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

Tractable ADMM Schemes for Computing KKT Points and Local Minimizers for ℓ_0-Minimization Problems

Yue Xie · **Uday V. Shanbhag**

Received: date / Accepted: date

Abstract We consider an ℓ_0-minimization problem where $f(x)+\gamma\|x\|_0$ is minimized over a polyhedral set and the ℓ_0-norm regularizer implicitly emphasizes sparsity of the solution. Such a setting captures a range of problems in image processing and statistical learning. Given the nonconvex and discontinuous nature of this norm, convex regularizers are often employed as substitutes. Therefore, far less is known about directly solving the ℓ_0-minimization problem. Inspired by [19], we consider resolving an equivalent formulation of the ℓ_0-minimization problem as a mathematical program with complementarity constraints (MPCC) and make the following contributions towards the characterization and computation of its KKT points: (i) First, we show that feasible points of this formulation satisfy the relatively weak Guignard constraint qualification. Furthermore, under the suitable convexity assumption on $f(x)$, an equivalence is derived between first-order KKT points and local minimizers of the MPCC formulation. (ii) Next, we apply two alternating direction method of multiplier (ADMM) algorithms to exploit special structure of the MPCC formulation: ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$) and ($\mathrm{ADMM}_{\mathrm{cf}}$). These two ADMM schemes both have tractable subproblems. Specifically, in spite of the overall nonconvexity, we show that the first of the ADMM updates can be effectively reduced to a closed-form expression by recognizing a hidden convexity property while the second necessitates solving a convex program. In ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$), we prove subsequential convergence to a perturbed KKT point under mild assumptions. Our preliminary numerical experiments suggest that the tractable ADMM schemes are more scalable than their standard counterpart and $\mathrm{ADMM}_{\mathrm{cf}}$ compares well with its competitors to solve the ℓ_0-minimization problem.

Y. Xie
Penn State University
E-mail: yux111@psu.edu

U. V. Shanbhag
Penn State University
E-mail: udaybag@psu.edu

Keywords Nonconvex sparse recovery · Constraint qualifications and KKT conditions · Alternating direction method of multiplier (ADMM) · Tractability · Convergence analysis

1 Introduction

In this paper, we consider the ℓ_0-minimization problem:

$$\min_{x} \; f(x) + \gamma \|x\|_0 \quad \text{subject to } Ax \geq b, \tag{1}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\gamma > 0$. Suppose $f(x) \triangleq f_Q(x) + g(x)$, where $f_Q : \mathbb{R}^n \to \mathbb{R}$ is a quadratic function and $g : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function. The ℓ_0-norm of a vector captures the number of nonzero entries while an ℓ_0-norm regularizer implicitly emphasizes the sparsity of the resulting minimizer. ℓ_0-minimization problems of the form (1) assume relevance in applications in image processing and statistical learning (cf. [13, 15, 30]). The nonconvexity and discontinuity of the ℓ_0-norm has prompted the usage of convex ℓ_1 or ℓ_2-norm regularizers or other tractable variants [2, 30]. While relatively less is known about directly solving problem (1), a solution of (1) may have better statistical property. In fact, global solutions of (1) achieve model selection consistency and are known to be sparse under weaker conditions than when utilizing the ℓ_1-norm (cf. [37]). Therefore, despite the computational challenges in addressing the ℓ_0-norm penalty, resolution of the ℓ_0-minimization problem is still desirable. In this work, we focus on direct resolution of (1).

Related work. To solve (1), Feng, Mitchell, Pang, Shen, and Wächter [19] introduced two complementarity-based formulations equivalent with (1) and processed them by standard nonlinear programming solvers. Blumensath and Davis proposed an iterative hard-thresholding (IHT) algorithm, applicable when $f(x)$ is a least-squares metric and the constraint $Ax \geq b$ is absent [8]. Convergence to a local minimizer may be claimed and performance of the scheme can be improved if warm-started from a point computed by matching pursuit.

A problem class closely related to (1) is the ℓ_0-constrained problem (2). Although they are not equivalent due to nonconvexity of ℓ_0-norm, solution method of (2) may inspire efficient algorithms to tackle (1).

$$\min_{x} \; f(x) \quad \text{subject to } Ax \geq b, \|x\|_0 \leq M. \tag{2}$$

This problem finds application in best subset regression [5, 6], cardinality constrained portfolio optimization [6], and graphical model estimation [18]. To solve (2), combining first-order methods and mixed-integer optimization [5] was seen to be promising. By considering an equivalent complementarity formulation of (2), Burdakov et al. [10] developed a regularization scheme. Moreover, a relatively weak constraint qualification was shown to hold at every feasible point of this reformulation and consequently KKT conditions are necessary at local minima.

In addition to ℓ_0-norm penalization, related work has examined the usage of the ℓ_p-norm ($p \in (0,1)$) [20, 21], the smoothly clipped absolute deviation (SCAD) penalty [17, 26], the minimax concave penalty (MCP) [36], and the capped-ℓ_1 penalty [38]. More recently, a generalization of the ℓ_0-norm constraint was considered in the form of an *affine sparsity constraint [14]*.

Nonconvex ADMM schemes. Since our focus lies in developing an ADMM framework to exploit the structure of an equivalent nonconvex formulation of (1), we provide a brief review of the available convergence statements in the context of ADMM schemes for nonconvex programs.

Encouraged by the success of ADMM on convex problems, researchers have tried to implement and analyze ADMM on nonconvex problems. Table 1 lists some of the main theoretical findings regarding variants of ADMM schemes employed to address different types of nonconvex problems [9, 24, 25, 33]. In the second column, we include the assumptions necessary for these authors to prove convergence or derive complexity bounds. Note that these assumptions pertain to the problem itself but may not be sufficient to guarantee the final result. More assumptions on the parameter settings or iterates of the algorithms may well be needed. Moreover, for some of the findings, it is shown that if the KŁ property (See Definition 6 in Appendix) is assumed, convergence can be guaranteed [9, 33]. Also note that all of the papers in Table 1 assume global resolution of each subproblem of ADMM, even when the subproblem is nonconvex. Specifically, in [33], it is explained that the proposed ADMM scheme can address MPCC but requires globally resolving an MPCC at each step; this is in sharp contrast with the tractable structure of each update in our scheme in this paper (in other words, we do not require resolving a nonconvex problem globally at each step).

There have also been extensions of nonconvex ADMM schemes to the linearized regime [27], nonlinear equality-constrained settings [32], amongst others [22, 31, 34, 35]. Despite all of these theoretical achievements on nonconvex ADMM, we point out that no scheme introduced above can guarantee the convergence or even the boundedness of the iterates when applying ADMM or its variants to the following formulation with both blocks constrained and one being nonconvex:

$$\begin{aligned} \min \quad & F(x) + G(y) \\ \text{subject to} \quad & x - y = 0, \\ & x \in X \subsetneqq \mathbb{R}^N, y \in Y \subsetneqq \mathbb{R}^N, \end{aligned} \tag{3}$$

where F is a quadratic function, G is smooth and convex, X is a nonconvex set defined by a quadratic equality constraint, while Y is a convex set. Formulation (3) is our focus in this paper because by reformulating the problem of interest in this way and applying ADMM-type schemes, each subproblem will be tractable and may possibly allow for a closed-form solution. Jiang et al. [25]

Table 1: Main results on convergence of nonconvex ADMM

Problem	Necessary Assumptions	Result	Lit.
$\min\ \phi(x_0, x_1, \ldots, x_p, y)$ s.t. $\sum_{i=0}^{p} A_i x_i + By = 0$.	$\nabla_y \phi(\cdot, y)$ is Lipschitz continuous in y. $\mathrm{Im}([A_1, \ldots, A_p]) \subseteq \mathrm{Im}(B)$.	Subsequential convergence to stationary points.	[33]
$\min\ f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^{N-1} r_i(x_i)$ s.t. $\sum_{i=1}^{N} A_i x_i = b, x_i \in \mathcal{X}_i$, $\forall i = 1, \ldots, N-1$.	f is differentiable. For any i, $\mathcal{X}_i$ is convex. A_N has full row rank.	Iter. complexity of $\mathcal{O}(1/\epsilon^2)$ to obtain an ϵ-stationary point.	[25]
$\min \sum_{k=1}^{K} g_k(x_k) + h(x_0)$ s.t. $x_k = x_0, \forall k = 1, \ldots, K$, $x_0 \in \mathcal{X}$.	For any $k = 1, \ldots, K$, $\nabla g_k(x)$ is Lipschitz continuous. $h(\cdot)$ is convex. $\mathcal{X}$ is convex and compact.	Subsequential convergence to stationary points.	[24]
$\min \sum_{k=1}^{K} g_k(x_k) + \ell(x_0)$ s.t. $\sum_{k=1}^{K} A_k x_k = x_0$, $x_k \in \mathcal{X}_k, \forall k = 1, \ldots, K$.	$g_k(\cdot)$ is either convex or Lipschitz continuously differentiable. $\nabla l(x)$ is Lipschitz continuous. $\mathcal{X}_k$ is convex and compact. A_k has full column rank.	Subsequential convergence to stationary points.	[24]
$\min\ F(z) + G(y) + H(x, y)$ s.t. $Ax - z = 0$.	F and G are proper lower semicontinuous. ∇H is Lipschitz continuous. A is surjective.	Subsequential convergence to KKT points.	[9]

discussed how ADMM schemes may be applied to (3) to allow for deriving convergence guarantees. Yet it requires changing the formulation of (3) through the addition of an unconstrained auxiliary block and requires penalizing the auxiliary variable in the objective function.

Motivation and contributions. Despite the breadth of prior research, less is known regarding the nature of solutions and tractable convergent schemes for continuous reformulations of (1). Motivated by this gap and inspired by [19], we consider an equivalent MPCC reformulation of (1):

$$\begin{array}{ll} \min\limits_{x^{\pm},\xi} & f(x^+ - x^-) + \gamma e^T(e - \xi) \\ \text{subject to} & A(x^+ - x^-) \geq b, \quad (x^+ + x^-)^T \xi = 0, \\ & x^+, x^- \geq 0, \quad 0 \leq \xi_i \leq 1, \text{ for } i = 1, \ldots, n. \end{array} \tag{4}$$

In particular, we focus on characterizing stationary points of (4) as well as developing **tractable** convergent scheme that may recover such solutions.
(i) Regularity properties and characterization of KKT points. In Section 2, we show that a feasible point of the MPCC reformulation satisfies the Guignard constraint qualification (GCQ). Under convexity of f, we derive an equivalence between first-order KKT points and local minimizers.
(ii) ADMM schemes with tractable subproblems. In Sections 3 and 4, we propose two ADMM schemes to exploit the special structure of the MPCC: ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$) and ($\mathrm{ADMM}_{\mathrm{cf}}$). In particular, we reformulate the MPCC (4) in the form of (3) and apply the ADMM frameworks. The algorithms require resolving two subproblems at each iteration where, one is convex and the other, while nonconvex, is shown to possess a hidden convexity property [4], and allow for closed-form solutions. In the perturbed proximal ADMM scheme ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$), the perturbation technique (inspired by Hajinezhad and Hong

[23]) allows us to show subsequential convergence. We also show that a limit point of this scheme is a perturbed KKT point where the inexactness depends on the choice of the perturbation parameters of the algorithm.
(iii) Numerics. In Section 5, we present some preliminary numerical experiments showing that the tractable ADMM schemes are more scalable than their standard counterpart and ADMM_{cf} competes well with other solution methods for a special case of the ℓ_0-minimization problem.

Notation. We let e denote $(1;\ldots;1)$ for an appropriate dimension. Given a set Z and a vector z, $\mathbb{1}_Z(z) = 0$ if $z \in Z$ and ∞ otherwise. The requirement $a \perp b$ is equivalent to $a_i b_i = 0$ for $i = 1,\ldots,n$. The matrix I_n denotes the $n-$dimensional identity matrix. $[1,n] \triangleq \{1,2,\ldots,n\}$. $|S|$ denotes the cardinality of set S. $(a)_i$ or $[a]_i$ denote the ith entry of vector a. We may also use a_i to denote ith entry of vector a, but often a_i may have other connotations such as the ith iterate in an algorithm, which will be specified. Let the support set of x be defined as $\text{supp}(x) \triangleq \{i \in \{1,\ldots,n\} \mid x_i \neq 0\}$. For any vector $z \in \mathbb{R}^N$, positive semidefinite matrix $M \in \mathbb{R}^{N\times N}$, $\|z\|_M^2 \triangleq z^T M z$.

2 Properties of the MPCC reformulation

In Section 2.1, we revisit the MPCC formulation (4) and study both its regularity properties (Section 2.2) and the relation between KKT points and local minimizers (Section 2.3).

2.1 Complementarity-based reformulations

In [19], several complementarity-based reformulations of (1) are introduced:
Half-complementarity

$$\begin{array}{rl} \min\limits_{x,\xi} & f(x) + \gamma e^T(e-\xi) \\ \text{subject to} & Ax \geq b,\ x_i\xi_i = 0, \\ & 0 \leq \xi_i \leq 1,\ \text{for } i = 1,\ldots,n. \end{array} \tag{5}$$

The term "half-complementarity" arises from noting that the equality constraint may be recast as $x \perp \xi \geq 0$.
Full-complementarity

$$\begin{array}{rl} \min\limits_{x,x^{\pm},\xi} & f(x) + \gamma e^T(e-\xi) \\ \text{subject to} & Ax \geq b, x^+ - x^- = x, \\ & (x^+)^T x^- = 0,\ (x^+ + x^-)^T\xi = 0, \\ & x_i^+, x_i^- \geq 0,\ 0 \leq \xi_i \leq 1,\ \text{for } i = 1,\ldots,n. \end{array} \tag{6}$$

where $x^+, x^-, \xi \in \mathbb{R}^n$. (6) may be further simplified by relaxing $(x^+)^T x^- = 0$, resulting in (4). It can be formally shown that (4) is a tight relaxation of (6) implying that a solution of (4) is a minimizer of (6) (See Lemma 8 in the Appendix). Since equivalence between (6) and (1) has been established [19], the tightness of relaxation indicates equivalence between (4) and (1). Moreover, the following result shows that local minimizers of (1) can also be recovered by local minimizers of (4).

Lemma 1 Given $\hat{x}, \hat{x}^+, \hat{x}^-, \hat{\xi} \in \mathbb{R}^n$ such that $\hat{x} = \hat{x}^+ - \hat{x}^-$ and $(\hat{x}^+; \hat{x}^-; \hat{\xi})$ is a local minimum of (4). Then $\hat{x}$ is a local minimum of (1).

Proof Suppose $\mathcal{Z}$ denotes the feasible region of (4). Since $\hat{z} \triangleq (\hat{x}^+; \hat{x}^-; \hat{\xi})$ is a local minimum of (4), $\hat{z} \in \mathcal{Z}$ and there exists an open neighbourhood $\mathcal{N} \triangleq B(\hat{z}, r) \triangleq \{z \in \mathbb{R}^{3n} \mid \|z - \hat{z}\| < r\}$ such that for all $(x^+; x^-; \xi) \in \mathcal{N} \cap \mathcal{Z}$, $f(x^+ - x^-) + \gamma e^T(e - \xi) \geq f(\hat{x}^+ - \hat{x}^-) + \gamma e^T(e - \hat{\xi})$. Let $X \triangleq \{x \mid Ax \geq b\}$. It suffices to show that (a) $\hat{x} \in X$ and (b) there exists an open neighbourhood $\mathcal{U} \ni \hat{x}$ such that for all $x \in \mathcal{U} \cap X$, $f(x) + \gamma\|x\|_0 \geq f(\hat{x}) + \gamma\|\hat{x}\|_0$. Of these, (a) holds immediately by noting that $A\hat{x} = A(\hat{x}^+ - \hat{x}^-) \geq b$ where the inequality follows from the feasibility of $(\hat{x}^+; \hat{x}^-; \hat{\xi})$ with respect to (4). Suppose $\mathcal{U}$ is defined as a sufficiently small set such that the following hold: (i) For all $x \in \mathcal{U}$, $f(x) \geq f(\hat{x}) - \gamma$, a consequence of the continuity of f;(ii) For all $x \in \mathcal{U} \cap X$, $\hat{x}_i \neq 0 \Rightarrow x_i \neq 0$, $\forall i = 1, \ldots, n$; (iii) $\mathcal{U} \subseteq B(\hat{x}, r)$. Then (ii) implies $\mathrm{supp}(x) \supseteq \mathrm{supp}(\hat{x})$ for all $x \in \mathcal{U} \cap X$ (or $\|\hat{x}\|_0 \leq \|x\|_0$). Therefore, the local optimality of $\hat{x}$ can be shown through the following two cases: (I). If $\bar{x} \in \{x \in \mathcal{U} \cap X \mid \mathrm{supp}(x) \supsetneq \mathrm{supp}(\hat{x})\}$, then $\|\bar{x}\|_0 \geq \|\hat{x}\|_0 + 1$ implying that $f(\bar{x}) + \gamma\|\bar{x}\|_0 \overset{(i)}{\geq} f(\hat{x}) - \gamma + \gamma(\|\hat{x}\|_0 + 1) = f(\hat{x}) + \gamma\|\hat{x}\|_0$; (II). If $\bar{x} \in \{x \in \mathcal{U} \cap X \mid \mathrm{supp}(x) = \mathrm{supp}(\hat{x})\}$, then let $\bar{x}_i^+ \triangleq \hat{x}_i^+ + \max\{\bar{x}_i - \hat{x}_i, 0\}$, $\bar{x}_i^- \triangleq \hat{x}_i^- - \min\{\bar{x}_i - \hat{x}_i, 0\}$ for $i = 1, \ldots, n$. Then we see that $\bar{x} = \bar{x}^+ - \bar{x}^-$ and $(\bar{x}^+; \bar{x}^-; \hat{\xi}) \in \mathcal{N} \cap \mathcal{Z}$. Therefore, $f(\bar{x}^+ - \bar{x}^-) + \gamma e^T(e - \hat{\xi}) \geq f(\hat{x}^+ - \hat{x}^-) + \gamma e^T(e - \hat{\xi})$ implying $f(\bar{x}^+ - \bar{x}^-) \geq f(\hat{x}^+ - \hat{x}^-)$ or $f(\bar{x}) \geq f(\hat{x})$. It follows that $f(\bar{x}) + \gamma\|\bar{x}\|_0 \geq f(\hat{x}) + \gamma\|\hat{x}\|_0$. ∎

While (1) can now be reformulated as a continuous problem, (4) is still an MPCC. It may be recalled that MPCCs are ill-posed nonconvex nonlinear programs in that standard regularity conditions (such as LICQ or MFCQ) may fail to hold at any feasible point [28]. Moreover, global resolution of such problems is generally challenging. We now discuss what constraint qualifications do hold at a feasible point of (4).

2.2 Constraint Qualifications

In this subsection, we analyze whether regularity conditions hold at feasible points for the simplified full complementarity formulation (4). This allows for stating necessary conditions of optimality. Recall that some common CQs are related as follows.

$$\textbf{(I)}\ \text{LICQ} \Rightarrow \text{CRCQ} \quad \text{and} \quad \textbf{(II)}\ \text{LICQ} \Rightarrow \text{MFCQ} \Rightarrow \text{ACQ} \Rightarrow \text{GCQ} \tag{7}$$

The first relation is obvious from the definition of LICQ and CRCQ (See [16, Page 262]) while the proof of the second relation may be found in [12]. In the context of the half-complementarity formulation (5), the constant rank constraint qualification (CRCQ) is proven to hold at points satisfying certain nondegeneracy property while LICQ may fail [19]. In this section, we focus on the simplified full-complementarity formulation (4). It can be shown that GCQ may hold at every feasible point while ACQ may fail.

We begin our discussion with some definitions. Suppose $g : \mathbb{R}^n \to \mathbb{R}^p$ and $h : \mathbb{R}^n \to \mathbb{R}^q$ are continuously differentiable functions while Ω is a set defined as follows.

$$\Omega \triangleq \{x \in \mathbb{R}^n : g(x) \le 0, h(x) = 0\}. \tag{8}$$

Then the tangent cone $T_\Omega(x^*)$ and linearized cone $L_\Omega(x^*)$ of Ω at x^* and the ACQ and the GCQ are defined as follows:

Definition 1 (Abadie and Guignard CQ (ACQ, GCQ)) If $\mathcal{I}(x^*) = \{i : g_i(x^*) = 0\}$, then $T_\Omega(x^*)$ and $L_\Omega(x^*)$ of Ω at x^* are defined as follows:

$$T_\Omega(x^*) \triangleq \left\{ d : \exists \{x_k\} \subseteq \Omega, \{t_k\} \downarrow 0, \text{ s.t. } x_k \to x^* \text{and } d = \lim_{k\to\infty} \frac{x_k - x^*}{t_k} \right\} \tag{9}$$

$$L_\Omega(x^*) \triangleq \left\{ d : \nabla g_i(x^*)^T d \le 0, \forall i \in \mathcal{I}(x^*), \nabla h_j^T(x^*) d = 0, j = 1, \ldots, q \right\}. \tag{10}$$

Then x^* satisfies the **Abadie Constraint Qualification** (ACQ) iff $T_\Omega(x^*) = L_\Omega(x^*)$. Further, x^* satisfies the **Guignard Constraint Qualification** (GCQ) iff $(T_\Omega(x^*))^* = (L_\Omega(x^*))^*$, where for a cone $C \subseteq \mathbb{R}^n$, $C^* \triangleq \{v : d^T v \le 0, \ \forall d \in C\}$.

Next, we prove that the GCQ holds at every feasible point of (4).

Lemma 2 (GCQ holds at feasible points) Consider the problem (4) and consider a feasible point $x = (x^+; x^-; \xi)$. Then the GCQ holds at this point.

Proof For the point $x = (x^+; x^-; \xi)$, define

$$A^T \triangleq (a_1, \ldots, a_m) \text{ and } E(x) \triangleq \{i : a_i^T (x^+ - x^-) = b_i\}, \tag{11}$$

$$\begin{aligned} & S(x) = \left\{ i : x_i^+ = x_i^- = 0 \right\}, \\ & S_0(x) = \{i \in S(x) : \xi_i = 0\}, S_1(x) = \{i \in S(x) : \xi_i = 1\}. \end{aligned} \tag{12}$$

In addition, define cones $C_1(x)$ and $C_2(x)$ as

$$C_2(x) \triangleq \left\{ \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} : \begin{array}{l} (d_1)_i = 0, (d_2)_i = 0, \forall i \in S(x) \setminus S_0(x); \\ (d_1)_i \ge 0, \forall i \in S_0(x) \cup (S(x)^c \cap \{i : x_i^+ = 0\}); \\ (d_2)_i \ge 0, \forall i \in S_0(x) \cup (S(x)^c \cap \{i : x_i^- = 0\}); \\ (d_3)_i \ge 0, \forall i \in S_0(x); (d_3)_i \le 0, \forall i \in S_1(x); \\ (d_3)_i = 0, \forall i \in S(x)^c; a_j^T d_1 - a_j^T d_2 \ge 0, \forall j \in E(x) \end{array} \right\},$$

$$C_1(x) \triangleq C_2(x) \cap \{d = (d_1; d_2; d_3) : [(d_1)_i + (d_2)_i] (d_3)_i = 0, \forall i \in S_0(x)\}, \tag{13}$$

respectively where it may be noted that $C_1(x)$ is characterized by an extra constraint $[(d_1)_i + (d_2)_i]\,(d_3)_i = 0$ for all $i \in S_0(x)$. Further, denote

$$X \triangleq \left\{ (y^+; y^-; \zeta) : \begin{array}{l} y^+, y^-, \zeta \in \mathbb{R}^n, (y^+ + y^-)^T \zeta = 0, A(y^+ - y^-) \geq b, \\ y^+ \geq 0, y^- \geq 0, 0 \leq \zeta_i \leq 1, \forall i = 1, \ldots, n, \end{array} \right\}.$$

We proceed to show the following.
(i). $T_X(x) = C_1(x)$: Suppose $d \in T_X(x)$. Then there exist sequences $\{x_k\}$ and $\{t_k\}$ such that $\{x_k\} \subseteq X, x_k \to x$, $\{t_k\} \downarrow 0$ and $d = \lim_{k\to\infty} \frac{x_k - x}{t_k}$. Denote $x_k \triangleq (x^+_{(k)}; x^-_{(k)}; \xi_{(k)})$, where $x^+_{(k)}, x^-_{(k)}, \xi_{(k)} \in \mathbb{R}^n$. Suppose that $d \triangleq (d_1; d_2; d_3), d_1, d_2, d_3 \in \mathbb{R}^n$. Based on feasibility of $x_k, \forall k \geq 1$ and the fact that $x_k \to x$, we may claim the following:

$$\begin{aligned}
&\forall i \in S(x) \setminus S_0(x), \exists K_1, \text{s.t.}, \forall k \geq K_1, (x^+_{(k)})_i = (x^-_{(k)})_i = 0 \\
&\quad \Longrightarrow (d_1)_i = (d_2)_i = 0, \forall i \in S(x) \setminus S_0(x) \\
&\text{if } i \in S_1(x), \text{ then } \xi_i = 1 \text{ and } (\xi_{(k)})_i \leq 1, \forall k \\
&\quad \Longrightarrow (\xi_{(k)})_i - \xi_i \leq 0, \forall k, (d_3)_i \leq 0, \forall i \in S_1(x).
\end{aligned}$$

Similarly we may claim the following:

$$\begin{aligned}
&\forall i \in S(x)^c, \exists K_2, \text{ s.t.} \forall k \geq K_2, (\xi_{(k)})_i = 0, (x^+_{(k)})_i \geq 0, (x^-_{(k)})_i \geq 0 \\
&\Longrightarrow (d_3)_i = 0, \forall i \in S(x)^c;\ (d_1)_i \geq 0, \forall i \in S(x)^c \cap \{i : x^+_i = 0\}; \\
&\text{and } (d_2)_i \geq 0, \forall i \in S(x)^c \cap \{i : x^-_i = 0\}.
\end{aligned}$$

For indices $i \in S_0(x)$, the following holds:

$$\begin{aligned}
&x^+_i = x^-_i = \xi_i = 0 \Rightarrow (x^+_{(k)})_i - x^+_i \geq 0, (x^-_{(k)})_i - x^-_i \geq 0, (\xi_{(k)})_i - \xi_i \geq 0, \forall k, \\
&\Longrightarrow (d_1)_i \geq 0, (d_2)_i \geq 0, (d_3)_i \geq 0. \\
&\left[(x^+_{(k)})_i + (x^-_{(k)})_i\right] (\xi_{(k)})_i = 0, \forall k \\
&\Longrightarrow (x^+_{(k)})_i + (x^-_{(k)})_i = 0, \text{ or } (\xi_{(k)})_i = 0, \text{ inf. often}; \\
&\Longrightarrow (d_1)_i + (d_2)_i = 0, \text{ or } (d_3)_i = 0 \iff [(d_1)_i + (d_2)_i]\,(d_3)_i = 0 \\
&\text{Furthermore,} \\
&\forall j \in E(x), a_j^T x^+ - a_j^T x^- = b_j, a_j^T x^+_{(k)} - a_j^T x^-_{(k)} \geq b_j, \text{ for all } k \geq 1 \\
&\Longrightarrow a_j^T (x^+_{(k)} - x^+) - a_j^T (x^-_{(k)} - x^-) \geq 0, \forall j \in E(x) \text{ and } k \geq 1 \\
&\Longrightarrow a_j^T d_1 - a_j^T d_2 \geq 0, \forall j \in E(x).
\end{aligned}$$

Therefore, we may conclude from (13) that $d \in C_1(x)$ and $T_X(x) \subseteq C_1(x)$.

We now proceed to show that $C_1(x) \subseteq T_X(x)$. Choose any $d \in C_1(x)$. Then based on property of $C_1(x)$, it is easy to see that we may choose λ large enough such that $x + d/(k\lambda) \in X, \forall k \geq 1$. Let $x_k \triangleq x + d/(k\lambda), t_k \triangleq 1/(k\lambda)$ for all $k \geq 1$, implying that $\{x_k\} \subseteq X, x_k \to x, t_k \downarrow 0, d = \lim_{k\to\infty} \frac{x_k - x}{t_k}$ implying

that $d \in T_X(x)$ which further implies $C_1(x) \subseteq T_X(x)$.
(ii). $L_X(x) = C_2(x)$: The set X contains the following active constraints.

$$\begin{aligned}
&-y_i^+ \leq 0, \forall i \in S(x) \cup \{i \in S(x)^c : x_i^+ = 0\};\\
&-y_i^- \leq 0, \forall i \in S(x) \cup \{i \in S(x)^c : x_i^- = 0\};\\
&-\zeta_i \leq 0, \forall i \in S_0(x) \cup S(x)^c; \quad \zeta_i \leq 1, \forall i \in S_1(x);\\
&a_i^T(y^+ - y^-) \geq b_i, \forall i \in E(x); \quad (y^+ + y^-)^T\zeta = 0.
\end{aligned}$$

This allows for defining the linearized cone $L_X(x)$ at $x \in X$.

$$L_X(x) \triangleq \left\{ \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} : \begin{array}{l} -(d_1)_i \leq 0, \forall i \in S(x) \cup \{i \in S(x)^c : x_i^+ = 0\};\\ -(d_2)_i \leq 0, \forall i \in S(x) \cup \{i \in S(x)^c : x_i^- = 0\};\\ (d_3)_i \leq 0, \forall i \in S_1(x); -(d_3)_i \leq 0, \forall i \in S_0(x) \cup S(x)^c;\\ a_i^T(d_1 - d_2) \geq 0, \forall i \in E(x);\\ \xi^T(d_1 + d_2) + (x^+ + x^-)^T d_3 = 0. \end{array} \right\} \tag{14}$$

Suppose $d \in L_X(x)$. Then the following holds:

$$\begin{aligned}
&\xi^T(d_1 + d_2) + (x^+ + x^-)^T d_3 = 0\\
\iff &\sum_{i \in S(x) \setminus S_0(x)} \xi_i[(d_1)_i + (d_2)_i] + \sum_{i \in S(x)^c} (d_3)_i (x_i^+ + x_i^-) = 0\\
\iff &(d_1)_i = (d_2)_i = 0, \forall i \in S(x) \setminus S_0(x); \quad (d_3)_i = 0, \forall i \in S(x)^c,
\end{aligned} \tag{15}$$

where the first equivalence follows from the definition of $S_0(x)$ and $S(x)$ while the second follows from noting that $(d_1)_i \geq 0, (d_2)_i \geq 0, \xi_i > 0, \forall i \in S(x) \setminus S_0(x)$ and $(d_3)_i \geq 0, x_i^+ + x_i^- > 0, \forall i \in S(x)^c$. Therefore, by replacing $\xi^T(d_1 + d_2) + (x^+ + x^-)^T d_3 = 0$ with (15) in the representation (14), we observe that $L_X(x) = C_2(x)$.
(iii). We conclude the proof by showing that $C_2(x) = \mathrm{cl}(\mathrm{conv}(C_1(x)))$. Since $C_2(x)$ is a polyhedral cone, it is closed and convex. Furthermore, by definition, $C_2(x) \supseteq C_1(x)$, implying that $C_2(x) \supseteq \mathrm{cl}(\mathrm{conv}(C_1(x)))$. To prove the reverse direction, choose any vector $d \triangleq (d_1; d_2; d_3) \in C_2(x)$ where $d_1, d_2, d_3 \in \mathbb{R}^n$. It is easy to verify that both vectors $\tilde{d} \triangleq (\mathbf{0_{n\times 1}}; \mathbf{0_{n\times 1}}; \mathbf{2d_3})$ and $\hat{d} \triangleq (2d_1; 2d_2; \mathbf{0_{n\times 1}})$ are in $C_1(x)$. Note that $d = \frac{1}{2}\tilde{d} + \frac{1}{2}\hat{d} \in \mathrm{cl}(\mathrm{conv}(C_1(x)))$. Therefore, $C_2(x) \subseteq \mathrm{cl}(\mathrm{conv}(C_1(x)))$.

By (iii) $L_X(x) = \mathrm{cl}(\mathrm{conv}(T_X(x)))$, implying that $T_X(x)^* = L_X(x)^*$. ∎

Remark 1 (i) At a feasible point $x = (x^+; x^-; \xi)$ such that $[x^+ + x^-]_i = 0$ and $\xi_i = 0$ for some index i, ACQ may fail to hold. In fact, it is very likely that $T_X(x) = C_1(x) \subsetneq C_2(x) = L_X(x)$. On the other hand, at all other points, $S_0(x) = \emptyset$ and ACQ holds. (ii) KKT conditions are necessary at local minimum.

2.3 KKT conditions and local optimality

In this subsection, we discuss the relation between (first-order) KKT conditions and local optimality. We begin with the definition of KKT conditions.

Definition 2 (KKT conditions) Consider the problem $\{\min_{x\in\Omega} F(x)\}$, where $F(x)$ is a continuously differentiable function and Ω is defined in (8). Suppose x^* denotes a feasible solution of Ω. Then x^* satisfies the first-order KKT conditions if and only if there exists $\lambda \in \mathbb{R}^p_+, \mu \in \mathbb{R}^q$ such that

$$\begin{aligned} \nabla F(x^*) + \sum_{i=1}^p \lambda_i \nabla g_i(x^*) + \sum_{j=1}^q \mu_j \nabla h_j(x^*) = 0, \\ \lambda_i g_i(x^*) = 0, \qquad \forall i = 1,\ldots,p. \end{aligned} \tag{16}$$

Then by Definition 2, a point $x \triangleq (x^+; x^-; \xi)$ satisfies the first-order KKT conditions of problem (4) if there exist multipliers $(\mu, \beta_1, \beta_2, \beta_3, \beta_4, \pi) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ such that the following conditions hold:

$$0 = \begin{pmatrix} \nabla_x f(x^+ - x^-) \\ -\nabla_x f(x^+ - x^-) \\ -\gamma e \end{pmatrix} + \mu \begin{pmatrix} \xi \\ \xi \\ x^+ + x^- \end{pmatrix} + \begin{pmatrix} -\beta_1 - A^T\pi \\ -\beta_2 + A^T\pi \\ \beta_4 - \beta_3 \end{pmatrix}, \tag{17a}$$

$$0 \le \beta_1 \perp x^+ \ge 0, \tag{17b}$$

$$0 \le \beta_2 \perp x^- \ge 0, \tag{17c}$$

$$0 \le \beta_3 \perp \xi \ge 0, \tag{17d}$$

$$0 \le \beta_4 \perp e - \xi \ge 0, \tag{17e}$$

$$0 \le \pi \perp A(x^+ - x^-) - b \ge 0, \tag{17f}$$

$$(x^+ + x^-)^T \xi = 0. \tag{17g}$$

Before presenting the main result, we point out a non-degeneracy property of KKT points.

Lemma 3 (Nondegeneracy of first-order KKT points) Consider a point $x = (x^+; x^-; \xi)$ and a set of multipliers $(\mu, \beta_1, \beta_2, \beta_3, \beta_4, \pi)$ that satisfy the first-order KKT conditions (17) of (4). Then x satisfies the nondegeneracy property:

$$[x^+ + x^-]_i = 0 \Rightarrow \xi_i = 1. \tag{18}$$

Proof Suppose that $(x^+; x^-; \xi)$ verifies KKT conditions (17) with multipliers $\mu, \beta_1, \beta_2, \beta_3, \beta_4, \pi$. Then, by (17a), we have that $(x^+ + x^-)_i = 0 \Rightarrow (\beta_4 - \beta_3)_i = \gamma > 0$. But for a given i, for both $[\beta_4]_i$ and $[\beta_3]_i$ to be positive, we require that both $[\xi]_i = 0$ and $[1 - \xi]_i = 0$ hold, which is impossible. It follows that the only possibility is that $[\beta_4]_i = \gamma$ and $[\beta_3]_i = 0$, implying that $[\xi]_i = 1$. It follows that $(x^+; x^-; \xi)$ satisfies the property (18). ∎

Lemma 3 leads to the rather surprising equivalence between local minimizers and (first-order) KKT points.

Theorem 1 (Equivalence between local minimizers and KKT points) Consider problem (4), and let $x = (x^+; x^-; \xi)$ denote a feasible point. Assume that f in (4) is convex. Then the following statements are equivalent:

(a) x is a local minimizer of (4);
(b) There exist $\mu \in \mathbb{R}$, $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}^n$, and $\pi \in \mathbb{R}^m$ such that the first-order KKT conditions (17) hold;

Proof **(a)⇒(b).** This is true because GCQ holds at every feasible point by Lemma 2.
(b)⇒(a). Suppose that $x = (x^+; x^-; \xi)$ satisfies KKT conditions (17) with multipliers $(\mu, \beta_1, \beta_2, \beta_3, \beta_4, \pi)$. Then by the nondegeneracy property of a KKT point (Lemma 3), the set $\{1, \ldots, n\}$ can be partitioned into the following two sets, as in the same fashion when proving the CQ: $S(x) \triangleq \{i \in \{1, \ldots, n\} : x_i^+ = x_i^- = 0, \xi_i = 1\}$ and $S^c(x) \triangleq \{i \in \{1, \ldots, n\} : x_i^+ + x_i^- > 0, \xi_i = 0\}$. We denote that $A = (a_1, \ldots, a_n)$ (Note different notation from (11)). Then (17a) implies

$$\left.\begin{aligned} (\nabla_x f(x^+ - x^-))_i - a_i^T \pi = (\beta_1)_i \geq 0 \\ -(\nabla_x f(x^+ - x^-))_i + a_i^T \pi = (\beta_2)_i \geq 0 \end{aligned}\right\} \quad \forall i \in S^c(x).$$

because $\xi_i = 0$ for all $i \in S^c(x)$. Consequently, $(\beta_1)_i = -(\beta_2)_i$ where β_1 and β_2 are nonnegative. It follows that $(\beta_1)_i = (\beta_2)_i = 0$, and

$$(\nabla_x f(x^+ - x^-))_i = a_i^T \pi, \qquad \forall i \in S^c(x). \tag{19}$$

We proceed to prove that $(x^+; x^-)$ is a global minimizer of the following program:

$$\min \tilde{f}(z) \triangleq f(z^+ - z^-), \quad \text{subject to } z = (z^+; z^-) \in \tilde{X}(x), \tag{20}$$

where

$$\tilde{X}(x) \triangleq \left\{(z^+; z^-) \mid z^+, z^- \in \mathbb{R}^n_+, A(z^+ - z^-) \geq b, z_i^+ = z_i^- = 0, \forall i \in S(x)\right\}.$$

Consider any feasible point $(\tilde{x}^+; \tilde{x}^-)$ of (20). By applying (19) and noticing $x_i^\pm, \tilde{x}_i^\pm = 0 \ \forall i \in S(x)$ (by def.), $\pi^T A(x^+ - x^-) = \pi^T b, \pi \geq 0$ (by (17)), and $A(\tilde{x}^+ - \tilde{x}^-) - b \geq 0$,

$$\begin{aligned}
&\begin{pmatrix} \nabla_x f(x^+ - x^-) \\ -\nabla_x f(x^+ - x^-) \end{pmatrix}^T \left[\begin{pmatrix} \tilde{x}^+ \\ \tilde{x}^- \end{pmatrix} - \begin{pmatrix} x^+ \\ x^- \end{pmatrix} \right] \\
&= \nabla_x f(x^+ - x^-)^T [(\tilde{x}^+ - \tilde{x}^-) - (x^+ - x^-)] \\
&= \sum_{i \in S^c(x)} (\nabla_x f(x^+ - x^-))_i [(\tilde{x}_i^+ - \tilde{x}_i^-) - (x_i^+ - x_i^-)] \\
&= \sum_{i \in S^c(x)} a_i^T \pi (\tilde{x}_i^+ - \tilde{x}_i^-) - \sum_{i \in S^c(x)} a_i^T \pi (x_i^+ - x_i^-)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i \in S(x) \cup S^c(x)} a_i^T \pi(\tilde{x}_i^+ - \tilde{x}_i^-) - \sum_{i \in S(x) \cup S^c(x)} a_i^T \pi(x_i^+ - x_i^-) \\
&= \pi^T [A(\tilde{x}^+ - \tilde{x}^-) - b] \geq 0.
\end{aligned}
$$

It follows that $(x^+; x^-)$ is a solution of $\mathrm{VI}(\tilde{X}(x), \nabla_x \tilde{f})$. By convexity of f (thus $\tilde{f}$) and $\tilde{X}(x)$, $(x^+; x^-)$ is a global minimizer of (20). Since $\xi_i = 1$ for $i \in S(x)$, by the separability of the objective and the structure of the constraint sets, it follows that $(x^+; x^-; \xi)$ is a minimizer of the tightened (4) as follow:

$$
\min \ f(\tilde{x}_1 - \tilde{x}_2) + \gamma e^T(e - \tilde{x}_3) \quad \text{subject to } (\tilde{x}_1; \tilde{x}_2; \tilde{x}_3) \in X_{\text{tight}}(x),
$$

where

$$
X_{\text{tight}}(x) \triangleq \left\{ \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{pmatrix} : \begin{array}{rl} \tilde{x}_1, \tilde{x}_2 \geq 0, \ 0 \leq \tilde{x}_3 \leq e, & A(\tilde{x}_1 - \tilde{x}_2) \geq b, \\ (\tilde{x}_1)_i = (\tilde{x}_2)_i = 0, & \forall i \in S(x), \\ (\tilde{x}_3)_i = 0, & \forall i \in S^c(x) \end{array} \right\}.
$$

If X denotes the feasible region in (4), then we can take a sufficiently small neighborhood of x, denoted by $\mathcal{N}(x)$, such that $X \cap \mathcal{N}(x) = X_{\text{tight}}(x) \cap \mathcal{N}(x)$. Since $x = (x^+; x^-; \xi)$ is a global minimizer of $f(\tilde{x}_1 - \tilde{x}_2) + \gamma e^T(e - \tilde{x}_3)$ over $X_{\text{tight}}(x)$, it is a global minimizer of $f(\tilde{x}_1 - \tilde{x}_2) + \gamma e^T(e - \tilde{x}_3)$ over the smaller set $\mathcal{N}(x) \cap X_{\text{tight}}(x)$. Since $\mathcal{N}(x) \cap X_{\text{tight}}(x) = \mathcal{N}(x) \cap X$, it follows that x is a local minimizer of (4). ∎

Remark 2 Note that while convexity of f is observed for many loss functions, it does not guarantee the overall convexity of the problem and (4) is still a nonconvex problem.

3 Tractable ADMM frameworks

In this section we discuss how to use ADMM to efficiently address MPCC (4). In Section 3.1, we present a perturbed proximal ADMM framework for obtaining a suitably defined solution of (4) and show in Section 3.2 that both of the ADMM subproblems can be solved tractably, of which, one can be recast as a convex program, while the other can be resolved in closed form. In Section 3.3, a basic ADMM framework will be presented, along with a discussion regarding why we consider its perturbed proximal variant. A standard ADMM applied to an alternative formulation of (4) is introduced in Section 3.4. Note that ADMM applied to this formulation is easier to analyze but does have computational disadvantages arising from the intractability of the subproblem.

3.1 A perturbed proximal ADMM framework

We may reformulate (4) as follows.

$$
\min \quad f(x^+ - x^-) + \gamma \sum_{i=1}^{n} (1 - \xi_i) + \mathbb{1}_{Z_1}(w) + \mathbb{1}_{Z_2}(w). \tag{21}
$$

Recall that $f(x) = f_Q(x) + g(x)$, where $f_Q(x) \triangleq x^T Mx + d^T x$, $g(x)$ is convex and smooth, $M \in \mathbb{R}^{n\times n}$ is a symmetric matrix, and $d \in \mathbb{R}^n$. Let Z_1, Z_2, and w be defined as

$$\begin{aligned} Z_1 &\triangleq \left\{(x^+; x^-; \xi) : (x^+ + x^-)^T \xi = 0\right\}, \\ Z_2 &\triangleq \left\{ \begin{pmatrix} x^+ \\ x^- \\ \xi \end{pmatrix} : \begin{array}{l} 0 \le \xi_i \le 1, \forall i \\ 0 \le x^+, x^- \\ b \le A(x^+ - x^-) \end{array} \right\}, \end{aligned} \tag{22}$$

and $w \triangleq \left(x^+; x^-; \xi\right)$, respectively. We introduce separability into the objective by adding a variable $y \triangleq (y^+; y^-; \zeta), y^+, y^-, \zeta \in \mathbb{R}^n$ and imposing an additional linear constraint.

$$\min_{w=y} \; f_Q(x^+ - x^-) + \gamma \sum_{i=1}^n (1 - \xi_i) + \mathbb{1}_{Z_1}(w) + g(y^+ - y^-) + \mathbb{1}_{Z_2}(y). \tag{23}$$

Note that (23) is in the form of (3). The intuition behind this formulation is that by separating the nonconvex set Z_1 from the convex polytope Z_2, we may potentially obtain easier subproblems when applying a splitting method. We now define a perturbed augmented Lagrangian function as follows.

$$\begin{aligned} \tilde{\mathcal{L}}_{\rho,\alpha}(w, y, \lambda) \triangleq{}& f_Q(x^+ - x^-) + \gamma \sum_{i=1}^n (1 - \xi_i) + g(y^+ - y^-) \\ &+ (1 - \rho\alpha)\lambda^T (w - y - \alpha\lambda) + \frac{\rho}{2}\|w - y\|^2, \end{aligned}$$

where $w \triangleq (x^+; x^-; \xi), \alpha > 0, \rho > 0$. The perturbed proximal ADMM algorithm is presented as Algorithm 1, denoted as $\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$, where “cf” stands for “complementarity formulation”, and μ, α, ρ are algorithm parameters. The perturbation technique is inspired by Hajinezhad and Hong [23]. Note that ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$) reduces to a basic ADMM when $\mu = \alpha = 0$, which will be discussed in Section 3.3. We refer the readers to Remark 4 and Remark 6 for discussion of the stopping criteria.

Algorithm 1 A perturbed proximal ADMM scheme: $\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$

(0) Given w_0, y_0, λ_0; Choose $\alpha, \rho, \mu, \epsilon_0 > 0$ such that $\rho\alpha \in (0, 1), (\rho + \mu)I + 4M \succ 0$, and set $k := 0$.

(1) Let $w_{k+1}, y_{k+1}, \lambda_{k+1}$ be given by the following:

$$w_{k+1} := \arg\min_{w\in Z_1} \quad \tilde{\mathcal{L}}_{\rho,\alpha}(w, y_k, \lambda_k) + \frac{\mu}{2}\|w - w_k\|^2, \tag{Update-1}$$

$$y_{k+1} := \arg\min_{y\in Z_2} \quad \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y, \lambda_k), \tag{Update-2}$$

$$\lambda_{k+1} := (1 - \rho\alpha)\lambda_k + \rho\,(w_{k+1} - y_{k+1}). \tag{Update-3}$$

(3) If $\max\{\|\rho(y_{k+1} - y_k) + \mu(w_{k+1} - w_k)\|, \|\lambda_{k+1} - \lambda_k\|/\rho\} < \epsilon_0$, STOP; else $k := k + 1$ and return to (1).

We observe that there are indeed some benefits by considering decomposition (23) that separates the nonconvex domain Z_1 and the convex polytope Z_2. It turns out that this approach reduces the difficulty of both subproblems of the ADMM framework. Next, (Update-1) and (Update-2) are shown to be tractable[1].

3.2 Tractable resolution of ADMM Updates

We now show that (Update-1) possesses a hidden convexity property [4], allowing for claiming tractability of (Update-1) and obtaining its closed form solution.

Proposition 1 (Tractability of Update-1) Recall that $f_Q(x) = x^T Mx + d^T x$ where M may be a symmetric indefinite matrix with real eigenvalues given by $s_1, \ldots, s_n$. Consider (Update-1) in scheme ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$) at iteration $k+1$. Then $(\rho+\mu)I + 4M \succ 0$ implies $\rho + \mu + 4s_i > 0$, $\forall i = 1, \ldots, n$, and the following hold:
(i) The solution $w_{k+1} \triangleq (x^+_{k+1}; x^-_{k+1}; \xi_{k+1})$ can be obtained as a solution to a tractable convex program.
(ii) The solution w_{k+1} is available in closed form. In particular, let $h \triangleq (d; -d; -\gamma e) + (1-\rho\alpha)\lambda_k - \rho y_k - \mu w_k$, and let V be an orthogonal matrix such that $V^T MV = \mathrm{diag}(s_1, \ldots, s_n) \triangleq S$, and let

$$G \triangleq \begin{pmatrix} \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ -\frac{\sqrt{2}}{2}I_n & & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix}. \tag{24}$$

Also let $q \triangleq G^T h \triangleq (q_1; q_2; q_3)$, $z \triangleq \big(z_1; z_2; z_3\big)$, $q_1, q_2, q_3, z_1, z_2, z_3 \in \mathbb{R}^n$, and

$$z_1 \triangleq \begin{cases} \frac{-(\|q_1\|+\|q_3\|)q_1}{2(\rho+\mu)\|q_1\|}, & \|q_1\| > 0 \\ \frac{\|q_3\|}{2(\rho+\mu)}u, \|u\| = 1, & \|q_1\| = 0 \end{cases}, \quad z_3 \triangleq \begin{cases} \frac{-(\|q_1\|+\|q_3\|)q_3}{2(\rho+\mu)\|q_3\|}, & \|q_3\| > 0 \\ \frac{\|q_1\|}{2(\rho+\mu)}v, \|v\| = 1, & \|q_3\| = 0 \end{cases}$$

$$(z_2)_i \triangleq -(q_2)_i/(\rho+\mu+4s_i), \forall i = 1, \ldots, n.$$

Then $w_{k+1} \triangleq Gz$ is the solution to (Update-1).

Proof **(i).** The first subproblem in ($\mathrm{ADMM}_{\mathrm{cf}}$) is equivalent to the following:

$$\min_{w\in Z_1} \ \tilde{\mathcal{L}}_{\rho,\alpha}(w, y_k, \lambda_k) + \frac{\mu}{2}\|w - w_k\|^2 \tag{25}$$

$$\equiv \min_{(x^+ + x^-)^T\xi = 0} \left\{ f_Q(x^+ - x^-) + \gamma \sum_{i=1}^n (1-\xi_i) + (1-\rho\alpha)\lambda_k^T w + \frac{\rho}{2}\|w - y_k\|^2 \right.$$

[1] By saying that an optimization problem is tractable we mean that it either has a closed-form solution or lies in the range of convex programs that are polynomially solvable. We refer the readers to [3] for detailed discussion.

$$
+\frac{\mu}{2}\|w-w_k\|^2\Big\}
$$

$$
\equiv \min_{w^T\tilde{Q}w=0} \left\{w^T H w + h^T w\right\}, \tag{26}
$$

where $H \triangleq \begin{pmatrix} M+\frac{\rho+\mu}{2}I & -M & \\ -M & M+\frac{\rho+\mu}{2}I & \\ & & \frac{\rho+\mu}{2}I \end{pmatrix}, \tilde{Q} \triangleq \begin{pmatrix} & & I \\ & & I \\ I & I & \end{pmatrix}$. In fact, $H, \tilde{Q}$ can be simultaneously orthogonally diagonalized by using G defined in (24) (See Lemma 9 in Appendix for the linear algebra). Therefore, by leveraging the hidden convexity (See discussion in Section 7.1 in Appendix), a global solution to this nonconvex QCQP (26) can be obtained by solving a tractable convex program (In [4], it is described how a polynomial time interior point method can be applied to solve this convex program).

(ii). By substituting $z = G^T w, z \triangleq (z_1; z_2; z_3), z_1, z_2, z_3 \in \mathbb{R}^n$, (25) is equivalent to a simple QCQP,

$$
\begin{aligned}
\min \quad & \frac{\rho+\mu}{2}\|z_1\|^2 + \sum_{i=1}^n \left(\frac{\rho+\mu}{2} + 2s_i\right)(z_2)_i^2 + \frac{\rho+\mu}{2}\|z_3\|^2 + q^T z \\
\text{subject to} \quad & \|z_1\|_2 = \|z_3\|_2.
\end{aligned} \tag{27}
$$

Again, this is a result by leveraging Lemma 9. To obtain an optimal solution of (27), we require that the objective value is bounded below. By completing squares, a sufficient condition for boundedness of (27) is $\frac{\rho+\mu}{2} + 2s_i > 0, \forall i = 1,\ldots,n$ because z_2 is unconstrained. This is implied by the condition $(\rho+\mu)I_n + 4M \succ 0$. The result of (ii) follows by noting that **all** optimal solutions $(z_1^*; z_2^*; z_3^*)$ of (27) can be characterized as follows:

$$
z_1^* = \begin{cases} \frac{-(\|q_1\|+\|q_3\|)q_1}{2(\rho+\mu)\|q_1\|}, & \|q_1\|>0 \\ \frac{\|q_3\|}{2(\rho+\mu)}u, \|u\|=1, & \|q_1\|=0 \end{cases}, \quad z_3^* = \begin{cases} \frac{-(\|q_1\|+\|q_3\|)q_3}{2(\rho+\mu)\|q_3\|}, & \|q_3\|>0 \\ \frac{\|q_1\|}{2(\rho+\mu)}v, \|v\|=1, & \|q_3\|=0 \end{cases},
$$

$$
(z_2^*)_i = -(q_2)_i/(\rho+\mu+4s_i), \quad \text{for } i=1,\ldots,n. \tag{28}
$$

Next we show that this is true. Note that $(z_2^*)_i = -(q_2)_i/(\rho+\mu+4s_i), \forall i$, because z_2 is unconstrained. Since the problem is separable with respect to z_2, it may be removed, leading to the problem of

$$
\min_{z_1,z_3} \frac{\rho+\mu}{2}\|z_1\|^2 + \frac{\rho+\mu}{2}\|z_3\|^2 + q_1^T z_1 + q_3^T z_3 \quad \text{subject to } \|z_1\|^2 - \|z_3\|^2 = 0.
$$

Since z_1 and z_3 have the same magnitude, let $z_1 \triangleq r d_1$ and $z_3 \triangleq r d_3$, where $\|d_1\| = \|d_3\| = 1$. Then the constraint may be removed and the problem is further simplified as

$$
\min_{r,d_1,d_3} (\rho+\mu)r^2 + r q_1^T d_1 + r q_3^T d_3 \quad \text{subject to } r \geq 0, \|d_1\|=1, \|d_3\|=1.
$$

It follows that $r^* = \operatorname{argmin}_{r\geq 0}\{(\rho+\mu)r^2 - (\|q_1\|+\|q_3\|)r\} = (\|q_1\|+\|q_3\|)/(2(\rho+\mu))$. This leads to concluding that if $\|q_1\| > 0, \|q_3\| > 0$, $z_1^* = -(\|q_1\| +$

$\|q_3\|)q_1/(2(\rho+\mu)\|q_1\|), z_3^* = -(\|q_1\|+\|q_3\|)q_3/(2(\rho+\mu)\|q_3\|)$. If $\|q_1\| = 0, \|q_3\| > 0$, then $\|z_1^*\| = \|q_3\|/(2(\rho+\mu))$ and $z_3^* = -q_3/(2(\rho+\mu))$ and z_1^* can take any direction. If $\|q_3\| = 0, \|q_1\| > 0$, then $\|z_3^*\| = \|q_1\|/(2(\rho+\mu))$ and $z_1^* = -q_1/(2(\rho+\mu))$ and z_3^* can take any direction. If $\|q_1\| = \|q_3\| = 0$, then $z_1^* = z_3^* = 0$. ∎

Remark 3 Note that in order to compute the closed form solution, we do need eigendecomposition $M = VSV^{-1}$. However, this only needs to be done once instead of in every iteration.

Next, (Update-2) is shown to be tractable and its solution may be available in closed-form.

Proposition 2 (Tractability of Update-2) Consider (Update-2) at iteration $k+1$ in ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu;\alpha,\rho}$). Then the following hold: (i) (Update-2) is a convex program and can be computed tractably. (ii) If $g(x) \equiv 0$ and the constraints $Ax \geq b$ are absent, (Update-2) reduces to

$$y_{k+1} = \begin{pmatrix} y_{k+1,1} \\ y_{k+1,2} \\ y_{k+1,3} \end{pmatrix}, \quad \left.\begin{array}{l} (y_{k+1,1})_i := \max\{(x_{k+1}^+)_i + (1-\rho\alpha)(\lambda_{k,1})_i/\rho, 0\} \\ (y_{k+1,2})_i := \max\{(x_{k+1}^-)_i + (1-\rho\alpha)(\lambda_{k,2})_i/\rho, 0\} \\ (y_{k+1,3})_i := \Pi_{[0,1]}((\xi_{k+1})_i + (1-\rho\alpha)(\lambda_{k,3})_i/\rho) \end{array}\right\} \forall i, \tag{29}$$

where $\lambda_k \triangleq (\lambda_{k,1};\lambda_{k,2};\lambda_{k,3})$, $y_{k+1,1}, y_{k+1,2}, y_{k+1,3}, \lambda_{k,1}, \lambda_{k,2}, \lambda_{k,3} \in \mathbb{R}^n$, and $\Pi_Z(z)$ denotes the projection of z onto set Z.

Proof **(i).** (Update-2) can be cast as a linearly constrained convex smooth program: $\min_{y\in Z_2}\left\{g(y^+-y^-) - (1-\rho\alpha)\lambda_k^T y + \frac{\rho}{2}\|(x_{k+1}^+; x_{k+1}^-; \xi_{k+1}) - y\|^2\right\}$, which may be tractably resolved [3].
(ii). When $g \equiv 0$ and the constraints $Ax \geq b$ are absent, then (Update-2) can be viewed as a projection of $(x_{k+1}^+; x_{k+1}^-; \xi_{k+1}) + (1-\rho\alpha)\lambda_k/\rho$ onto a Cartesian set: $\hat{Z}_2 \triangleq \{(y_1;y_2;y_3) \mid y_1, y_2, y_3 \in \mathbb{R}^n, y_1 \geq 0, y_2 \geq 0, 0 \leq (y_3)_i \leq 1, \forall i\}$. Consequently, the projection onto this set reduces to the update given by (29). ∎

3.3 A basic ADMM framework with tractable subproblems

In Algorithm 1, a perturbation parameter and a proximal term are introduced. In fact, as we see in Section 4, an explicit bound can be derived for the multiplier sequence generated by Algorithm 1. Therefore, we may show subsequential convergence and estimate a norm of the limit point. In this subsection, we present the vanilla ADMM framework (denoted by $\mathrm{ADMM}_{\mathrm{cf}}$ and defined in Algorithm 2) applied to the tractable decomposition (23). In ($\mathrm{ADMM}_{\mathrm{cf}}$), the augmented Lagrangian function $\mathcal{L}_\rho$ is defined as follows.

$$\mathcal{L}_\rho(x^+, x^-, \xi, y, \lambda) \triangleq f_Q(x^+ - x^-) + \gamma\sum_{i=1}^n (1-\xi_i) + g(y^+ - y^-)$$

Algorithm 2 $\mathrm{ADMM_{cf}}$

(0) Given y_0, λ_0, $\epsilon > 0$, $k := 0$; Choose ρ_0, s.t. $\rho_0 I_n + 4M \succ 0$;
(1) Let $x^+_{k+1}, x^-_{k+1}, \xi_{k+1}, y_{k+1}, \lambda_{k+1}$ be given by the following:

$$(x^+_{k+1}; x^-_{k+1}; \xi_{k+1}) \in \underset{(x^{\pm},\xi)}{\operatorname{argmin}} \quad \mathcal{L}_{\rho_k}(x^+, x^-, \xi, y_k, \lambda_k), \tag{Update-1}$$

$$y_{k+1} := \underset{y}{\operatorname{argmin}} \quad \mathcal{L}_{\rho_k}(x^+_{k+1}, x^-_{k+1}, \xi_{k+1}, y, \lambda_k), \tag{Update-2}$$

$$\lambda_{k+1} := \lambda_k + \rho_k \left((x^+_{k+1}; x^-_{k+1}; \xi_{k+1}) - y_{k+1} \right). \tag{Update-3}$$

(2) Update ρ_k and let $\rho_{k+1} \leftarrow \rho_k$;
(3) If $\max\{\|(x^+_{k+1}; x^-_{k+1}; \xi_{k+1}) - y_{k+1}\|, \rho_k \|y_{k+1} - y_k\|\} < \epsilon$, STOP; else $k := k+1$, return to (1).

$$+ \lambda^T (w - y) + \frac{\rho}{2}\|w - y\|^2 + \mathbb{1}_{Z_1}((x^+; x^-; \xi)) + \mathbb{1}_{Z_2}(y).$$

We will specify in Section 5 the update rule for ρ_k in Step 2. Note that if we let $\mu = 0$, $\alpha = 0$ and replace ρ by ρ_k, then $\mathrm{ADMM}^{\mu,\alpha,\rho}_{\mathrm{cf}}$ reduces to ($\mathrm{ADMM_{cf}}$). The similarity between these two algorithms allows for ($\mathrm{ADMM_{cf}}$) to maintain the property of tractability of the subproblems (A special case of proposition 1 and 2 when $\alpha = 0$, $\mu = 0$). However, convergence analysis of ($\mathrm{ADMM_{cf}}$) is by no means straightforward. Since (23) is in the form of (3), we have discussed in Section 1 that no existing convergence theory for ADMM schemes in nonconvex regimes is applicable. It turns out that even to show boundedness of the multiplier sequence is challenging. On the other hand, if we assume boundedness of a subsequence of the multiplier together with other assumptions, subsequential convergence of the algorithm may be obtained. Convergence properties can be further enhanced given the KŁ property. Details of these discussions are kept in Section 7.4 in the Appendix while an investigation of the numerical behavior of this algorithm is presented in Section 5.

3.4 A standard ADMM framework on an alternative formulation

In section 3.1, we consider reformulation (23) of (4), which allows for efficient resolution of the subproblem; Note that an alternative formulation of (4) exists as specified next.

$$\min \ \mathbb{1}_Z(w) + f(y^+ - y^-) + \gamma e^T(e - \zeta) \quad \text{subject to } w - y = 0, \tag{30}$$

where $w = (x^+; x^-; \xi), y = (y^+; y^-; \zeta), Z \triangleq Z_1 \cap Z_2$, Z_1 and Z_2 are defined as in (22). Note that in (30), the y block is unconstrained and has a smooth objective function. Such a reformulation of optimization over complementarity constraints is considered in [33] and an ADMM scheme can be applied (See Algorithm 3 below). We referred to this framework as a standard ADMM framework (or ($\mathrm{ADMM_0}$)), since this type of ADMM scheme is favored and most studied in literature due to clear convergence guarantee. As indicated

Algorithm 3 A standard ADMM framework: ADMM_0

(0) Given $y_0, \lambda_0, \rho_0 > 0, \epsilon > 0$, set $k := 0$.
(1) Let $w_{k+1}, y_{k+1}, \lambda_{k+1}$ be given by the following:

$$w_{k+1} \in \underset{w \in Z}{\operatorname{argmin}} \quad \|w - y_k + \lambda_k/\rho_k\|^2, \tag{Update-1}$$

$$y_{k+1} := \underset{y}{\operatorname{argmin}} \quad f(y^+ - y^-) + \gamma e^T(e - \zeta) + \frac{\rho_k}{2}\|y - w_{k+1} - \lambda_k/\rho_k\|^2, \tag{Update-2}$$

$$\lambda_{k+1} := \lambda_k + \rho_k(w_{k+1} - y_{k+1}). \tag{Update-3}$$

(2) Update ρ_k and let $\rho_{k+1} \leftarrow \rho_k$;
(3) If $\max(\|w_{k+1} - y_{k+1}\|, \rho_k\|y_{k+1} - y_k\|) < \epsilon$, stop; else $k := k+1$ and return to (1).

in [33, Corollary 3], if $\rho_k \equiv \rho$ is large enough and the augmented Lagrangian function has the KŁ property, (ADMM_0) generates a sequence convergent to a stationary point. However, such a framework is potentially slow because (Update-1) requires globally resolving an MPCC and may render the scheme impractical. We will further explain with numerical experiments in Section 5.

4 Convergence Analysis

In the prior section, we consider the formulation (23) to resolve (4) and present a perturbed proximal ADMM framework ($\text{ADMM}_{\text{cf}}^{\mu,\alpha,\rho}$) reliant on tractable updates at each iteration. In this section, we analyze the convergence property of this framework. Specifically, we show that under mild assumptions, the sequence $\{\lambda_k\}$ is bounded and a subsequence of $\{(w_k, y_k, \lambda_k)\}_{k\geq 1}$ converges to a perturbed KKT point of (23). The main results are Theorem 2, Corollary 1 and Theorem 3. First we present some definitions used in this section. We refer interested readers to [1] and [29] for more details.

Definition 3 ((Limiting) subdifferential and critical point) Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function and let $\bar{\partial} F(x)$ denote the Fréchet subdifferential of F at x, i.e.,

$$\bar{\partial} F(x) \triangleq \left\{ d : \liminf_{z \neq x, z \to x} \frac{1}{\|z - x\|}[F(z) - F(x) - d^T(z - x)] \geq 0 \right\},$$

for $x \in \text{dom}F = \{x : F(x) < +\infty\}$ and $\bar{\partial} F(x) = \emptyset$ if $x \notin \text{dom}F$. Then,

(i). The (limiting) subdifferential of F at $x \in \text{dom}F$, is defined as follows.

$$\partial F(x) \triangleq$$
$$\{d \in \mathbb{R}^n : \exists \{x_k\}_{k\geq 1}, \text{ s.t. } x_k \to x,\ F(x_k) \to F(x),\ d_k \to d,\ d_k \in \bar{\partial} F(x_k)\}.$$

(ii). x is a critical point of F if and only if $0 \in \partial F(x)$.

Definition 4 ((Limiting) normal cone) Suppose Z is a nonempty closed subset of $\mathbb{R}^n$, and $\bar{N}_Z(x)$ denotes the Fréchet normal cone to Z at x, so

$$\bar{N}_Z(x) = \{v \in \mathbb{R}^n : v^T(z-x) \le o(x-z), \forall z \in Z\},$$

if $x \in Z$ and $\bar{N}_Z(x) = \emptyset$ if $x \notin Z$. Then the (limiting) normal cone to Z at $x \in Z$, denoted as $N_Z(x)$, is defined as follows.

$$N_Z(x) \triangleq \{v \in \mathbb{R}^n : \exists \{x_k\}_{k\ge1}, \text{ s.t. } x_k \to x,\ x_k \in Z,\ v_k \to v,\ v_k \in \bar{N}_Z(x_k)\}.$$

These concepts have the following properties:

(i). (Closedness of ∂F) If $d_k \to d$, $x_k \to x$ and $d_k \in \partial F(x_k)$, $F(x_k) \to F(x)$, then $d \in \partial F(x)$.
(ii). Let $F = F_0 + F_1$. If F_0 is finite at x and F_1 is smooth in a neighborhood of x, then $\partial F = \partial F_0 + \nabla F_1$.
(iii). If $x^* \in \operatorname{argmin} F(x)$, then x^* is a critical point of F.
(iv). $\partial \mathbb{1}_Z(z) = N_Z(z), \forall z \in Z$.

To simplify the notation, we rewrite (23) as the following structured program:

$$\min_{w\in Z_1, y\in Z_2} \ h(w) + p(y) \quad \text{subject to } w - y = 0, \tag{31}$$

where Z_1, Z_2 are defined by (22), $w \triangleq (x^+; x^-; \xi), y \triangleq (y^+; y^-; \zeta), h(w) \triangleq f_Q(x^+ - x^-) + \gamma \sum_{i=1}^n (1-\xi_i)$, $f_Q(x^+ - x^-) \triangleq (x^+ - x^-)M(x^+ - x^-) + d^T(x^+ - x^-)$, $p(y) \triangleq g(y^+ - y^-)$. In addition, the perturbed augmented Lagrangian function is rewritten as follows.

$$\tilde{\mathcal{L}}_{\rho,\alpha}(w, y, \lambda) \triangleq h(w) + p(y) + (1-\rho\alpha)\lambda^T(w - y - \alpha\lambda) + \frac{\rho}{2}\|w - y\|^2.$$

Let $r_k \triangleq w_k - y_k$ and $\Delta\lambda_{k+1} \triangleq \lambda_{k+1} - \lambda_k$ for all $k \ge 0$. We define a Lyapunov function P_τ^k for any $\tau > 0$ and $k \ge 1$.

$$P_\tau^k \triangleq \tilde{\mathcal{L}}_{\rho,\alpha}(w_k, y_k, \lambda_k) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_k\|^2 + \tau\left(\frac{1-\rho\alpha}{2\rho}\right)\|\lambda_k - \lambda_{k-1}\|^2. \tag{32}$$

We intend to show that the sequence $\{P_\tau^k\}_{k\ge1}$ is nonincreasing and the following two lemmas are needed.

Lemma 4 Consider the sequence $\{w_k, y_k, \lambda_k\}$ generated by (ADMM$_{cf}^{\mu,\alpha,\rho}$). Then the following holds for any $\nu > 0$, and any $k \ge 1$,

$$\begin{aligned}&\frac{1-\rho\alpha}{2\rho}\left(\|\lambda_{k+1} - \lambda_k\|^2 - \|\lambda_k - \lambda_{k-1}\|^2\right) \\ &\le -\left(\alpha - \frac{\nu}{2}\right)\|\lambda_{k+1} - \lambda_k\|^2 + \frac{1}{2\nu}\|w_{k+1} - w_k\|^2. \end{aligned}\tag{33}$$

Proof Let $G_{k+1} \triangleq \nabla_y p(y_{k+1})$. By (Update-2), for all $y \in Z_2$ and $k \geq 0$,

$$\begin{aligned} 0 &\geq \left(G_{k+1} - (1-\rho\alpha)\lambda_k - \rho r_{k+1}\right)^T (y_{k+1} - y) \\ &= \left(G_{k+1} - \lambda_{k+1}\right)^T (y_{k+1} - y). \end{aligned} \tag{34}$$

Consequently, we have that $\forall k \geq 1$,

$$\left(G_k - \lambda_k\right)^T (y_k - y) \leq 0, \qquad \forall y \in Z_2. \tag{35}$$

By choosing $y = y_k$ in (34), $y = y_{k+1}$ in (35), then adding (34) and (35), we have that for any $k \geq 1$,

$$\begin{aligned} & (G_{k+1} - G_k - \lambda_{k+1} + \lambda_k)^T (y_{k+1} - y_k) \leq 0, \\ \Longrightarrow\ & \left(G_{k+1} - G_k\right)^T (y_{k+1} - y_k) - (\lambda_{k+1} - \lambda_k)^T (y_{k+1} - y_k) \leq 0, \\ \Longrightarrow\ & -(\lambda_{k+1} - \lambda_k)^T (y_{k+1} - y_k) \leq 0, \end{aligned} \tag{36}$$

where the last step follows from convexity of $p(y)$. Recall $\Delta\lambda_k \triangleq \lambda_k - \lambda_{k-1}, \forall k \geq 1$. Then by adding $\Delta\lambda_{k+1}^T(w_{k+1} - w_k)$ on both sides, (36) can be rewritten as follows for $\forall k \geq 1$,

$$\begin{aligned} & \Delta\lambda_{k+1}^T (w_{k+1} - w_k) \\ & \overset{(36)}{\geq} \Delta\lambda_{k+1}^T (w_{k+1} - y_{k+1} - w_k + y_k) \\ & = \Delta\lambda_{k+1}^T (r_{k+1} - r_k) \\ & = \Delta\lambda_{k+1}^T (r_{k+1} - \alpha\lambda_k - r_k + \alpha\lambda_{k-1}) + \Delta\lambda_{k+1}^T (\alpha\lambda_k - \alpha\lambda_{k-1}) \\ & = \Delta\lambda_{k+1}^T \left(\frac{\Delta\lambda_{k+1}}{\rho} - \frac{\Delta\lambda_k}{\rho} \right) + \alpha\Delta\lambda_{k+1}^T \Delta\lambda_k \\ & = \frac{1-\rho\alpha}{\rho} \Delta\lambda_{k+1}^T (\Delta\lambda_{k+1} - \Delta\lambda_k) + \alpha\|\Delta\lambda_{k+1}\|^2. \end{aligned} \tag{37}$$

Note that $\Delta\lambda_{k+1}^T(\Delta\lambda_{k+1} - \Delta\lambda_k) = \frac{1}{2}(\|\Delta\lambda_{k+1}\|^2 - \|\Delta\lambda_k\|^2 + \|\Delta\lambda_{k+1} - \Delta\lambda_k\|^2)$. Let $\Delta w_{k+1} \triangleq w_{k+1} - w_k$. Then by (37),

$$\begin{aligned} & \frac{1-\rho\alpha}{2\rho} \cdot (\|\Delta\lambda_{k+1}\|^2 - \|\Delta\lambda_k\|^2 + \|\Delta\lambda_{k+1} - \Delta\lambda_k\|^2) + \alpha\|\Delta\lambda_{k+1}\|^2 \\ & \leq \Delta\lambda_{k+1}^T \Delta w_{k+1} \\ \Longrightarrow\ & \frac{1-\rho\alpha}{2\rho} \cdot (\|\Delta\lambda_{k+1}\|^2 - \|\Delta\lambda_k\|^2) \\ & \leq -\alpha\|\Delta\lambda_{k+1}\|^2 + \Delta\lambda_{k+1}^T \Delta w_{k+1} \\ & \leq -\alpha\|\Delta\lambda_{k+1}\|^2 + \frac{\nu\|\Delta\lambda_{k+1}\|^2}{2} + \frac{\|\Delta w_{k+1}\|^2}{2\nu} \\ & = \frac{\|\Delta w_{k+1}\|^2}{2\nu} - \left(\alpha - \frac{\nu}{2}\right) \|\Delta\lambda_{k+1}\|^2. \end{aligned}$$

Then the proof is complete. ∎

Lemma 5 *Consider* $\{w_k, y_k, \lambda_k\}$ *generated by* (ADMM$_{\text{cf}}^{\mu,\alpha,\rho}$). *Then*

$$\left(\tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_{k+1}) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_{k+1}\|^2\right)$$
$$-\left(\tilde{\mathcal{L}}_{\rho,\alpha}(w_k, y_k, \lambda_k) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_k\|^2\right)$$
$$\leq -\frac{\mu}{2}\|w_{k+1}-w_k\|^2 - \frac{\rho}{2}\|y_{k+1}-y_k\|^2 + \frac{(1-\rho\alpha)(2-\rho\alpha)}{2\rho}\|\Delta\lambda_{k+1}\|^2. \quad (38)$$

Proof From (Update-1),

$$\tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_k, \lambda_k) + \frac{\mu}{2}\|w_{k+1}-w_k\|^2 - \tilde{\mathcal{L}}_{\rho,\alpha}(w_k, y_k, \lambda_k) \leq 0$$
$$\Longrightarrow \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_k, \lambda_k) - \tilde{\mathcal{L}}_{\rho,\alpha}(w_k, y_k, \lambda_k) \leq -\frac{\mu}{2}\|w_{k+1}-w_k\|^2. \quad (39)$$

Also, by the optimality condition of (Update-2), if $\tilde{G}_{k+1} \triangleq \nabla_y \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_k)$, then $\tilde{G}_{k+1}^T(y - y_{k+1}) \geq 0, \forall y \in Z_2$. Using this fact and the strong convexity of $\tilde{\mathcal{L}}_{\rho,\alpha}$ in terms of y with constant ρ,

$$\tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_k) - \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_k, \lambda_k)$$
$$\leq -\tilde{G}_{k+1}^T(y_k - y_{k+1}) - \frac{\rho}{2}\|y_{k+1}-y_k\|^2 \leq -\frac{\rho}{2}\|y_{k+1}-y_k\|^2. \quad (40)$$

The fact that $\Delta\lambda_{k+1} = \rho r_{k+1} - \rho\alpha\lambda_k$ and $\lambda_{k+1}^T\Delta\lambda_{k+1} = \frac{1}{2}(\|\lambda_{k+1}\|^2 - \|\lambda_k\|^2 + \|\Delta\lambda_{k+1}\|^2)$ imply:

$$\begin{aligned}
&\tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_{k+1}) - \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_k)\\
&= (1-\rho\alpha)\lambda_{k+1}^T(r_{k+1} - \alpha\lambda_{k+1}) - (1-\rho\alpha)\lambda_k^T(r_{k+1} - \alpha\lambda_k)\\
&= (1-\rho\alpha)\lambda_{k+1}^T(r_{k+1} - \alpha\lambda_k - \alpha\Delta\lambda_{k+1}) - (1-\rho\alpha)\lambda_k^T(r_{k+1} - \alpha\lambda_k)\\
&= (1-\rho\alpha)(\lambda_{k+1} - \lambda_k)^T(r_{k+1} - \alpha\lambda_k) - (1-\rho\alpha)\lambda_{k+1}^T\alpha\Delta\lambda_{k+1}\\
&= \frac{1-\rho\alpha}{\rho}\|\Delta\lambda_{k+1}\|^2 - \frac{(1-\rho\alpha)\alpha}{2}(\|\lambda_{k+1}\|^2 - \|\lambda_k\|^2 + \|\Delta\lambda_{k+1}\|^2)\\
&= \frac{(1-\rho\alpha)(2-\rho\alpha)}{2\rho}\|\Delta\lambda_{k+1}\|^2 - \frac{(1-\rho\alpha)\alpha}{2}(\|\lambda_{k+1}\|^2 - \|\lambda_k\|^2). \quad (41)
\end{aligned}$$

Finally, by adding (39), (40) and (41), the following holds $\forall k \geq 0$,

$$\begin{aligned}
&\tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_{k+1}) - \tilde{\mathcal{L}}_{\rho,\alpha}(w_k, y_k, \lambda_k)\\
&= \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_{k+1}) - \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_k) + \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_{k+1}, \lambda_k)\\
&- \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_k, \lambda_k) + \tilde{\mathcal{L}}_{\rho,\alpha}(w_{k+1}, y_k, \lambda_k) - \tilde{\mathcal{L}}_{\rho,\alpha}(w_k, y_k, \lambda_k)\\
&\leq -\frac{\mu}{2}\|w_{k+1}-w_k\|^2 - \frac{\rho}{2}\|y_{k+1}-y_k\|^2\\
&+ \frac{(1-\rho\alpha)(2-\rho\alpha)}{2\rho}\|\Delta\lambda_{k+1}\|^2 - \frac{(1-\rho\alpha)\alpha}{2}(\|\lambda_{k+1}\|^2 - \|\lambda_k\|^2).
\end{aligned}$$

Then the result follows. ∎

We now impose a requirement on $h(w)+p(y)+\frac{\rho}{2}\|w-y\|^2$ and define several constants to be used later.

Assumption 1 $h(w) + p(y) + \frac{\rho}{2}\|w - y\|^2 \geq \bar{L}$ for all $w \in Z_1, y \in Z_2$.

Definition 5 Recall that α, τ, μ, ρ are nonnegative parameters from Algorithm 1 and the definition (32). Let ν and R be nonnegative constants. Then $c_1(\nu) \triangleq \frac{\mu}{2} - \frac{\tau}{2\nu}$, $c_2 \triangleq \frac{\rho}{2}$, $c_3(\nu) \triangleq \tau\left(\alpha - \frac{\nu}{2}\right) - \frac{(1-\rho\alpha)(2-\rho\alpha)}{2\rho}$,
$c_4(R) \triangleq \frac{(1-\rho\alpha)[(R+1)\rho\alpha-1]}{2\rho R}$, $c_5(R) \triangleq \frac{1-\rho\alpha}{2\rho}\left[\tau - (1-\rho\alpha)R\right]$.

Assumption 2 *$\exists \nu > 0$, $R > 0$ such that $c_1(\nu), c_3(\nu), c_4(R), c_5(R) > 0$.*

In the next Lemma, we prove that $\{P_\tau^k\}_{k\geq 1}$ is a nonincreasing sequence.

Lemma 6 Consider $\{w_k, y_k, \lambda_k\}$ generated by ($\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$). Then,
(i). $P_\tau^{k+1} - P_\tau^k \leq -c_1(\nu)\|w_{k+1} - w_k\|^2 - c_2\|y_{k+1} - y_k\|^2 - c_3(\nu)\|\lambda_{k+1} - \lambda_k\|^2$, $\forall k \geq 1$.
(ii). Suppose that Assumption 2 holds. Then $\{P_\tau^k\}_{k\geq 1}$ is non-increasing. If Assumption 1 also holds, then P_τ^k is bounded from below.
(iii). If Assumption 1 and 2 hold, then $\lim_{k\to\infty}(w_{k+1} - w_k) = \lim_{k\to\infty}(y_{k+1} - y_k) = \lim_{k\to\infty}(\lambda_{k+1} - \lambda_k) = 0$.

Proof (i). Take $\tau \times (33) + (38)$ and the result follows.
(ii). When $c_1(\nu), c_2, c_3(\nu) > 0$ for certain ν, we conclude from (i) that $P_\tau^{k+1} \leq P_\tau^k$ for all $k \geq 1$. Further,

$$
\begin{aligned}
&(1-\rho\alpha)\lambda_k^T(r_k - \alpha\lambda_k) \\
&= (1-\rho\alpha)\lambda_k^T(r_k - \alpha\lambda_{k-1} - \alpha\Delta\lambda_k) \\
&= (1-\rho\alpha)\lambda_k^T\left[\Delta\lambda_k/\rho - \alpha\Delta\lambda_k\right] \\
&= \frac{(1-\rho\alpha)^2}{\rho}\lambda_k^T(\lambda_k - \lambda_{k-1}) \\
&= \frac{(1-\rho\alpha)^2}{2\rho}(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2 + \|\lambda_k - \lambda_{k-1}\|^2) \qquad (42) \\
&\geq [(1-\rho\alpha)^2/(2\rho)](\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2), \qquad k \geq 1. \qquad (43)
\end{aligned}
$$

Then,

$$
\begin{aligned}
P_\tau^k &= h(w_k) + p(y_k) + \frac{\rho}{2}\|w_k - y_k\|^2 \\
&\quad + (1-\rho\alpha)\lambda_k^T(r_k - \alpha\lambda_k) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_k\|^2 + \tau\left(\frac{1-\rho\alpha}{2\rho}\right)\|\Delta\lambda_k\|^2 \\
&\geq \bar{L} + (1-\rho\alpha)\lambda_k^T(r_k - \alpha\lambda_k) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_k\|^2 + \tau\left(\frac{1-\rho\alpha}{2\rho}\right)\|\Delta\lambda_k\|^2 \qquad (44) \\
&\geq \bar{L} + (1-\rho\alpha)\lambda_k^T(r_k - \alpha\lambda_k) \overset{(43)}{\geq} \bar{L} + \frac{(1-\rho\alpha)^2}{2\rho}(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2).
\end{aligned}
$$

Then, $\sum_{k=1}^{K}(P_\tau^k - \bar{L}) \geq \frac{(1-\rho\alpha)^2}{2\rho}\sum_{k=1}^{K}(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2) \geq -\frac{(1-\rho\alpha)^2}{2\rho}\|\lambda_0\|^2$, $\forall K \geq 1$. Since $\{P_\tau^k - \bar{L}\}_{k\geq 1}$ is a non-increasing sequence and the above inequality holds, $\{P_\tau^k - \bar{L}\}_{k\geq 1}$ is nonnegative. Thus $\{P_\tau^k\}_{k\geq 1}$ is bounded from below. (iii). This may be concluded based on (i) and (ii). ∎

Remark 4 By Lemma 6(iii) and the stopping criterion, Algorithm $\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$ may terminate in finite time.

The following Lemma provides an inequality related to $\|\lambda_k\|$, which helps in showing boundedness of $\|\lambda_k\|$.

Lemma 7 Consider $\{w_k, y_k, \lambda_k\}$ generated by ($\mathrm{ADMM}_{cf}^{\mu,\alpha,\rho}$). Suppose Assumption 1 holds. Then $P_\tau^k \geq \bar{L} + c_4(R)\|\lambda_k\|^2 + c_5(R)\|\lambda_k - \lambda_{k-1}\|^2$ for all $R > 0$ and $k \geq 1$. Therefore, if Assumption 2 holds, $\|\lambda_k\|^2 \leq \frac{1}{c_4(R)}(P_\tau^k - \bar{L})$ for all $k \geq 1$.

Proof We may use the following result for any $R > 0$:

$$\begin{aligned}\|\lambda_{k-1}\|^2 &= \|\lambda_{k-1} - \lambda_k + \lambda_k\|^2 \\ &\leq (1+R)\|\lambda_{k-1} - \lambda_k\|^2 + (1+1/R)\|\lambda_k\|^2.\end{aligned} \tag{45}$$

From the definition of P_τ^k, we have that:

$$\begin{aligned}&P_\tau^k \\ &\overset{(44),(42)}{\geq} \bar{L} + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_k\|^2 + \tau\left(\frac{1-\rho\alpha}{2\rho}\right)\|\lambda_k - \lambda_{k-1}\|^2 \\ &+ \frac{(1-\rho\alpha)^2}{2\rho}(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2 + \|\lambda_k - \lambda_{k-1}\|^2) \\ &\overset{(45)}{\geq} \bar{L} + \frac{(1-\rho\alpha)\alpha\|\lambda_k\|^2}{2} + \tau\left(\frac{(1-\rho\alpha)\|\lambda_k - \lambda_{k-1}\|^2}{2\rho}\right) \\ &- \frac{(1-\rho\alpha)^2\left(\frac{\|\lambda_k\|^2}{R} + R\|\lambda_k - \lambda_{k-1}\|^2\right)}{2\rho} \\ &\geq \bar{L} + \frac{(1-\rho\alpha)[(R+1)\rho\alpha - 1]}{2\rho R}\|\lambda_k\|^2 + \frac{1-\rho\alpha}{2\rho}\left[\tau - (1-\rho\alpha)R\right]\|\lambda_k - \lambda_{k-1}\|^2.\end{aligned}$$

Then the result follows from definitions of $c_4(R)$ and $c_5(R)$. ∎

Boundedness of $\|\lambda_k\|$ and subsequential convergence are proved in the next theorem.

Theorem 2 Suppose $\{(w_k; y_k; \lambda_k)\}_{k\geq 0}$ is generated by ($\mathrm{ADMM}_{cf}^{\mu,\alpha,\rho}$). Assume that the sequences $\{w_k\}$ and $\{y_k\}$ are bounded. Suppose Assumptions 1 and 2 hold. Then the sequence $\{\lambda_k\}_{k\geq 1}$ is bounded and a subsequence of $\{(w_k; y_k; \lambda_k)\}$ converges to $(w^*; y^*; \lambda^*)$ such that

$$0 \in \partial(h + \mathbb{1}_{Z_1})(w^*) + \lambda^*, \quad 0 \in \partial(p + \mathbb{1}_{Z_2})(y^*) - \lambda^*, \quad w^* - y^* = \alpha\lambda^*. \tag{46}$$

Proof Since $c_1(\nu) > 0, c_3(\nu) > 0$, then by Lemma 6, $\{P_\tau^k\}$ is a non-increasing sequence. Futhermore, since $c_4(R) > 0, c_5(R) > 0$, Lemma 7 indicates that

$$\|\lambda_k\|^2 \leq \frac{1}{c_4(R)}(P_\tau^k - \bar{L}) \leq \frac{1}{c_4(R)}(P_\tau^1 - \bar{L}) < +\infty. \tag{47}$$

Therefore, the sequence $\{\lambda_k\}$ is bounded, implying that $\{(w_k; y_k; \lambda_k)\}$ is bounded. Suppose $\{(w_{n_k}; y_{n_k}; \lambda_{n_k})\}$ denotes a convergent subsequence of $\{(w_k; y_k; \lambda_k)\}$ such that $(w_{n_k}; y_{n_k}; \lambda_{n_k}) \to (w^*; y^*; \lambda^*)$ as $k \to \infty$. Based on the optimality conditions of (Update-1), (Update-2) translated using property of critical points, and the multiplier update, the following hold:

$$\begin{aligned} &0 \in \partial(h + \mathbb{1}_{Z_1})(w_{n_k}) + \lambda_{n_k} + \rho(y_{n_k} - y_{n_k-1}) + \mu(w_{n_k} - w_{n_k-1}), \\ &0 \in \partial(p + \mathbb{1}_{Z_2})(y_{n_k}) - \lambda_{n_k}, \quad w_{n_k} - y_{n_k} - \alpha\lambda_{n_k-1} = (\lambda_{n_k} - \lambda_{n_k-1})/\rho. \end{aligned} \tag{48}$$

By Lemma 6, $w_{n_k} - w_{n_k-1} \to 0$, $y_{n_k} - y_{n_k-1} \to 0$, $\lambda_{n_k} - \lambda_{n_k-1} \to 0, k \to +\infty$, so we also have $\lambda_{n_k-1} \to \lambda^*, k \to +\infty$. Therefore, by taking limits and the closedness of a subdifferential map, we may conclude the result. ∎

Remark 5 (i). Boundedness of $\{y_k\}$ and $\{w_k\}$ is a mild assumption. First, boundedness of $\{y_k\}$ can be obtained from adding constraints such as $x^+ \leq ub^+$, $x^- \leq ub^-$ to Z_2. If ub^+ and ub^- are large enough, Z_2 will still include the optimal solution. Second, since $r_k = w_k - y_k = \frac{1}{\rho}\lambda_k - \frac{1-\rho\alpha}{\rho}\lambda_{k-1}, \forall k \geq 1$ and $\{\lambda_k\}$ is bounded, r_k is also a bounded sequence. Thus, boundedness of $\{w_k\}$ is implied by boundedness of $\{y_k\}$.
(ii). It can be shown that the conditions (46) are equivalent to KKT conditions with a feasibility error (See Theorem 3 below).
(iii). Denote $\mathcal{H}_\tau(w, y, \lambda)$ as

$$\begin{aligned} &\mathcal{H}_\tau(w, y, \lambda) \\ &\triangleq \tilde{\mathcal{L}}_{\rho,\alpha}(w, y, \lambda) + \mathbb{1}_{Z_1}(w) + \mathbb{1}_{Z_2}(y) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda\|^2 + \frac{\rho\|w - y - \alpha\lambda\|^2}{2(1-\rho\alpha)/\tau}. \end{aligned}$$

Then $\mathcal{H}_\tau(w_k, y_k, \lambda_k) = P_\tau^k, \forall k \geq 1, \tau > 0$. If the assumptions in Theorem 2 hold, and in addition, $\mathcal{H}_\tau(w, y, \lambda)$ satisfies the KŁ property at (w^*, y^*, λ^*), then $\{(w_k, y_k, \lambda_k)\}$ converges to (w^*, y^*, λ^*). The proof is similar to [1, Theorem 3.1] and is omitted.
(iv). The KŁ property assumption on $\mathcal{H}_\tau$ indeed holds when $p(y)$ is semialgebraic (See Definition 7 in the Appendix). In fact, $\mathcal{H}_\tau$ is a sum of semialgebraic functions and is therefore semialgebraic. Then the result follows from the fact that a semialgebraic function satisfies the KŁ property at every point in its domain [1].
(v). Although in the context of this paper we focus on problem (23), it should be noted that Theorem 2 may be generalized. Specifically, Theorem 2 may hold if we apply $\mathrm{ADMM}_{\mathrm{cf}}^{\mu,\alpha,\rho}$ to tackle a more general class of problem:

$$\min f(x) + g(y) \quad \text{subject to } Ax + By = b, x \in X, y \in Y,$$

where g is smooth and convex, Y is convex, but f could be nonsmooth and nonconvex, X could be nonconvex, A, B, b are matrices and vectors with appropriate dimensions and do not need to be $I, -I, 0$. The analysis will basically remain the same.

Note that (46) are not the precise conditions for $(w^*; y^*; \lambda^*)$ to be a critical point of the Lagrangian $\mathcal{L}(w, y, \lambda) \triangleq h(w) + \mathbb{1}_{Z_1}(w) + p(y) + \mathbb{1}_{Z_2}(y) + \lambda^T(w - y)$, i.e. $0 \in \partial\mathcal{L}(w^*, y^*, \lambda^*)$. There exists an infeasibility error $\alpha\lambda^*$ and the following corollary discusses how to choose the parameters such that this error can be made arbitrarily small.

Corollary 1 Suppose that sequences $\{w_k\}$ and $\{y_k\}$ are bounded. In addition, assume that $\exists\rho_- > 0$ such that $h(w) + p(y) + \frac{\rho_-}{2}\|w - y\|^2 \geq l$ for all $w \in Z_1$, $y \in Z_2$. Then for any $\epsilon > 0$ such that $\epsilon \leq 1/(4\rho_- + 2)$, if the parameters in Algorithm 1 satisfy $\alpha = \epsilon, \rho = \frac{1}{2\epsilon}, \mu > \frac{2}{\epsilon}, w_0 = y_0$, then a subsequence of $\{(w_k; y_k; \lambda_k)\}_{k\geq 1}$ converges to $(w^*; y^*; \lambda^*)$ such that

$$\begin{aligned} 0 \in \partial(h + \mathbb{1}_{Z_1})(w^*) + \lambda^*, \ 0 \in \partial(p + \mathbb{1}_{Z_2})(y^*) - \lambda^*, \\ \|w^* - y^*\|^2 \leq \alpha^2\|\lambda^*\|^2 \leq (64(h(w_0) + p(y_0)) - 64l + (14 + 5\epsilon)\|\lambda_0\|^2)\epsilon. \end{aligned}$$

Proof Let $\nu = \alpha = \epsilon, R = 2$ and $\tau = 2$, then

$$\begin{aligned} c_1(\nu) &= \frac{\mu}{2} - \frac{\tau}{2\nu} = \frac{\mu}{2} - \frac{1}{\epsilon} > 0, \ c_2 = \frac{\rho}{2} = \frac{1}{4\epsilon}, \\ c_3(\nu) &= \tau\left(\alpha - \frac{\nu}{2}\right) - \frac{(1-\rho\alpha)(2-\rho\alpha)}{2\rho} = 2\left(\epsilon - \frac{\epsilon}{2}\right) - \frac{(1-\frac{1}{2})(2-\frac{1}{2})}{1/\epsilon} = \frac{\epsilon}{4}, \\ c_4(R) &= \frac{(1-\rho\alpha)[(R+1)\rho\alpha - 1]}{2\rho R} = \frac{(1-1/2)[(2+1)/2-1]}{2/\epsilon} = \frac{\epsilon}{8}, \\ c_5(R) &= \frac{1-\rho\alpha}{2\rho}\left[\tau - (1-\rho\alpha)R\right] = \frac{1-1/2}{1/\epsilon}[2 - (1-1/2)\cdot 2] = \frac{\epsilon}{2}. \end{aligned}$$

Therefore, Assumption 2 holds. Since $\rho = \frac{1}{2\epsilon} \geq 2\rho_- + 1$, we have that

$$h(w) + p(y) + \frac{\rho}{2}\|w - y\|^2 \geq h(w) + p(y) + \frac{\rho_-}{2}\|w - y\|^2 \geq l, \forall w \in Z_1, y \in Z_2.$$

Thus Assumption 1 holds. Based on Theorem 2, it suffices to show that $\alpha^2\|\lambda^*\|^2 \leq (64(h(w_0) + p(y_0)) - 64l + (14 + 5\epsilon)\|\lambda_0\|^2)\epsilon$. By (47) in Theorem 2, for $k \geq 1$,

$$\begin{aligned} \|\lambda_k\|^2 &\leq \frac{1}{c_4(R)}(P_\tau^1 - l) = \left(\frac{2}{1-\rho\alpha}\right)\left(\frac{\rho R}{(R+1)\rho\alpha - 1}\right)(P_\tau^1 - l) \\ \Longrightarrow \ \alpha^2\|\lambda_k\|^2 &\leq \left(\frac{2}{1-\rho\alpha}\right)\left(\frac{\alpha R}{R + 1 - 1/(\rho\alpha)}\right)(P_\tau^1 - l). \end{aligned} \tag{49}$$

Since $P_\tau^1 = \tilde{\mathcal{L}}_{\rho,\alpha}(w_1, y_1, \lambda_1) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_1\|^2 + \tau\left(\frac{1-\rho\alpha}{2\rho}\right)\|\lambda_1 - \lambda_0\|^2$,

$$P_\tau^1 \overset{(38)}{\leq} \tilde{\mathcal{L}}_{\rho,\alpha}(w_0, y_0, \lambda_0) + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_0\|^2 + \tau\left(\frac{1-\rho\alpha}{2\rho}\right)\|\lambda_1 - \lambda_0\|^2$$

$$
\begin{aligned}
& - \frac{\mu}{2}\|w_1 - w_0\|^2 - \frac{\rho}{2}\|y_1 - y_0\|^2 + \frac{(1-\rho\alpha)(2-\rho\alpha)}{2\rho}\|\lambda_1 - \lambda_0\|^2 \\
& \le h(w_0) + p(y_0) + \frac{\rho}{2}\|r_0\|^2 + (1-\rho\alpha)\lambda_0^T(r_0 - \alpha\lambda_0) \\
& + \frac{(1-\rho\alpha)\alpha}{2}\|\lambda_0\|^2 + \frac{(1-\rho\alpha)(2+\tau-\rho\alpha)}{2\rho}\|\lambda_1 - \lambda_0\|^2 \\
& = h(w_0) + p(y_0) + 0 - (1-1/2)\epsilon\|\lambda_0\|^2 + \frac{(1-1/2)\epsilon}{2}\|\lambda_0\|^2 \\
& + \frac{(1-1/2)(2+2-1/2)}{1/\epsilon}\|(1-1/2)\lambda_0 + \rho(w_1 - y_1) - \lambda_0\|^2 \\
& \le h(w_0) + p(y_0) - \frac{\epsilon}{4}\|\lambda_0\|^2 + \frac{7\epsilon}{4}\cdot\left(\frac{1}{2}\|\lambda_0\|^2 + 2\rho^2\|w_1 - y_1\|^2\right) \\
& \le h(w_0) + p(y_0) + \frac{5\epsilon}{8}\|\lambda_0\|^2 + \frac{7}{8\epsilon}\|w_1 - y_1\|^2.
\end{aligned} \tag{50}
$$

Adding (39) and (40) and letting $k = 0$, we obtain the following.

$$
\tilde{L}_{\rho,\alpha}(w_1, y_1, \lambda_0) - \tilde{L}_{\rho,\alpha}(w_0, y_0, \lambda_0) \le -\frac{\mu}{2}\|w_1 - w_0\|^2 - \frac{\rho}{2}\|y_1 - y_0\|^2,
$$

which indicates that

$$
\begin{aligned}
& h(w_1) + p(y_1) + (1-\rho\alpha)\lambda_0^T(w_1 - y_1 - \alpha\lambda_0) + \frac{\rho}{2}\|w_1 - y_1\|^2 \\
& \le h(w_0) + p(y_0) + (1-\rho\alpha)\lambda_0^T(w_0 - y_0 - \alpha\lambda_0) + \frac{\rho}{2}\|w_0 - y_0\|^2 \\
& = h(w_0) + p(y_0) - (1-\rho\alpha)\alpha\|\lambda_0\|^2 \\
\Longrightarrow\ & (1-\rho\alpha)\lambda_0^T(w_1 - y_1) + \frac{\rho - \rho_-}{2}\|w_1 - y_1\|^2 \\
& \le h(w_0) + p(y_0) - h(w_1) - p(y_1) - \frac{\rho_-}{2}\|w_1 - y_1\|^2 \\
\Longrightarrow\ & \frac{\rho - \rho_-}{2}\|w_1 - y_1\|^2 \le h(w_0) + p(y_0) - l - (1-\rho\alpha)\lambda_0^T(w_1 - y_1) \\
& \le h(w_0) + p(y_0) - l + (1-\rho\alpha)\|\lambda_0\|^2/2 + (1-\rho\alpha)\|w_1 - y_1\|^2/2 \\
\Longrightarrow\ & \frac{\rho - \rho_- - (1-\rho\alpha)}{2}\|w_1 - y_1\|^2 \le h(w_0) + p(y_0) - l + (1-\rho\alpha)\|\lambda_0\|^2/2 \\
\Longrightarrow\ & \frac{1/\epsilon - 2\rho_- - 1}{4}\|w_1 - y_1\|^2 \le h(w_0) + p(y_0) - l + \|\lambda_0\|^2/4. \\
\Longrightarrow\ & \|w_1 - y_1\|^2 \le 8\epsilon(h(w_0) + p(y_0) - l + \|\lambda_0\|^2/4),
\end{aligned} \tag{51}
$$

where the last inequality holds because $\epsilon \le 1/(4\rho_- + 2)$. By (50) and (51),

$$
P_\tau^1 \le 8(h(w_0) + p(y_0)) - 7l + \left(\frac{7}{4} + \frac{5\epsilon}{8}\right)\|\lambda_0\|^2 \tag{52}
$$

By combining (49) and (52), we have for any $k \ge 1$,

$$
\alpha^2\|\lambda_k\|^2
$$

$$\leq \frac{2}{1-1/2} \cdot \frac{2}{2+1-2} \left(8(h(w_0) + p(y_0)) - 7l + \left(\frac{7}{4} + \frac{5\epsilon}{8} \right) \|\lambda_0\|^2 - l \right) \epsilon$$
$$= (64(h(w_0) + p(y_0)) - 64l + (14 + 5\epsilon)\|\lambda_0\|^2)\epsilon. \quad (53)$$

This implies that $\alpha^2 \|\lambda^*\|^2 \leq (64(h(w_0) + p(y_0)) - 64l + (14 + 5\epsilon)\|\lambda_0\|^2)\epsilon$. ∎

Remark 6 Based on the optimality conditions of (Update-1), (Update-2), and the multiplier update, the following holds for any $k \geq 0$:

$$0 \in \partial(h + \mathbb{1}_{Z_1})(w_{k+1}) + \lambda_{k+1} + \rho(y_{k+1} - y_k) + \mu(w_{k+1} - w_k),$$
$$0 \in \partial(p + \mathbb{1}_{Z_2})(y_{k+1}) - \lambda_{k+1}, \; w_{k+1} - y_{k+1} - \alpha\lambda_k = (\lambda_{k+1} - \lambda_k)/\rho.$$

According to (53), if we choose the parameters as in Corollary 1, the stopping criteria in Algorithm 1 indicates that:

$$\operatorname{dist}(0, \partial(h + \mathbb{1}_{Z_1})(w_{k+1}) + \lambda_{k+1}) < \epsilon_0, \; 0 \in \partial(p + \mathbb{1}_{Z_2})(y_{k+1}) - \lambda_{k+1},$$
$$\|w_{k+1} - y_{k+1}\| < \sqrt{(64(h(w_0) + p(y_0)) - 64l + (14 + 5\epsilon)\|\lambda_0\|^2)\epsilon} + \epsilon_0.$$

Finally we will show that the conditions (46) in Theorem 2 are equivalent to KKT conditions of (23) with a feasibility error $\alpha\lambda^*$.

Denote $w^* \triangleq (x_+^*; x_-^*; \xi^*), w \triangleq (x^+; x^-; \xi), y^* \triangleq (y_1^*; y_2^*; y_3^*), y \triangleq (y_1; y_2; y_3)$. By Definition 2, $(w^*; y^*)$ satisfies first-order KKT conditions of (23) if there exist $\mu \in \mathbb{R}$, $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}^n$, $\pi \in \mathbb{R}^m$ such that

$$\begin{pmatrix} \nabla f_Q(x_+^* - x_-^*) + \nabla g(y_1^* - y_2^*) \\ -\nabla f_Q(x_+^* - x_-^*) - \nabla g(y_1^* - y_2^*) \\ -\gamma e \end{pmatrix} + \begin{pmatrix} \mu\xi^* - \beta_1 - A^T\pi \\ \mu\xi^* - \beta_2 + A^T\pi \\ \mu(x_+^* + x_-^*) + \beta_4 - \beta_3 \end{pmatrix} = 0, \quad (54a)$$
$$0 \leq \beta_1 \perp y_1^* \geq 0, \quad (54b)$$
$$0 \leq \beta_2 \perp y_2^* \geq 0, \quad (54c)$$
$$0 \leq \beta_3 \perp y_3^* \geq 0, \quad (54d)$$
$$0 \leq \beta_4 \perp e - y_3^* \geq 0, \quad (54e)$$
$$0 \leq \pi \perp A(y_1^* - y_2^*) - b \geq 0, \quad (54f)$$
$$(x_+^* + x_-^*)^T \xi^* = 0, \quad (54g)$$
$$w^* - y^* = 0. \quad (54h)$$

It can be easily seen that (54) is equivalent to (17) by merging w^* and y^*. Recall that according to discussions in Section 2, point satisfying (17) is exactly the local minimum of (4), thus the local minimum of ℓ_0-minimization (1).

Theorem 3 Suppose that $(w^*; y^*; \lambda^*)$ satisfies (46), and recall that $h(w) = f_Q(x^+ - x^-) + \gamma e^T(e - \xi)$, $p(y) = g(y_1 - y_2)$ are smooth functions. Assume that vector $(\xi^*; \xi^*; x_+^* + x_-^*) \neq 0$. Then $\exists \mu \in \mathbb{R}, \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}^n, \pi \in \mathbb{R}^m$ such that (54a) - (54g) hold and $w^* - y^* = \alpha\lambda^*$.

Proof We know that $\partial \mathbb{1}_{Z_1}(w) = \mathcal{N}_{Z_1}(w)$, $\partial \mathbb{1}_{Z_2}(y) = \mathcal{N}_{Z_2}(y)$. Due to (46) and the smoothness of function h, $0 \in \partial(h + \mathbb{1}_{Z_1})(w^*) + \lambda^* \Rightarrow 0 \in \nabla_w h(w^*) + \lambda^* + \partial \mathbb{1}_{Z_1}(w^*) \Rightarrow -\nabla_w h(w^*) - \lambda^* \in \mathcal{N}_{Z_1}(w^*)$. Recall $Z_1 = \{(x^+; x^-; \xi) \in \mathbb{R}^{3n} \mid \xi^T(x^+ + x^-) = 0\}$. Then by Lemma 10 in the Appendix and the assumption $(\xi^*; \xi^*; x_+^* + x_-^*) \neq 0$, we have $\mathcal{N}_{Z_1}(w^*) = \{\mu(\xi^*; \xi^*; x_+^* + x_-^*) \mid \mu \in \mathbb{R}\}$. Therefore, $\exists \mu \in \mathbb{R}$ s.t.

$$\nabla_w h(w^*) + \lambda^* + \mu(\xi^*; \xi^*; x_+^* + x_-^*) = 0. \tag{55}$$

On the other hand, (46) and smoothness of function p imply $0 \in \partial(p + \mathbb{1}_{Z_2})(y^*) - \lambda^* \Rightarrow 0 \in \nabla_y p(y^*) - \lambda^* + \partial \mathbb{1}_{Z_2}(y^*) \Rightarrow -\nabla_y p(y^*) + \lambda^* \in \mathcal{N}_{Z_2}(y^*)$. Since Z_2 is a convex set, $\mathcal{N}_{Z_2}(y^*) = \{v \mid v^T(y - y^*) \le 0, \forall y \in Z_2\}$. Therefore, $(\nabla_y p(y^*) - \lambda^*)^T(y - y^*) \ge 0, \forall y \in Z_2$. This indicates that y^* is the optimal solution of the linear program: $\min_{y \in Z_2}\{(\nabla_y p(y^*) - \lambda^*)^T y\}$. Thus the KKT conditions are satisfied at y^*, i.e. $\exists \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}^n, \pi \in \mathbb{R}^m$ s.t.

$$\begin{aligned}
\nabla_y p(y^*) - \lambda^* + (-\beta_1 - A^T\pi; -\beta_2 + A^T\pi; \beta_4 - \beta_3) = 0,\\
0 \le \beta_1 \perp y_1^* \ge 0, 0 \le \beta_2 \perp y_2^* \ge 0, 0 \le \beta_3 \perp y_3^* \ge 0,\\
0 \le \beta_4 \perp e - y_3^* \ge 0, 0 \le \pi \perp A(y_1^* - y_2^*) - b \ge 0.
\end{aligned} \tag{56}$$

From (55) and (56), utilizing the def. of h and p, and adding the feasibility constraints $(x_+^* + x_-^*)^T\xi^* = 0$, $w^* - y^* = \alpha\lambda^*$, we obtain the perturbed KKT conditions. ∎

5 Preliminary numerics

In Section 5.1, we describe the test problem of interest while in Section 5.2, we study the impact of tractability by comparing tractable ADMM frameworks to their standard counterpart. In Section 5.3, performance of ($\mathrm{ADMM_{cf}}$) is examined by comparing it with other methods.[2]

5.1 Least squares regression with ℓ_0-norm

Suppose $f_Q(x) \triangleq \|Cx - d\|^2$, $C \in \mathbb{R}^{p \times n}, g(x) \equiv 0$, and there is no linear constraint $Ax \ge b$ in (1), leading to the following ℓ_0-regularized least-squares regression:

$$\min \|Cx - d\|^2 + \gamma\|x\|_0. \tag{ℓ_0-LSR}$$

This special case finds application in signal recovery and regression problems. The rows of C are generated from a multivariate normal $N(0, I_n)$ while the true coefficients x^{true} are created as follows: (1) Generate x_i^{true} for $i = 1, \ldots, n$ from uniform distribution $U(-60, 60)$. (2) If $|x_i^{\text{true}}| \ge \frac{60\kappa}{n}$ then $x_i^{\text{true}} \leftarrow 0$ for $i = 1, \ldots, n$. Then x^{true} is approximately κ-sparse (or $\|x^{\text{true}}\|_0 \approx \kappa$). The observation vector $d = Cx^{\text{true}} + \epsilon$, where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma^2 = \|x^{\text{true}}\|^2/10$.

[2] All experiments are conducted on Matlab and the code is uploaded to `https://github.com/yue-xie/l0-minimization`.

5.2 Impact of tractable subproblems

In this subsection, we will compare the ADMM based algorithms (Algorithms 1-3) proposed in Section 3 to resolve (ℓ_0-LSR).

Algorithm descriptions and settings. We start all algorithms from an initial point $w_0 = y_0 = (e; \mathbf{0}_n; \mathbf{0}_n)$, $\lambda_0 = \mathbf{0}_{3n}$. The maximum runtime allowed is 200s. Experiments are run on CPU of 1.3GHz Intel Core i5 with 8GB memory. Other settings are as follows.

($\mathbf{ADMM_{cf}}$): Please refer to Algorithm 2 for details. Specifically, we use the following rule for updating ρ_k:[3]

$$\text{If } (\rho_k - \delta)\|y_{k+1} - y_k\| < \sqrt{2}\|\lambda_{k+1} - \lambda_k\| \text{ and } \rho_k \le \rho_{\max}, \text{ then } \rho_{k+1} := \delta_\rho \rho_k;$$
$$\text{else } \rho_{k+1} := \rho_k.$$

In addition, $\rho_0 = 1$, $\rho_{\max} = 2000$, $\delta = \rho_0/2$, $\delta_\rho = 1.01$, $\epsilon = 10^{-4}$;

($\mathbf{ADMM_{cf}^{\mu,\alpha,\rho}}$): Algorithm 1. $\alpha = 10^{-3}$, $\rho = 1/(2\alpha)$, $\mu = 3/\alpha$, $\epsilon_0 = 10^{-2}$.

($\mathbf{ADMM_0}$): Algorithm 3 with Update-1 solved by `Baron`. The maximum runtime for `Baron` is set to 200s. Note that we do not fix the penalty parameter at a value suggested by theory, which more often than not is impractical and involves problem parameter estimation. Instead, we update it adaptively. The update rule for ρ_k, inspired by augmented Lagrangian schemes [7], is as follows:

$$\text{If } k = 0 \text{ or } (h_{k+1} \ge 10^{-2} \text{ and } h_{k+1} > 0.9h_k \text{ and } \rho_k < 500), \rho_{k+1} = 1.01\rho_k;$$
$$\text{otherwise } \rho_{k+1} = \rho_k;$$

where $h_k = \|w_k - y_k\|$ for all $k \ge 0$. $\rho_0 = 1$, $\epsilon = 10^{-4}$.

Stopping criteria. The stopping criteria for ($\mathrm{ADMM_{cf}}$) and ($\mathrm{ADMM_0}$) guarantee that the KKT residual is below ϵ if terminated within time limit. For ($\mathrm{ADMM_{cf}^{\mu,\alpha,\rho}}$), it is guaranteed that the KKT residual is below $\epsilon_0 + \mathcal{O}(\sqrt{\epsilon})$ (See Remark 6). Thus, stopping criteria of the three algorithms are related to the optimality conditions.

Metric. In Table 2, KKT residual for ($\mathrm{ADMM_{cf}}$), ($\mathrm{ADMM_{cf}^{\mu,\alpha,\rho}}$) and ($\mathrm{ADMM_0}$) are $\max\{\rho_{K-1}\|y_K - y_{K-1}\|, \|(x_K^+; x_K^-; \xi_K) - y_K\|\}$, $\max\{\|\rho(y_K - y_{K-1}) + \mu(x_K - x_{K-1})\|, \|w_K - y_K\|\}$ and $\max\{\rho_{K-1}\|y_K - y_{K-1}\|, \|w_K - y_K\|\}$, respectively. K is the last iteration.

Results. In Table 2, we provide a comparison of ($\mathrm{ADMM_{cf}}$), ($\mathrm{ADMM_{cf}^{\mu,\alpha,\rho}}$) and ($\mathrm{ADMM_0}$) to address ($\ell_0$-LSR). Note that ($\mathrm{ADMM_{cf}}$), ($\mathrm{ADMM_{cf}^{\mu,\alpha,\rho}}$) are designed for formulation (23), which renders tractable subproblems, while ($\mathrm{ADMM_0}$) is for formulation (30), which requires global resolution of an MPCC as the subproblem. Therefore, even though ($\mathrm{ADMM_0}$) can be efficient when the dimension is relatively low, but becomes less so when n is larger. This is because the subproblem solver does not scale well with problem size and requires significant time for larger dimensions. This is supported by the drastically reduced number of outer loop iterations and large KKT residual upon termination when n is larger. Meanwhile, ($\mathrm{ADMM_{cf}}$) and ($\mathrm{ADMM_{cf}^{\mu,\alpha,\rho}}$) appear to be

[3] This update rule for ρ_k is related to the convergence property of ($\mathrm{ADMM_{cf}}$). Please refer to Section 7.4 for details.

scale far better with n due to tractability of the subproblem. Furthermore, it can be seen that (ADMM_{cf}) is far more efficient than the other two methods. It spends far less time to find solutions with lower KKT residual. In fact, it also provides better objective function value than the other two methods as observed during the experiment. Further exploration of (ADMM_{cf}) through comparison with other ℓ_0-minimization solvers is presented in the next subsection.

$(n, \|x^{\text{tr.}}\|_0, \gamma)$	(ADMM_{cf})			($\text{ADMM}_{\text{cf}}^{\mu,\alpha,\rho}$)			(ADMM_0)		
	t(s)	res.	iter.	t(s)	res.	iter.	t(s)	res.	iter.
(20, 1, 1)	0.38	9.4e-5	587	2.2e0	1.0e-2	4.90e3	201	3.0e-3	135
(20, 1, 10)	0.46	7.9e-5	515	2.0e2	3.5e-1	2.05e5	200	1.0e-2	158
(20, 4, 1)	0.26	7.2e-5	214	2.0e2	7.6e-2	2.67e5	200	8.7e-3	197
(20, 4, 10)	0.70	9.9e-5	798	2.0e2	1.8e-1	2.18e5	200	2.1e-2	150
(50, 10, 1)	1.10	1.0e-4	781	6.3e-1	9.9e-3	551	209	7.2e0	2
(50, 10, 10)	2.10	9.9e-5	1093	2.0e0	1.6e-1	1.18e5	201	2.5e-1	101
(50, 18, 1)	0.62	7.7e-5	283	2.0e0	1.8e-1	1.42e5	215	1.3e1	2
(50, 18, 10)	4.20	9.9e-5	1802	2.0e0	6.9e-1	1.29e5	201	4.8e-1	47
(100, 6, 1)	0.58	9.3e-5	617	2.5e-1	9.6e-3	200	209	9.8e0	2
(100, 6, 10)	0.66	9.8e-5	593	2.0e2	1.4e-1	2.10e5	399	4.8e0	14
(100, 19, 1)	0.76	9.2e-5	483	4.0e-1	9.9e-3	376	211	1.0e1	2
(100, 19, 10)	1.80	9.0e-5	535	2.0e2	1.3e-1	1.72e5	289	5.5e0	5

Table 2: Comparison of methods on (ℓ_0-LSR), $p = 10$

5.3 Comparison among (ADMM_{cf}) and other methods

In this set of experiments, we test (ADMM_{cf}) on (ℓ_0-LSR) with higher dimensions ($p = 256, n = 1024$) and compare it with other known methods directly addressing ℓ_0-minimization: iterative hard thresholding (IHT) and iterative hard thresholding with warm start (IHTWS) [8]. We again test the schemes on (ℓ_0-LSR) and choose almost the same settings as in Section 5.1, the only difference being that $\epsilon \in \mathbb{R}^p, \epsilon_i \sim N(0, \sigma^2), i.i.d., \sigma^2 = \frac{(x^{\text{true}})^T I_n x^{\text{true}}}{\text{SNR}}$, where SNR refers to the signal-to-noise ratio. All experiments are conducted on CPU of 3.4GHz Intel Core i7 with 16GB memory.

Algorithm descriptions and settings.

(IHT) and (IHTWS): (IHT) is implemented with 50 initial points (including the origin and points drawn from normal distribution $N(0, I_n)$), and the best solution is chosen. (IHTWS) is warm-started from a point computed by matching pursuit. The termination condition for both (IHT) and (IHTWS) is $\|x_{k+1} - x_k\| < 1 \times 10^{-6}$.

(ADMM_{cf}): Implementation of (ADMM_{cf}) is almost the same with the last experiment in Section 5.2: Initial point is selected as $y_0 = (e_n, \mathbf{0}_n, \mathbf{0}_n)$, $\lambda_0 = \mathbf{0}_{3n}$, and the parameters are chosen as $\rho_0 = \gamma$, $\epsilon = 10^{-4}$, $\delta_\rho = 1.01$, $\delta = \rho_0/2$, $\rho_{\max} = 2000$, $\text{time}_{\max} = 300$ for all cases.

Metric. In Table 3, RDF $\triangleq \frac{f_{\text{method}} - f_{\text{ADMM}_{\text{cf}}}}{f_{\text{ADMM}_{\text{cf}}}}$, where $f_{\text{ADMM}_{\text{cf}}}$ is calculated as follows. Suppose Algorithm 2 terminates when $k = T$. Let $(\bar{x}^+; \bar{x}^-; \bar{\xi}) = y_{T+1}$. Then the solution given by (ADMM$_{\text{cf}}$) is $\bar{x}^+ - \bar{x}^-$ and

$$f_{\text{ADMM}_{\text{cf}}} = \begin{cases} \|C(\bar{x}^+ - \bar{x}^-) - d\|^2 + \gamma(n - e^T\bar{\xi}), \\ \qquad \text{if } \max(\|w_{T+1} - y_{T+1}\|, \rho_T\|y_{T+1} - y_T\|) \leq \epsilon \\ \|C(\bar{x}^+ - \bar{x}^-) - d\|^2 + \gamma\|\bar{x}^+ - \bar{x}^-\|_0, \\ \qquad \text{if } \max(\|w_{T+1} - y_{T+1}\|, \rho_T\|y_{T+1} - y_T\|) > \epsilon \end{cases}$$

	RDF		time(s)			$\|x^{\text{sol.}}\|_0$		
(SNR, $\|x^{\text{tr.}}\|_0$, γ)	(I)	(W)	(I)	(W)	(A)	(I)	(W)	(A)
(5, 16, 0.10)	1.92	-0.02	1.2	1.2	95.4	624	194	98
(5, 16, 1.00)	0.83	-0.03	1.3	0.3	17.2	135	77	62
(5, 16, 10.00)	0.58	-0.05	1.5	0.1	41.7	7	15	12
(5, 16, 50.00)	0.93	-0.10	2.0	0.1	42.5	0	10	7
(5, 87, 0.10)	2.93	0.62	1.2	2.2	102.1	981	404	180
(5, 87, 1.00)	3.37	0.59	1.3	1.7	19.5	893	326	161
(5, 87, 10.00)	2.94	0.46	1.5	4.5	75.8	639	230	133
(5, 87, 50.00)	0.82	0.33	20.1	1.0	92.4	261	168	103
(5, 253, 0.10)	2.75	0.80	1.4	2.5	236.3	1006	483	268
(5, 253, 1.00)	2.69	0.59	1.5	2.0	234.2	990	425	268
(5, 253, 10.00)	2.49	0.34	1.6	1.8	300.0	932	357	267
(5, 253, 50.00)	2.15	0.11	1.9	3.1	63.7	831	293	264
(10, 20, 0.10)	1.65	-0.18	1.3	0.5	118.7	529	128	68
(10, 20, 1.00)	0.29	-0.18	1.5	0.2	15.4	76	53	22
(10, 20, 10.00)	2.09	-0.03	1.6	0.1	43.2	4	16	14
(10, 20, 50.00)	0.34	-0.18	2.1	0.1	58.7	0	10	4
(10, 82, 0.10)	2.82	0.42	1.3	1.7	96.1	978	365	164
(10, 82, 1.00)	3.83	0.56	1.4	4.3	20.9	881	285	150
(10, 82, 10.00)	2.68	0.35	1.7	1.8	73.8	599	209	117
(10, 82, 50.00)	0.74	0.23	3.8	0.7	72.8	212	138	93
(10, 249, 0.10)	2.32	0.66	1.4	2.6	12.4	1010	506	304
(10, 249, 1.00)	2.26	0.47	1.4	2.1	8.7	990	446	304
(10, 249, 10.00)	2.11	0.25	1.5	1.8	6.1	945	381	304
(10, 249, 50.00)	1.91	0.05	1.8	2.4	6.6	877	316	301

Table 3: Comparison of methods on (ℓ_0-LSR) ($p \times n = 256 \times 1024$)
(I): (IHT) (W): (IHTWS) (A): (ADMM$_{\text{cf}}$)

Results. From Table 3, we conclude the following:

(1) Although (ADMM$_{\text{cf}}$) takes more time, it generally produces solutions that are superior to (IHT) in objective function value and provides better values than (IHTWS) in most cases. Note that (ADMM$_{\text{cf}}$) is cold started.
(2) (ADMM$_{\text{cf}}$) generally produces sparser solution than (IHTWS) and (IHT), which indicates that (ADMM$_{\text{cf}}$) scheme is potentially more favorable from a statistical standpoint.

6 Concluding remarks and future work

We consider a full complementarity reformulation of a general class of ℓ_0-norm minimization problems. The focus of this paper lies on the characterization and efficient computation of KKT points for this formulation. In particular, we show that a suitable (Guignard) constraint qualification holds at every feasible point. Moreover, when f is a convex function, a point satisfies the first-order KKT conditions if and only if it is a local minimizer. Next, two tractable ADMM schemes are presented for resolution. In these schemes, a hidden convexity property is leveraged to allow for tractable resolution of ADMM subproblems. For the perturbed proximal ADMM framework, subsequential convergence to KKT points with arbitrarily small error under mild assumptions can be shown. Preliminary empirical studies show the significance of having tractable subproblems in ADMM schemes and that the tractable ADMM framework compares well with its competitors. In future work, we may consider characterization and computation of KKT points of problems complicated by cardinality constraints (2) and affine sparsity constraints [14].

Acknowledgements. The authors would like to acknowledge an early discussion with Dr. Ankur Kulkarni of IIT, Mumbai, as well as the inspiration provided by Dr. J. S. Pang during his visit to Penn. State University, and suggestion by Dr. Mingyi Hong in INFORMS 2018, Denver.

References

1. 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*, Mathematics of Operations Research, 35 (2010), pp. 438–457.
2. F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with sparsity-inducing penalties*, Foundations and Trends® in Machine Learning, 4 (2012), pp. 1–106.
3. A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, vol. 2, SIAM, 2001, ch. 5. Computational tractability of convex programs.
4. A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Mathematical Programming, 72 (1996), pp. 51–63.
5. D. Bertsimas, A. King, and R. Mazumder, *Best subset selection via a modern optimization lens*, The Annals of Statistics, 44 (2016), pp. 813–852.
6. D. Bertsimas and R. Shioda, *Algorithm for cardinality-constrained quadratic optimization*, Computational Optimization and Applications, 43 (2009), pp. 1–22.
7. E. G. Birgin, C. A. Floudas, and J. M. Martínez, *Global minimization using an augmented Lagrangian method with variable lower-level constraints*, Mathematical Programming, 125 (2010), pp. 139–162.
8. T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 629–654.
9. R. Boţ, E. Csetnek, and D. Nguyen, *A proximal minimization algorithm for structured nonconvex and nonsmooth problems*, SIAM Journal on Optimization, 29 (2019), pp. 1300–1328.
10. O. P. Burdakov, C. Kanzow, and A. Schwartz, *Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method*, SIAM Journal on Optimization, 26 (2016), pp. 397–425.

11. J. Burke, *Fundamentals of optimization, Chapter 5, Langrange multipliers.* Course Notes, AMath/Math 515, University of Washington.
12. ———, *Numerical optimization.* Course Notes, AMath/Math 516, University of Washington, Spring Term 2012.
13. E. J. Candès and M. B. Wakin, *An introduction to compressive sampling*, IEEE signal processing magazine, 25 (2008), pp. 21–30.
14. H. Dong, M. Ahn, and J.-S. Pang, *Structural properties of affine sparsity constraints*, Mathematical Programming, 176 (2019), pp. 95–135.
15. D. L. Donoho, *Compressed sensing*, IEEE Transactions on information theory, 52 (2006), pp. 1289–1306.
16. F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems Volumn I*, Springer Science & Business Media, 2007.
17. J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), pp. 1348–1360.
18. E. X. Fang, H. Liu, and M. Wang, *Blessing of massive scale: spatial graphical model estimation with a total cardinality constraint approach*, Mathematical Programming, 176 (2019), pp. 175–205.
19. M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter, *Complementarity formulations of ℓ_0-norm optimization problems*, Pacific Journal of Optimization, 14 (2018).
20. G. Fung and O. Mangasarian, *Equivalence of minimal ℓ_0 and ℓ_p norm solutions of linear equalities, inequalities and linear programs for sufficiently small p*, Journal of Optimization Theory and Applications, 151 (2011), pp. 1–10.
21. D. Ge, X. Jiang, and Y. Ye, *A note on the complexity of L_p minimization*, Mathematical Programming, 129 (2011), pp. 285–299.
22. M. L. Gonçalves, J. G. Melo, and R. D. Monteiro, *Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems*, arXiv preprint arXiv:1702.01850, (2017).
23. D. Hajinezhad and M. Hong, *Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization*, Mathematical Programming, 176 (2019), pp. 207–245.
24. M. Hong, Z. Luo, and M. Razaviyayn, *Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems*, SIAM Journal on Optimization, 26 (2016), pp. 337–364.
25. B. Jiang, T. Lin, S. Ma, and S. Zhang, *Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis*, Computational Optimization and Applications, 72 (2019), pp. 115–157.
26. H. Liu, T. Yao, and R. Li, *Global solutions to folded concave penalized nonconvex learning*, Annals of statistics, 44 (2016), p. 629.
27. Q. Liu, X. Shen, and Y. Gu, *Linearized ADMM for non-convex non-smooth optimization with convergence analysis*, CoRR, abs/1705.02502 (2017).
28. Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical programs with equilibrium constraints*, Cambridge University Press, Cambridge, 1996.
29. R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317, Springer Science & Business Media, 2009.
30. R. Tibshirani, *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.
31. F. Wang, W. Cao, and Z. Xu, *Convergence of multi-block Bregman ADMM for nonconvex composite problems*, Science China Information Sciences, 61 (2018), p. 122101.
32. J. Wang and L. Zhao, *Nonconvex generalizations of ADMM for nonlinear equality constrained problems*, CoRR, abs/1705.03412 (2017).
33. Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, Journal of Scientific Computing, (2018).
34. Z. Xu, S. De, M. A. T. Figueiredo, C. Studer, and T. Goldstein, *An empirical study of ADMM for nonconvex problems*, CoRR, abs/1612.03349 (2016).
35. L. Yang, T. K. Pong, and X. Chen, *Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 74–110.
36. C.-H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics, 38 (2010), pp. 894–942.

37. C.-H. ZHANG AND T. ZHANG, *A general theory of concave regularization for high-dimensional sparse estimation problems*, Statistical Science, 27 (2012), pp. 576–593.
38. T. ZHANG, *Analysis of multi-stage convex relaxation for sparse regularization*, Journal of Machine Learning Research, 11 (2010), pp. 1081–1107.

7 Appendix

7.1 Hidden Convexity

Consider a QCQP defined as follows:

$$\min \left\{ x^T H x + h^T x \mid \ell \le x^T Q x \le r \right\}, \tag{57}$$

where $x \in \mathbb{R}^n, H \in \mathbb{R}^{n\times n}, Q \in \mathbb{R}^{n\times n}, h \in \mathbb{R}^n, \ell \in \mathbb{R}, r \in \mathbb{R}$. Suppose that (57) is feasible and the two matrices H and Q can be simultaneously diagonalized. Recall from [4] that H and Q can be simultaneously diagonalized if there exists a nonsingular matrix $P \in \mathbb{R}^{n\times n}$ such that

$$P^T H P = D \triangleq \mathrm{diag}(d_1, \ldots, d_n) \text{ and } P^T Q P = S \triangleq \mathrm{diag}(s_1, \ldots, s_n).$$

Let $d \triangleq (d_1; \ldots; d_n), s \triangleq (s_1; \ldots; s_n)$. Then, by utilizing a transformation $x = Py$ and $c = P^T h$, (57) can be written as follows:

$$\min \left\{ y^T D y + c^T y \mid \ell \le y^T S y \le r \right\}, \tag{58}$$

Further, consider the following convex program:

$$\min_{w_1,\ldots,w_n} \left\{ \sum_{i=1}^n (d_i w_i - |c_i|\sqrt{w_i}) \mid \ell \le \sum_{i=1}^n s_i w_i \le r,\ w \ge 0 \right\}, \tag{59}$$

which is defined using c, d, ℓ, and r. In fact, problems (58) and (59) are equivalent in the following sense: (i) if one is unbounded, so is the other; (ii) if both are finite, the infimum of (58) is equal to the the infimum of (59); (iii) the optimal solution y^* of (58) can be constructed from the optimal solution w^* of (59) through the following equations:

$$y_j^* = -\,\mathrm{sgn}(c_j)\sqrt{w_j^*}, \quad j = 1, \ldots, n \tag{60}$$

where $\mathrm{sign}(c_j) = 1$ if $c_j \ge 0$ and $\mathrm{sign}(c_j) = -1$ if $c_j < 0$. Through the above arguments, a global minimizer of the solution to nonconvex program (58) may be obtained by the solution of a suitably defined convex program (59). Thus (58) may have *hidden convexity*. This property of QCQP is first discovered by Ben-Tal and Teboulle in 1996 [4].

7.2 Proofs

Lemma 8 (Tightness of relaxation) Consider the problem (4) and suppose a global minimizer to this problem is denoted by $(x^{\pm}, \xi)$. Let $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$ be defined as follows:

$$\tilde{\xi} \triangleq \xi, \tilde{x}_i^+ \triangleq \begin{cases} x_i^+ - x_i^-, & \text{if } x_i^+ \geq x_i^- \\ 0, & \text{otherwise} \end{cases}, \tilde{x}_i^- \triangleq \begin{cases} 0, & \text{if } x_i^+ \geq x_i^- \\ x_i^- - x_i^+. & \text{otherwise} \end{cases}, \tag{61}$$
$$\forall i = 1, \ldots, n. \tilde{x} \triangleq \tilde{x}^+ - \tilde{x}^-.$$

Then $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$ is a a global minimizer of (6).

Proof Consider a solution $(x^{\pm}, \xi)$ to (4). We first prove that the constructed solution $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$ is feasible with respect to (6) and then prove that it is optimal.

Feasibility of $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$. By definition (61), we have that $\tilde{x}^{\pm} \geq 0$ and $\min(\tilde{x}_i^+, \tilde{x}_i^-) = 0$ for $i = 1, \ldots, n$. Consequently, $\tilde{x}^+ \perp \tilde{x}^-$. Furthermore, $\tilde{x} = \tilde{x}^+ - \tilde{x}^- = x^+ - x^-$ and $\tilde{\xi} = \xi$. This implies that $A\tilde{x} = A\tilde{x}^+ - A\tilde{x}^- = Ax^+ - Ax^- \geq b$. Finally, it suffices to show that $(\tilde{x}^+ + \tilde{x}^-)^T \tilde{\xi} = 0$. By the feasibility of $(x^{\pm}, \xi)$ with respect to (4), we have that

$$\begin{aligned} 0 = \sum_{i=1}^n (x_i^+ + x_i^-)\xi_i &= \sum_{i \in \mathcal{I}_+} (\underbrace{x_i^+ - x_i^-}_{\triangleq \tilde{x}_i^+} + 2x_i^-)\xi_i + \sum_{i \in \mathcal{I}_-} (2x_i^+ + \underbrace{x_i^- - x_i^+}_{\triangleq \tilde{x}_i^-})\xi_i \\ &= \sum_{i \in \mathcal{I}_+} (\tilde{x}_i^+ + \underbrace{\tilde{x}_i^-}_{=0} + 2x_i^-)\xi_i + \sum_{i \in \mathcal{I}_-} (\underbrace{\tilde{x}_i^+}_{=0} + \tilde{x}_i^- + 2x_i^+)\xi_i, \end{aligned} \tag{62}$$

where $\mathcal{I}_+ \triangleq \{i : x_i^+ \geq x_i^-\}$ and $\mathcal{I}_- \triangleq \{i : x_i^+ < x_i^-\}$. Since, (62) can be expressed as follows:

$$\begin{aligned} 0 &= \sum_{i \in \mathcal{I}_+} (\tilde{x}_i^+ + \tilde{x}_i^- + 2x_i^-)\xi_i + \sum_{i \in \mathcal{I}_-} (\tilde{x}_i^+ + \tilde{x}_i^- + 2x_i^+)\xi_i \\ &= (\tilde{x}^+ + \tilde{x}^-)^T \xi + \sum_{i \in \mathcal{I}_+} 2x_i^- \xi_i + \sum_{i \in \mathcal{I}_-} 2x_i^+ \xi_i, \end{aligned} \tag{63}$$

and $x^{\pm}, \tilde{x}^{\pm}, \xi \geq 0$ implying that

$$\begin{aligned} &(\tilde{x}^+ + \tilde{x}^-)^T \xi \geq 0, \sum_{i \in \mathcal{I}_+} 2x_i^- \xi_i \geq 0, \sum_{i \in \mathcal{I}_-} 2x_i^- \xi_i \geq 0 \\ &\implies (\tilde{x}^+ + \tilde{x}^-)^T \xi = 0, \overset{\xi = \tilde{\xi}}{\implies} (\tilde{x}^+ + \tilde{x}^-)^T \tilde{\xi} = 0. \end{aligned}$$

Optimality of $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$. We observe that the $f(\tilde{x}) + \gamma \sum_{i=1}^n (1 - \tilde{\xi}_i) = f(x^+ - x^-) + \gamma \sum_{i=1}^n (1 - \xi_i)$ for $\tilde{x} = x^+ - x^-$ and $\tilde{\xi} = \xi$. But since (4) is a relaxation of (6), the optimality of $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$ follows from feasibility of $(\tilde{x}, \tilde{x}^{\pm}, \tilde{\xi})$ with respect to the tightened optimization problem with an identical objective value. ∎

Lemma 9 Given $H \triangleq \begin{pmatrix} M+\frac{\rho+\mu}{2}I_n & -M & \\ -M & M+\frac{\rho+\mu}{2}I_n & \\ & & \frac{\rho+\mu}{2}I_n \end{pmatrix}$, $\tilde{Q} \triangleq \begin{pmatrix} & & I_n \\ & & I_n \\ I_n & I_n & \end{pmatrix}$, and $G \triangleq \begin{pmatrix} \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ -\frac{\sqrt{2}}{2}I_n & & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix}$, where V is orthogonal such that $S \triangleq V^TMV$ is diagonal. Then G is an orthogonal matrix, and $H, \tilde{Q}$ can be simultaneously diagonalized through G.

Proof

$$
\begin{aligned}
&G^TG \\
&= \begin{pmatrix} I_n & & \\ & V^T & \\ & & I_n \end{pmatrix} \begin{pmatrix} \frac{1}{2}I_n & \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n \\ \frac{\sqrt{2}}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \\ \frac{1}{2}I_n & \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ -\frac{\sqrt{2}}{2}I_n & & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix} \\
&= \begin{pmatrix} I_n & & \\ & V^T & \\ & & I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & I_n & \\ & & I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix} = \begin{pmatrix} I_n & & \\ & V^TV & \\ & & I_n \end{pmatrix} = I_{3n}
\end{aligned}
$$

Therefore, G is orthogonal. Note that

$$
\begin{aligned}
&\begin{pmatrix} \frac{1}{2}I_n & \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n \\ \frac{\sqrt{2}}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \\ \frac{1}{2}I_n & \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} M+\frac{\rho+\mu}{2}I_n & -M & \\ -M & M+\frac{\rho+\mu}{2}I_n & \\ & & \frac{\rho+\mu}{2}I_n \end{pmatrix} \begin{pmatrix} \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ -\frac{\sqrt{2}}{2}I_n & & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2}I_n & \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n \\ \frac{\sqrt{2}}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \\ \frac{1}{2}I_n & \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} \frac{\rho+\mu}{4}I_n & \sqrt{2}M+\frac{\sqrt{2}}{4}(\rho+\mu)I_n & \frac{\rho+\mu}{4}I_n \\ \frac{\rho+\mu}{4}I_n & -\sqrt{2}M-\frac{\sqrt{2}}{4}(\rho+\mu)I_n & \frac{\rho+\mu}{4}I_n \\ -\frac{\sqrt{2}}{4}(\rho+\mu)I_n & & \frac{\sqrt{2}}{4}(\rho+\mu)I_n \end{pmatrix} \\
&= \begin{pmatrix} \frac{\rho+\mu}{2}I_n & & \\ & 2M+\frac{\rho+\mu}{2}I_n & \\ & & \frac{\rho+\mu}{2}I_n \end{pmatrix}.
\end{aligned}
$$

This fact implies:

$$
\begin{aligned}
&G^THG \\
&= \begin{pmatrix} I_n & & \\ & V^T & \\ & & I_n \end{pmatrix} \begin{pmatrix} \frac{\rho+\mu}{2}I_n & & \\ & 2M+\frac{\rho+\mu}{2}I_n & \\ & & \frac{\rho+\mu}{2}I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix} \\
&= \begin{pmatrix} \frac{\rho+\mu}{2}I_n & & \\ & 2V^TMV+\frac{\rho+\mu}{2}V^TV & \\ & & \frac{\rho+\mu}{2}I_n \end{pmatrix} = \begin{pmatrix} \frac{\rho+\mu}{2}I_n & & \\ & 2S+\frac{\rho+\mu}{2}I_n & \\ & & \frac{\rho+\mu}{2}I_n \end{pmatrix}
\end{aligned}
$$

Meanwhile,

$$
G^T\tilde{Q}G
$$

$$
\begin{aligned}
&= \begin{pmatrix} I_n & & \\ & V^T & \\ & & I_n \end{pmatrix} \begin{pmatrix} \frac{1}{2}I_n & \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n \\ \frac{\sqrt{2}}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \\ \frac{1}{2}I_n & \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} & & I_n \\ & & I_n \\ I_n & I_n & \end{pmatrix} \cdot \\
&\quad \begin{pmatrix} \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \frac{1}{2}I_n \\ -\frac{\sqrt{2}}{2}I_n & & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix} \\
&= \begin{pmatrix} I_n & & \\ & V^T & \\ & & I_n \end{pmatrix} \begin{pmatrix} \frac{1}{2}I_n & \frac{1}{2}I_n & -\frac{\sqrt{2}}{2}I_n \\ \frac{\sqrt{2}}{2}I_n & -\frac{\sqrt{2}}{2}I_n & \\ \frac{1}{2}I_n & \frac{1}{2}I_n & \frac{\sqrt{2}}{2}I_n \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2}I_n & 0 & \frac{\sqrt{2}}{2}I_n \\ -\frac{\sqrt{2}}{2}I_n & 0 & \frac{\sqrt{2}}{2}I_n \\ I_n & 0 & I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix} \\
&= \begin{pmatrix} I_n & & \\ & V^T & \\ & & I_n \end{pmatrix} \begin{pmatrix} -\sqrt{2}I_n & & \\ & 0 & \\ & & \sqrt{2}I_n \end{pmatrix} \begin{pmatrix} I_n & & \\ & V & \\ & & I_n \end{pmatrix} = \begin{pmatrix} -\sqrt{2}I_n & & \\ & 0 & \\ & & \sqrt{2}I_n \end{pmatrix}
\end{aligned}
$$

Thus, H and $\tilde{Q}$ can be simultaneously diagonalized through G. ∎

7.3 Miscellaneous

Definition 6 (Kurdyka-Łojasiewicz (KŁ) property) A proper lower semi-continuous function $\mathcal{L} : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ has the KŁ property at $\bar{x} \in \mathrm{dom}(\partial\mathcal{L})$, if there exists $\eta \in (0, +\infty)$, a neighborhood U of $\bar{x}$, and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}_+$ such that the following hold: (i) $\phi(0) = 0$, and ϕ is continuously differentiable on $(0, \eta)$. For all $s \in (0, \eta)$, $\phi'(s) > 0$; (ii) For all x in $U \cap \{x \in \mathbb{R}^N : \mathcal{L}(\bar{x}) < \mathcal{L}(x) < \mathcal{L}(\bar{x}) + \eta]$, the Kurdyka-Łojasiewicz (KŁ) inequality holds: $\phi'(\mathcal{L}(x) - \mathcal{L}(\bar{x}))\mathrm{dist}(0, \partial\mathcal{L}(x)) \geq 1$.

Definition 7 (Semialgebraic function) Recall that a semialgebraic set $S \subseteq \mathbb{R}^n$ is defined as:

$$S \triangleq \{x \in \mathbb{R}^n : p_i(x) = 0, q_i(x) < 0, i = 1, \ldots, m\},$$

where p_i and q_i are real polynomial functions. A function $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a semialgebraic function if and only if its graph $\{(x; y) \in \mathbb{R}^n \times \mathbb{R} : y = F(x)\}$ is a semialgebraic subset in $\mathbb{R}^{n+1}$.

Remark 7 Semialgebraic function has the following properties: (i). It satisfies KŁ property with $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1) \cap \mathbb{Q}$ and $c > 0$. (ii). Finite sums and products of semialgebraic functions are semialgebraic. See [1, Section 4.3] for more details.

Lemma 10 (Theorem 10 [11]) In $\mathbb{R}^{n_1}$, let $C = \{x \in X \mid F(x) \in D\}$, for closed convex sets $X \subset \mathbb{R}^{n_1}, D \subset \mathbb{R}^{n_2}$, and a $\mathcal{C}^1$ mapping $F : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$, written componentwise as $F(x) = (f_1(x); \ldots ; f_{n_2}(x))$. Suppose the following constraint qualification is satisfied at a point $\bar{x} \in C$:

$$\sum_{j=1}^{n_2} y_j \nabla f_j(\bar{x}) + z = 0, y = (y_1; \ldots ; y_{n_2}) \in \mathcal{N}_D(F(\bar{x})), z \in \mathcal{N}_X(\bar{x})$$
$$\implies y = \mathbf{0}, z = 0.$$

Then the normal cone $\mathcal{N}_C(\bar{x})$ consists of all vectors v of the form

$$v = y_1 \nabla f_1(\bar{x}) + \ldots + y_{n_2} \nabla f_{n_2}(\bar{x}) + z \text{ with } y = (y_1; \ldots ; y_{n_2}) \in \mathcal{N}_D(F(\bar{x})),$$
$$z \in \mathcal{N}_X(\bar{x}).$$

Note: When $X = \mathbb{R}^{n_1}$, the normal cone $\mathcal{N}_X(\bar{x}) = \{0\}$, so the z terms here drop out. When D is a singleton, $\mathcal{N}_D(F(\bar{x})) = \mathbb{R}^{n_2}$.

7.4 Convergence analysis of ADMM$_{\mathrm{cf}}$

We now analyze convergence of (ADMM$_{\mathrm{cf}}$) under the following update rule of ρ_k:

If $(\rho_k - \delta)\|y_{k+1} - y_k\| < \sqrt{2}\|\lambda_{k+1} - \lambda_k\|$ and $\rho_k \le \rho_{\max}$, then $\rho_{k+1} := \delta_\rho \rho_k$; else $\rho_{k+1} := \rho_k$.

Further, we make the following assumption.

Assumption 3 *The penalty parameter sequence $\{\rho_k\}$ in (ADMM$_{cf}$) never exceeds the prescribed upper bound, i.e. $\rho_k \le \rho_{\max}$, $\forall k \ge 0$.*

Proposition 3 (Limit points of (ADMM$_{\mathrm{cf}}$) are KKT Points) Consider problem (4) with $f_Q(x) = x^T M x + d^T x$. Suppose (ADMM$_{\mathrm{cf}}$) generates a sequence $\{w_k \triangleq (x_k^+; x_k^-; \xi_k), y_k, \lambda_k\}$. Assume that this sequence converges to a limit point denoted by $(\bar{w}, \bar{y}, \bar{\lambda})$. Then, we may claim the following: (a) The point $\bar{w} = (\bar{x}^+; \bar{x}^-; \bar{\xi})$ satisfies first-order KKT conditions of (4). (b) If f is convex, then $\bar{w}$ is a local minimum of (4).

Proof **(a)**. By the update rule of (ADMM$_{\mathrm{cf}}$), $\exists K > 0, s.t.\ \rho_k \equiv \rho$, $\forall k \ge K$. Consequently, suppose $\rho_k \equiv \rho$ for all k without loss of generality. (Otherwise, we may initialize using $w_K, y_K, \lambda_K, \rho_K$) By (Update-1) at iteration $k+1$, we have that

$$w_{k+1} \in \operatorname*{argmin}_{(x^+ + x^-)^T \xi = 0} \left[f_Q(x^+ - x^-) - \gamma e^T \xi + \frac{\rho}{2}\|w - y_k + \lambda_k/\rho\|^2 \right].$$

Let $z_{k+1} \triangleq G^T w_{k+1}$ and $q_k \triangleq G^T \left((d; -d; -\gamma e) + \lambda_k - \rho y_k \right)$, where G is defined in (24). Denote $z_{k+1} \triangleq (z_{k+1,1}; z_{k+1,2}; z_{k+1,3})$, $z_{k+1,1}, z_{k+1,2}, z_{k+1,3} \in \mathbb{R}^n$. From (28), we have $\exists u_k, v_k \in \mathbb{R}^n$ such that

$$z_{k+1,1} = \begin{cases} \frac{-(\|q_{k,1}\| + \|q_{k,3}\|) q_{k,1}}{2\rho \|q_{k,1}\|}, \ \|q_{k,1}\| > 0 \\ \frac{\|q_{k,3}\|}{2\rho} u_k, \ \|u_k\| = 1, \ \|q_{k,1}\| = 0 \end{cases},$$

$$z_{k+1,3} = \begin{cases} \frac{-(\|q_{k,1}\|+\|q_{k,3}\|)q_{k,3}}{2\rho\|q_{k,3}\|}, \ \|q_{k,3}\| > 0 \\ \frac{\|q_{k,1}\|}{2\rho} v_k, \|v_k\| = 1, \ \|q_{k,3}\| = 0 \end{cases},$$

$$(z_{k+1,2})_i = -(q_{k,2})_i/(\rho + 4s_i), \forall i = 1, \ldots, n, \tag{64}$$

where $q_k \triangleq (q_{k,1}; q_{k,2}; q_{k,3}), q_{k,1}, q_{k,2}, q_{k,3} \in \mathbb{R}^n$. Since $w_{k+1} \to \bar{w}$, $y_k \to \bar{y}$, $\lambda_k \to \bar{\lambda}$ as $k \to \infty$, $z_{k+1} \to \bar{z} = G^T \bar{w}$ and as $k \to +\infty$, $q_k \to \bar{q} \triangleq G^T \left(\left(d; -d; -\gamma e\right) + \bar{\lambda} - \rho \bar{y} \right)$. We proceed to show that $\bar{z}$ and $\bar{q}$ also satisfy the following: $\exists \bar{u}, \bar{v} \in \mathbb{R}^n$ such that

$$\bar{z}_1 = \begin{cases} \frac{-(\|\bar{q}_1\|+\|\bar{q}_3\|)\bar{q}_1}{2\rho\|\bar{q}_1\|}, \ \|\bar{q}_1\| > 0 \\ \frac{\|\bar{q}_3\|}{2\rho} \bar{u}, \|\bar{u}\| = 1, \|\bar{q}_1\| = 0 \end{cases}; \bar{z}_3 = \begin{cases} \frac{-(\|\bar{q}_1\|+\|\bar{q}_3\|)\bar{q}_3}{2\rho\|\bar{q}_3\|}, & \|\bar{q}_3\| > 0 \\ \frac{\|\bar{q}_1\|}{2\rho} \bar{v}, \|\bar{v}\| = 1, & \|\bar{q}_3\| = 0 \end{cases} \tag{65}$$

$$(\bar{z}_2)_i = -(\bar{q}_2)_i/(\rho + 4s_i), \quad \forall i = 1, \ldots, n, \tag{66}$$

where $\bar{z} = (\bar{z}_1; \bar{z}_2; \bar{z}_3)$ and $\bar{q} = (\bar{q}_1; \bar{q}_2; \bar{q}_3)$. First, we prove that $\bar{z}_1$ and $\bar{q}$ satisfy (65).
(i) Case 1. $\|\bar{q}_1\| > 0$. Then $\exists K$ s.t. $\forall k \geq K, \|q_{k,1}\| > 0$,

$$z_{k+1,1} = \frac{-(\|q_{k,1}\| + \|q_{k,3}\|)q_{k,1}}{2\rho\|q_{k,1}\|}.$$

Therefore,

$$\bar{z}_1 = \lim_{k\to+\infty} z_{k+1,1} = \lim_{k\to+\infty} \frac{-(\|q_{k,1}\| + \|q_{k,3}\|)q_{k,1}}{2\rho\|q_{k,1}\|} = \frac{-(\|\bar{q}_1\| + \|\bar{q}_3\|)\bar{q}_1}{2\rho\|\bar{q}_1\|}.$$

(ii) Case 2. $\|\bar{q}_1\| = 0$. Then

$$\|\bar{z}_1\| = \lim_{k\to+\infty} \|z_{k+1,1}\| = \|\bar{q}_3\|/(2\rho) \implies \exists \bar{u}, \|\bar{u}\| = 1, \ s.t., \ \bar{z}_1 = \frac{\|\bar{q}_3\|}{2\rho}\bar{u}.$$

Note that expression of $\bar{z}_3$ can be proven similarly and we omit proof of (66). Therefore,

$$\bar{w} = G\bar{z} \in \operatorname*{argmin}_{(x^+ + x^-)^T\xi=0} \left[f_Q(x^+ - x^-) - \gamma e^T \xi + \frac{\rho}{2} \|(x^+; x^-; \xi) - \bar{y} + \bar{\lambda}/\rho\|^2 \right]. \tag{67}$$

In particular, it follows that $(\bar{x}^+ + \bar{x}^-)^T \bar{\xi} = 0$. Next, we consider whether such a limit point satisfies the first-order KKT conditions of (67) by examining two cases:
(i) Suppose $(\bar{\xi}; \bar{\xi}; \bar{x}^+ + \bar{x}^-) \neq 0$. Then the linear independence constraint qualification (LICQ) holds at $(\bar{x}^+, \bar{x}^-, \bar{\xi})$. Consequently, there exists a scalar μ such that

$$\begin{pmatrix} \nabla f_Q(\bar{x}^+ - \bar{x}^-) \\ -\nabla f_Q(\bar{x}^+ - \bar{x}^-) \\ -\gamma e \end{pmatrix} + \rho(\bar{w} - \bar{y} + \bar{\lambda}/\rho) + \mu \begin{pmatrix} \bar{\xi} \\ \bar{\xi} \\ \bar{x}^+ + \bar{x}^- \end{pmatrix} = 0. \tag{68}$$

(ii) Suppose $(\bar{\xi}, \bar{\xi}, \bar{x}^+ + \bar{x}^-) = 0$, implying $\bar{\xi} = 0$ and $x^+ + x^- = 0$. Since $(\bar{x}^+; \bar{x}^-; \bar{\xi})$ is a global optimizer of (67), when we fix $\xi \equiv \bar{\xi} = 0$, the following must hold:

$$\begin{aligned}
&\begin{pmatrix} \bar{x}^+ \\ \bar{x}^- \end{pmatrix} \in \underset{x^+, x^-}{\operatorname{argmin}} \ f_Q(x^+ - x^-) + \frac{\rho}{2} \left(\left\| x^+ - \bar{y}_1 + \frac{\bar{\lambda}_1}{\rho} \right\|^2 + \left\| x^- - \bar{y}_2 + \frac{\bar{\lambda}_2}{\rho} \right\|^2 \right) \\
&\implies 0 = \begin{pmatrix} \nabla f_Q(\bar{x}^+ - \bar{x}^-) \\ -\nabla f_Q(\bar{x}^+ - \bar{x}^-) \end{pmatrix} + \rho \begin{pmatrix} \bar{x}^+ - \bar{y}_1 + \bar{\lambda}_1/\rho \\ \bar{x}^- - \bar{y}_2 + \bar{\lambda}_2/\rho \end{pmatrix}.
\end{aligned}$$

If we fix $x^+ \equiv \bar{x}^+, x^- \equiv \bar{x}^-$, then

$$\bar{\xi} \in \underset{\xi \in \mathbb{R}^n}{\operatorname{argmin}} \left[-\gamma e^T \xi + \frac{\rho}{2} \| \xi - \bar{y}_3 + \bar{\lambda}_3/\rho \|^2 \right] \implies 0 = -\gamma e + \rho(\bar{\xi} - \bar{y}_3 + \bar{\lambda}_3/\rho),$$

where $\bar{y} = (\bar{y}_1; \bar{y}_2; \bar{y}_3), \bar{y}_1, \bar{y}_2, \bar{y}_3 \in \mathbb{R}^n$, $\bar{\lambda} = (\bar{\lambda}_1; \bar{\lambda}_2; \bar{\lambda}_3), \bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3 \in \mathbb{R}^n$. Thus, (68) holds for every $\mu \in \mathbb{R}$. Next, in (Update-2), we need to compute y_{k+1}, where $y_{k+1} = \operatorname{argmin}_{y \in Z_2} \quad g(y^+ - y^-) + \frac{\rho}{2} \| y - w_{k+1} - \lambda_k/\rho \|^2$. Note that the following first order condition holds because it is a convex program:

$$\begin{aligned}
([\nabla g(y_{k+1,1} - y_{k+1,2}); -\nabla g(y_{k+1,1} - y_{k+1,2}); \mathbf{0}_n] + \rho(y_{k+1} - w_{k+1} - \lambda_k/\rho))^T& \\
(y - y_{k+1}) \geq 0,& \\
\forall y \in Z_2, y_{k+1} = (y_{k+1,1}; y_{k+1,2}; y_{k+1,3}), y_{k+1,i} \in \mathbb{R}^n, i = 1, 2, 3.&
\end{aligned}$$

By continuity of $\nabla g(\bullet)$, since $w_{k+1} \to \bar{w}, y_{k+1} \to \bar{y}, \lambda_k \to \bar{\lambda}$, we have that

$$\begin{aligned}
([\nabla g(\bar{y}_1 - \bar{y}_2); -\nabla g(\bar{y}_1 - \bar{y}_2); \mathbf{0}_{n \times 1}] + \rho(\bar{y} - \bar{w} - \bar{\lambda}/\rho))^T (y - \bar{y}) \geq 0, \forall y \in Z_2,& \\
\implies \bar{y} \in \operatorname{argmin}_{y \in Z_2} \left[g(y^+ - y^-) + \frac{\rho}{2} \| y - \bar{w} - \bar{\lambda}/\rho \|^2 \right].&
\end{aligned}$$

Thus, by the definition of Z_2 in (22), $\exists \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}^n, \pi \in \mathbb{R}^m$ such that

$$\begin{pmatrix} \nabla g(\bar{y}_1 - \bar{y}_2) \\ -\nabla g(\bar{y}_1 - \bar{y}_2) \\ \mathbf{0}_{n \times 1} \end{pmatrix} + \rho(\bar{y} - \bar{w} - \bar{\lambda}/\rho) + \begin{pmatrix} -\beta_1 - A^T \pi \\ -\beta_2 + A^T \pi \\ \beta_4 - \beta_3 \end{pmatrix} = 0, \tag{69}$$

$$0 \leq \beta_i \perp \bar{y}_i \geq 0, i = 1, 2, 3, \ 0 \leq e - \bar{y}_3 \perp \beta_4 \geq 0, \ 0 \leq A(\bar{y}_1 - \bar{y}_2) - b \perp \pi \geq 0.$$

Note that $\lambda_k \to \bar{\lambda}$ implies that $w_{k+1} - y_{k+1} = (\lambda_{k+1} - \lambda_k)/\rho \to 0$, which further implies that $\bar{w} = \bar{y}$. By combining (68) and (69), letting $\bar{y} = \bar{w}$, and by adding $(\bar{x}^+ + \bar{x}^-)^T \bar{\xi} = 0$, we have exactly the KKT conditions ((17)) at $(\bar{x}^+; \bar{x}^-; \bar{\xi})$ for (4).

(b). If f is convex, then by Theorem 1, $\bar{w} = (\bar{x}^+; \bar{x}^-; \bar{\xi})$ is a local minimum of (4). ∎

Next, suppose $h(w) \triangleq f_Q(x^+ - x^-) + \gamma \sum_{i=1}^n (1 - \xi_i)$, $p(y) \triangleq g(y^+ - y^-)$. Define:

$$\begin{aligned}
\mathcal{L}(w, y, \lambda, \rho) &\triangleq h(w) + p(y) + \lambda^T (w - y) + \frac{\rho}{2} \| w - y \|^2 \\
\mathcal{H}(w, y, \lambda) &\triangleq h(w) + \mathbb{1}_{Z_1}(w) + p(y) + \mathbb{1}_{Z_2}(y) + \lambda^T (w - y)
\end{aligned}$$

$$\mathcal{H}_\rho(w,y,\lambda) \triangleq h(w) + \mathbb{1}_{Z_1}(w) + p(y) + \mathbb{1}_{Z_2}(y) \\ + \lambda^T(w-y) + \frac{\rho}{2}\|w-y\|^2. \tag{70}$$

Then the updates of ($\mathrm{ADMM_{cf}}$) can be rewritten as follows:

$$w_{k+1} := \operatorname*{argmin}_{w \in Z_1} \ \mathcal{L}(w, y_k, \lambda_k, \rho_k) = \operatorname*{argmin}_{w} \ \mathcal{H}_{\rho_k}(w, y_k, \lambda_k), \tag{71}$$

$$y_{k+1} := \operatorname*{argmin}_{y \in Z_2} \ \mathcal{L}(w_{k+1}, y, \lambda_k, \rho_k) = \operatorname*{argmin}_{y} \ \mathcal{H}_{\rho_k}(w_{k+1}, y, \lambda_k), \tag{72}$$

$$\lambda_{k+1} := \lambda_k + \rho_k(w_{k+1} - y_{k+1}).$$

Deriving convergence statements of Algorithm 2 necessitates the following lemma.

Lemma 11 Consider the sequence $\{w_k, y_k, \lambda_k, \rho_k\}$ generated by ($\mathrm{ADMM_{cf}}$). Then for all $k \geq 0$, where $\Delta\mathcal{L}_k \triangleq \mathcal{L}(w_{k+1}, y_{k+1}, \lambda_{k+1}, \rho_{k+1}) - \mathcal{L}(w_k, y_k, \lambda_k, \rho_k)$,

$$\Delta\mathcal{L}_k \leq \left(\frac{1}{\rho_k} + \frac{\rho_{k+1} - \rho_k}{2\rho_k^2}\right)\|\lambda_{k+1} - \lambda_k\|^2 - \frac{\rho_k}{2}\|y_{k+1} - y_k\|^2. \tag{73}$$

Proof From the definition of augmented Lagrangian function,

$$\begin{aligned} &\mathcal{L}(w_{k+1}, y_{k+1}, \lambda_k, \rho_k) - \mathcal{L}(w_{k+1}, y_k, \lambda_k, \rho_k) \\ &\leq -\nabla_y \mathcal{L}(w_{k+1}, y_{k+1}, \lambda_k, \rho_k)^T (y_k - y_{k+1}) - \frac{\rho_k}{2}\|y_{k+1} - y_k\|^2 \\ &\leq -\frac{\rho_k}{2}\|y_{k+1} - y_k\|^2, \end{aligned} \tag{74}$$

where the first inequality follows since $\mathcal{L}(w_{k+1}, y, \lambda_k, \rho_k)$ is ρ_k-strongly convex in y with constant ρ_k, while the second inequality may be derived from the optimality conditions of update (72) whereby $\nabla_y \mathcal{L}(w_{k+1}, y_{k+1}, \lambda_k, \rho_k)^T(y - y_{k+1}) \geq 0, \forall y \in Z_2$. Since w_{k+1} is a minimizer associated with (71), we have that

$$\mathcal{L}(w_{k+1}, y_k, \lambda_k, \rho_k) - \mathcal{L}(w_k, y_k, \lambda_k, \rho_k) \leq 0. \tag{75}$$

By invoking the definition of the augmented Lagrangian function, and utilizing the update rule for λ_{k+1}, i.e. $\lambda_{k+1} = \lambda_k + \rho_k(w_{k+1} - y_{k+1})$,

$$\begin{aligned} &\mathcal{L}(w_{k+1}, y_{k+1}, \lambda_{k+1}, \rho_k) - \mathcal{L}(w_{k+1}, y_{k+1}, \lambda_k, \rho_k) \\ &= (\lambda_{k+1} - \lambda_k)^T (w_{k+1} - y_{k+1}) \\ &= \|\lambda_{k+1} - \lambda_k\|^2/\rho_k \end{aligned} \tag{76}$$

$$\begin{aligned} &\mathcal{L}(w_{k+1}, y_{k+1}, \lambda_{k+1}, \rho_{k+1}) - \mathcal{L}(w_{k+1}, y_{k+1}, \lambda_{k+1}, \rho_k) \\ &= \frac{\rho_{k+1} - \rho_k}{2}\|w_{k+1} - y_{k+1}\|^2 \\ &= \frac{\rho_{k+1} - \rho_k}{2\rho_k^2}\|\lambda_{k+1} - \lambda_k\|^2. \end{aligned} \tag{77}$$

By adding (74), (75), (76), and (77), we obtain the required result. ∎

Proving convergence requires the Kurdyka-Łojasiewicz property and a requirement on the multiplier sequence.

Assumption 4 $\liminf_{k\to+\infty}\|\lambda_k\| < +\infty$.

Theorem 4 Consider the sequence $\{w_k, y_k, \lambda_k, \rho_k\}$ generated by (ADMM$_{\mathrm{cf}}$). Suppose that Assumption 3, 4 hold and $\{y_k\}$ is bounded. Then the following hold:
(i) $\exists K_0 \in \mathbb{N}, s.t. \rho_k \equiv \rho, \forall k \geq K_0$. $\|\lambda_{k+1} - \lambda_k\| \leq C\|y_{k+1} - y_k\|, \forall k \geq K_0$, $C \triangleq \frac{\rho-\delta}{\sqrt{2}}$.
(ii) $\{\mathcal{L}(w_k, y_k, \lambda_k, \rho_k)\}_{k\geq K_0}$ is a non-increasing sequence satisfying

$$\begin{aligned}
&\mathcal{L}(w_{k+1}, y_{k+1}, \lambda_{k+1}, \rho) - \mathcal{L}(w_k, y_k, \lambda_k, \rho)\\
&\leq \left(C^2/\rho - \rho/2\right)\|y_{k+1} - y_k\|^2\\
&\leq -(\delta/2)\|y_{k+1} - y_k\|^2, \quad \forall k \geq K_0.
\end{aligned} \tag{78}$$

(iii) $\{\mathcal{L}(w_k, y_k, \lambda_k, \rho_k)\}_{k\geq K_0}$ is bounded from below. Furthermore, $y_{k+1} - y_k \to 0$, $y_k - w_k \to 0$ as $k \to \infty$ and $\{w_k\}$ is a bounded sequence. Therefore, $\{(w_k; y_k; \lambda_k)\}$ has a convergent subsequence with limit point given by $z^* \triangleq (w^*; y^*; \lambda^*)$.
(iv) Suppose $\mathcal{H}_\rho$ satisfies the KŁ property at z^*. Then $\sum_{k=0}^{\infty}\|y_{k+1} - y_k\| < \infty$.
(v) Suppose $\mathcal{H}_\rho$ satisfies the KŁ property at z^*. Then $(w_k; y_k; \lambda_k)$ converges to z^*, and z^* satisfying the first-order KKT conditions of (4).

Proof **(i)**. By Assumption 3, ρ_k remains unchanged for sufficiently large k, so we denote ρ, K_0 such that $\rho_k \equiv \rho \leq \rho_{\max}, \forall k \geq K_0$. Moreover, by the update rule in step 2 of Alg. 2, $\|\lambda_{k+1} - \lambda_k\| \leq \frac{\rho-\delta}{\sqrt{2}}\|y_{k+1} - y_k\|, \forall k \geq K_0$.
(ii). From Lemma 11 and (i), for $\forall k \geq K_0, \rho_k = \rho$, and

$$\begin{aligned}
&\mathcal{L}(w_{k+1}, y_{k+1}, \lambda_{k+1}, \rho_{k+1}) - \mathcal{L}(w_k, y_k, \lambda_k, \rho_k)\\
&\leq \frac{1}{\rho}\|\lambda_{k+1} - \lambda_k\|^2 - \frac{\rho}{2}\|y_{k+1} - y_k\|^2\\
&\leq \frac{(\rho-\delta)^2}{2\rho}\|y_{k+1} - y_k\|^2 - \frac{\rho}{2}\|y_{k+1} - y_k\|^2\\
&= \frac{\delta^2 - 2\delta\rho}{2\rho}\|y_{k+1} - y_k\|^2\\
&\overset{(\delta<\rho)}{\leq} \left(-\delta + \frac{\delta}{2}\right)\|y_{k+1} - y_k\|^2\\
&= -\frac{\delta}{2}\|y_{k+1} - y_k\|^2.
\end{aligned} \tag{79}$$

Thus, $\{\mathcal{L}(w_k, y_k, \lambda_k, \rho_k)\}_{k\geq K_0}$ is a non-increasing sequence.
(iii). We first show that $\inf_{k\geq 0}\{h(w_k) + p(y_k) + \frac{\rho}{2}\|w_k - y_k\|^2\}$ is finite. Note that

$$h(w) - n\gamma + p(y) + \frac{\rho}{2}\|w - y\|^2$$

$$
\begin{aligned}
&= f_Q(x^+ - x^-) - \gamma e^T \xi + p(y) + \frac{\rho}{2}\|w - y\|^2 \\
&= (x^+ - x^-)^T M (x^+ - x^-) + d^T(x^+ - x^-) - \gamma e^T \xi + p(y) + \frac{\rho}{2}\|w - y\|^2 \\
&= w^T H w + [(d; -d; -\gamma e) - \rho y]^T w + \frac{\rho}{2}\|y\|^2 + p(y) \\
&= \left\| w + \frac{1}{2} H^{-1}[(d; -d; -\gamma e) - \rho y] \right\|_H^2 + \frac{\rho}{2}\|y\|^2 + p(y) \\
&- \frac{1}{4}\|(d; -d; -\gamma e) - \rho y\|_{H^{-1}}^2 \\
&\geq \frac{\rho}{2}\|y\|^2 + p(y) - \frac{1}{4}\|(d; -d; -\gamma e) - \rho y\|_{H^{-1}}^2,
\end{aligned}
$$

where $H \triangleq \begin{pmatrix} M + \frac{\rho}{2}I & -M & \\ -M & M + \frac{\rho}{2}I & \\ & & \frac{\rho}{2}I \end{pmatrix}$ and $H \succ 0$ (Note that $\rho_0 I + 4M \succ 0$ leading to $\rho I + 4M \succ 0$, further implying $H \succ 0$). Since $\{y_k\}$ is bounded by assumption, and $p(y) = g(y^+ - y^-)$ is smooth,

$$
\begin{aligned}
&\inf_{k \geq 0} \left[h(w_k) - n\gamma + p(y_k) + (\rho/2)\|w_k - y_k\|^2 \right] \\
&\geq \inf_{k \geq 0} [(\rho/2)\|y_k\|^2 + p(y_k) - (1/4)\|(d; -d; -\gamma e) - \rho y_k\|_{H^{-1}}^2] > -\infty.
\end{aligned}
$$

If $\bar{L} \triangleq \inf_{k \geq 0}\{h(w_k) + p(y_k) + \frac{\rho}{2}\|w_k - y_k\|^2\}$, then

$$
\begin{aligned}
\mathcal{L}(w_k, y_k, \lambda_k, \rho_k) &\geq \bar{L} + \lambda_k^T (w_k - y_k) \\
&= \bar{L} + \lambda_k^T (\lambda_k - \lambda_{k-1})/\rho \\
&= \bar{L} + \frac{1}{2\rho}(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2 + \|\lambda_k - \lambda_{k-1}\|^2),
\end{aligned}
$$

implying that

$$
\sum_{k=K_0}^{N} (\mathcal{L}(w_k, y_k, \lambda_k, \rho_k) - \bar{L}) \geq \frac{\|\lambda_N\|^2 - \|\lambda_{K_0-1}\|^2}{2\rho} \geq \frac{-\|\lambda_{K_0-1}\|^2}{2\rho} > -\infty,
$$

for all $N \geq K_0$. Since $\{\mathcal{L}(w_k, y_k, \lambda_k, \rho_k) - \bar{L}\}_{k \geq K_0}$ is a non-increasing sequence from (ii), it's nonnegative. Otherwise, $\lim_{N \to +\infty} \sum_{k=K_0}^{N} (\mathcal{L}(w_k, y_k, \lambda_k, \rho_k) - \bar{L}) = -\infty$. This is a contradiction. Therefore, $\{\mathcal{L}(w_k, y_k, \lambda_k, \rho_k)\}_{k \geq K_0}$ is bounded from below. Consequently, $h_k \triangleq \mathcal{H}_\rho(w_k, y_k, \lambda_k)$ is a convergent sequence because $\mathcal{H}_\rho(w_k, y_k, \lambda_k) = \mathcal{L}(w_k, y_k, \lambda_k, \rho) = \mathcal{L}(w_k, y_k, \lambda_k, \rho_k), \forall k \geq K_0$. Without loss of generality, suppose $h_k \to 0$. Then, by summing up (79) for $k \geq K_0$, we have $\sum_{k=K_0}^{\infty} \|y_{k+1} - y_k\|^2 \leq \frac{h_{K_0}}{\delta/2} < \infty$. It follows that $y_{k+1} - y_k \to 0$ as $k \to \infty$. From (i), we also have $\|\lambda_{k+1} - \lambda_k\| \to 0$ as $k \to \infty$. In other words, $\rho\|w_k - y_k\| \to 0$ as $k \to \infty$. But $\{y_k\}$ is a bounded sequence, implying that $\{w_k\}$ is a bounded sequence. By Assumption 4, there exists a convergent subsequence of $\{\lambda_k\}$. Therefore, there exists a subsequence of $\{w_k, y_k, \lambda_k\}$ converging to a point denoted by $\{w^*, y^*, \lambda^*\} \triangleq z^*$.

(iv). Next we prove $\|y_{k+1}-y_k\|$ is summable by using the KŁ inequality. By assumption, $\mathcal{H}_\rho$ admits the KŁ property at z^* and suppose the concave function ψ, a neighborhood U, and a scalar $\eta > 0$ are associated with the KŁ property. Further, suppose $B(z^*, r) \subseteq U$ and denote $z_k = (w_k; y_k; \lambda_k)$. We know that $h_k \to 0$. If for some $k_0 \geq K_0$, $h_{k_0} = 0$, then by (78), $y_k = y_{k+1}, \forall k \geq k_0$, the proof is complete. Therefore, let $h_k > 0, \forall k \geq K_0$. Since a subsequence of $\{z_k\}$ converges to z^*, and $h_k \to 0$, $\exists K \geq K_0 + 1$ such that:

$$\left(\frac{2C}{\rho} + C + 2\right)\sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}} + \left(\frac{C}{\rho} + \frac{C}{2} + 1\right) \times \\ \left[\frac{\psi(h_K)}{C_0} + \left[\frac{\psi(h_K)}{C_0}\left(\frac{\psi(h_K)}{C_0} + 4\sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}}\right)\right]^{1/2}\right] + \|z_K - z^*\| < r, \tag{80}$$

where $h_K < \eta$ and $C_0 = \frac{\frac{\rho}{2} - \frac{C^2}{\rho}}{\frac{C}{\rho} + 2C + \rho}$. Then we inductively prove (81) for $k \geq K + 1$:

$$z_k \in B(z^*, r), \quad \|y_{k-1} - y_{k-2}\| > 0, \quad \frac{C_0\|y_k - y_{k-1}\|^2}{\|y_{k-1} - y_{k-2}\|} \leq \psi(h_{k-1}) - \psi(h_k). \tag{81}$$

We first prove two useful inequalities. From (78), we have (82) for $k \geq K - 1$:

$$\begin{aligned} \|y_{k+1} - y_k\|^2 &\leq \frac{h_k - h_{k+1}}{\rho/2 - C^2/\rho} \leq \frac{h_{K-1}}{\rho/2 - C^2/\rho} \\ \Longrightarrow \|y_{k+1} - y_k\| &\leq \sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}}. \end{aligned} \tag{82}$$

Furthermore, $\|z_{k+1} - z_k\|$ may be bounded as follows for $\forall k \geq K$.

$$\begin{aligned} \|z_{k+1} - z_k\| &= \sqrt{\|w_{k+1} - w_k\|^2 + \|y_{k+1} - y_k\|^2 + \|\lambda_{k+1} - \lambda_k\|^2} \\ &\leq \|w_{k+1} - w_k\| + \|y_{k+1} - y_k\| + \|\lambda_{k+1} - \lambda_k\| \\ &\leq \|w_{k+1} - y_{k+1}\| + \|y_k - w_k\| + 2\|y_{k+1} - y_k\| + \|\lambda_{k+1} - \lambda_k\| \\ &\leq (1/\rho + 1)\|\lambda_{k+1} - \lambda_k\| + \|\lambda_k - \lambda_{k-1}\|/\rho + 2\|y_{k+1} - y_k\| \\ &\overset{(i)}{\leq} C\|y_k - y_{k-1}\|/\rho + (C/\rho + C + 2)\|y_{k+1} - y_k\|. \end{aligned} \tag{83}$$

We utilize these inequalities to show (81) by induction.
$\boldsymbol{K+1}$: Through (80), (82), (83), the following holds

$$\begin{aligned} &\|z_{K+1} - z^*\| \leq \|z_{K+1} - z_K\| + \|z_K - z^*\| \\ &\overset{(83)}{\leq} C\|y_K - y_{K-1}\|/\rho + (C/\rho + C + 2)\|y_{K+1} - y_K\| + \|z_K - z^*\| \end{aligned}$$

$$\overset{(82)}{\leq} \left(\frac{2C}{\rho} + C + 2\right)\sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}} + \|z_K - z^*\| \overset{(80)}{<} r,$$

implying $z_{K+1} \in B(z^*, r)$. From the optimality conditions of (71) and (72),

$$\begin{aligned}
&0 \in \nabla_w h(w_K) + \lambda_{K-1} + \rho(w_K - y_{K-1}) + \partial \mathbb{1}_{Z_1}(w_K) \\
\Longrightarrow\ & \Delta\lambda_K - \rho\Delta y_K \in \nabla_w h(w_K) + \lambda_K + \rho(w_K - y_K) + \partial \mathbb{1}_{Z_1}(w_K) \qquad (84) \\
&0 \in \nabla_y p(y_K) - \lambda_{K-1} + \rho(y_K - w_K) + \partial \mathbb{1}_{Z_2}(y_K) \\
\Longrightarrow\ & -\Delta\lambda_K \in \nabla_y p(y_K) - \lambda_K + \rho(y_K - w_K) + \partial \mathbb{1}_{Z_2}(y_K), \qquad (85)
\end{aligned}$$

where $\Delta\lambda_K \triangleq \lambda_K - \lambda_{K-1}, \Delta y_K \triangleq y_K - y_{K-1}$. So (84), (85) and the fact that

$$\begin{aligned}
\partial\mathcal{H}_\rho(z_K) = &(\nabla_w h(w_K) + \lambda_K + \rho(w_K - y_K) + \partial \mathbb{1}_{Z_1}(w_K)) \\
&\times (\nabla_y p(y_K) - \lambda_K + \rho(y_K - w_K) + \partial \mathbb{1}_{Z_2}(y_K)) \times (w_K - y_K)
\end{aligned}$$

imply that $\partial\mathcal{H}_\rho(z_K) \ni \big(\Delta\lambda_K - \rho\Delta y_K; -\Delta\lambda_K; w_K - y_K\big)$ and

$$\begin{aligned}
&\mathrm{dist}(0, \partial\mathcal{H}_\rho(z_K)) \leq \sqrt{\|\Delta\lambda_K - \rho\Delta y_K\|^2 + \|\Delta\lambda_K\|^2 + \|w_K - y_K\|^2} \\
&\leq \|\Delta\lambda_K - \rho\Delta y_K\| + \|\Delta\lambda_K\| + \|w_K - y_K\| \\
&\leq (1/\rho + 2)\|\Delta\lambda_K\| + \rho\|\Delta y_K\| \overset{(i)}{\leq} (C/\rho + 2C + \rho)\|\Delta y_K\| \qquad (86) \\
\Longrightarrow & -\Delta\psi(h_{K+1}) \geq \psi'(h_K)(h_K - h_{K+1}) \overset{(78)}{\geq} \psi'(h_K)\left(\frac{\rho}{2} - \frac{C^2}{\rho}\right)\|\Delta y_{K+1}\|^2 \\
&\geq \frac{(\frac{\rho}{2} - \frac{C^2}{\rho})\|\Delta y_{K+1}\|^2}{\mathrm{dist}(0, \partial\mathcal{H}_\rho(z_K))} \overset{(86)}{\geq} \frac{(\frac{\rho}{2} - \frac{C^2}{\rho})\|\Delta y_{K+1}\|^2}{(\frac{C}{\rho} + 2C + \rho)\|\Delta y_K\|} = \frac{C_0\|\Delta y_{K+1}\|^2}{\|\Delta y_K\|}, \qquad (87)
\end{aligned}$$

where $\Delta\psi(h_{K+1}) \triangleq \psi(h_{K+1}) - \psi(h_K)$, the first inequality of (87) follows from the concavity of ψ, and the third inequality is due to the KŁ inequality $\psi'(h_K)$ $\mathrm{dist}(0, \partial\mathcal{H}_\rho(z_K)) \geq 1$ because $\|z_K - z^*\| < r$. Moreover, the KŁ inequality indicates that $\mathrm{dist}(0, \partial\mathcal{H}_\rho(z_K)) > 0$, thus $\|\Delta y_K\| > 0$ by (86). Therefore, the inductive hypothesis holds for $K+1$. Assume that it holds for $K+2, \ldots, k$ and consider $k+1$.

$\boldsymbol{k+1}$: We begin by showing that:

$$\sum_{i=K}^{k-1} \|y_{i+1} - y_i\| \leq \frac{1}{2}\left[\frac{\psi(h_K)}{C_0} + \left[\frac{\psi(h_K)}{C_0}\left(\frac{\psi(h_K)}{C_0} + 4\sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}}\right)\right]^{1/2}\right]. \qquad (88)$$

Combining inductive hypothesis (81) for $K+1, \ldots, k$, we have that

$$\begin{aligned}
&\psi(h_K) - \psi(h_k) \geq C_0 \sum_{i=K}^{k-1} \frac{\|y_{i+1} - y_i\|^2}{\|y_i - y_{i-1}\|} \\
\Longrightarrow\ & \frac{\psi(h_K) - \psi(h_k)}{C_0} \sum_{i=K}^{k-1} \|y_i - y_{i-1}\| \geq \left(\sum_{i=K}^{k-1} \frac{\|y_{i+1} - y_i\|^2}{\|y_i - y_{i-1}\|}\right)\left(\sum_{i=K}^{k-1} \|y_i - y_{i-1}\|\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(*)}{\geq} \left(\sum_{i=K}^{k-1} \|y_{i+1} - y_i\| \right)^2 \\
\implies &\left(\sum_{i=K}^{k-1} \|y_{i+1} - y_i\| \right)^2 \leq \frac{\psi(h_K)}{C_0} \sum_{i=K-1}^{k-2} \|y_{i+1} - y_i\| \\
&\leq \frac{\psi(h_K)}{C_0} \sum_{i=K-1}^{k-1} \|y_{i+1} - y_i\| \\
&= \frac{\psi(h_K)}{C_0} \left(\sum_{i=K}^{k-1} \|\Delta y_{i+1}\| + \|\Delta y_K\| \right) \\
&\overset{(82)}{\leq} \frac{\psi(h_K)}{C_0} \left(\sum_{i=K}^{k-1} \|\Delta y_{i+1}\| + \sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}} \right),
\end{aligned}
$$

where $(*)$ holds because of Hölder's inequality. If $x \triangleq \sum_{i=K}^{k-1} \|y_{i+1} - y_i\|$, $C_1 \triangleq \frac{\psi(h_K)}{C_0}$, and $C_2 \triangleq \frac{\psi(h_K)}{C_0}\sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}}$, then the above inequality is equivalent to $x^2 - C_1 x - C_2 \leq 0 \implies x \leq \frac{1}{2}\left(C_1 + \sqrt{C_1^2 + 4C_2}\right)$. This is exactly (88). Therefore,

$$
\begin{aligned}
&\|z_{k+1} - z_K\| \leq \sum_{i=K}^{k} \|z_{i+1} - z_i\| \\
&\overset{(83)}{\leq} \sum_{i=K}^{k} \left(\frac{\|y_i - y_{i-1}\|}{\rho/C} + \left(\frac{C}{\rho} + C + 2\right) \|y_{i+1} - y_i\| \right) \\
&= \frac{C\|y_K - y_{K-1}\|}{\rho} + \sum_{i=K}^{k-1} \left(\frac{2C}{\rho} + C + 2\right) \|y_{i+1} - y_i\| \\
&+ \left(\frac{C}{\rho} + C + 2\right) \|y_{k+1} - y_k\| \\
&\overset{(82)(88)}{\leq} \left(\frac{2C}{\rho} + C + 2\right) \sqrt{\frac{h_{K-1}}{\rho/2 - C^2/\rho}} + \left(\frac{C}{\rho} + \frac{C}{2} + 1\right) \left(C_1 + \sqrt{C_1^2 + 4C_2}\right) \\
&\implies \|z_{k+1} - z^*\| \leq \|z_{k+1} - z_K\| + \|z_K - z^*\| \overset{(80)}{<} r.
\end{aligned}
$$

Thus, $z_{k+1} \in B(z^*, r)$. Since $z_k \in B(z^*, r)$, the KŁ inequality holds for z_k, and in a fashion similar to (87), we obtain that $\|y_k - y_{k-1}\| > 0$ and $\frac{C_0\|y_{k+1} - y_k\|^2}{\|y_k - y_{k-1}\|} \leq \psi(h_k) - \psi(h_{k+1})$, completing the proof of the inductive hypothesis. By the hypothesis, (88) holds for $k \geq K + 1$. This indicates that $\sum_{i=K}^{+\infty} \|y_{i+1} - y_i\| < +\infty$, implying that $\sum_{i=0}^{+\infty} \|y_{i+1} - y_i\| < +\infty$.

(v). From (iv) and by recalling that $\|\lambda_{k+1} - \lambda_k\| \leq C\|y_{k+1} - y_k\|$ for k sufficiently large, we have that $\{y_k\}$ and $\{\lambda_k\}$ are Cauchy sequences, convergent to y^* and λ^*, respectively. Since $w_k - y_k \to 0$, $\{w_k\}$ also converges to $w^* = y^*$. By Proposition 3, (w^*, y^*, λ^*) is a KKT point. ∎

Remark 8 (i). To derive convergence of the sequence, we leverage the KŁ property of $\mathcal{H}_\rho$. When $p(y)$ is semialgebraic [1, Sec. 4.3], $\mathcal{H}_\rho$ is a sum of semialgebraic functions and is therefore semialgebraic. Then the result follows from [1, Sec. 4.3] whereby a semialgebraic function $\mathcal{L}$ satisfies the KŁ property at every point in $\mathrm{dom}(\partial\mathcal{L})$.
(ii). If we cannot invoke the KŁ property to show convergence of $\{(w_k, y_k, \lambda_k)\}$, we may merely conclude that any cluster point of $\{(w_k, y_k, \lambda_k)\}$ satisfies first order KKT conditions of (4). The proof is similar to Proposition 3 thus omitted.
(iii). Boundedness of $\{y_k\}$ holds by assuming compactness of Z_2 (obtainable by adding constraints $x^+ \le ub^+, x^- \le ub^-$).