let r_i denote the expected return per unit of asset s_i, r_0 a desirable profit, Q the covariance matrix, and x_i the proportion of the total funds invested on asset s_i. The following Markowitz seminal mean-variance (MV) model,

$$\begin{array}{ll} \min & \frac{1}{2}x^T Q x \\ \text{s.t.} & e^T x = 1, \quad r^T x = r_0, \;\; x \geq 0, \end{array}$$

has been widely used in portfolio selections. However, it is observed that some solutions of the MV model contain extremely small elements [16]. In practice, a portfolio with a large number of assets with very small holdings for some assets is clearly not desirable due to transaction costs and management complexity. Thus, investors would be willing to sacrifice a small degree of performance for a more manageable sparse portfolio. It is known that using the ℓ_q quasinorm $(0 < q < 1)$ as a regularizer often leads to a sparser solution (see, e.g., [10, 15]) than using theℓ_1 norm. Motivated by the ℓ_1-regularized Markowitz model [8] and the ℓ_2-norm constrained minimum-variance model [18], Chen et al. [12] recently proposed several ℓ_q-regularized Markowitz models. One of their models is of the form

$$\begin{array}{ll} \min & \frac{1}{2}x^T Q x - \alpha r^T x + \lambda \|x\|_q^q \\ \text{s.t.} & e^T x = 1, \quad\;\; x \geq 0, \end{array} \tag{4.1}$$

where $\lambda > 0$, $q \in (0, 1)$, e is the all-ones vector, and the scale $1/\alpha$ with $\alpha > 0$ is the risk-aversion coefficient to measure the tradeoff between the risk and the return of the portfolio. They also proposed the ℓ_2-ℓ_q double regularized Markowitz model

$$\begin{array}{ll} \min & \frac{1}{2}x^T Q x - \alpha r^T x + \varrho\|x\|_2^2 + \lambda \|x\|_q^q \\ \text{s.t.} & e^T x = 1, \end{array} \tag{4.2}$$

where $\varrho > 0$. Clearly, models (4.1) and (4.2) are special cases of problem (1.1).

Interior point methods are an important class of optimization methods, and have been used for solving optimization problems where the objective function involves the ℓ_q quasinorm [5, 12, 21]. In particular, Chen et al. [12] proposed an IP method for solving (4.1) and (4.2). In this section, we apply the AL method and the IP method [12] for solving models (4.1) and (4.2), and compare their performance. For the AL method, when applied to solve models (4.1) and (4.2), the constraint $e^T x = 1$ is treated as a hard constraint, while the constraint $x \geq 0$ is viewed as an easy one. The parameters of the AL method are chosen as follows. We set $\rho_0 = 1$, $\mu^0 = 0$, $\gamma = 10$,

$\tau = 10^{-2}$, $\eta = 0.9$, and

$$\Upsilon = \max\{f(x^{\mathbf{feas}}) + \Phi(x^{\mathbf{feas}}), \mathcal{L}(x^0_{\text{init}}, \mu^0, \rho_0)\},$$

where $x^{\mathbf{feas}} = e/n$ is a feasible solution of (4.1) and (4.2), and $x^0_{\text{init}} = e/n$ is the initial point of the AL method. In addition, in step (1) of the AL method, we apply the NPG method (Algorithm 3.2) to solve the AL subproblem (3.3) with $\mu = \mu^k$, $\nu = \nu^k$, and $\rho = \rho_k$. As discussed in the second paragraph after Algorithm 3.2, (3.35) is a separable minimization problem with n one-dimensional problems. By using the monotonicity of objective functions, these one-dimensional problems can be easily solved by Newton's method. For the NPG method, we set $L_{\min} = 1$, $L_{\max} = 10^8$, $\theta = 5$, $M = 10$, and $\sigma = 10^{-4}$. The NPG method is terminated once condition (3.5) is satisfied with $\epsilon_k = 10^{-5}$. We terminate the AL method once the approximate solution x^k of the kth AL subproblem satisfies $\max\{|e^T x^k - 1|, \|x^k - x^{k-1}\|\} \leq 10^{-5}$. For the IP method, the parameters are the same as those in [12], and the initial point and the termination condition are set the same as the AL method; that is, the IP method started from e/n and terminated when the distance between the points generated in two consecutive iterations is not greater than 10^{-5} (the feasibility condition $e^T x = 1$ is automatically satisfied at every iteration point for the IP method). The AL method and the IP method are both coded in MATLAB, and all computations are performed on a Lenovo laptop (1.80 GHz–2.40 GHz, 7.92GB RAM) with MATLAB R2015b.

In the first experiment, we apply the AL and IP methods to solve model (4.1) with $q = 1/2$, in which Q and r are randomly generated. In particular, we generate the matrix Q by setting $Q = \hat{Q}^T \hat{Q}$, where the entries of matrix $\hat{Q}$ are randomly chosen from the standard normal distribution. The vector r is also randomly generated by the standard normal distribution.

The numerical results of the AL and IP methods for the above instances are reported in Tables 4.1 and 4.2 for $\alpha = 0.05$ and $\alpha = 0.2$, respectively. In detail, the problem size n and the regularization parameter λ are listed in the first two columns, respectively. We report the objective function value of model (4.1) in the columns labeled **ObjVal**, where x is an approximate solution found by these two methods. The CPU time (in seconds) of both methods is given in the columns labeled **CPU**. As observed in our experiment, the approximate solution x found by the IP method is *fully dense*. For a fair comparison with the AL method, similarly as in [12], we only count the number of entries of $|x|$ greater than 10^{-5}, which is referred to as the number of truncated nonzeros (NTNZ). We report NTNZ for both methods in the columns labeled **ntnz**. Besides, in the **nnz** column we present the number of hard (actual) zeros of the approximate solution of the AL method. From Tables 4.1 and 4.2, we observe that the approximate solutions obtained by these two methods have similar NTNZ and also objective function values, but the AL method is substantially faster than the IP method. Moreover, for both methods, the NTNZ decreases and the objective function value increases as λ increases.

In the second experiment, we apply the AL and IP methods to solve model (4.1), in which Q and r are estimated from some samples collected from stock market. In particular, we set $q = 1/2$ and $\alpha = 0.1, 0.2, 0.3, 0.4$ for model (4.1). To estimate Q and r, we collect historical daily stock return data of A-shares in Shanghai and Shenzhen Stock Exchanges from China Stock Market Trading Database provided by the China Stock Market and Accounting Research (CSMAR) center, spanning from January 1, 2007 to December 31, 2014. It shall be mentioned that we only concentrate on the stocks which have at least 80% observations during the entire data period. This results in 1468 stocks with 1944 observations. We then estimate Q and r as the sample

Table 4.1
Comparison of the AL and IP methods for (4.1) *with random instances and* $\alpha = 0.05$.

Data		IP method			AL method			
n	λ	**ObjVal**	**ntnz**	**CPU**	**ObjVal**	**ntnz**	**nnz**	**CPU**
500	1e-5	9.3e-02	331	8.1	9.3e-02	329	329	1.4
	1e-4	9.5e-02	326	7.9	9.5e-02	324	324	0.9
	1e-3	1.1e-01	308	6.4	1.1e-01	305	305	1.2
	1e-2	2.5e-01	251	4.8	2.5e-01	236	236	2.0
1000	1e-5	9.5e-02	669	39.8	9.9e-02	663	663	2.3
	1e-4	9.7e-02	660	49.7	1.0e-01	650	650	2.8
	1e-3	1.2e-01	624	41.3	1.2e-01	602	602	4.3
	1e-2	3.2e-01	476	38.6	3.2e-01	439	439	4.1
1500	1e-5	1.1e-01	1026	120.6	1.1e-01	1007	1008	2.8
	1e-4	1.1e-01	1012	150.4	1.1e-01	994	994	5.7
	1e-3	1.4e-01	925	91.4	1.4e-01	898	898	7.7
	1e-2	3.8e-01	695	151.4	3.8e-01	601	601	7.9
2000	1e-5	1.1e-01	1355	300.0	1.0e-01	1347	1352	4.2
	1e-4	1.1e-01	1342	322.0	1.0e-01	1318	1318	6.7
	1e-3	1.4e-01	1228	256.7	1.4e-01	1206	1206	10.3
	1e-2	4.2e-01	896	348.1	4.1e-01	788	788	7.4

Table 4.2
Comparison of the AL and IP methods for (4.1) *with random instances and* $\alpha = 0.2$.

Data		IP method			AL method			
n	λ	**ObjVal**	**ntnz**	**CPU**	**ObjVal**	**ntnz**	**nnz**	**CPU**
500	1e-5	5.3e-02	315	9.6	5.3e-02	315	315	0.8
	1e-4	5.5e-02	314	8.6	5.5e-02	313	313	0.8
	1e-3	7.0e-02	294	6.1	6.9e-02	288	288	1.7
	1e-2	2.1e-01	248	8.8	2.1e-01	229	229	1.1
1000	1e-5	6.1e-02	654	31.9	7.4e-02	643	643	1.9
	1e-4	6.4e-02	647	58.5	7.6e-02	633	633	2.7
	1e-3	8.5e-02	599	46.0	9.7e-02	598	598	4.1
	1e-2	2.8e-01	470	54.2	2.9e-01	419	419	3.7
1500	1e-5	7.0e-02	970	117.7	8.0e-02	965	966	3.0
	1e-4	7.3e-02	953	185.8	8.2e-02	950	950	3.6
	1e-3	9.9e-02	892	157.1	1.1e-01	879	879	8.4
	1e-2	3.4e-01	674	209.9	3.5e-01	592	592	7.6
2000	1e-5	7.9e-02	1301	279.6	6.5e-02	1283	1285	3.4
	1e-4	8.2e-02	1288	449.7	6.9e-02	1255	1255	8.0
	1e-3	1.1e-01	1196	366.8	9.8e-02	1138	1138	11.0
	1e-2	3.9e-01	890	352.6	3.6e-01	785	785	10.9

covariance and the sample mean of these samples. The initial point, the termination criterion, and all the parameters for the AL method and the IP method are the same as those mentioned above.

The numerical results of the AL and IP methods on model (4.1) with real-world data are reported in Table 4.3. In detail, the coefficients λ and α are listed in the first two columns, respectively. The other columns of Table 4.3 are the same as those in Tables 4.1 and 4.2. From Table 4.3, we observe that the approximate solutions obtained by these two methods have similar NTNZ and objective function values but the AL method is much faster than the IP method. Moreover, for both methods, the NTNZ and objective function value decrease as α increases.

Table 4.3

Comparison of the AL and IP methods for (4.1) *with real-world data.*

Data		IP method			AL method			
λ	α	**ObjVal**	**ntnz**	**CPU**	**ObjVal**	**ntnz**	**nnz**	**CPU**
5e-8	0.1	-1.2e-04	52	139.0	-1.2e-04	55	55	35.1
	0.2	-4.1e-04	40	209.9	-4.1e-04	41	41	34.5
	0.3	-7.3e-04	37	191.7	-7.3e-04	37	37	38.7
	0.4	-1.1e-03	30	173.2	-1.1e-03	32	32	31.6
1e-7	0.1	-1.2e-04	47	113.0	-1.2e-04	55	55	33.8
	0.2	-4.1e-04	40	165.8	-4.1e-04	41	41	27.8
	0.3	-7.3e-04	37	187.7	-7.3e-04	37	37	31.4
	0.4	-1.1e-03	30	175.3	-1.1e-03	31	31	33.4
1e-6	0.1	-1.1e-04	39	207.8	-1.1e-04	46	46	27.9
	0.2	-4.0e-04	36	168.4	-4.0e-04	38	38	46.9
	0.3	-7.3e-04	32	167.6	-7.3e-04	34	34	20.4
	0.4	-1.1e-03	27	181.7	-1.1e-03	30	30	20.2
5e-6	0.1	-9.2e-05	30	226.2	-9.2e-05	35	35	22.2
	0.2	-3.8e-04	28	252.4	-3.8e-04	33	33	18.3
	0.3	-7.1e-04	28	222.5	-7.1e-04	30	30	21.2
	0.4	-1.1e-03	23	234.0	-1.1e-03	26	26	18.8

In the third experiment, we apply the AL and IP methods to solve model (4.2), where $q = 1/2$ and $\alpha = 0.1$, where Q and r are the same as those in the second experiment. The initial point, the termination criterion, and all the parameters for the AL method and the IP method are the same as those mentioned above.

The numerical results of the AL and IP methods on model (4.2) with real-world data are reported in Table 4.4. In detail, the coefficients ϱ and λ are listed in the first and second columns, respectively, while the other columns are the same as those in Table 4.3. From Table 4.4, we observe that the AL method is comparable to the IP method in terms of solution quality. However, the AL method is substantially faster than the IP method.

Table 4.4

Comparison of the AL and IP methods for (4.2) *with real-world data.*

Data		IP method			AL method			
ϱ	λ	**ObjVal**	**ntnz**	**CPU**	**ObjVal**	**ntnz**	**nnz**	**CPU**
0.01	5e-8	-1.4e-04	1390	713.3	-1.4e-04	697	1468	8.7
	1e-7	-1.3e-04	1311	811.9	-1.3e-04	648	1095	14.9
	1e-6	-8.1e-05	782	1028.1	-4.8e-05	222	229	24.0
	5e-6	4.2e-06	157	2001.7	6.7e-06	149	149	8.2
0.005	5e-8	-2.3e-04	1410	731.7	-2.3e-04	643	1222	17.0
	1e-7	-2.3e-04	1319	820.0	-2.2e-04	578	1058	21.7
	1e-6	-1.6e-04	826	1261.9	-9.2e-05	173	186	23.2
	5e-6	-4.0e-05	149	1952.2	-2.5e-05	100	100	9.9
0.002	5e-8	-4.1e-04	1402	928.6	-4.0e-04	592	1174	41.1
	1e-7	-4.0e-04	1333	1294.1	-3.8e-04	526	998	32.0
	1e-6	-3.0e-04	815	1861.1	-1.7e-04	152	174	24.8
	5e-6	-1.1e-04	125	2080.6	-5.4e-05	70	70	13.2
0.001	5e-8	-6.0e-04	1416	1052.8	-5.8e-04	569	1140	68.0
	1e-7	-5.8e-04	1362	1292.9	-5.6e-04	493	961	78.9
	1e-6	-4.6e-04	783	2580.9	-2.4e-04	135	160	38.0
	5e-6	-1.6e-04	105	2159.2	-7.0e-05	52	52	15.0

To end this section, we consider the ℓ_2-ℓ_q double regularized Markowitz model with a noise tolerance:

$$\begin{array}{ll} \min & \frac{1}{2}x^TQx + \varrho\|x\|_2^2 + \lambda\|x\|_q^q \\ \text{s.t.} & \|Ax-b\|^2 \le \varepsilon^2, \end{array} \tag{4.3}$$

where $A := (e, r)^T$, $b := (1, r_0)^T$, and $\varepsilon > 0$. Problem (4.3) has quadratic and non-Lipschitz functions in the objective and a nonlinear function in the constraint. Algorithms for solving such a problem are hard to find in the literature. We now apply our AL method to solve problem (4.3), in which $r_0 = 0.0005$, $\varrho = 0.005$, $q = 1/2$, and Q and r are the same as those in the second experiment. The initial point, the termination condition, and all the parameters for the AL method are the same as those mentioned above except $\nu^0 = 0$, $x^{\mathbf{feas}} = A^\dagger b$, and

$$\Upsilon = \max\{f(x^{\mathbf{feas}}) + \Phi(x^{\mathbf{feas}}), \mathcal{L}(x^0_{\text{init}}, \nu^0, \rho_0)\},$$

where $A^\dagger$ denotes the Moore–Penrose pseudoinverse of matrix A. It is clear to see that such $A^\dagger b$ is a feasible solution to problem (4.3). The numerical results of the AL method for model (4.3) with real-world data are reported in Table 4.5. In detail, the noise tolerance ε is listed in the first column while the other columns are similar to those presented in Table 4.3. The numerical results demonstrate that our AL method is capable of solving non-Lipschitz programming with nonlinear constraints.

Table 4.5

Results of the AL method for (4.3) *with real-world data.*

ε	λ	**ObjVal**	**ntnz**	**nnz**	**CPU**	λ	**ObjVal**	**ntnz**	**nnz**	**CPU**
1e-1	5e-8	6.3e-05	653	1151	6.7	1e-7	6.6e-05	587	977	5.3
1e-2		7.6e-05	658	1246	4.2		7.8e-05	602	1089	4.0
1e-3		7.7e-05	657	1239	4.4		8.0e-05	603	1097	4.4
1e-4		7.7e-05	657	1249	4.3		8.0e-05	599	1073	4.2
1e-1	1e-6	9.6e-05	181	181	3.6	5e-6	1.4e-04	111	111	3.7
1e-2		1.1e-04	182	183	2.5		1.6e-04	111	111	2.6
1e-3		1.2e-04	182	183	2.7		1.6e-04	111	111	2.6
1e-4		1.1e-04	196	201	2.6		1.6e-04	111	111	2.6

Acknowledgments. The authors would like to express their gratitude to the referees for their helpful comments and constructive suggestions. The authors are also grateful to Jiming Peng, Ting Kei Pong, and Caihua Chen for their helpful comments.

REFERENCES

[1] R. Andreani, G. Haeser, M.L. Schuverdt, and P.J.S. Silva, *A relaxed constant positive linear dependence constraint qualification and applications*, Math. Program., 135 (2012), pp. 255–273, https://doi.org/10.1007/s10107-011-0456-0.

[2] J. Barzilai and J.M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148, https://doi.org/10.1093/imanum/8.1.141.

[3] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, Belmont, MA, 1996.

[4] W. Bian and X. Chen, *Linearly constrained non-Lipschitz optimization for image restoration*, SIAM J. Imaging Sci., 8 (2015), pp. 2294–2322, https://doi.org/10.1137/140985639.

[5] W. Bian, X. Chen, and Y. Ye, *Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization*, Math. Program., 149 (2012), pp. 301–327, https://doi.org/10.1007/s10107-014-0753-5.

[6] T. Blumensath and M.E. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654, https://doi.org/10.1007/s00041-008-9035-z.

[7] J.M. Borwein, J.S. Treiman, and Q.J. Zhu, *Necessary conditions for constrained optimization problems with semicontinuous and continuous data*, Trans. Amer. Math. Soc., 350 (1998), pp. 2409–2429.

[8] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris, *Sparse and stable Markowitz portfolios*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 12267–12272, https://doi.org/10.1073/pnas.0904287106.

[9] A.M. Bruckstein, D.L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81, https://doi.org/10.1137/060657704.

[10] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Process. Lett., 14 (2007), pp. 707–710, https://doi.org/10.1109/LSP.2007.898300.

[11] S.S. Chen, D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61, https://doi.org/10.1137/S1064827596304010.

[12] C. Chen, X. Li, C. Tolman, S. Wang, and Y. Ye, *Sparse portfolio selection via quasi-norm regularization*, preprint, arXiv:1312.6350, 2014, https://arxiv.org/abs/1312.6350.

[13] X. Chen, L. Guo, Z. Lu, and J.J. Ye, *Supplementary material of this paper: Lower bounds of nonzero entries in solutions*, http://www.polyu.edu.hk/ama/staff/xjchen/ChenXJ.htm.

[14] X. Chen, Z. Lu, and T.K. Pong, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM J. Optim., 26 (2016), pp. 1465–1492, https://doi.org/10.1137/15M1028054.

[15] X. Chen, F. Xu, and Y. Ye, *Lower bound theory of nonzero entries in solutions of l_2-l_p minimization*, SIAM J. Sci. Comput., 32 (2010), pp. 2832–2852, https://doi.org/10.1137/090761471.

[16] G. Cornuejols and R. Tütüncü, *Optimization Methods in Finance*, Cambridge University Press, Cambridge, UK, 2007.

[17] F.E. Curtis, H. Jiang, and D.P. Robinson, *An adaptive augmented Lagrangian method for large-scale constrained optimization*, Math. Program., 152 (2015), pp. 201–245, https://doi.org/10.1007/s10107-014-0784-y.

[18] V. DeMiguel, L. Garlappi, F.J. Nogales, and R. Uppal, *A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms*, Manage. Sci., 55 (2009), pp. 798–812, https://doi.org/10.1287/mnsc.1080.0986.

[19] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360, https://doi.org/10.1198/016214501753382273.

[20] W.J. Fu, *Penalized regression: The bridge versus the lasso*, J. Comput. Graph. Statist., 7 (1998), pp. 397–416, https://doi.org/10.1080/10618600.1998.10474784.

[21] D. Ge, X. Jiang, and Y. Ye, *A note on complexity of L_p minimization*, Math. Program., 129 (2011), pp. 285–299, https://doi.org/10.1007/s10107-011-0470-2.

[22] D. Geman and G. Reynolds, *Constrained restoration and the recovery of discontinuities*, IEEE Trans. Pattern Anal. Mach. Intell., 14 (1992), pp. 367–383, https://doi.org/10.1109/34.120331.

[23] L. Guo, J. Zhang, and G.-H. Lin, *New results on constraint qualifications for nonlinear extremum problems and extensions*, J. Optim. Theory Appl., 163 (2014), pp. 737–754, https://doi.org/10.1007/s10957-013-0510-6.

[24] Y. Liu, Y.-H. Dai, and S. Ma, *Joint power and admission control: Non-convex L_q approximation and an effective polynomial time deflation approach*, IEEE Trans. Signal Process., 63 (2015), pp. 3641–3656, https://doi.org/10.1109/TSP.2015.2428224.

[25] Y.-F. Liu, S. Ma, Y.-H. Dai, and S. Zhang, *A smoothing SQP framework for a class of composite ℓ_q minimization over polyhedron*, Math. Program., 158 (2016), pp. 467–500, https://doi.org/10.1007/s10107-015-0939-5.

[26] Z. Lu, *Iterative hard thresholding methods for l_0 regularized convex cone programming*, Math. Program., 147 (2014), pp. 125–154, https://doi.org/10.1007/s10107-013-0714-4.

[27] Z. Lu and Y. Zhang, *An augmented Lagrangian approach for sparse principal component analysis*, Math. Program., 135 (2012), pp. 149–193, https://doi.org/10.1007/s10107-011-0452-4.

[28] B.S. Mordukhovich, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526–530.

[29] A.Y. Ng, *Feature selection, ℓ_1 vs ℓ_2 regularization and rotational invariance*, in Proceedings of the Twenty-First International Conference on Machine Learning, ACM, New York, 2004, 78, https://doi.org/10.1145/1015330.1015435.

[30] M. Nikolova, M.K. Ng, S. Zhang, and W.-K. Ching, *Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization*, SIAM J. Imaging Sci., 1 (2008), pp. 2–25, https://doi.org/10.1137/070692285

[31] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[32] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal Statist. Soc. Ser. B, 58 (1996), pp. 267–288.

[33] S.J. Wright, D.N. Robert, and M.A.T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Proces., 57 (2009), pp. 2479–2493, https://doi.org/10.1109/TSP.2009.2016892.

[34] T. Zhang, *Analysis of multi-stage convex relaxation for sparse regularization*, J. Mach. Learn. Res., 11 (2010), pp. 1081–1107.