Global convergence of splitting methods for nonconvex composite optimization

Guoyin Li * Ting Kei Pong †

November 30, 2014

Abstract

We consider the problem of minimizing the sum of a smooth function h with a bounded Hessian, and a nonsmooth function. We assume that the latter function is a composition of a proper closed function P and a surjective linear map $\mathcal{M}$, with the proximal mappings of τP, $\tau > 0$, simple to compute. This problem is nonconvex in general and encompasses many important applications in engineering and machine learning. In this paper, we examined two types of splitting methods for solving this nonconvex optimization problem: alternating direction method of multipliers and proximal gradient algorithm. For the direct adaptation of the alternating direction method of multipliers, we show that, if the penalty parameter is chosen sufficiently large and the sequence generated has a cluster point, then it gives a stationary point of the nonconvex problem. We also establish convergence of the whole sequence under an additional assumption that the functions h and P are semi-algebraic. Furthermore, we give simple sufficient conditions to guarantee boundedness of the sequence generated. These conditions can be satisfied for a wide range of applications including the least squares problem with the $\ell_{1/2}$ regularization. Finally, when $\mathcal{M}$ is the identity so that the proximal gradient algorithm can be efficiently applied, we show that any cluster point is stationary under a slightly more flexible constant step-size rule than what is known in the literature for a nonconvex h.

1 Introduction

In this paper, we consider the following optimization problem:

$$\min_{x} \quad h(x) + P(\mathcal{M}x), \tag{1}$$

where $\mathcal{M}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, P is a proper closed function on $\mathbb{R}^m$ and h is twice continuously differentiable on $\mathbb{R}^n$ with a bounded Hessian. We also assume that the proximal (set-valued) mappings

$$u \mapsto \operatorname*{Arg\,min}_{y} \left\{ \tau P(y) + \frac{1}{2}\|y-u\|^2 \right\}$$

are well-defined and are simple to compute for all u and for any $\tau > 0$. Here, Arg min denotes the set of minimizers, and the simplicity is understood in the sense that *at least one* element of the set of minimizers can be obtained efficiently. Concrete examples of such P that arise in applications

*Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia. E-mail: g.li@unsw.edu.au. This author was partially supported by a research grant from Australian Research Council.

†Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong. This author was also supported as a PIMS Postdoctoral Fellow at Department of Computer Science, University of British Columbia, Vancouver, Canada, during the early stage of preparation of this manuscript. E-mail: tk.pong@polyu.edu.hk.

include functions listed in [18, Table 1], the $\ell_{1/2}$ regularization [31], the ℓ_0 regularization, and the indicator functions of the set of vectors with cardinality at most s [4], matrices with rank at most r and s-sparse vectors in simplex [21], etc. Moreover, for a large class of nonconvex functions, a general algorithm has been proposed recently in [19] for computing the proximal mapping.

The model problem (1) with h and P satisfying the above assumptions encompasses many important applications in engineering and machine learning; see, for example, [4, 10, 11, 18, 22]. In particular, many sparse learning problems are in the form of (1) with h being a loss function, $\mathcal{M}$ being the identity map and P being a regularizer; see, for example, [4] for the use of the ℓ_0 norm as a regularizer, [11] for the use of the ℓ_1 norm, [10] for the use of the nuclear norm, and [18] and the references therein for the use of various continuous difference-of-convex functions with simple proximal mappings. For the case when $\mathcal{M}$ is not the identity map, an application in stochastic realization where h is a multiple of the trace inner product, P is the rank function and $\mathcal{M}$ is the linear map that takes the variable x into a block Hankel matrix was discussed in [22, Section II].

When $\mathcal{M}$ is the identity map, the proximal gradient algorithm [15, 16, 25] (also known as forward-backward splitting algorithm) can be applied whose subproblem involves a computation of the proximal mapping of τP for some $\tau > 0$. It is known that when h and P are convex, the sequence generated from this algorithm is convergent to a globally optimal solution if the step-size is chosen from $(0, \frac{2}{L})$, where L is any number larger than the Lipschitz continuity modulus of ∇h. For nonconvex h and P, the step-size can be chosen from $(0, \frac{1}{L})$ so that any cluster point of the sequence generated is stationary [7, Proposition 2.3] (see Section 2 for the definition of stationary points), and convergence of the whole sequence is guaranteed if the sequence generated is bounded and $h + P$ satisfies the Kurdyka-Łojasiewicz (KL) property [3, Theorem 5.1, Remark 5.2(a)]. On the other hand, when $\mathcal{M}$ is a general linear map so that the computation of the proximal mapping of $\tau P \circ \mathcal{M}$, $\tau > 0$, is not necessarily simple, the proximal gradient algorithm cannot be applied efficiently. In the case when h and P are both convex, one feasible approach is to apply the alternating direction method of multipliers (ADMM) [13, 14, 17]. This has been widely used recently; see, for example [8, 9, 27, 28, 30]. While it is tempting to directly apply the ADMM to the nonconvex problem (1), convergence has only been shown under specific assumptions. In particular, in [29], the authors studied an application that can be modeled as (1) with $h = 0$, P being some risk measures and $\mathcal{M}$ typically being an injective linear map coming from data. They showed that any cluster point gives a stationary point, assuming square summability of the successive changes in the dual iterates. More recently, in [1], the authors considered the case when h is a nonconvex quadratic and P is the sum of the ℓ_1 norm and the indicator function of the Euclidean norm ball. They showed that if the penalty parameter is chosen sufficiently large (with an explicit lower bound) and the dual iterates satisfy a particular assumption, then any cluster point gives a stationary point. In particular, their assumption is satisfied if $\mathcal{M}$ is surjective.

Motivated by the findings in [1], in this paper, we focus on the case when $\mathcal{M}$ is surjective and consider both the ADMM (for a general surjective $\mathcal{M}$) and the proximal gradient algorithm (for $\mathcal{M}$ being the identity). The contributions of this paper are as follows:

- First, we characterize cluster points of the sequence generated from the ADMM. In particular, we show that if the (fixed) penalty parameter in the ADMM is chosen sufficiently large (with a computable lower bound), and a cluster point of the sequence generated exists, then it gives a stationary point of problem (1). This extends the result in [1] to the more general model problem (1) in the case where $\mathcal{M}$ is surjective.

  Moreover, our analysis allows replacing h in the ADMM subproblems by its local quadratic approximations so that in each iteration of this variant, the subproblems only involve computing the proximal mapping of τP for some $\tau > 0$ and solving an unconstrained convex quadratic minimization problem. Furthermore, we also give simple sufficient conditions to guarantee the boundedness of the sequence generated. These conditions are satisfied in a wide range of applications; see Examples 4, 5 and 6.

- Second, under the additional assumption that h and P are semi-algebraic functions, we show that if a cluster point of the sequence generated from the ADMM exists, it is actually convergent. Our assumption on semi-algebraicity not only can be easily verified or recognized, but also covers a broad class of optimization problems such as problems involving quadratic functions, polyhedral norms and the cardinality function.

- Finally, for the particular case when $\mathcal{M}$ equals the identity map, we show that the proximal gradient algorithm can be applied with a slightly more flexible step-size rule when h is nonconvex (see Theorem 5 for the precise statement).

The rest of the paper is organized as follows. We discuss notation and preliminary materials in the next section. Convergence of the ADMM is analyzed in Section 3, and Section 4 is devoted to the analysis of the proximal gradient algorithm. Some numerical results are presented in Section 5 to illustrate the algorithms. We give concluding remarks and discuss future research directions in Section 6.

2 Notation and preliminaries

We denote the n-dimensional Euclidean space as $\mathbb{R}^n$, and use $\langle \cdot, \cdot \rangle$ to denote the inner product and $\|\cdot\|$ to denote the norm induced from the inner product. Linear maps are denoted by scripted letters. The identity map is denoted by $\mathcal{I}$. For a linear map $\mathcal{M}$, $\mathcal{M}^*$ denotes the adjoint linear map with respect to the inner product and $\|\mathcal{M}\|$ is the induced operator norm of $\mathcal{M}$. A linear self-map $\mathcal{T}$ is called symmetric if $\mathcal{T} = \mathcal{T}^*$. For a symmetric linear self-map $\mathcal{T}$, we use $\|\cdot\|_{\mathcal{T}}^2$ to denote its induced quadratic form given by $\|x\|_{\mathcal{T}}^2 = \langle x, \mathcal{T}x \rangle$ for all x, and use $\lambda_{\max}$ (resp., $\lambda_{\min}$) to denote the maximum (resp., minimum) eigenvalue of $\mathcal{T}$. A symmetric linear self-map $\mathcal{T}$ is called positive semidefinite, denoted by $\mathcal{T} \succeq 0$ (resp., positive definite, $\mathcal{T} \succ 0$) if $\|x\|_{\mathcal{T}}^2 \geq 0$ (resp., $\|x\|_{\mathcal{T}}^2 > 0$) for all nonzero x. For two symmetric linear self-maps $\mathcal{T}_1$ and $\mathcal{T}_2$, we use $\mathcal{T}_1 \succeq \mathcal{T}_2$ (resp., $\mathcal{T}_1 \succ \mathcal{T}_2$) to denote $\mathcal{T}_1 - \mathcal{T}_2 \succeq 0$ (resp., $\mathcal{T}_1 - \mathcal{T}_2 \succ 0$).

An extended-real-valued function f is called proper if it is finite somewhere and never equals $-\infty$. Such a function is called closed if it is lower semicontinuous. Given a proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$, we use the symbol $z \xrightarrow{f} x$ to indicate $z \to x$ and $f(z) \to f(x)$. The domain of f is denoted by $\mathrm{dom} f$ and is defined as $\mathrm{dom} f = \{x \in \mathbb{R}^n : f(x) < +\infty\}$. Our basic *subdifferential* of f at $x \in \mathrm{dom}\, f$ (known also as the limiting subdifferential) is defined by

$$\partial f(x) := \left\{ v \in \mathbb{R}^n : \exists x^t \xrightarrow{f} x,\ v^t \to v \ \text{ with } \liminf_{z \to x^t} \frac{f(z) - f(x^t) - \langle v^t, z - x^t \rangle}{\|z - x^t\|} \geq 0 \text{ for each } t \right\}. \tag{2}$$

It follows immediately from the above definition that this subdifferential has the following robustness property:

$$\left\{ v \in \mathbb{R}^n : \exists x^t \xrightarrow{f} x,\ v^t \to v\ , v^t \in \partial f(x^t) \right\} \subseteq \partial f(x). \tag{3}$$

For a convex function f the subdifferential (2) reduces to the classical subdifferential in convex analysis (see, for example, [23, Theorem 1.93])

$$\partial f(x) = \{ v \in \mathbb{R}^n : \langle v, z - x \rangle \leq f(z) - f(x) \ \ \forall\, z \in \mathbb{R}^n \}.$$

Moreover, for a continuously differentiable function f, the subdifferential (2) reduces to the derivative of f denoted by ∇f. For a function f with more than one group of variables, we use $\partial_x f$

(resp., $\nabla_x f$) to denote the subdifferential (resp., derivative) of f with respect to the variable x. Furthermore, we write $\mathrm{dom}\,\partial f = \{x \in \mathbb{R}^n : \partial f(x) \neq \emptyset\}$.

In general, the subdifferential set (2) can be nonconvex (e.g., for $f(x) = -|x|$ at $0 \in \mathbb{R}$) while ∂f enjoys comprehensive calculus rules based on *variational/extremal principles* of variational analysis [24]. In particular, when $\mathcal{M}$ is a surjective linear map, using [24, Exercise 8.8(c)] and [24, Exercise 10.7], we see that

$$\partial(h + P \circ \mathcal{M})(x) = \nabla h(x) + \mathcal{M}^* \partial P(\mathcal{M}x)$$

for any $x \in \mathrm{dom}(P \circ \mathcal{M})$. Hence, at an optimal solution $\bar{x}$, the following necessary optimality condition always holds:

$$0 \in \partial(h + P \circ \mathcal{M})(\bar{x}) = \nabla h(\bar{x}) + \mathcal{M}^* \partial P(\mathcal{M}\bar{x}). \tag{4}$$

Throughout this paper, we say that $\widetilde{x}$ is a stationary point of (1) if $\widetilde{x}$ satisfies (4) in place of $\bar{x}$.

For a continuously differentiable function ϕ on $\mathbb{R}^n$, the Bregman distance D_ϕ is defined as

$$D_\phi(x_1, x_2) := \phi(x_1) - \phi(x_2) - \langle \nabla\phi(x_2), x_1 - x_2\rangle$$

for any x_1, $x_2 \in \mathbb{R}^n$. If ϕ is twice continuously differentiable and there exists $\mathcal{Q}$ so that the Hessian $\nabla^2\phi$ satisfies $[\nabla^2\phi(x)]^2 \preceq \mathcal{Q}$ for all x, then for any x_1 and x_2 in $\mathbb{R}^n$, we have

$$\begin{aligned}
&\|\nabla\phi(x_1) - \nabla\phi(x_2)\|^2 = \left\| \int_0^1 \nabla^2\phi(x_2 + t(x_1 - x_2)) \cdot [x_1 - x_2] dt \right\|^2 \\
&\le \left(\int_0^1 \left\| \nabla^2\phi(x_2 + t(x_1 - x_2)) \cdot [x_1 - x_2] \right\| dt \right)^2 \\
&= \left(\int_0^1 \sqrt{\langle x_1 - x_2, [\nabla^2\phi(x_2 + t(x_1 - x_2))]^2 \cdot [x_1 - x_2]\rangle} dt \right)^2 \le \|x_1 - x_2\|^2_{\mathcal{Q}}.
\end{aligned} \tag{5}$$

On the other hand, if there exists $\mathcal{Q}$ so that $\nabla^2\phi(x) \succeq \mathcal{Q}$ for all x, then

$$\begin{aligned}
&D_\phi(x_1, x_2) = \int_0^1 \langle \nabla\phi(x_2 + t(x_1 - x_2)) - \nabla\phi(x_2), x_1 - x_2\rangle dt \\
&= \int_0^1 \int_0^1 t\langle x_1 - x_2, \nabla^2\phi(x_2 + st(x_1 - x_2)) \cdot [x_1 - x_2]\rangle ds\ dt \ge \frac{1}{2}\|x_1 - x_2\|^2_{\mathcal{Q}}
\end{aligned} \tag{6}$$

for any x_1 and x_2 in $\mathbb{R}^n$.

A semi-algebraic set $S \subseteq \mathbb{R}^n$ is a finite union of sets of the form

$$\{x \in \mathbb{R}^n : h_1(x) = \cdots = h_k(x) = 0, g_1(x) < 0, \ldots, g_l(x) < 0\},$$

where $h_1, \ldots, h_k$ and $g_1, \ldots, g_l$ are polynomials with real coefficients in n variables. In other words, S is a union of finitely many sets, each defined by finitely many polynomial equalities and strict inequalities. A map $F : \mathbb{R}^n \to \mathbb{R}$ is semi-algebraic if $\mathrm{gph} F \in \mathbb{R}^{n+1}$ is a semi-algebraic set. Semi-algebraic sets and semi-algebraic mappings enjoy many nice structural properties. One important property which we will use later on is the Kurdyka-Łojasiewicz (KL) property.

Definition 1. (KL property & KL function) *A proper function f is said to have the Kurdyka-Łojasiewicz (KL) property at $\widehat{x} \in \mathrm{dom}\,\partial f$ if there exist $\eta \in (0, \infty]$, a neighborhood V of $\widehat{x}$ and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:*

(i) *$\varphi(0) = 0$ and φ is continuously differentiable on $(0, \eta)$ with positive derivatives;*

(ii) *for all* $x \in V$ *satisfying* $f(\widehat{x}) < f(x) < f(\widehat{x}) + \eta$, *it holds that*

$$\varphi'(f(x) - f(\widehat{x}))\,\mathrm{dist}(0, \partial f(x)) \geq 1.$$

A proper closed function f *satisfying the KL property at all points in* $\mathrm{dom}\,\partial f$ *is called a KL function.*

It is known that a proper closed semi-algebraic function is a KL function as such a function satisfies the KL property for all points in $\mathrm{dom}\,\partial f$ with $\varphi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and some $c > 0$ (for example, see [2, Section 4.3]; further discussion can be found in [6, Corollary 16] and [5, Section 2]).

3 Alternating direction method of multipliers

In this section, we study the alternating direction method of multipliers for finding a stationary point of (1). To describe the algorithm, we first reformulate (1) as

$$\begin{array}{ll} \min\limits_{x,y} & h(x) + P(y) \\ \text{s.t.} & y = \mathcal{M}x, \end{array}$$

to decouple the linear map and the nonsmooth part. Recall that the augmented Lagrangian function for the above problem is defined, for each $\beta > 0$, as:

$$L_\beta(x, y, z) := h(x) + P(y) - \langle z, \mathcal{M}x - y\rangle + \frac{\beta}{2}\|\mathcal{M}x - y\|^2.$$

Our algorithm is then presented as follows:

Proximal ADMM

Step 0. Input (x^0, y^0, z^0), $\beta > 0$ and a twice continuously differentiable convex function $\phi(x)$.

Step 1. Set

$$\begin{cases} y^{t+1} \in \mathop{\mathrm{Arg\,min}}\limits_{y} L_\beta(x^t, y, z^t), \\ x^{t+1} \in \mathop{\mathrm{Arg\,min}}\limits_{x} \{L_\beta(x, y^{t+1}, z^t) + D_\phi(x, x^t)\}, \\ z^{t+1} = z^t - \beta(\mathcal{M}x^{t+1} - y^{t+1}). \end{cases} \tag{7}$$

Step 2. If a termination criterion is not met, go to Step 1.

Notice that the first subproblem is essentially computing the proximal mapping of τP for some $\tau > 0$. The above algorithm is called the proximal ADMM since, in the second subproblem, we allow a proximal term D_ϕ and hence a choice of ϕ to simplify this subproblem. If $\phi = 0$, then this algorithm reduces to the usual ADMM described in, for example, [13]. For other popular non-trivial choices of ϕ, see Remark 1 below.

We next study global convergence of the above algorithm under suitable assumptions. Specifically, we consider the following assumption.

Assumption 1. (i) $\mathcal{M}\mathcal{M}^* \succeq \sigma\mathcal{I}$ *for some* $\sigma > 0$*; and there exist* $\mathcal{Q}_1$, $\mathcal{Q}_2$ *such that for all* x, $\mathcal{Q}_1 \succeq \nabla^2 h(x) \succeq \mathcal{Q}_2$.

(ii) *$\beta > 0$ and ϕ are chosen so that*

- *there exist $\mathcal{T}_1 \succeq \mathcal{T}_2 \succeq 0$ so that $\mathcal{T}_1^2 \succeq [\nabla^2\phi(x)]^2 \succeq \mathcal{T}_2^2$ for all x;*
- *$\mathcal{Q}_2 + \beta\mathcal{M}^*\mathcal{M} + \mathcal{T}_2 \succeq \delta\mathcal{I}$ for some $\delta > 0$;*
- *with $\mathcal{Q}_3 \succeq [\nabla^2 h(x) + \nabla^2\phi(x)]^2$ for all x, it holds that*

$$\begin{cases} \delta\mathcal{I} + \mathcal{T}_2 \succ \frac{4}{\sigma\beta}\mathcal{Q}_3 + \frac{4}{\sigma\beta}\mathcal{T}_1^2 & \text{if } \mathcal{T}_1 \neq 0, \\ \delta\mathcal{I} \succ \frac{2}{\sigma\beta}\mathcal{Q}_3 & \text{if } \mathcal{T}_1 = 0. \end{cases}$$

Remark 1. (Comments on Assumption 1) *Point (i) says $\mathcal{M}$ is surjective. The first and second points in (ii) would be satisfied if $\phi(x)$ is chosen to be $\frac{L}{2}\|x\|^2 - h(x)$, where L is at least as large as the Lipschitz continuity modulus of $\nabla h(x)$. In this case, one can pick $\mathcal{T}_1 = 2L\mathcal{I}$ and $\mathcal{T}_2 = 0$. This choice is of particular interest since it simplifies the x-update in (7) to a convex quadratic programming problem; see [26, Section 2.1]. Indeed, under this choice, we have*

$$D_\phi(x, x^t) = \frac{L}{2}\|x - x^t\|^2 - h(x) + h(x^t) + \langle \nabla h(x^t), x - x^t \rangle,$$

and hence the second subproblem becomes

$$\min_x \ \frac{L}{2}\|x - x^t\|^2 + \langle \nabla h(x^t) - \mathcal{M}^* z^t, x - x^t \rangle + \frac{\beta}{2}\|\mathcal{M}x - y^{t+1}\|^2.$$

Finally, point 3 in (ii) can always be enforced by picking β sufficiently large if ϕ, $\mathcal{T}_1$ and $\mathcal{T}_2$, are chosen independently of β.

Before stating our convergence results, we note first that from the optimality conditions, the iterates generated satisfy

$$\begin{aligned} &0 \in \partial P(y^{t+1}) + z^t - \beta(\mathcal{M}x^t - y^{t+1}), \\ &0 = \nabla h(x^{t+1}) - \mathcal{M}^* z^t + \beta\mathcal{M}^*(\mathcal{M}x^{t+1} - y^{t+1}) + (\nabla\phi(x^{t+1}) - \nabla\phi(x^t)). \end{aligned} \tag{8}$$

Hence, if

$$\lim_{t\to\infty} \|y^{t+1} - y^t\|^2 + \|x^{t+1} - x^t\|^2 + \|z^{t+1} - z^t\|^2 = 0, \tag{9}$$

and if for a cluster point (x^*, y^*, z^*) of the sequence $\{(x^t, y^t, z^t)\}$, we have

$$\lim_{i\to\infty} P(y^{t_i+1}) = P(y^*) \tag{10}$$

along a convergent subsequence $\{(x^{t_i}, y^{t_i}, z^{t_i})\}$ that converges to (x^*, y^*, z^*), then x^* is a stationary point of (1). To see this, notice from (8) and the definition of z^{t+1} that

$$\begin{cases} -z^{t+1} - \beta\mathcal{M}(x^{t+1} - x^t) \in \partial P(y^{t+1}), \\ \nabla h(x^{t+1}) - \mathcal{M}^* z^{t+1} = -\nabla\phi(x^{t+1}) + \nabla\phi(x^t), \\ \mathcal{M}x^{t+1} - y^{t+1} = \dfrac{1}{\beta}(z^t - z^{t+1}). \end{cases} \tag{11}$$

Passing to the limit in (11) along the subsequence $\{(x^{t_i}, y^{t_i}, z^{t_i})\}$ and invoking (9), (10) and (3), it follows that

$$\nabla h(x^*) = \mathcal{M}^* z^*, \quad -z^* \in \partial P(y^*), \quad y^* = \mathcal{M}x^*. \tag{12}$$

In particular, x^* is a stationary point of the model problem (1).

We now state our global convergence result. The first conclusion establishes (9) under Assumption 1, and so, any cluster point of the sequence generated from the proximal ADMM produces a stationary point of our model problem (1) such that (12) holds. The second conclusion states that if the algorithm is suitably initialized, we can get a strict improvement in objective values. In the special case where h is a nonconvex quadratic and P is the sum of the ℓ_1 norm and the indicator function of the Euclidean norm ball, this convergence analysis has been established for the ADMM (i.e., proximal ADMM with $\phi = 0$) in [1]. Moreover, the proof of our convergence result is inspired from the recent work [1, Section 3.3] and [29], and uses similar line of arguments therein.

Theorem 1. *Suppose that Assumption 1 holds. Then we have the following results.*

(i) **(Global subsequential convergence)** *If the sequence $\{(x^t, y^t, z^t)\}$ generated from the proximal ADMM has a cluster point (x^*, y^*, z^*), then* (9) *holds. Moreover, x^* is a stationary point of* (1) *such that* (12) *holds.*

(ii) **(Strict improvement in objective values)** *Suppose that the algorithm is initialized at a non-stationary x^0 with $h(x^0) + P(\mathcal{M}x^0) < \infty$, $y^0 = \mathcal{M}x^0$ and z^0 satisfying $\mathcal{M}^* z^0 = \nabla h(x^0)$. Then for any cluster point (x^*, y^*, z^*) of the sequence $\{(x^t, y^t, z^t)\}$, if exists, we have*

$$h(x^*) + P(\mathcal{M}x^*) < h(x^0) + P(\mathcal{M}x^0).$$

Remark 2. *The proximal ADMM does not necessarily guarantee that the objective value of* (1) *is decreasing along the sequence $\{x^t\}$ generated. However, under the assumptions in Theorem 1, any cluster point of the sequence generated from the proximal ADMM improves the starting (non-stationary) objective value.*

We now describe one way of choosing the initialization as suggested in (ii) when P is nonconvex. In this case, it is common to approximate P by a proper closed convex function $\widetilde{P}$ and obtain a relaxation to (1)*, i.e.,*

$$\min_x \; h(x) + \widetilde{P}(\mathcal{M}x).$$

Then any stationary point $\widetilde{x}$ of this relaxed problem, if exists, satisfies $-\nabla h(\widetilde{x}) \in \mathcal{M}^* \partial \widetilde{P}(\mathcal{M}\widetilde{x})$. Thus, if $P(\mathcal{M}\widetilde{x}) < \infty$, then one can initialize the proximal ADMM by taking $x^0 = \widetilde{x}$, $y^0 = \mathcal{M}\widetilde{x}$ and $z^0 \in -\partial \widetilde{P}(\mathcal{M}\widetilde{x})$ with $\nabla h(\widetilde{x}) = \mathcal{M}^* z^0$, so that the conditions in (ii) are satisfied.

Proof. We first focus on the case when $\mathcal{T}_1 \neq 0$. We will comment on the case when $\mathcal{T}_1 = 0$ at the end of the proof.

We start by showing that (9) holds. First, observe from the second relation in (11) that

$$\mathcal{M}^* z^{t+1} = \nabla h(x^{t+1}) + \nabla \phi(x^{t+1}) - \nabla \phi(x^t). \tag{13}$$

Consequently, we have

$$\mathcal{M}^*(z^{t+1} - z^t) = \nabla h(x^{t+1}) - \nabla h(x^t) + (\nabla \phi(x^{t+1}) - \nabla \phi(x^t)) - (\nabla \phi(x^t) - \nabla \phi(x^{t-1})).$$

Taking norm on both sides, squaring and making use of (i) in Assumption 1, we obtain further that

$$\begin{aligned}
&\sigma \|z^{t+1} - z^t\|^2 \le \|\mathcal{M}^*(z^{t+1} - z^t)\|^2 \\
&= \|\nabla h(x^{t+1}) - \nabla h(x^t) + (\nabla \phi(x^{t+1}) - \nabla \phi(x^t)) - (\nabla \phi(x^t) - \nabla \phi(x^{t-1}))\|^2 \\
&\le 2\|\nabla h(x^{t+1}) - \nabla h(x^t) + \nabla \phi(x^{t+1}) - \nabla \phi(x^t)\|^2 + 2\|\nabla \phi(x^t) - \nabla \phi(x^{t-1})\|^2 \\
&\le 2\|x^{t+1} - x^t\|^2_{\mathcal{Q}_3} + 2\|x^t - x^{t-1}\|^2_{\mathcal{T}_1^2},
\end{aligned} \tag{14}$$

where the last inequality follows from points 1 and 3 in (ii) of Assumption 1, and (5). On the other hand, from the definition of z^{t+1}, we have

$$y^{t+1} = \mathcal{M}x^{t+1} + \frac{1}{\beta}(z^{t+1} - z^t),$$

which implies

$$\|y^{t+1} - y^t\| \le \|\mathcal{M}(x^{t+1} - x^t)\| + \frac{1}{\beta}\|z^{t+1} - z^t\| + \frac{1}{\beta}\|z^t - z^{t-1}\|. \tag{15}$$

In view of (14) and (15), to establish (9), it suffices to show that

$$\lim_{t\to\infty} \|x^{t+1} - x^t\| = 0. \tag{16}$$

To prove (16), consider the difference

$$\begin{aligned} L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - L_\beta(x^t, y^t, z^t) &= (L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - L_\beta(x^{t+1}, y^{t+1}, z^t)) \\ &\quad + (L_\beta(x^{t+1}, y^{t+1}, z^t) - L_\beta(x^t, y^{t+1}, z^t)) \\ &\quad + (L_\beta(x^t, y^{t+1}, z^t) - L_\beta(x^t, y^t, z^t)). \end{aligned}$$

We estimate the three terms on the right hand side one by one. For the first term, we have

$$\begin{aligned} &L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - L_\beta(x^{t+1}, y^{t+1}, z^t) = -(z^{t+1} - z^t)^T(\mathcal{M}x^{t+1} - y^{t+1}) \\ &= \frac{1}{\beta}\|z^{t+1} - z^t\|^2 \le \frac{2}{\sigma\beta}(\|x^{t+1} - x^t\|^2_{\mathcal{Q}_3} + \|x^t - x^{t-1}\|^2_{\mathcal{T}_1^2}). \end{aligned} \tag{17}$$

We next estimate the second term. Recall from [20, Page 553, Ex.17] that the operation of taking positive square root preserves the positive semidefinite ordering. Thus, point 1 in (ii) of Assumption 1 implies that $\nabla^2\phi(x) \succeq \mathcal{T}_2$ for all x. From this and point 2 in (ii) of Assumption 1, we see further that the function $x \mapsto L_\beta(x, y^{t+1}, z^t) + D_\phi(x, x^t)$ is strongly convex with modulus at least δ. Using this, the definition of x^{t+1} (as a minimizer) and (6), we have

$$L_\beta(x^{t+1}, y^{t+1}, z^t) - L_\beta(x^t, y^{t+1}, z^t) \le -\frac{\delta}{2}\|x^{t+1} - x^t\|^2 - \frac{1}{2}\|x^{t+1} - x^t\|^2_{\mathcal{T}_2}. \tag{18}$$

Moreover, for the third term, using the definition of y^{t+1} as a minimizer, we have

$$L_\beta(x^t, y^{t+1}, z^t) - L_\beta(x^t, y^t, z^t) \le 0. \tag{19}$$

Summing (17), (18) and (19), we obtain that

$$\begin{aligned} &L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - L_\beta(x^t, y^t, z^t) \\ &\le \frac{1}{2}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2} + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}. \end{aligned} \tag{20}$$

Summing the above relation from $t = M, ..., N-1$ with $M \ge 1$, we see that

$$\begin{aligned} &L_\beta(x^N, y^N, z^N) - L_\beta(x^M, y^M, z^M) \\ &\le \frac{1}{2}\sum_{t=M}^{N-1}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2} + \frac{1}{2}\sum_{t=M}^{N-1}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &= \frac{1}{2}\sum_{t=M}^{N-1}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2} + \frac{1}{2}\sum_{t=M-1}^{N-2}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &= \frac{1}{2}\sum_{t=M}^{N-2}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2 + \frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &\quad + \frac{1}{2}\|x^N - x^{N-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2} + \frac{1}{2}\|x^M - x^{M-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &\le -\frac{1}{2}\sum_{t=M}^{N-2}\|x^{t+1} - x^t\|^2_{\mathcal{R}} + \frac{1}{2}\|x^M - x^{M-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}, \end{aligned} \tag{21}$$

where $\mathcal{R} := -\frac{4}{\sigma\beta}\mathcal{Q}_3 + \delta\mathcal{I} + \mathcal{T}_2 - \frac{4}{\sigma\beta}\mathcal{T}_1^2 \succ 0$ due to point 3 in (ii) of Assumption 1; and the last inequality also follows from the same point.

Now, suppose that (x^*, y^*, z^*) is a cluster point of the sequence $\{(x^t, y^t, z^t)\}$ and consider a convergent subsequence, i.e.,

$$\lim_{i\to\infty} (x^{t_i}, y^{t_i}, z^{t_i}) = (x^*, y^*, z^*). \tag{22}$$

From lower semicontinuity of L, we see that

$$\liminf_{i\to\infty} L_\beta(x^{t_i}, y^{t_i}, z^{t_i}) \geq h(x^*) + P(y^*) - \langle z^*, \mathcal{M}x^* - y^*\rangle + \frac{\beta}{2}\|\mathcal{M}x^* - y^*\|^2 > -\infty, \tag{23}$$

where the last inequality follows from the properness assumption on P. On the other hand, putting $M = 1$ and $N = t_i$ in (21), we see that

$$L_\beta(x^{t_i}, y^{t_i}, z^{t_i}) - L_\beta(x^1, y^1, z^1) \leq -\frac{1}{2}\sum_{t=1}^{t_i-2} \|x^{t+1} - x^t\|^2_{\mathcal{R}} + \frac{1}{2}\|x^1 - x^0\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}. \tag{24}$$

Passing to the limit in (24) and making use of (23) and (ii) in Assumption 1, we conclude that

$$0 \geq -\frac{1}{2}\sum_{t=1}^{\infty} \|x^{t+1} - x^t\|^2_{\mathcal{R}} > -\infty$$

The desired relation (16) now follows from this and the fact that $\mathcal{R} \succ 0$. Consequently, (9) holds.

We next show that (10) holds along the convergent subsequence in (22). Indeed, from the definition of y^{t_i} (as a minimizer), we have

$$L_\beta(x^{t_i}, y^{t_i+1}, z^{t_i}) \leq L_\beta(x^{t_i}, y^*, z^{t_i}).$$

Taking limit and using (22), we see that

$$\limsup_{i\to\infty} L_\beta(x^{t_i}, y^{t_i+1}, z^{t_i}) \leq h(x^*) + P(y^*) - \langle z^*, \mathcal{M}x^* - y^*\rangle + \frac{\beta}{2}\|\mathcal{M}x^* - y^*\|^2.$$

On the other hand, from lower semicontinuity, (22) and (9), we have

$$\liminf_{i\to\infty} L_\beta(x^{t_i}, y^{t_i+1}, z^{t_i}) \geq h(x^*) + P(y^*) - \langle z^*, \mathcal{M}x^* - y^*\rangle + \frac{\beta}{2}\|\mathcal{M}x^* - y^*\|^2.$$

The above two relations show that $\lim_{i\to\infty} P(y^{t_i+1}) = P(y^*)$. This together with (9) and the discussions preceding this theorem shows that x^* is a stationary point of (1) and that (12) holds. This proves (i) for $\mathcal{T}_1 \neq 0$.

Next, we suppose that the algorithm is initialized at a non-stationary x^0 with $h(x^0)+P(\mathcal{M}x^0) < \infty$, $y^0 = \mathcal{M}x^0$ and z^0 chosen with $\mathcal{M}^*z^0 = \nabla h(x^0)$. We first show that $x^1 \neq x^0$. To this end, we notice that

$$\begin{aligned}\mathcal{M}^*(z^1 - z^0) &= \nabla h(x^1) + \nabla\phi(x^1) - \nabla\phi(x^0) - \mathcal{M}^*z^0 \\ &= \nabla h(x^1) - \nabla h(x^0) + \nabla\phi(x^1) - \nabla\phi(x^0).\end{aligned}$$

Proceeding as in (14), we have

$$\sigma\|z^1 - z^0\|^2 \leq 2\|x^1 - x^0\|^2_{\mathcal{Q}_3}. \tag{25}$$

On the other hand, combining the relations $z^1 = z^0 - \beta(\mathcal{M}x^1 - y^1)$ and $y^0 = \mathcal{M}x^0$, we see that

$$y^1 - y^0 = \mathcal{M}(x^1 - x^0) + \frac{1}{\beta}(z^1 - z^0). \tag{26}$$

Consequently, if $x^1 = x^0$, then it follows from (25) and (26) that $z^1 = z^0$ and $y^1 = y^0$. This together with (11) implies that

$$0 \in \nabla h(x^0) + \mathcal{M}^* \partial P(\mathcal{M}x^0),$$

i.e., x^0 is a stationary point. Since x^0 is non-stationary by assumption, we must have $x^1 \neq x^0$.

We now derive an upper bound on $L_\beta(x^N, y^N, z^N) - L_\beta(x^0, y^0, z^0)$ for any $N > 1$. To this end, using the definition of augmented Lagrangian function, the z-update and (25), we have

$$L_\beta(x^1, y^1, z^1) - L_\beta(x^1, y^1, z^0) = \frac{1}{\beta}\|z^1 - z^0\|^2 \leq \frac{2}{\sigma\beta}\|x^1 - x^0\|^2_{\mathcal{Q}_3} .$$

Combining this relation with (18) and (19), we obtain the following estimate

$$L_\beta(x^1, y^1, z^1) - L_\beta(x^0, y^0, z^0) \leq \frac{1}{2}\|x^1 - x^0\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2}. \tag{27}$$

On the other hand, by specializing (21) to $N > M = 1$ and recalling that $\mathcal{R} \succ 0$, we see that

$$\begin{aligned} L_\beta(x^N, y^N, z^N) - L_\beta(x^1, y^1, z^1) &\leq -\frac{1}{2}\sum_{t=1}^{N-2}\|x^{t+1} - x^t\|^2_{\mathcal{R}} + \frac{1}{2}\|x^1 - x^0\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &\leq \frac{1}{2}\|x^1 - x^0\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}. \end{aligned} \tag{28}$$

Combining (27), (28) and the definition of $\mathcal{R}$, we obtain

$$L_\beta(x^N, y^N, z^N) - L_\beta(x^0, y^0, z^0) \leq -\frac{1}{2}\|x^1 - x^0\|^2_{\mathcal{R}} < 0,$$

where the strictly inequality follows from the fact that $x^1 \neq x^0$, and the fact that $\mathcal{R} \succ 0$. The conclusion of the theorem for the case when $\mathcal{T}_1 \neq 0$ now follows by taking limit in the above inequality along any convergent subsequence, and noting that $y^0 = \mathcal{M}x^0$ by assumption, and that $y^* = \mathcal{M}x^*$.

In the case when $\mathcal{T}_1 = 0$, we must have $\mathcal{T}_2 = 0$ and $\phi = 0$. Hence, (14) can be replaced by

$$\sigma\|z^{t+1} - z^t\|^2 \leq \|x^{t+1} - x^t\|^2_{\mathcal{Q}_3}. \tag{29}$$

The rest of the proof follows similarly by using this estimate in place of (14). □

We illustrate in the following examples how the parameters can be chosen in special cases.

Example 1. *Suppose that $\mathcal{M} = \mathcal{I}$ and that ∇h is Lipschitz continuous with modulus bounded by L. Then one can take $\mathcal{Q}_1 = L\mathcal{I}$ and $\mathcal{Q}_2 = -L\mathcal{I}$. Moreover, Assumption 1(i) holds with $\sigma = 1$. Furthermore, one can take $\phi(x) = \frac{L}{2}\|x\|^2 - h(x)$ so that $\mathcal{T}_1 = 2L\mathcal{I}$ and $\mathcal{T}_2 = 0$. For the second and third points of Assumption 1(ii) to hold, β can be chosen so that $\beta - L = \delta > 0$ and that*

$$\delta > \frac{4}{\beta}L^2 + \frac{4}{\beta}(2L)^2 = \frac{20}{\beta}L^2.$$

These can be achieved by picking $\beta > 5L$.

Example 2. *Suppose again that $\mathcal{M} = \mathcal{I}$ and $h(x) = \frac{1}{2}\|\mathcal{A}x - b\|^2$ for some linear map $\mathcal{A}$ and vector b. Then one can take $\phi = 0$ so that $\mathcal{T}_1 = \mathcal{T}_2 = 0$, and $\mathcal{Q}_1 = L\mathcal{I}$, $\mathcal{Q}_2 = 0$, where $L = \lambda_{\max}(\mathcal{A}^*\mathcal{A})$. Observe that Assumption 1(i) holds with $\sigma = 1$. For the second and third points of Assumption 1(ii) to hold, we only need to pick β so that $\beta = \delta > \frac{2}{\beta}L^2$, i.e., $\beta > \sqrt{2}L$.*

Example 3. *Suppose that $\mathcal{M}$ is a general surjective linear map and h is strongly convex. Specifically, assume that $h(x) = \frac{1}{2}\|x - \widehat{x}\|^2$ for some $\widehat{x}$ so that $\mathcal{Q}_1 = \mathcal{Q}_2 = \mathcal{I}$. Then we can take $\phi = 0$ and hence $\mathcal{T}_1 = \mathcal{T}_2 = 0$. Assumption 1(i) holds with $\sigma = \lambda_{\min}(\mathcal{M}\mathcal{M}^*)$. The second point of Assumption 1(ii) holds with $\delta = 1$. For the third point to hold, it suffices to pick $\beta > 2/\sigma$.*

We next give some sufficient conditions under which the sequence $\{(x^t, y^t, z^t)\}$ generated from the proximal ADMM under Assumption 1 is bounded. This would guarantee the existence of cluster point, which is the assumption required in Theorem 1. We start with the following theorem that concerns the ADMM, i.e., $\phi = 0$. The case for the proximal ADMM, i.e., $\phi \neq 0$, will be discussed a little bit later.

Theorem 2. (Boundedness of sequence generated from ADMM) *Suppose that Assumption 1 holds with $\mathcal{T}_1 = 0$ (and thus $\phi = 0$), and β is further chosen so that there exists $0 < \zeta < \beta$ with*

$$\inf_x \left\{ h(x) - \frac{1}{2\sigma\zeta}\|\nabla h(x)\|^2 \right\} =: h_* > -\infty. \tag{30}$$

Suppose that either

(i) *$\mathcal{M}$ is invertible and $\liminf_{\|y\|\to\infty} P(y) = \infty$; or*

(ii) *$\liminf_{\|x\|\to\infty} \|\nabla h(x)\| = \infty$ and $\inf_y P(y) > -\infty$.*

Then the sequence $\{(x^t, y^t, z^t)\}$ generated from the ADMM is bounded.

Proof. Since $\mathcal{T}_1 = 0$, proceeding as in (20) by using (29) instead of (14), we obtain that

$$L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - L_\beta(x^t, y^t, z^t) \le \frac{1}{2}\|x^{t+1} - x^t\|^2_{\frac{2}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I}} \le 0$$

where the last inequality follows from point 3 in (ii) of Assumption 1. In particular, $\{L_\beta(x^t, y^t, z^t)\}$ is decreasing and thus

$$h(x^t) + P(y^t) + \frac{\beta}{2}\left\|\mathcal{M}x^t - y^t - \frac{z^t}{\beta}\right\|^2 - \frac{1}{2\beta}\|z^t\|^2 = L_\beta(x^t, y^t, z^t) \le L_\beta(x^1, y^1, z^1) \tag{31}$$

whenever $t \ge 1$. Next, from the second relation in (11), we see that $\mathcal{M}^* z^t = \nabla h(x^t)$. Thus

$$\sigma\|z^t\|^2 \le \|\mathcal{M}^* z^t\|^2 = \|\nabla h(x^t)\|^2. \tag{32}$$

Plugging (32) into (31) and making use of (30), we obtain further that

$$\begin{aligned} L_\beta(x^1, y^1, z^1) &\ge h(x^t) - \frac{1}{2\sigma\beta}\|\nabla h(x^t)\|^2 + P(y^t) + \frac{\beta}{2}\left\|\mathcal{M}x^t - y^t - \frac{z^t}{\beta}\right\|^2 \\ &\ge h_* + \frac{1}{2\sigma}\left(\frac{1}{\zeta} - \frac{1}{\beta}\right)\|\nabla h(x^t)\|^2 + P(y^t) + \frac{\beta}{2}\left\|\mathcal{M}x^t - y^t - \frac{z^t}{\beta}\right\|^2. \end{aligned} \tag{33}$$

Now suppose that the conditions in (i) hold. Since P is proper and closed, the condition that $\liminf_{\|y\|\to\infty} P(y) = \infty$ implies $\inf_y P(y) > -\infty$. Hence, we see immediately from (33) and $\beta > \zeta > 0$ that $\{y^t\}$ and $\{\nabla h(x^t)\}$ are bounded. Boundedness of $\{z^t\}$ then follows from this and (32), whereas the boundedness of $\{x^t\}$ follows from the boundedness of $\{y^t\}$, $\{z^t\}$, the third relation in (7) and the invertibility of $\mathcal{M}$. We next consider the conditions in (ii). We conclude from (33) and the coerciveness of $\|\nabla h(x)\|$ that $\{x^t\}$ is bounded. This together with (32) shows that $\{z^t\}$ is also bounded. The boundedness of $\{y^t\}$ follows from these and the third relation in (7). This completes the proof. □

Notice that in order to guarantee boundedness of the sequence generated from the ADMM (i.e., $\phi = 0$), we have to choose β to satisfy both Assumption 1 and (30). We illustrate the conditions in Theorem 2 in the next few examples. In particular, we shall see that such a choice of β does exist in the following examples.

Example 4. *Consider the problem in Example 2, and suppose in addition that P is coercive, i.e., $\liminf_{\|y\|\to\infty} P(y) = \infty$. This covers the model of $\ell_{\frac{1}{2}}$ regularization considered in [31]. We show that $\{(x^t, y^t, z^t)\}$ is bounded by verifying the conditions in Theorem 2. In particular, we will argue that* (30) *holds for our choice of β. The conclusion will then follow from Theorem 2* (i)*.*

To this end, note that $h(x) = \frac{1}{2}\|\mathcal{A}x - b\|^2$ and thus

$$h(x) - \frac{1}{2\sqrt{2}L}\|\nabla h(x)\|^2 = \frac{1}{2}\|\mathcal{A}x - b\|^2 - \frac{1}{2\sqrt{2}L}\|\mathcal{A}^*(\mathcal{A}x - b)\|^2 \geq \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)\|\mathcal{A}x - b\|^2 \geq 0. \quad (34)$$

Thus, (30) *holds with $\sigma = 1$ and $\zeta = \sqrt{2}L < \beta$; recall that $L = \lambda(\mathcal{A}^*\mathcal{A})$ in this example.*

Example 5. *Consider the problem in Example 3, and assume in addition that $\inf_y P(y) > -\infty$. We show that $\{(x^t, y^t, z^t)\}$ is bounded by showing that* (30) *holds for our choice of β. The conclusion will then follow from Theorem 2* (ii)*.*

To this end, note that $h(x) = \frac{1}{2}\|x - \widehat{x}\|^2$ and thus

$$h(x) - \frac{1}{4}\|\nabla h(x)\|^2 = \frac{1}{4}\|x - \widehat{x}\|^2 \geq 0.$$

Thus, (30) *holds with $\zeta = 2/\sigma < \beta$.*

Remark 3. *We further comment on the condition* (30)*. In particular, we shall argue that for a fairly large class of twice continuously differentiable function h with a bounded Hessian, there exists $\alpha > 0$ so that*

$$\inf_x \left\{ h(x) - \frac{1}{2\alpha}\|\nabla h(x)\|^2 \right\} > -\infty.$$

Indeed, let q be a convex twice continuously differentiable function with a bounded Hessian and a minimizer x^*. Then it is well known that

$$q(x) - q(x^*) \geq \frac{1}{2L}\|\nabla q(x)\|^2, \quad \forall x \in \mathbb{R}^n,$$

where L is a Lipschitz continuity modulus of $\nabla q(x)$. Thus, if a twice continuously differentiable function h agrees with any such q outside a compact set $\mathcal{D}$, then

$$\begin{aligned} &\inf_x \left\{ h(x) - \frac{1}{2L}\|\nabla h(x)\|^2 \right\} \\ &= \min\left\{ \inf_{x\in\mathcal{D}} \left\{ h(x) - \frac{1}{2L}\|\nabla h(x)\|^2 \right\}, \inf_{x\notin\mathcal{D}} \left\{ q(x) - \frac{1}{2L}\|\nabla q(x)\|^2 \right\} \right\} \\ &\geq \min\left\{ \inf_{x\in\mathcal{D}} \left\{ h(x) - \frac{1}{2L}\|\nabla h(x)\|^2 \right\}, q(x^*) \right\} > -\infty, \end{aligned}$$

since the continuous function $x \mapsto h(x) - \frac{1}{2L}\|\nabla h(x)\|^2$ must be bounded below on $\mathcal{D}$.

We next consider the proximal ADMM. The assumptions are similar to those in Theorem 2; the only difference is that, a factor of 2 is missing in the condition (35).

Theorem 3. (Boundedness of sequence generated from the proximal ADMM) *Suppose that Assumption 1 holds with $\mathcal{T}_1 \neq 0$, and β is further chosen so that there exists $0 < \zeta < \beta$ with*

$$\inf_x \left\{ h(x) - \frac{1}{\sigma\zeta}\|\nabla h(x)\|^2 \right\} =: h_0 > -\infty. \quad (35)$$

Suppose that either

(i) $\mathcal{M}$ *is invertible and* $\liminf_{\|y\|\to\infty} P(y) = \infty$*; or*

(ii) $\liminf_{\|x\|\to\infty} \|\nabla h(x)\| = \infty$ *and* $\inf_y P(y) > -\infty$.

Then the sequence $\{(x^t, y^t, z^t)\}$ *generated from the proximal ADMM is bounded.*

Proof. First, observe from (20) that

$$\left(L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) + \frac{1}{2}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}\right) - \left(L_\beta(x^t, y^t, z^t) + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}\right)$$
$$\leq \frac{1}{2}\|x^{t+1} - x^t\|^2_{\frac{4}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I} - \mathcal{T}_2 + \frac{4}{\sigma\beta}\mathcal{T}_1^2} \leq 0,$$

where the last inequality follows from point 3 in (ii) of Assumption 1. In particular, the sequence $\{L_\beta(x^t, y^t, z^t) + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}\}$ is decreasing and consequently, we have, for all $t \geq 1$, that

$$L_\beta(x^t, y^t, z^t) + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \leq L_\beta(x^1, y^1, z^1) + \frac{1}{2}\|x^1 - x^0\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2}, \tag{36}$$

Next, recall from (13) that

$$\begin{aligned} \sigma\|z^t\|^2 &\leq \|\mathcal{M}^* z^t\|^2 = \|\nabla h(x^t) + \nabla\phi(x^t) - \nabla\phi(x^{t-1})\|^2 \\ &\leq 2\|\nabla h(x^t)\|^2 + 2\|\nabla\phi(x^t) - \nabla\phi(x^{t-1})\|^2 \\ &\leq 2\|\nabla h(x^t)\|^2 + 2\|x^t - x^{t-1}\|^2_{\mathcal{T}_1^2}. \end{aligned} \tag{37}$$

Plugging this into (36), we see further that

$$\begin{aligned} & L_\beta(x^1, y^1, z^1) + \frac{1}{2}\|x^1 - x^0\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \geq L_\beta(x^t, y^t, z^t) + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &= h(x^t) + P(y^t) + \frac{\beta}{2}\left\|\mathcal{M}x^t - y^t - \frac{z^t}{\beta}\right\|^2 - \frac{1}{2\beta}\|z^t\|^2 + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{4}{\sigma\beta}\mathcal{T}_1^2} \\ &\geq h(x^t) + P(y^t) + \frac{\beta}{2}\left\|\mathcal{M}x^t - y^t - \frac{z^t}{\beta}\right\|^2 - \frac{1}{\sigma\beta}\|\nabla h(x^t)\|^2 + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{2}{\sigma\beta}\mathcal{T}_1^2} \\ &\geq h_0 + \frac{1}{\sigma}\left(\frac{1}{\zeta} - \frac{1}{\beta}\right)\|\nabla h(x^t)\|^2 + P(y^t) + \frac{\beta}{2}\left\|\mathcal{M}x^t - y^t - \frac{z^t}{\beta}\right\|^2 + \frac{1}{2}\|x^t - x^{t-1}\|^2_{\frac{2}{\sigma\beta}\mathcal{T}_1^2}. \end{aligned} \tag{38}$$

Now, suppose that the conditions in (i) hold. Note that $\liminf_{\|y\|\to\infty} P(y) = \infty$ implies $\inf_y P(y) > -\infty$. This together with (38) and $\beta > \zeta > 0$ implies that $\{y^t\}$, $\{\nabla h(x^t)\}$, and $\{\|x^t - x^{t-1}\|_{\mathcal{T}_1^2}\}$ are bounded. Boundedness of $\{z^t\}$ follows from these and (37). Moreover, the boundedness of $\{x^t\}$ follows from the boundedness of $\{y^t\}$, $\{z^t\}$, the invertibility of $\mathcal{M}$ and the third relation in (7). Next, consider the conditions in (ii). Since P is bounded below, (38) and the coerciveness of $\|\nabla h(x)\|$ give the boundedness of $\{x^t\}$. The boundedness of $\{z^t\}$ follows from this and (37). Finally, the boundedness of $\{y^t\}$ follows from these and the third relation in (7). This completes the proof. □

Example 6. *Consider the problem in Example 1, and suppose in addition that* $h(x) = \frac{1}{2}\|\mathcal{A}x - b\|^2$ *for some linear map* $\mathcal{A}$ *and vector* b*, and that* P *is coercive, i.e.,* $\liminf_{\|y\|\to\infty} P(y) = \infty$*. This includes the model of* $\ell_{\frac{1}{2}}$ *regularization considered in [31]. Recall from* (34) *that for* $h(x) = \frac{1}{2}\|\mathcal{A}x - b\|^2$*, we have*

$$h(x) - \frac{1}{2\sqrt{2}L}\|\nabla h(x)\|^2 \geq \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)\|\mathcal{A}x - b\|^2 \geq 0,$$

where $L = \lambda_{\max}(\mathcal{A}^*\mathcal{A})$*. Thus,* (35) *holds with* $\sigma = 1$ *and* $\zeta = 2\sqrt{2}L < 5L < \beta$*. Hence, the sequence generated from the proximal ADMM is bounded, according to Theorem 3.*

Finally, before ending this section, we study convergence of the whole sequence generated by the ADMM (i.e., proximal ADMM with $\phi = 0$) when the objective function is semi-algebraic. The proof of this theorem relies heavily on the KL property. For recent applications of KL property to convergence analysis of a broad class of optimization methods, see [3].

Theorem 4. (Global convergence for the whole sequence) *Suppose that Assumption 1 holds with $\mathcal{T}_1 = 0$ (and hence $\phi = 0$), and that h and P are semi-algebraic functions. Suppose further that the sequence $\{(x^t, y^t, z^t)\}$ generated from the ADMM has a cluster point (x^*, y^*, z^*). Then the sequence $\{(x^t, y^t, z^t)\}$ converges to (x^*, y^*, z^*) and x^* is a stationary point of* (1)*. Moreover,*

$$\sum_{t=1}^{\infty} \|x^{t+1} - x^t\| < \infty. \tag{39}$$

Proof. The conclusion that x^* is a stationary point of (1) follows from Theorem 1. Moreover, (9) holds. We now establish convergence.

First, consider the subdifferential of L_β at $(x^{t+1}, y^{t+1}, z^{t+1})$. Specifically, we have

$$\begin{aligned} \nabla_x L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) &= \nabla h(x^{t+1}) - \mathcal{M}^* z^{t+1} + \beta \mathcal{M}^*(\mathcal{M}x^{t+1} - y^{t+1}) \\ &= \beta \mathcal{M}^*(\mathcal{M}x^{t+1} - y^{t+1}) = -\mathcal{M}^*(z^{t+1} - z^t), \end{aligned}$$

where the last two equalities follow from the second and third relations in (11). Similarly,

$$\begin{aligned} \nabla_z L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) &= -(\mathcal{M}x^{t+1} - y^{t+1}) = \frac{1}{\beta}(z^{t+1} - z^t). \\ \partial_y L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) &= \partial P(y^{t+1}) + z^{t+1} - \beta(\mathcal{M}x^{t+1} - y^{t+1}) \\ &\ni z^{t+1} - z^t - \beta\mathcal{M}(x^{t+1} - x^t), \end{aligned}$$

since $0 \in \partial P(y^{t+1}) + z^t - \beta(\mathcal{M}x^t - y^{t+1})$ from (8). The above relations together with the assumption that $\mathcal{T}_1 = 0$ and (29) imply the existence of a constant $C > 0$ so that

$$\mathrm{dist}(0, \partial L_\beta(x^{t+1}, y^{t+1}, z^{t+1})) \le C\|x^{t+1} - x^t\|. \tag{40}$$

Moreover, proceed similarly as in (20) where we use (29) in place of (14), and invoke Point 3 in (ii) of Assumption 1, we see also that

$$L_\beta(x^t, y^t, z^t) - L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) \ge -\frac{1}{2}\|x^{t+1} - x^t\|^2_{\frac{2}{\sigma\beta}\mathcal{Q}_3 - \delta\mathcal{I}} \ge D\|x^{t+1} - x^t\|^2 \tag{41}$$

for some $D > 0$. In particular, $\{L_\beta(x^t, y^t, z^t)\}$ is decreasing. Since L_β is also bounded below along the subsequence in (22), we conclude that $\lim_{t\to\infty} L_\beta(x^t, y^t, z^t)$ exists.

We now show that $\lim_{t\to\infty} L_\beta(x^t, y^t, z^t) = l^*$; here, we write $l^* := L_\beta(x^*, y^*, z^*)$ for notational simplicity. To this end, notice from the definition of y^{t+1} as a minimizer that

$$L_\beta(x^t, y^{t+1}, z^t) \le L_\beta(x^t, y^*, z^t).$$

Using this relation, (9) and the continuity of L_β with respect to the x and z variables, we have

$$\limsup_{j\to\infty} L_\beta(x^{t_j+1}, y^{t_j+1}, z^{t_j+1}) \le L_\beta(x^*, y^*, z^*), \tag{42}$$

where $\{(x^{t_j}, y^{t_j}, z^{t_j})\}$ is a subsequence that converges to (x^*, y^*, z^*). On the other hand, from (9), we see that $\{(x^{t_j+1}, y^{t_j+1}, z^{t_j+1})\}$ also converges to (x^*, y^*, z^*). This together with the lower semicontinuity of L_β imply

$$\liminf_{j\to\infty} L_\beta(x^{t_j+1}, y^{t_j+1}, z^{t_j+1}) \ge L_\beta(x^*, y^*, z^*). \tag{43}$$

Combining (42), (43) and the existence of $\lim L_\beta(x^t, y^t, z^t)$, we conclude that

$$\lim_{t\to\infty} L_\beta(x^t, y^t, z^t) = l^*, \tag{44}$$

as claimed. Furthermore, if $L_\beta(x^t, y^t, z^t) = l^*$ for some $t \geq 1$, since the sequence is decreasing, we must have $L_\beta(x^t, y^t, z^t) = L_\beta(x^{t+k}, y^{t+k}, z^{t+k})$ for all $k \geq 0$. From (41), we see that $x^t = x^{t+k}$ and hence $z^t = z^{t+k}$ from (29), for all $k \geq 0$. Consequently, we conclude from (15) that $y^{t+1} = y^{t+k}$ for all $k \geq 1$, meaning that the algorithm terminates finitely. Since the conclusion of this theorem holds trivially if the algorithm terminates finitely, from now on, we only consider the case where $L_\beta(x^t, y^t, z^t) > l^*$ for all $t \geq 1$.

Next, notice that the function $(x, y, z) \mapsto L_\beta(x, y, z)$ is semi-algebraic due to the semi-algebraicity of h and P. Thus, it is a KL function from [2, Section 4.3]. From the property of KL functions, there exist $\eta > 0$, a neighborhood V of (x^*, y^*, z^*) and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ as described in Definition 1 so that for all $(x, y, z) \in V$ satisfying $l^* < L_\beta(x, y, z) < l^* + \eta$, we have

$$\varphi'(L_\beta(x, y, z) - l^*)\,\mathrm{dist}(0, \partial L_\beta(x, y, z)) \geq 1. \tag{45}$$

Pick $\rho > 0$ so that

$$\mathbf{B}_\rho := \left\{(x, y, z) :\ \|x - x^*\| < \rho, \|y - y^*\| < (\|\mathcal{M}\| + 1)\rho, \|z - z^*\| < \sqrt{\frac{\lambda_{\max}(\mathcal{Q}_3)}{\sigma}}\rho\right\} \subseteq V$$

and set $B_\rho := \{x :\ \|x - x^*\| < \rho\}$. From the second relation in (11) and (12), we obtain for any $t \geq 1$ that

$$\sigma\|z^t - z^*\|^2 \leq \|\mathcal{M}^*(z^t - z^*)\|^2 = \|\nabla h(x^t) - \nabla h(x^*)\|^2 \leq \lambda_{\max}(\mathcal{Q}_3)\|x^t - x^*\|^2.$$

Hence $\|z^t - z^*\| < \sqrt{\frac{\lambda_{\max}(\mathcal{Q}_3)}{\sigma}}\rho$ whenever $x^t \in B_\rho$ and $t \geq 1$. Moreover, from the definition of z^{t+1} and (12), we see that whenever $t \geq 1$,

$$\|y^t - y^*\| = \left\|\mathcal{M}(x^t - x^*) + \frac{1}{\beta}(z^t - z^{t-1})\right\| \leq \|\mathcal{M}\|\|x^t - x^*\| + \frac{1}{\beta}\|z^t - z^{t-1}\|.$$

Since there exists $N_0 \geq 1$ so that for all $t \geq N_0$, we have $\|z^t - z^{t-1}\| < \beta\rho$ (such an N_0 exists due to (9)), it follows that $\|y^t - y^*\| < (\|\mathcal{M}\| + 1)\rho$ whenever $x^t \in B_\rho$ and $t \geq N_0$. Thus, if $x^t \in B_\rho$ and $t \geq N_0$, we have $(x^t, y^t, z^t) \in \mathbf{B}_\rho \subseteq V$. Moreover, it is not hard to see that there exists (x^N, y^N, z^N) with $N \geq N_0$ such that

(i) $x^N \in B_\rho$;

(ii) $l^* < L_\beta(x^N, y^N, z^N) < l^* + \eta$;

(iii) $\|x^N - x^*\| + 2\sqrt{\frac{L_\beta(x^N, y^N, z^N) - l^*}{D}} + \frac{C}{D}\varphi(L_\beta(x^N, y^N, z^N) - l^*) < \rho$.

Indeed, these properties follow from the fact that (x^*, y^*, z^*) is a cluster point, (44) and that $L_\beta(x^t, y^t, z^t) > l^*$ for all $t \geq 1$.

We next show that, if $x^t \in B_\rho$ and $l^* < L_\beta(x^t, y^t, z^t) < l^* + \eta$ for some fixed $t \geq N_0$, then

$$\begin{aligned}&\|x^{t+1} - x^t\| + (\|x^{t+1} - x^t\| - \|x^t - x^{t-1}\|)\\ &\leq \frac{C}{D}[\varphi(L_\beta(x^t, y^t, z^t) - l^*) - \varphi(L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - l^*)].\end{aligned} \tag{46}$$

To see this, notice that $x^t \in B_\rho$ and $t \geq N_0$ implies $(x^t, y^t, z^t) \in \mathbf{B}_\rho \subseteq V$. Hence, (45) holds for (x^t, y^t, z^t). Combining (40), (41), (45) and the concavity of ϕ, we conclude that for all such t

$$\begin{aligned}
&C\|x^t - x^{t-1}\| \cdot [\varphi(L_\beta(x^t, y^t, z^t) - l^*) - \varphi(L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - l^*)] \\
&\geq \operatorname{dist}(0, \partial L_\beta(x^t, y^t, z^t)) \cdot [\varphi(L_\beta(x^t, y^t, z^t) - l^*) - \varphi(L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) - l^*)] \\
&\geq \operatorname{dist}(0, \partial L_\beta(x^t, y^t, z^t)) \cdot \varphi'(L_\beta(x^t, y^t, z^t) - l^*) \cdot [L_\beta(x^t, y^t, z^t) - L_\beta(x^{t+1}, y^{t+1}, z^{t+1})] \\
&\geq D\|x^{t+1} - x^t\|^2.
\end{aligned}$$

Dividing both sides by D, taking square root, using the inequality $2\sqrt{ab} \leq a + b$ as in the proof of [3, Lemma 2.6], and rearranging terms, we conclude that (46) holds.

We now show that $x^t \in B_\rho$ whenever $t \geq N$. We establish this claim by induction, and our proof is similar to the proof of [3, Lemma 2.6]. The claim is true for $t = N$ by construction. For $t = N + 1$, we have

$$\begin{aligned}
\|x^{N+1} - x^*\| &\leq \|x^{N+1} - x^N\| + \|x^N - x^*\| \\
&\leq \sqrt{\frac{L_\beta(x^N, y^N, z^N) - L_\beta(x^{N+1}, y^{N+1}, z^{N+1})}{D}} + \|x^N - x^*\| \\
&\leq \sqrt{\frac{L_\beta(x^N, y^N, z^N) - l^*}{D}} + \|x^N - x^*\| < \rho,
\end{aligned}$$

where the first inequality follows from (41). Now, suppose the claim is true for $t = N, \ldots, N+k-1$ for some $k > 1$; i.e., $x^N, \ldots, x^{N+k-1} \in B_\rho$. We now consider the case when $t = N + k$:

$$\begin{aligned}
&\|x^{N+k} - x^*\| \leq \|x^N - x^*\| + \|x^N - x^{N+1}\| + \sum_{j=1}^{k-1} \|x^{N+j+1} - x^{N+j}\| \\
&= \|x^N - x^*\| + 2\|x^N - x^{N+1}\| - \|x^{N+k} - x^{N+k-1}\| \\
&\quad + \sum_{j=1}^{k-1} [\|x^{N+j+1} - x^{N+j}\| + (\|x^{N+j+1} - x^{N+j}\| - \|x^{N+j} - x^{N+j-1}\|)] \\
&\leq \|x^N - x^*\| + 2\|x^N - x^{N+1}\| \\
&\quad + \frac{C}{D} \sum_{j=1}^{k-1} [\varphi(L_\beta(x^{N+j}, y^{N+j}, z^{N+j}) - l^*) - \varphi(L_\beta(x^{N+j+1}, y^{N+j+1}, z^{N+j+1}) - l^*)] \\
&\leq \|x^N - x^*\| + 2\|x^N - x^{N+1}\| + \frac{C}{D} \varphi(L_\beta(x^{N+1}, y^{N+1}, z^{N+1}) - l^*),
\end{aligned}$$

where the first inequality follows from (46), the monotonicity of $\{L_\beta(x^t, y^t, z^t)\}$ from (41), and the induction assumption that $x^N, \ldots, x^{N+k-1} \in B_\rho$. Moreover, in view of (41) and the definition of ρ, we see that the last expression above is less than ρ. Hence, $\|x^{N+k} - x^*\| < \rho$ as claimed, and we have shown that $x^t \in B_\rho$ for $t \geq N$ by induction.

Since $x^t \in B_\rho$ for $t \geq N$, we can sum (46) from $t = N$ to $M \to \infty$. Invoking (9), we arrive at

$$\sum_{t=N}^{\infty} \|x^{t+1} - x^t\| \leq \frac{C}{D} \varphi(L_\beta(x^N, y^N, z^N) - l^*) + \|x^N - x^{N-1}\|,$$

which implies that (39) holds. Convergence of $\{x^t\}$ follows immediately from this. Convergence of $\{y^t\}$ follows from the convergence of $\{x^t\}$, the relation $y^{t+1} = \mathcal{M}x^{t+1} + \frac{1}{\beta}(z^{t+1} - z^t)$ from (7), and (9). Finally, the convergence of $\{z^t\}$ follows from the surjectivity of $\mathcal{M}$, and the relation $\mathcal{M}^* z^{t+1} = \nabla h(x^{t+1})$ from (11). This completes the proof. $\square$

Remark 4. (Comments on Theorem 4)

(1) *A close inspection of the above proof shows that the conclusion of Theorem 4 continues to hold as long as the augmented Lagrangian L_β is a KL-function. Here, we only state the case where h and P are semi-algebraic because this simple sufficient condition can be easily verified.*

(2) *Although a general convergence analysis framework was established in [3] for a broad class of optimization problems, it is not clear to us whether their results can be applied directly here. Indeed, to ensure convergence, three basic properties* **H1**, **H2** *and* **H3** *were imposed in [3, Page 99]. In particular, their property* **H1** *(sufficient descent property) in our case reads:*

$$L_\beta(x^t, y^t, z^t) - L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) \geq D(\|x^{t+1} - x^t\|^2 + \|y^{t+1} - y^t\|^2 + \|z^{t+1} - z^t\|^2),$$

for some $D > 0$. On the other hand, (41) in our proof only gives us that $L_\beta(x^t, y^t, z^t) - L_\beta(x^{t+1}, y^{t+1}, z^{t+1}) \geq D\|x^{t+1} - x^t\|^2$, which is not sufficient for property **H1** *to hold.*

(3) *In Theorem 4, we only discussed the case where $\phi = 0$. This condition is used to ensure that $\{L_\beta(x^t, y^t, z^t)\}$ is a decreasing sequence that is at least as large as $L_\beta(x^*, y^*, z^*)$. It would be interesting to see whether the analysis here can be further extended to the case where $\phi \neq 0$.*

4 Proximal gradient algorithm when $\mathcal{M} = \mathcal{I}$

In this section, we look at the model problem (1) in the case where $\mathcal{M} = \mathcal{I}$. Since the objective is the sum of a smooth and a possibly nonsmooth part with a simple proximal mapping, it is natural to consider the proximal gradient algorithm (also known as the forward-backward splitting algorithm). In this approach, one consider the update

$$x^{t+1} \in \operatorname*{Arg\,min}_x \left\{ \langle \nabla h(x^t), x - x^t \rangle + \frac{1}{2\beta}\|x - x^t\|^2 + P(x) \right\}. \tag{47}$$

From our assumption on P, the update can be performed efficiently via a computation of the proximal mapping of βP. When $\beta \in (0, \frac{1}{L})$, where $L \geq \sup\{\|\nabla^2 h(x)\| : x \in \mathbb{R}^n\}$, it is not hard to show that any cluster point x^* of the sequence generated above is a stationary point of (1); see, for example, [7]. In what follows, we analyze the convergence under a slightly more flexible step-size rule.

Theorem 5. *Suppose that there exists a twice continuously differentiable convex function q and $\ell > 0$ such that for all x,*

$$-\ell\mathcal{I} \preceq \nabla^2 h(x) + \nabla^2 q(x) \preceq \ell\mathcal{I}. \tag{48}$$

Let $\{x^t\}$ be generated from (47) *with $\beta \in (0, \frac{1}{\ell})$. Then the algorithm is a descent algorithm. Moreover, any cluster point x^* of $\{x^t\}$, if exists, is a stationary point.*

Remark 5. *For the algorithm to converge faster, intuitively, a larger step-size β should be chosen; see also Table 2. Condition* (48) *indicates that the "concave" part of the smooth objective h does not impose any restrictions on the choice of step-size. This could result in an ℓ smaller than the Lipschitz continuity modulus of $\nabla h(x)$, and hence allow a choice of a larger β. On the other hand, since the algorithm is a descent algorithm by Theorem 5, the sequence generated from* (47) *would be bounded under standard coerciveness assumptions on the objective function.*

Proof. Notice from assumption that $\nabla(h+q)$ is Lipschitz continuous with Lipschitz continuity modulus at most ℓ. Hence

$$(h+q)(x^{t+1}) \le (h+q)(x^t) + \langle \nabla h(x^t) + \nabla q(x^t), x^{t+1} - x^t\rangle + \frac{\ell}{2}\|x^{t+1} - x^t\|^2. \tag{49}$$

From this we see further that

$$\begin{aligned}
&h(x^{t+1}) + P(x^{t+1}) = (h+q)(x^{t+1}) + P(x^{t+1}) - q(x^{t+1}) \\
&\le (h+q)(x^t) + \langle \nabla h(x^t) + \nabla q(x^t), x^{t+1} - x^t\rangle + \frac{\ell}{2}\|x^{t+1} - x^t\|^2 + P(x^{t+1}) - q(x^{t+1}) \\
&= h(x^t) + \langle \nabla h(x^t), x^{t+1} - x^t\rangle + \frac{\ell}{2}\|x^{t+1} - x^t\|^2 + P(x^{t+1}) \\
&\quad + q(x^t) + \langle \nabla q(x^t), x^{t+1} - x^t\rangle - q(x^{t+1}) \\
&\le h(x^t) + P(x^t) + \left(\frac{\ell}{2} - \frac{1}{2\beta}\right)\|x^{t+1} - x^t\|^2,
\end{aligned} \tag{50}$$

where the first inequality follows from (49), the last inequality follows from the definition of x^{t+1} and the subdifferential inequality applied to the function q. Since $\beta \in (0, \frac{1}{\ell})$ implies $\frac{1}{2\beta} > \frac{\ell}{2}$, (50) shows that the algorithm is a descent algorithm.

Rearranging terms in (50) and summing from $t = 0$ to any $N - 1 > 0$, we see further that

$$\left(\frac{1}{2\beta} - \frac{\ell}{2}\right)\sum_{t=0}^{N-1}\|x^{t+1} - x^t\|^2 \le h(x^0) + P(x^0) - h(x^N) - P(x^N).$$

Now, let x^* be a cluster point and take any convergent subsequence $\{x^{t_i}\}$ that converges to x^*. Taking limit on both sides of the above inequality along the convergent subsequence, one can see that $\lim_{t\to\infty}\|x^{t+1} - x^t\| = 0$. Finally, we wish to show that $\lim_{i\to\infty} P(x^{t_i+1}) = P(x^*)$. To this end, note first that since $\lim_{t\to\infty}\|x^{t+1} - x^t\| = 0$, we also have $\lim_{i\to\infty} x^{t_i+1} = x^*$. Then it follows from lower semicontinuity of P that $\liminf_{i\to\infty} P(x^{t_i+1}) \ge P(x^*)$. On the other hand, from (47), we have

$$\langle \nabla h(x^{t_i}), x^{t_i+1} - x^{t_i}\rangle + \frac{1}{2\beta}\|x^{t_i+1} - x^{t_i}\|^2 + P(x^{t_i+1}) \le \langle \nabla h(x^{t_i}), x^* - x^{t_i}\rangle + \frac{1}{2\beta}\|x^* - x^{t_i}\|^2 + P(x^*),$$

which gives $\limsup_{i\to\infty} P(x^{t_i+1}) \le P(x^*)$. Hence, $\lim_{i\to\infty} P(x^{t_i+1}) = P(x^*)$. Now, using this, $\lim_{t\to\infty}\|x^{t+1} - x^t\| = 0$, (3) and taking limit along the convergent subsequence in the following relation obtained from (47)

$$0 \in \nabla h(x^t) + \frac{1}{\beta}(x^{t+1} - x^t) + \partial P(x^{t+1}), \tag{51}$$

we see that the conclusion concerning stationary point holds. □

We illustrate the above theorem in the following examples.

Example 7. *Suppose that h admits an explicit representation as a difference of two convex twice continuously differentiable functions $h = h_1 - h_2$, and that h_1 has a Lipschitz continuous gradient with modulus at most L_1. Then (48) holds with $q = h_2$ and $\ell = L_1$. Hence, the step-size can be chosen from $(0, 1/L_1)$.*

A concrete example of this kind is given by $h(x) = \frac{1}{2}\langle x, Qx\rangle$, where Q is a symmetric indefinite matrix. Then (48) holds with $q(x) = -\frac{1}{2}\langle x, Q_- x\rangle$, where Q_- is the projection of Q onto the cone of nonpositive semidefinite matrices, and $\ell = \lambda_{\max}(Q) > 0$. The step-size β can be chosen within the open interval $(0, 1/\lambda_{\max}(Q))$.

In the case when $h(x)$ is a concave quadratic, say, for example, $h(x) = -\frac{1}{2}\|\mathcal{A}x - b\|^2$ for some linear map $\mathcal{A}$, it is easy to see that (48) *holds with $q(x) = \frac{1}{2}\|\mathcal{A}x\|^2$ for* any *positive number ℓ. Thus, step-size can be chosen to be any positive number.*

Example 8. *Suppose that h has a Lipschitz continuous gradient and it is known that all the eigenvalues of $\nabla^2 h(x)$, for any x, lie in the interval $[-\lambda_2, \lambda_1]$ with $-\lambda_2 < 0 < \lambda_1$. If $\lambda_1 \geq \lambda_2$, it is clear that ∇h is Lipschitz continuous with modulus bounded by λ_1, and hence the step-size for the proximal gradient algorithm can be chosen from $(0, 1/\lambda_1)$. On the other hand, if $\lambda_1 < \lambda_2$, then it is easy to see that* (48) *holds with $q(x) = \frac{\lambda_2 - \lambda_1}{4}\|x\|^2$ and $\ell = (\lambda_2 + \lambda_1)/2$. Hence, the step-size can be chosen from $(0, 2/(\lambda_1 + \lambda_2))$.*

We next comment on the convergence of the whole sequence. We consider the conditions **H1** through **H3** on [3, Page 99]. First, it is easy to see from (50) that **H1** is satisfied with $a = \frac{1}{2\beta} - \frac{\ell}{2}$. Next, notice from (51) that if $w^{t+1} := \nabla h(x^{t+1}) - \nabla h(x^t) - \frac{1}{\beta}(x^{t+1} - x^t)$, then $w^{t+1} \in \nabla h(x^{t+1}) + \partial P(x^{t+1})$. Moreover, from the definition of w^{t+1}, we have

$$\|w^{t+1}\| \leq \left(L + \frac{1}{\beta}\right)\|x^{t+1} - x^t\|$$

for any $L \geq \sup\{\|\nabla^2 h(x)\| : \ x \in \mathbb{R}^n\}$. This shows that the condition **H2** is satisfied with $b = L + \frac{1}{\beta}$. Finally, [3, Remark 5.2] shows that **H3** is satisfied. Thus, we conclude from [3, Theorem 2.9] that if $h + P$ is a KL-function and a cluster point x^* of the sequence $\{x^t\}$ exists, then the whole sequence converges to x^*.

A line-search strategy can also be incorporated to possibly speed up the above algorithm; see [18] for the case when P is a continuous difference-of-convex function. The convergence analysis there can be directly adapted. The result of Theorem 5 concerning the interval of viable step-sizes can be used in designing the initial step-size for backtracking in the line-search procedure.

5 Numerical simulations

In this section, we perform numerical experiments to illustrate our algorithms. All codes are written in MATLAB. All experiments are performed on a 32-bit desktop machine with an Intel® i7-3770 CPU (3.40 GHz) and a 4.00 GB RAM, equipped with MATLAB 7.13 (2011b).

5.1 Nonconvex ADMM

We consider the problem of finding the closest point to a given $\widehat{x} \in \mathbb{R}^n$ that violates at most r out of m equations. The problem is presented as follows:

$$\begin{array}{ll} \min\limits_{x} & \frac{1}{2}\|x - \widehat{x}\|^2 \\ \text{s.t.} & \|\mathcal{M}x - b\|_0 \leq r, \end{array} \tag{52}$$

where $\mathcal{M} \in \mathbb{R}^{m \times n}$ has full row rank, $b \in \mathbb{R}^m$, $n \geq m \geq r$. This can be seen as a special case of (1) by taking $h(x) = \frac{1}{2}\|x - \widehat{x}\|^2$ and $P(y)$ to be the indicator function of the set $\{y : \ \|y - b\|_0 \leq r\}$, which is a proper closed function; here, $\|y\|_0$ is the ℓ_0 norm that counts the number of nonzero entries in the vector y.

We apply the ADMM (i.e., proximal ADMM with $\phi = 0$) with parameters specified as in Example 3, and pick $\beta = 1.01 \cdot (2/\sigma)$ so that $\beta > 2/\sigma$. From Example 5, the sequence generated from the ADMM is always bounded and hence convergence of the sequence is guaranteed by Theorem 4. We compare our model against the standard convex model with the ℓ_0 norm replaced by the ℓ_1 norm. This latter model is solved by SDPT3 (Version 4.0), called via CVX (Version 1.22), using default settings.

For the ADMM, we consider two initializations: setting all variables at the origin (0 init.), or setting x^0 to be the approximate solution $\widetilde{x}$ obtained from solving the convex model, $y^0 = \mathcal{M}x^0$ and $z^0 = (\mathcal{M}\mathcal{M}^*)^{-1}\mathcal{M}(x^0 - \widehat{x})$ (ℓ_1 init.). As discussed in Remark 2, when $\widetilde{x}$ is feasible for (52), this latter initialization satisfies the conditions in Theorem 1(ii). We terminate the ADMM when the sum of successive changes is small, i.e., when

$$\frac{\|x^t - x^{t-1}\| + \|y^t - y^{t-1}\| + \|z^t - z^{t-1}\|}{\|x^t\| + \|y^t\| + \|z^t\| + 1} < 10^{-8}.$$

In our experiments, we consider random instances. In particular, to guarantee that the problem (52) is feasible for a fixed r, we generate the matrix $\mathcal{M}$ and the right hand side b using the following MATLAB codes:

```
M = randn(m,n);
x_orig = randn(n,1);
J = randperm(m);
b = randn(m,1);
b(J(1:m-r)) = M(J(1:m-r),:)*x_orig; % subsystem has a solution
```

We then generate $\widehat{x}$ with i.i.d. standard Gaussian entries.

We consider $n = 1000$, 2000, 3000, 4000 and 5000, $m = 500$, $r = 100$, 200 and 300. We generate one random instance for each (n, m, r) and solve (52) and the corresponding ℓ_1 relaxation. The computational results are shown in Table 1, where we report the number of violated constraints (vio) by the approximate solution x obtained, defined as $\#\{i : \ |(\mathcal{M}x - b)_i| > 10^{-4}\}$, and the distance from $\widehat{x}$ (dist) defined as $\|x - \widehat{x}\|$. We also report the number of iterations the ADMM takes, as well as the CPU time of both the ADMM initialized at the origin and SDPT3 called using CVX.[1] We see that the model (52) allows an explicit control on the number of violated constraints. In addition, comparing with the ℓ_1 model, the ℓ_0 model solved using the ADMM always gives a solution closer to $\widehat{x}$. Finally, the solution obtained from the ADMM initialized from an approximate solution of the ℓ_1 model can be slightly closer to $\widehat{x}$ than the solution obtained from the zero initialization, depending on the particular problem instance.

5.2 Proximal gradient algorithm

In this section, we consider the following concave minimization problem:

$$\begin{array}{ll} \min\limits_{x} & -\frac{1}{2}\|\mathcal{A}x - b\|^2 \\ \text{s.t.} & x \in \mathcal{C}, \end{array} \tag{53}$$

where $\mathcal{C}$ is a compact convex set whose projection is easy to compute, $\mathcal{A} \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We apply the proximal gradient algorithm and illustrate how the more flexible stepsize rule introduced via Theorem 5 affects the solution quality and the computational time. Specifically, we apply the proximal gradient algorithms with various step-size parameters $\beta > 0$. Since the objective in (53) is concave and $\mathcal{C}$ is compact, we see from Theorem 5 that for any $\beta > 0$, the sequence generated from the proximal gradient algorithm is bounded with cluster points being stationary points of (53).

[1] We include the preprocessing time by CVX in the CPU time.

Table 1: Computational results for perturbation with bounded number of violated equalities.

			ℓ_0-ADMM (0 init.)				ℓ_1-CVX			ℓ_0-ADMM (ℓ_1 init.)		
r	n	$\|x_{\rm orig} - \widehat{x}\|$	iter	CPU	vio	dist	CPU	vio	dist	iter	vio	dist
100	1000	4.70e+001	389	0.4	100	2.24e+001	10.1	13	3.25e+001	405	100	2.18e+001
100	2000	6.37e+001	158	0.4	100	2.05e+001	18.4	6	2.92e+001	150	100	1.89e+001
100	3000	7.72e+001	130	0.7	100	1.95e+001	27.7	8	2.97e+001	108	100	1.85e+001
100	4000	8.85e+001	101	0.8	100	2.01e+001	37.3	3	3.12e+001	95	100	1.89e+001
100	5000	1.00e+002	94	1.0	100	2.05e+001	49.7	3	2.96e+001	88	100	1.85e+001
200	1000	4.30e+001	518	0.4	200	1.50e+001	10.7	16	2.95e+001	577	200	1.38e+001
200	2000	6.35e+001	229	0.6	200	1.24e+001	21.1	12	2.91e+001	224	200	1.14e+001
200	3000	7.75e+001	146	0.8	200	1.22e+001	27.5	9	2.85e+001	136	200	1.21e+001
200	4000	9.14e+001	112	0.9	200	1.25e+001	37.2	5	2.78e+001	124	200	1.12e+001
200	5000	1.01e+002	113	1.2	200	1.17e+001	49.4	6	2.68e+001	97	200	1.06e+001
300	1000	4.65e+001	716	0.7	300	7.13e+000	9.2	22	2.81e+001	836	300	7.05e+000
300	2000	6.36e+001	219	0.6	300	5.95e+000	18.4	12	2.68e+001	232	300	6.33e+000
300	3000	7.88e+001	158	0.8	300	5.91e+000	29.3	12	2.58e+001	145	300	6.15e+000
300	4000	8.95e+001	142	1.1	300	5.61e+000	44.9	15	2.60e+001	140	300	6.27e+000
300	5000	1.01e+002	125	1.3	300	5.54e+000	49.4	7	2.73e+001	114	300	6.07e+000

We initialize the algorithm at the origin and terminate when the change between successive iterates is small, i.e., when

$$\frac{\|x^t - x^{t-1}\|}{\|x^t\| + 1} < 10^{-8}.$$

We consider random instances. Specifically, for $m = 1000$ and each $n = 3000$, 4000, 5000 and 6000, we generate a random matrix $\mathcal{A} \in \mathbb{R}^{m\times n}$ with i.i.d. standard Gaussian entries. We also generate $b \in \mathbb{R}^n$ with i.i.d. standard Gaussian entries.

The computational results are reported in Table 2, where we take $\mathcal{C}$ to be the unit ℓ_1 norm ball for the first 4 rows, and the unit ℓ_∞ norm ball for the rest. We report the quantity $\lambda_{\max}(\mathcal{A}^*\mathcal{A})$ for each of the random instances: the reciprocal of this quantity is typically used as an upper bound of the allowable step-size β in the usual proximal gradient algorithm. We consider $\beta = 1/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$, $2/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$, $10/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$ and $50/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$, and report the terminating function value and number of iterations. We observe that the number of iterations is typically less when β is larger. On the other hand, we can also observe that the terminating function values are not affected by the choice of step-size β for the easier problems corresponding to the ℓ_1 norm ball, but the solution quality concerning the ℓ_∞ norm ball does depend on the step-size β.

Table 2: Performance of the proximal gradient algorithm with varying β.

		$\beta = 1/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$		$\beta = 2/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$		$\beta = 10/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$		$\beta = 50/\lambda_{\max}(\mathcal{A}^*\mathcal{A})$	
n	$\lambda_{\max}(\mathcal{A}^*\mathcal{A})$	iter	fval	iter	fval	iter	fval	iter	fval
3000	7.41e+003	71	-1.108e+003	44	-1.108e+003	8	-1.189e+003	4	-1.189e+003
4000	8.97e+003	38	-1.205e+003	21	-1.205e+003	7	-1.205e+003	4	-1.205e+003
5000	1.04e+004	63	-1.102e+003	34	-1.102e+003	10	-1.102e+003	5	-1.102e+003
6000	1.19e+004	58	-1.135e+003	30	-1.135e+003	9	-1.135e+003	4	-1.135e+003
3000	7.44e+003	206	-7.259e+006	207	-7.180e+006	70	-7.005e+006	44	-6.829e+006
4000	8.96e+003	209	-1.154e+007	175	-1.148e+007	106	-1.136e+007	55	-1.122e+007
5000	1.05e+004	983	-1.722e+007	244	-1.709e+007	179	-1.713e+007	56	-1.694e+007
6000	1.18e+004	1068	-2.318e+007	377	-2.293e+007	166	-2.292e+007	43	-2.271e+007

6 Conclusion and future directions

In this paper, we study the proximal ADMM and the proximal gradient algorithm for solving problem (1) with a general surjective $\mathcal{M}$ and $\mathcal{M} = \mathcal{I}$, respectively. We prove that any cluster point of the sequence generated from the algorithms gives a stationary point by assuming merely a specific choice of parameters and the existence of a cluster point. We also show that if the functions h and P are in addition semi-algebraic and the sequence generated by the ADMM (i.e., proximal ADMM with $\phi = 0$) clusters, then the sequence is actually convergent. Furthermore, we give simple sufficient conditions for the boundedness of the sequence generated from the proximal ADMM.

Whether the proximal ADMM will return a stationary point when $\mathcal{M}$ is injective is still open. However, as suggested by the numerical experiments in [12] and our preliminary numerical tests, it is conceivable that the ADMM does *not* cluster at a stationary point in general when applied to solving problem (1) with an injective $\mathcal{M}$. One interesting research direction would be to adapt other splitting methods for convex problems to solve (1) and study their convergence properties.

Acknowledgement. The second author would like to thank Ernie Esser and Gabriel Goh for enlightening discussions.

References

[1] B. P. W. Ames and M. Hong. Alternating direction method of multipliers for sparse zero-variance discriminant analysis and principal component analysis. Preprint, January 2014. Available at `http://arxiv.org/abs/1401.5492`.

[2] H. Attouch, J. Bolte, P. Redont and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Lojasiewicz inequality. *Math. Oper. Res.* 35, pp. 438–457 (2010).

[3] H. Attouch, J. Bolte and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.* 137, Ser. A, pp. 91–129 (2013).

[4] T. Blumensath and M. Davies. Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.* 14, pp. 629–654 (2008).

[5] J. Bolte, A. Daniilidis and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* 17, pp. 1205–1223 (2007).

[6] J. Bolte, A. Daniilidis, A. Lewis and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18, pp. 556–572 (2007).

[7] K. Bredies and D. A. Lorenz. Minimization of non-convex, non-smooth functionals by iterative thresholding. Preprint (2009).

[8] R. H. Chan, J. Yang and X. Yuan. Alternating direction method for image inpainting in wavelet domain. *SIAM J. Imaging Sci.* 4, pp. 807–826 (2011).

[9] C. Chen, B. He and X. Yuan. Matrix completion via alternating direction methods. *IMA J. of Numer. Anal.* 32, pp. 227–245 (2012).

[10] E. J. Candés and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.* 9, pp. 717–772 (2009).

[11] E. J. Candés and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory* 51, pp. 4203–4215 (2005).

[12] B. Dong and Y. Zhang. An efficient algorithm for ℓ_0 minimization in wavelet frame based image restoration. *J. Sci. Comput.* 54, pp. 333–349 (2013).

[13] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Prog.* 55, pp. 293–318 (1992).

[14] M. Fortin and R. Glowinski. On decomposition-coordination methods using an augmented Lagrangian. In M. Fortin and R. Glowinski, eds., *Augmented Lagrangion Methods: Applications to the Solution of Boundary Problems.* North-Holland, Amsterdam, 1983.

[15] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *Int. J. Syst. Sci.* 12, pp. 989–1000 (1981).

[16] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, eds., *Augmented Lagrangion Methods: Applications to the Solution of Boundary Problems.* North-Holland, Amsterdam, 1983.

[17] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Comput. Math. Appl.* 2, pp. 17–40 (1976).

[18] P. Gong, C. Zhang, Z. Lu, J. Huang and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. The 30th International Conference on Machine Learning (ICML 2013).

[19] W. Hare and C. Sagastizábal. Computing proximal points of nonconvex functions. *Math. Program.* 116, pp. 221–258 (2009).

[20] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press (1991); 8th printing in 2007.

[21] A. Kyrillidis, S. Becker, V. Cevher and C. Koch. Sparse projections onto the simplex. Preprint, April 2013. Available at `http://arxiv.org/abs/1206.1529`.

[22] Z. Liu and L. Vandenberghe. Semidefinite programming methods for system realization and identification. *Proc. 48th IEEE Conference on Decision and Control*, pp. 4676–4681 (2009).

[23] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation, I: Basic Theory, II: Applications.* Springer, Berlin (2006).

[24] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer (1998).

[25] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.* 29, pp. 119–138 (1991).

[26] H. Wang and A. Banerjee. Bregman alternating direction method of multipliers. Preprint, June 2013. Available at `http://arxiv.org/abs/1306.3203`.

[27] Y. Wang, J. Yang, W. Yin and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sci.* 1, pp. 248–272 (2008).

[28] Z. Wen, D. Goldfarb and W. Yin. Alternating direction augmented Lagrangian methods for semidefinite programming. *Math. Prog. Comp.* 2, pp. 203–230 (2010).

[29] Z. Wen, X. Peng, X. Liu, X. Bai and X. Sun. Asset allocation under the basel accord risk measures. Preprint, January 2013. Available at http://www.optimization-online.org/DB_HTML/2013/01/3730.html.

[30] J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM J. Sci. Comput.* 33, pp. 250–278 (2011).

[31] J. Zeng, S. Lin, Y. Wang and Z. Xu. $L_{1/2}$ regularization: convergence of iterative half thresholding algorithm. *IEEE Trans. Signal Process.* 62, pp. 2317–2329 (2014).