Sparsity Constrained Minimization via Mathematical Programming with Equilibrium Constraints

Ganzhao Yuan [*]   Bernard Ghanem [†]

Abstract

Sparsity constrained minimization captures a wide spectrum of applications in both machine learning and signal processing. This class of problems is difficult to solve since it is NP-hard and existing solutions are primarily based on Iterative Hard Thresholding (IHT). In this paper, we consider a class of continuous optimization techniques based on Mathematical Programs with Equilibrium Constraints (MPECs) to solve general sparsity constrained problems. Specifically, we reformulate the problem as an equivalent biconvex MPEC, which we can solve using an exact penalty method or an alternating direction method. We elaborate on the merits of both proposed methods and analyze their convergence properties. Finally, we demonstrate the effectiveness and versatility of our methods on several important problems, including feature selection, segmented regression, MRF optimization, trend filtering and impulse noise removal. Extensive experiments show that our MPEC-based methods outperform state-of-the-art techniques, especially those based on IHT.

Keywords: Sparsity Constrained Optimization, Exact Penalty Method, Alternating Direction Method, MPEC, Convergence Analysis

1. Introduction

In this paper, we mainly focus on the following generalized sparsity constrained minimization problem:

$$\min_{\mathbf{x}} \ f(\mathbf{x}), \ s.t. \ \|\mathbf{Ax}\|_0 \leq k \qquad (1)$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and k $(0 < k < m)$ is a positive integer. $\|\cdot\|_0$ is a function that counts the number of nonzero elements in a vector. To guarantee convergence, we assume that $f(\cdot)$ is convex and L-Lipschitz continuous (but not necessarily smooth), $\mathbf{A}$ has right inverse (i.e. $rank(\mathbf{A}) = m$), and there always exists bounded solutions to Eq (1).

The optimization in Eq (1) describes many applications of interest in both machine learning and signal processing, including compressive sensing [19], Dantzig selector [11], feature selection [51], trend filtering [38], image restoration [65, 18], sparse classification and boosting [59, 61], subspace clustering [21], sparse coding [40, 4], portfolio selection, image smoothing [62], sparse inverse covariance estimation [26, 67], blind source separation[41], permutation problems [25, 36], joint power and admission control [42], Potts statistical functional [58], to name a few. Moreover, we notice that many binary optimization problems [57] can be reformulated as an ℓ_0 norm optimization problem, since $\mathbf{x} \in \{-1, +1\}^n \Leftrightarrow \|\mathbf{x} - \mathbf{1}\|_0 + \|\mathbf{x} + \mathbf{1}\|_0 \leq n$. Thus, it can be rewritten in a more compact form as $\|\mathbf{Ax} - \mathbf{b}\|_0 \leq n$ with $\mathbf{A} = [\mathbf{I}_n, \mathbf{I}_n]^T \in \mathbb{R}^{2n \times n}$ and $\mathbf{b} = [-\mathbf{1}^T, \mathbf{1}^T]^T \in \mathbb{R}^{2n}$.

A popular method to solve Eq (1) is to consider its ℓ_1 norm convex relaxation, which simply replaces the ℓ_0 norm function[1] with the ℓ_1 norm function. When $\mathbf{A}$ is identity and $f(\cdot)$ is a least squares objective function, it has been proven that when the sampling matrix in $f(\cdot)$ satisfies certain incoherence conditions and $\mathbf{x}$ is sparse at its optimal solution, the problem can be solved exactly by this convex method [12]. However, when these two assumptions do not hold, this convex method can be far from satisfactory [65, 42].

Recently, another breakthrough in sparse optimization is the multi-stage convex relaxation method [70, 13]. In the work of [70], the local solution obtained by this method is shown to be superior to the global solution of the standard ℓ_1 convex method, which is used as its initialization. However, this type of method seeks an approximate solution to the sparse *regularized* problem with a smooth objective, and cannot solve the sparsity *constrained* problem [2] considered here. Our method uses the same convex initialization strategy, but it is applicable to general ℓ_0 norm minimization problems.

Challenges and Contributions: We recognize three main challenges hindering existing work. **(a)** The general prob-

---

[*]King Abdullah University of Science & Technology (KAUST). Email: yuanganzhao@gmail.com.

[†]King Abdullah University of Science & Technology (KAUST). Email: bernard.ghanem@kaust.edu.sa.

[1]Strictly speaking, $\|\cdot\|_0$ is not a norm (but a pseudo-norm) since it is not positive homogeneous.

[2]Some may solve the constrained problem by tuning the continuous parameter of regularized problem, but it can be hard when m is large.

Table 1: ℓ_0 norm optimization techniques.

Method and Representative Reference	Description
greedy descent methods [46]	only for smooth (typically quadratic) objective
ℓ_1 norm relaxation [12]	$\|\mathbf{x}\|_0 \approx \|\mathbf{x}\|_1$
k-support norm relaxation [1]	$\|\mathbf{x}\|_0 \approx \|\mathbf{x}\|_{\text{k-sup}} \triangleq \max_{\mathbf{0} < \mathbf{v} \leq \mathbf{1}} \ (\sum_i \mathbf{x}_i^2/\mathbf{v}_i)^{1/2}, \ s.t. \ \langle \mathbf{v}, \mathbf{1} \rangle \leq k$
k-largest norm relaxation [64]	$\|\mathbf{x}\|_0 \approx \|\mathbf{x}\|_{\text{k-lar}} \triangleq \max_{-\mathbf{1} \leq \mathbf{v} \leq \mathbf{1}} \ \langle \mathbf{v}, \mathbf{x} \rangle, \ s.t. \ \|\mathbf{v}\|_1 \leq k$
SOCP convex relaxation [14]	$\|\mathbf{x}\|_0 \leq k \Rightarrow \|\mathbf{x}\|_1 \leq \sqrt{k}\|\mathbf{x}\|_2$
SDP convex relaxation [14]	$\|\mathbf{x}\|_0 \leq k \Rightarrow \mathbf{X} = \mathbf{x}\mathbf{x}^T, \ \|\mathbf{X}\|_1 \leq k tr(\mathbf{X})$
Schatten ℓ_p approximation [29]	$\|\mathbf{x}\|_0 \approx \|\mathbf{x}\|_p$
re-weighted ℓ_1 approximation [13]	$\|\mathbf{x}\|_0 \approx \langle \mathbf{1}, \ln(|\mathbf{x}| + \epsilon) \rangle$
$\ell_{1\text{-}2}$ DC approximation [63]	$\|\mathbf{x}\|_0 \approx \|\mathbf{x}\|_1 - \|\mathbf{x}\|_2$
0-1 mixed integer programming [8]	$\{\|\mathbf{x}\|_0 : \|\mathbf{x}\|_\infty \leq \lambda\} \Leftrightarrow \{\min_{\mathbf{v} \in \{0,1\}} \ \langle \mathbf{1}, \mathbf{v} \rangle : |\mathbf{x}| \leq \lambda \mathbf{v}\}$
iterative hard shreadholding [44]	$0.5\|\mathbf{x} - \mathbf{x}'\|_2^2, \ s.t. \ \|\mathbf{x}\|_0 \leq k$
non-separable MPEC [This paper]	$\|\mathbf{x}\|_0 = \min_{\mathbf{u}} \ \|\mathbf{u}\|_1, \ s.t. \ \|\mathbf{x}\|_1 = \langle \mathbf{x}, \mathbf{u} \rangle, \ -\mathbf{1} \leq \mathbf{u} \leq \mathbf{1}$
separable MPEC [65] [This paper]	$\|\mathbf{x}\|_0 = \min_{\mathbf{v}} \ \langle \mathbf{1}, \mathbf{1} - \mathbf{v} \rangle, \ s.t. \ |\mathbf{x}| \odot \mathbf{v} = \mathbf{0}, \ \mathbf{0} \leq \mathbf{v} \leq \mathbf{1}$

lem in Eq (1) is NP-hard. There is little hope of finding the global minimum efficiently in all instances. In order to deal with this issue, we reformulate the ℓ_0 norm minimization problem as an equivalent augmented optimization problem with a bilinear equality constraint using a variational characterization of the ℓ_0 norm function. Then, we propose two penalization/regularization methods (exact penalty and alternating direction) to solve it. The resulting algorithms seek a desirable exact solution to the original optimization problem. **(b)** Many existing convergence results for non-convex ℓ_0 minimization problems tend to be limited to unconstrained problems or inapplicable to constrained optimization. We carefully analyze both our proposed algorithms and prove that they always converge to a first-order KKT point. To the best of our knowledge, this is the first attempt to solve general ℓ_0 constrained minimization with guaranteed convergence. **(c)** Exact methods (e.g. methods using IHT) can produce unsatisfactory results, while approximation methods such as Schatten's ℓ_p norm method [29] and re-weighted ℓ_1 [13] fail to control the sparsity of the solution. In comparison, experimental results show that our MPEC-based methods are more effective than the state-of-the-art.

Organization and Notations: This paper is organized as follows. Section 2 provides a brief description of the related work. Section 3 and Section 4 present our proposed exact penalty method and alternating direction method optimization framework, respectively. Section 5 discusses some features of the proposed two algorithms. Section 6 summarizes the experimental results. Finally, Section 7 concludes this paper. Throughout this paper, we use lowercase and uppercase boldfaced letters to denote real vectors and matrices respectively. We use $\langle \mathbf{x}, \mathbf{y} \rangle$ and $\mathbf{x} \odot \mathbf{y}$ to denote the Euclidean inner product and elementwise product between

$\mathbf{x}$ and $\mathbf{y}$. "$\triangleq$" means define. $\sigma(\mathbf{A})$ denotes the smallest singular value of $\mathbf{A}$.

2. Related Work

There are mainly four classes of ℓ_0 norm minimization algorithms in the literature: **(i)** greedy descent methods, **(ii)** convex approximate methods, **(iii)** non-convex approximate methods, and **(iv)** exact methods. We summarize the main existing algorithms in Table 1.

Greedy descent methods. They have a monotonically decreasing property and optimality guarantees in some cases [55], but they are limited to solving problems with smooth objective functions (typically the square function). **(a)** Matching pursuit [46] selects at each step one atom of the variable that is the most correlated with the residual. **(b)** Orthogonal matching pursuit [54] uses a similar strategy but also creates an orthonormal set of atoms to ensure that selected components are not introduced in subsequent steps. **(c)** Gradient pursuit [9] is similar to matching pursuit, but it updates the sparse solution vector at each iteration with a directional update computed based on gradient information. **(d)** Gradient support pursuit [3] iteratively chooses the index with the largest magnitude as the pursuit direction while maintaining the stable restricted Hessian property. **(e)** Other greedy methods has been proposed, including basis pursuit [15], regularized orthogonal matching pursuit [50], compressive sampling matching pursuit [49], and forward-backward greedy method [71, 53].

Convex approximate methods. They seek convex approximate reformulations of the ℓ_0 norm function. **(a)** The ℓ_1 norm convex relaxation[12] provides a convex lower bound of the ℓ_0 norm function in the unit ℓ_∞ norm. It has been proven that under certain incoherence assumptions, this method leads to a near optimal sparse solution. How-

ever, such assumptions may be violated in real applications. **(b)** k-support norm provides the tightest convex relaxation of sparsity combined with an ℓ_2 penalty. Moreover, it is tighter than the elastic net by exactly a factor of $\sqrt{2}$ [1, 48]. **(c)** k-largest norm provides the tightest convex Boolean relaxation in the sense of minimax game [64, 52, 16]. **(d)** Other convex Second-Order Cone Programming (SOCP) relaxations and Semi-Definite Programming (SDP) relaxations [14] have been considered in the literature.

Non-convex approximate methods. They seek non-convex approximate reformulations of the ℓ_0 norm function. **(a)** Schatten ℓ_p norm with $p \in (0, 1)$ was considered by [29] to approximate the discrete ℓ_0 norm function. It results in a local gradient Lipschitz continuous function, to which some smooth optimization algorithms can be applied. **(b)** Re-weighted ℓ_1 norm [13] minimizes the first-order Taylor series expansion of the objective function iteratively to find a local minimum. It is expected to refine the ℓ_0 regularized problem, since its first iteration is equivalent to solving the ℓ_1 norm problem. **(c)** $\ell_{1\text{-}2}$ DC (difference of convex) approximation [63] was considered for sparse recovery. Exact stable sparse recovery error bounds were established under a restricted isometry property condition. **(d)** Other non-convex surrogate functions of the ℓ_0 function have been proposed, including Smoothly Clipped Absolute Deviation (SCAD) [23], Logarithm [27], Minimax Concave Penalty (MCP) [69], Capped ℓ_1 [70], Exponential-Type [28], Half-Quadratic [30], Laplace [56], and MC+ [47]. Please refer to [43, 31] for a detailed summary and discussion.

Exact methods. Despite the success of approximate methods, they are unappealing in cases when an exact solution is required. Therefore, many researchers have sought out exact reformulations of the ℓ_0 norm function. **(a)** 0-1 mixed integer programming [8, 6] assumes the solution has bound constraints. It can be solved by a tailored branch-and-bound algorithm, where the cardinality constraint is replaced by a surrogate constraint. **(b)** Hard thresholding [44, 5, 68, 35] iteratively sets the smallest (in magnitude) values to zero in a gradient descent format. It has been incorporated into the Quadratic Penalty Method (QPM) [44] and Mean Doubly Alternating Direction Method [18] (MD-ADM). However, we found they often converge to unsatisfactory results in practise. **(c)** The MPEC reformulation [65] considers the complementary system of the ℓ_0 norm problem by introducing additional dual variables and minimizing the complimentary error of the MPEC problem. However, it is limited to ℓ_0 norm regularized problems and the convergence results are weak.

From above, we observe that existing methods either produce approximate solutions or are limited to smooth objectives $f(.)$ and when $\mathbf{A} = \mathbf{I}$. The only existing general purpose exact method for solving Eq (1) is the quadratic penalty method [44] and the mean doubly alternating di-

rection method [18]. However, they often produce unsatisfactory results in practise. The unappealing shortcomings of existing solutions and renewed interests in MPECs [65, 24] motivate us to design new MPEC-based algorithms and convergence results for sparsity constrained minimization.

3. Exact Penalty Method

In this section, we present an exact penalty method [45, 34, 37, 7] for solving the problem in Eq (1). This method is based on a new equivalent non-separable MPEC reformulation of the ℓ_0 norm function.

3.1. Non-Separable MPEC Reformulation

First of all, we present a new non-separable MPEC reformulation.

Lemma 1. *For any given* $\mathbf{x} \in \mathbb{R}^n$*, it holds that*

$$\|\mathbf{x}\|_0 = \min_{-\mathbf{1} \leq \mathbf{u} \leq \mathbf{1}} \|\mathbf{u}\|_1, \ s.t. \ \|\mathbf{x}\|_1 = \langle \mathbf{u}, \mathbf{x} \rangle, \qquad (2)$$

and $\mathbf{u}^* = \mathrm{sign}(\mathbf{x})$ *is the unique optimal solution of Eq(2). Here, the standard signum function sign is applied componentwise, and* $\mathrm{sign}(0) = 0$*.*

Proof. Refer to the **Appendix** A. □

Using Lemma 1, we can rewrite Eq (1) in an equivalent form as follows.

$$\min_{\mathbf{x},\mathbf{u}} \ f(\mathbf{x}) + I(\mathbf{u}), \ s.t. \ \|\mathbf{A}\mathbf{x}\|_1 = \langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle \qquad (3)$$

where $I(\mathbf{u})$ is an indicator function on Θ with

$$I(\mathbf{u}) = \begin{cases} 0, \ \mathbf{u} \in \Theta \\ \infty, \ \mathbf{u} \notin \Theta, \end{cases} \ \Theta \triangleq \{\mathbf{u}| \ \|\mathbf{u}\|_1 \leq k, \ -\mathbf{1} \leq \mathbf{u} \leq \mathbf{1}\}.$$

3.2. Proposed Optimization Framework

We now give a detailed description of our solution algorithm to the optimization in Eq (3). Our solution is based on the exact penalty method, which penalizes the complementary error directly by a penalty function. The resulting objective $\mathcal{J} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined in Eq (4), where ρ is the tradeoff penalty parameter that is iteratively increased to enforce the constraints.

$$\mathcal{J}_\rho(\mathbf{x}, \mathbf{u}) \triangleq f(\mathbf{x}) + I(\mathbf{u}) + \rho(\|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle) \qquad (4)$$

In each iteration, ρ is fixed and we use a proximal point method [2, 10] to minimize over $\mathbf{x}$ and $\mathbf{u}$ in an alternating fashion. Details of this exact penalty method are in Algorithm 1. Note that the parameter T is the number of inner iterations for solving the bi-convex problem. We make the following observations about the algorithm.

3

(a) Initialization. We initialize $\mathbf{u}^0$ to $\mathbf{0}$. This is for the sake of finding a reasonable local minimum in the first iteration, as it reduces to a convex ℓ_1 norm minimization problem for the $\mathbf{x}$-subproblem.

(b) Exact property. The key feature of this method is the boundedness of the penalty parameter ρ (see Theorem 1). Therefore, we terminate the optimization when the threshold is reached (see Eq (9)). This distinguishes it from the quadratic penalty method [44], where the penalty may become arbitrarily large for non-convex problems.

(c) u-Subproblem. Variable $\mathbf{u}$ in Eq (8) is updated by solving the following problem:

$$\mathbf{u}^{t+1} = \arg\min_{-\mathbf{1} \leq \mathbf{u} \leq \mathbf{1}} \frac{1}{2}\|\mathbf{u} - \mathbf{z}\|_2^2, \ s.t. \ \|\mathbf{u}\|_1 \leq k, \qquad (5)$$

where $\mathbf{z} = \rho \mathbf{A}\mathbf{x}^{t+1}/\mu$ and μ is a non-negative proximal constant. Due to the symmetry in the objective and constraints, we can without loss of generality assume that $\mathbf{z} \geq 0$, and flip signs of the resulting solution at the end. As such, Eq (5) can be solved by

$$\bar{\mathbf{u}}^{t+1} = \arg\min_{\mathbf{0} \leq \mathbf{u} \leq \mathbf{1}} \frac{1}{2}\|\mathbf{u} - |\mathbf{z}|\|_2^2, \ s.t. \ \langle \mathbf{u}, \mathbf{1} \rangle \leq k, \qquad (6)$$

where the solution in Eq (5) can be recovered via $\mathbf{u}^{t+1} = sign(\mathbf{z}) \odot \bar{\mathbf{u}}^{t+1}$. Eq (6) can be solved *exactly* in $n\log(n)$ time using the breakpoint search algorithm [33]. Note that this algorithm includes the simplex projection [20] as a special case. For completeness, we also include a Matlab code in **Appendix** D.

(d) x-Subproblem. Variable $\mathbf{x}$ in Eq (7) is updated by solving a ℓ_1 norm convex problem which has no closed-form solution. However, it can be solved using existing ℓ_1 norm optimization solvers such as classical or linearized ADM [32] with convergence rate $\mathcal{O}(1/t)$.

3.3. Theoretical Analysis

In Theorem 1, we show that when the penalty parameter ρ is larger than a threshold, the biconvex objective function in Eq (4) is equivalent to the original constrained MPEC problem in Eq (1). Moreover, Theorem 2 characterizes the convergence rate of our exact penalty method and determines the number of iterations beyond which a certain accuracy is guaranteed.

Theorem 1. *Exactness of the Penalty Function. The penalty problem* $\min_{\mathbf{x},\mathbf{v}} \ \mathcal{J}_\rho(\mathbf{x}, \mathbf{v})$ *admits the same local and global optimal solutions as the original problem in Eq (3) when* $\rho > L/\sigma(\mathbf{A})$.

Proof. Refer to **Appendix** B. □

Theorem 2. *Convergence rate of Algorithm 1. Assume that* $\|\mathbf{x}^t\| \leq \delta$ *for all* t. *Algorithm 1 will converge to the*

Algorithm 1 MPEC-EPM: An Exact Penalty Method for Solving MPEC Problem (3)

(S.0) Initialize $\mathbf{x}^0 = 0 \in \mathbb{R}^n$, $\mathbf{u}^0 = \mathbf{0} \in \mathbb{R}^m$, $\rho^0 > 0$. Set $t = 0$ and $\mu > 0$.

(S.1) Solve the following $\mathbf{x}$-subproblem:

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \ \mathcal{J}(\mathbf{x}, \mathbf{u}^t) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 \qquad (7)$$

(S.2) Solve the following $\mathbf{u}$-subproblem:

$$\mathbf{u}^{t+1} = \arg\min_{\mathbf{u}} \ \mathcal{J}(\mathbf{x}^{t+1}, \mathbf{u}) + \frac{\mu}{2}\|\mathbf{u} - \mathbf{u}^t\|_2^2 \qquad (8)$$

(S.3) Update the penalty in every T iterations:

$$\rho^{t+1} = \min(L/\sigma(\mathbf{A}), 2\rho^t) \qquad (9)$$

(S.4) Set $t := t + 1$ and then go to Step (S.1)

---

first-order KKT point in at most $\lceil (\ln(L\delta) - \ln(\epsilon\rho^0))/\ln 2 \rceil$ *outer iterations* [3] *with the accuracy at least* $\|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{v} \rangle \leq \epsilon$.

Proof. Refer to **Appendix** B. □

4. Alternating Direction Method

This section presents a proximal alternating direction method (PADM) for solving Eq (1). This is mainly motivated by the recent popularity of ADM in the non-convex optimization literature. One direct solution is to apply ADM on the MPEC problem in Eq (3). However, this strategy may not be appealing, since it introduces a non-separable structure with no closed form solution for its sub-problems. Instead, we consider a separable MPEC reformulation used in [65, 66].

4.1. Separable MPEC Reformulations

Lemma 2. *[65] For any given* $\mathbf{x} \in \mathbb{R}^n$, *it holds that*

$$\|\mathbf{x}\|_0 = \min_{\mathbf{0} \leq \mathbf{v} \leq \mathbf{1}} \ \langle \mathbf{1}, \mathbf{1} - \mathbf{v} \rangle, \ s.t. \ \mathbf{v} \odot |\mathbf{x}| = \mathbf{0}, \qquad (10)$$

and $\mathbf{v}^* = \mathbf{1} - sign(|\mathbf{x}|)$ *is the unique solution of Eq(10).*

Using Lemma 2, we can rewrite Eq (1) in an equivalent form as follows.

$$\min_{\mathbf{x},\mathbf{v}} \ f(\mathbf{x}) + I(\mathbf{v}), \ s.t. \ |\mathbf{A}\mathbf{x}| \odot \mathbf{v} = \mathbf{0} \qquad (11)$$

where $I(\mathbf{v})$ is an indicator functionon Ω with

$$I(\mathbf{v}) = \begin{cases} 0, & \mathbf{v} \in \Omega \\ \infty, & \mathbf{v} \notin \Omega \end{cases}, \quad \Omega \triangleq \{\mathbf{v} | \langle \mathbf{1}, \mathbf{1} - \mathbf{v} \rangle \leq k, \ \mathbf{0} \leq \mathbf{v} \leq \mathbf{1}\},$$

---
[3]Every time we increase ρ, we call it one outer iteration.

4.2. Proposed Optimization Framework

To solve Eq (11) using PADM, we form the augmented Lagrangian function $\mathcal{L}(.)$ in Eq (12)

$$\mathcal{L}(\mathbf{x}, \mathbf{v}, \boldsymbol{\pi}) \triangleq f(\mathbf{x}) + I(\mathbf{v}) + \langle |\mathbf{Ax}| \odot \mathbf{v}, \boldsymbol{\pi} \rangle + \frac{\alpha}{2} \||\mathbf{Ax}| \odot \mathbf{v}\|^2 \qquad (12)$$

where $\boldsymbol{\pi}$ is the Lagrange multiplier associated with the constraint $|\mathbf{Ax}| \odot \mathbf{v} = \mathbf{0}$, and $\alpha > 0$ is the penalty parameter. We detail the PADM iteration steps for Eq (11) in Algorithm 2, which has the following properties.

(a) Initialization. We set $\mathbf{v}^0 = \mathbf{1}$ and $\boldsymbol{\pi}^0 = \eta\mathbf{1}$, where η is a small parameter. This finds a reasonable local minimum in the first iteration, as it reduces to an ℓ_1 norm minimization problem for the $\mathbf{x}$-subproblem.

(b) Monotone and boundedness property. For any feasible solution $\mathbf{v}$ in Eq (15), it holds that $|\mathbf{Ax}| \odot \mathbf{v} \geq \mathbf{0}$. Using the fact that $\alpha^t > 0$ and due to the $\boldsymbol{\pi}^t$ update rule, $\boldsymbol{\pi}^t$ is monotone increasing. If we initialize $\boldsymbol{\pi}^0 > 0$ in the first iteration, $\boldsymbol{\pi}$ is always positive. Another key feature of this method is the boundedness of the multiplier $\boldsymbol{\pi}$ (see Theorem 3). It reduces to alternating minimization algorithm for biconvex optimization problem [2, 10].

(c) v-Subproblem. Variable $\mathbf{v}$ in Eq (15) is updated by solving the following problem:

$$\mathbf{v}^{t+1} = \arg\min_{\mathbf{v}} \frac{1}{2}\mathbf{v}^T \mathbf{D}\mathbf{v} + \langle \mathbf{v}, \mathbf{b} \rangle, \; s.t. \; \mathbf{v} \in \Omega \qquad (13)$$

where $\mathbf{b} \triangleq \boldsymbol{\pi}^t \odot |\mathbf{Ax}^{t+1}| - \mu\mathbf{v}^t$ and $\mathbf{D}$ is a diagonal matrix with $\mathbf{d} \triangleq \alpha|\mathbf{Ax}^{t+1}| \odot |\mathbf{Ax}^{t+1}| + \mu$ in the main diagonal entries. Introducing the proximal term in the $\mathbf{v}$-subproblem leads to a strongly convex problem. Thus, it can be solved *exactly* using [33].

(d) x-Subproblem. Variable $\mathbf{x}$ in Eq (14) is updated by solving a re-weighted ℓ_1 norm optimization problem. Similar to Eq (8), it can be solved using classical/linearized ADM.

4.3. Theoretical Analysis

In Theorem 3, we show that when the monotone increasing multiplier $\boldsymbol{\pi}$ is larger than a threshold, the biconvex objective function in Eq (12) is equivalent to the original constrained MPEC problem in Eq (11). Moreover, based on the Kurdyka-Łojasiewicz inequality [2, 10], Theorem 4 establishes the convergence of the alternating direction method.

Theorem 3. *Boundedness of multiplier and exactness of the augmented Lagrangian Function. The augmented Lagrangian minimization problem $\min_{\mathbf{x},\mathbf{v}} \mathcal{L}(\mathbf{x}, \mathbf{v}, \bar{\boldsymbol{\pi}})$ admits the same local and global optimal solutions as the original problem in Eq (3) when $\bar{\boldsymbol{\pi}} > L/\sigma(\mathbf{A})$.*

Proof. Refer to **Appendix** C. $\qquad\qquad\square$

Algorithm 2   MPEC-ADM: An Alternating Direction Method for Solving MPEC Problem (11)

(S.0) Initialize $\mathbf{x}^0 = \mathbf{0} \in \mathbb{R}^n$, $\mathbf{v}^0 = \mathbf{1} \in \mathbb{R}^m$, $\boldsymbol{\pi}^0 = \eta \cdot \mathbf{1} \in \mathbb{R}^m$. Set $t = 0$ and $\mu > 0$.
(S.1) Solve the following $\mathbf{x}$-subproblem:

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \; \mathcal{L}(\mathbf{x}, \mathbf{v}^t, \boldsymbol{\pi}^t) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^t\|^2 \qquad (14)$$

(S.2) Solve the following $\mathbf{v}$-subproblem:

$$\mathbf{v}^{t+1} = \arg\min_{\mathbf{v}} \; \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{v}, \boldsymbol{\pi}^t) + \frac{\mu}{2}\|\mathbf{v} - \mathbf{v}^t\|^2 \qquad (15)$$

(S.3) Update the Lagrange multiplier:

$$\boldsymbol{\pi}^{t+1} = \boldsymbol{\pi}^t + \alpha(|\mathbf{Ax}^{t+1}| \odot \mathbf{v}^{t+1}) \qquad (16)$$

(S.4) Set $t := t + 1$ and then go to Step (S.1).

Theorem 4. *Convergence of Algorithm 2. Assume that $\mathbf{x}^t$ is bounded for all t. As $t \to \infty$, Algorithm 2 converges to a first order KKT point of the reformulated MPEC problem in Eq (11).*

Proof. Refer to **Appendix** C. $\qquad\qquad\square$

5. On MPEC Optimization

In this paper, MPEC reformulations are considered to solve the ℓ_0 norm problem. Mathematical programs with equilibrium constraints (MPEC) are optimization problems where the constraints include variational inequalities or complementarities. It is related to the Stackelberg game and it is used in the study of economic equilibrium and engineering design. MPECs are difficult to deal with because their feasible region is not necessarily convex or even connected.

General MPEC motivation. The basic idea behind MPEC-based methods is to transform the thin and nonsmooth, nonconvex feasible region into a thick and smooth one by introducing regularization/penalization on the complementary/equilibrium constraints. In fact, several other nonlinear methods solve the MPEC problem with similar motivations. For example, an exact penalty method where the complementarity term is moved to the objective in the form of an ℓ_1-penalty was proposed in [34]. The work in [22] suggested a smoothing family by replacing the complementarity system with the perturbed Fischer-Burmeister function. Also, a quadratic regularization technique to handle the complementarity constraints was proposed in [17].

In the defense of our methods. We propose two different penalization/regularization schemes to solve the equivalent MPECs. They have several merits. **(a)** Each reformulation is equivalent to the original ℓ_0 norm problem. **(b)** They are continuous reformulations, so they facilitate

KKT analysis and are amenable to the use of existing continuous optimization techniques to solve the convex sub-problems. We argue that, from a practical viewpoint, improved solutions to Eq (1) can be obtained by reformulating the problem using MPEC and focusing on the complementarity constraints. **(c)** They find a good initialization because they both reduce to a convex relaxation method in the first iteration. **(d)** Both methods exhibit strong convergence guarantees, this is because they essentially reduce to alternating minimization for bi-convex optimization problems [2, 10]. **(e)** They have a monotone/greedy property owing to the complimentarity constraints brought on by MPEC. The complimentary system characterizes the optimality of the KKT solution. We let $w \triangleq \{\mathbf{x}, \mathbf{u}\}$ (or $w \triangleq \{\mathbf{x}, \mathbf{v}\}$). Our solution directly handles the complimentary system of Eq (1) which takes the following form: $\langle p(w), q(w) \rangle = 0$, $p(w) \geq 0$, $q(w) \geq 0$. Here $p(\cdot)$ and $q(\cdot)$ are non-negative mappings of w which are vertical to each other. The complimentary constraints enable all the special properties of MPEC that distinguish it from general nonlinear optimization. We penalize the complimentary error of $\varepsilon \triangleq \langle p(w), q(w) \rangle$ (which is always non-negative) and ensure that the error ε is decreasing in every iteration. See Figure 1 for a geometric interpretation for MPEC optimization.

MPEC-EPM vs. MPEC-ADM. We consider two algorithms based on two variational characterizations of the ℓ_0 norm function (separable and non-separable MPEC[4]). Generally speaking, both methods have their own merits. **(a)** MPEC-EPM is more simple since it can directly use existing ℓ_1 norm minimization solvers/codes while MPEC-ADM involves additional computation for the quadratic term in the standard Lagrangian function. **(b)** MPEC-EPM is less adaptive in its per-iteration optimization, since the penalty parameter ρ is monolithically increased until a threshold is achieved. In comparison, MPEC-ADM is more adaptive, since a constant penalty also guarantees monotonically non-decreasing multipliers and convergence. **(c)** Regarding numerical robustness, while MPEC-EPM needs to solve the subproblem with certain accuracy, MPEC-ADM is more robust since the dual Lagrangian function provides a support function and the solution never degenerates even if the $\mathbf{x}$-problem is solved only approximately. **(d)** For matrix extensions, one can extend our results to the matrix rank problem for non-separable MPEC[5] and separable MPEC[6]. The former reformulation for the rank function is more interesting, since it results in a convex function with respect to $\mathbf{X}$.
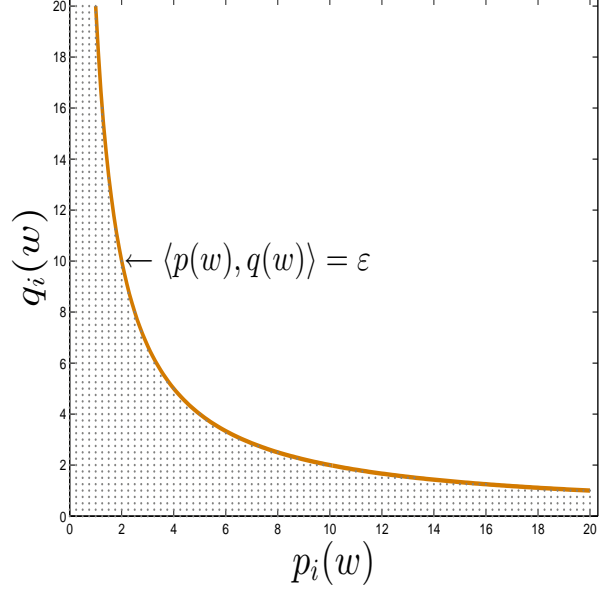

Figure 1: Effect of regularizing/smoothing the original optimization problem.

6. Experimental Validation

In this section, we demonstrate the effectiveness of our algorithms on five ℓ_0 norm optimization tasks, namely feature selection, segmented regression, MRF optimization, trend filtering, and impulse noise removal. We compare against the following six methods, (i) convex ℓ_1 minimization method (CVX)[7], (ii) Direct Alternating Direction Method (DI-ADM)[8], (iii) Mean Doubly Alternating Direction Method (MD-ADM) [18] [9], (iv) Quadratic Penalty Method (QPM) [44], (v) MPEC-EPM (Algorithm 1) and (vi) MPEC-ADM (Algorithm 2). We vary the sparsity parameter $k = \{0.01, 0.06, ..., 0.96\} \times n$ on different data sets and report the objective value of the optimization problem[10]. All algorithms are implemented in Matlab on an Intel 2.6 GHz CPU with 8 GB RAM.

---

[4]Note that the non-separable MPEC has one equilibrium constraint ($\|\mathbf{x}\|_1 = \langle \mathbf{x}, \mathbf{u} \rangle$) and the separable MPEC has n equilibrium constraints ($|\mathbf{x}| \odot \mathbf{v} = \mathbf{0}$). The terms separable and non-separable are related to whether the constraints can be decomposed to independent components.

[5]$rank(\mathbf{X}) = \min_{\|\mathbf{U}\| \leq 1} \|\mathbf{U}\|_*$, $s.t.$ $\|\mathbf{X}\|_* = \langle \mathbf{X}, \mathbf{U} \rangle$

[6]$rank(\mathbf{X}) = \min_{\mathbf{0} \leq \mathbf{v} \leq \mathbf{1}} \langle \mathbf{1}, \mathbf{1} - \mathbf{v} \rangle$, $s.t.$ $\sigma(\mathbf{X}) \odot \mathbf{v} = \mathbf{0}$.

[7]Since this approximate method can not control the sparsity of the solution, we solve a ℓ_1 regularized problem where the regulation parameter is swept over $\{2^{-10}, 2^{-8}, ..., 2^{10}\}$. Finally, the solution that leads to smallest objective value after a hard thresholding projection (which reduces to setting the smallest $n - k$ values of the solution in magnitude to 0) is selected.

[8]This is an ADM directly applied to Eq (1).

[9]This is an ADM that treats arithmetic means of the solution sequence as the actual output.

[10]All the algorithms generally terminate within 8 minutes, we do not include the CPU time here. In fact, our methods converge within reasonable time. They are only 3-7 times slower than the classical convex methods. This is expected, since they are alternating methods and have the same computational complexity as the convex methods.

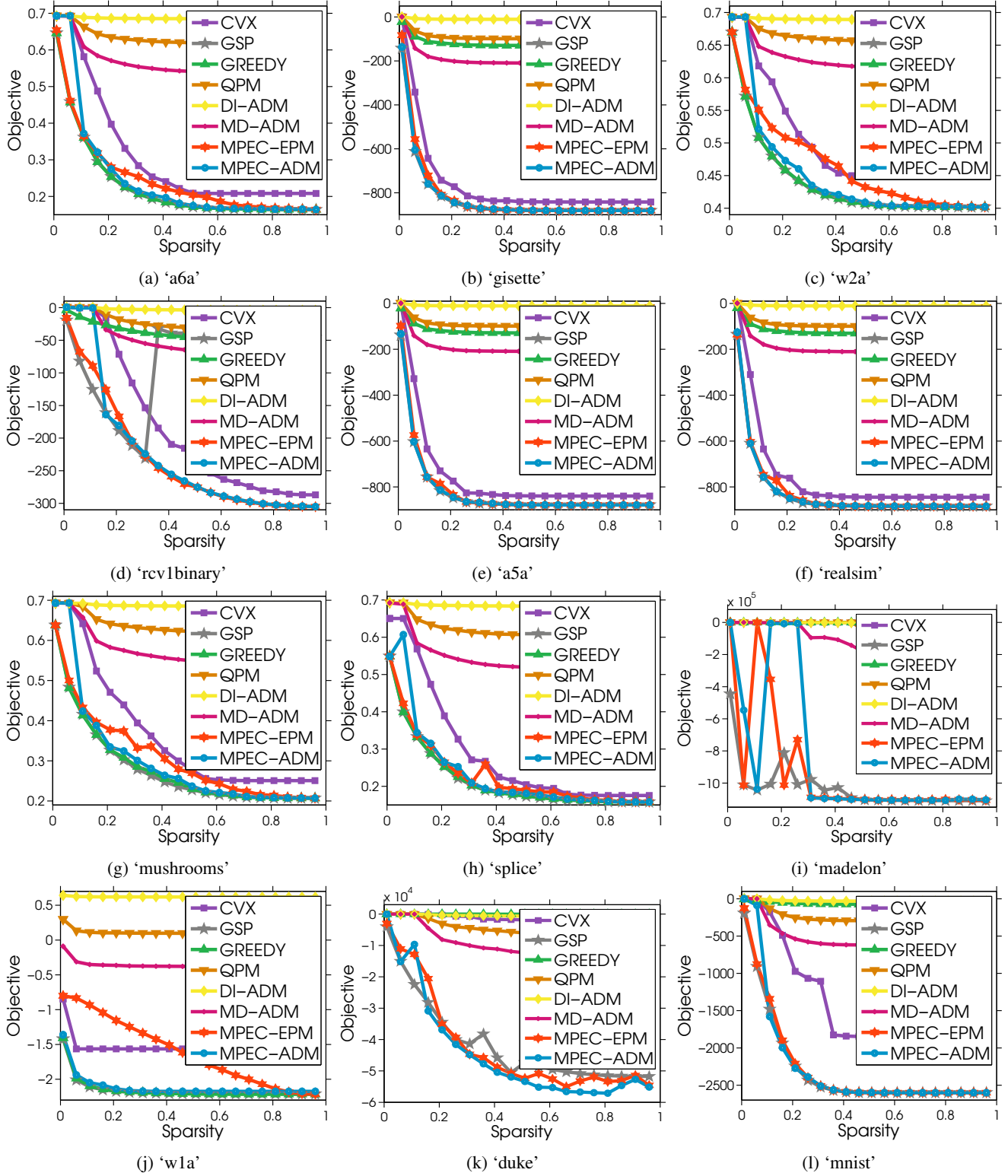(a) 'a6a' (b) 'gisette' (c) 'w2a'

(d) 'rcv1binary' (e) 'a5a' (f) 'realsim'

(g) 'mushrooms' (h) 'splice' (i) 'madelon'

(j) 'w1a' (k) 'duke' (l) 'mnist'

Figure 2: Feature selection based on logistic loss minimization. We set $\lambda = 0.01$.

(a) 'a6a' (b) 'gisette' (c) 'w2a'

(d) 'rcv1binary' (e) 'a5a' (f) 'realsim'

(g) 'mushrooms' (h) 'splice' (i) 'madelon'

(j) 'w1a' (k) 'duke' (l) 'mnist'
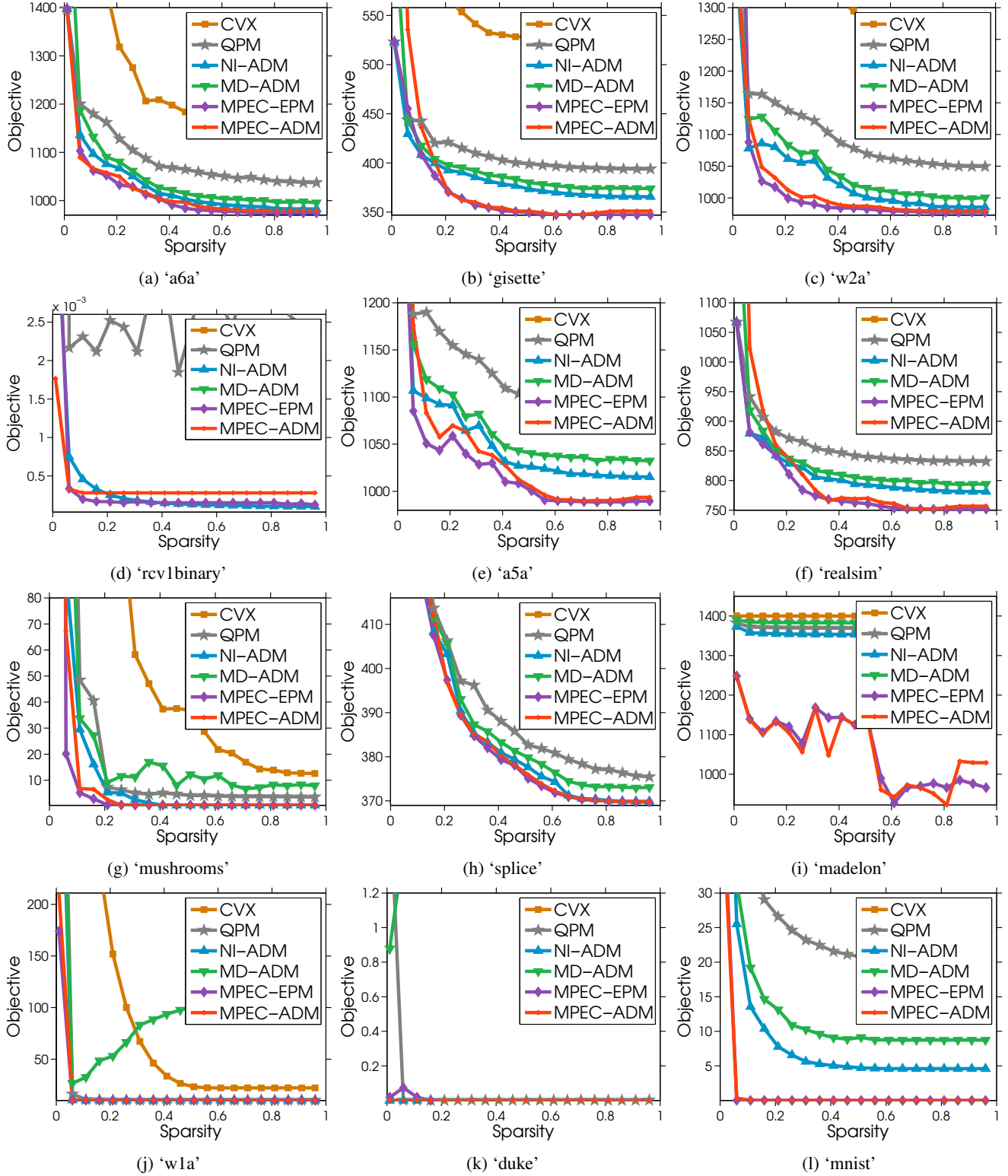
Figure 3: Feature selection based on hinge loss minimization. We set $\lambda = 0.01$.

(a) $n = 1024$, $\sigma = 10$ (b) $n = 2048$, $\sigma = 10$ (c) $n = 4096$, $\sigma = 10$

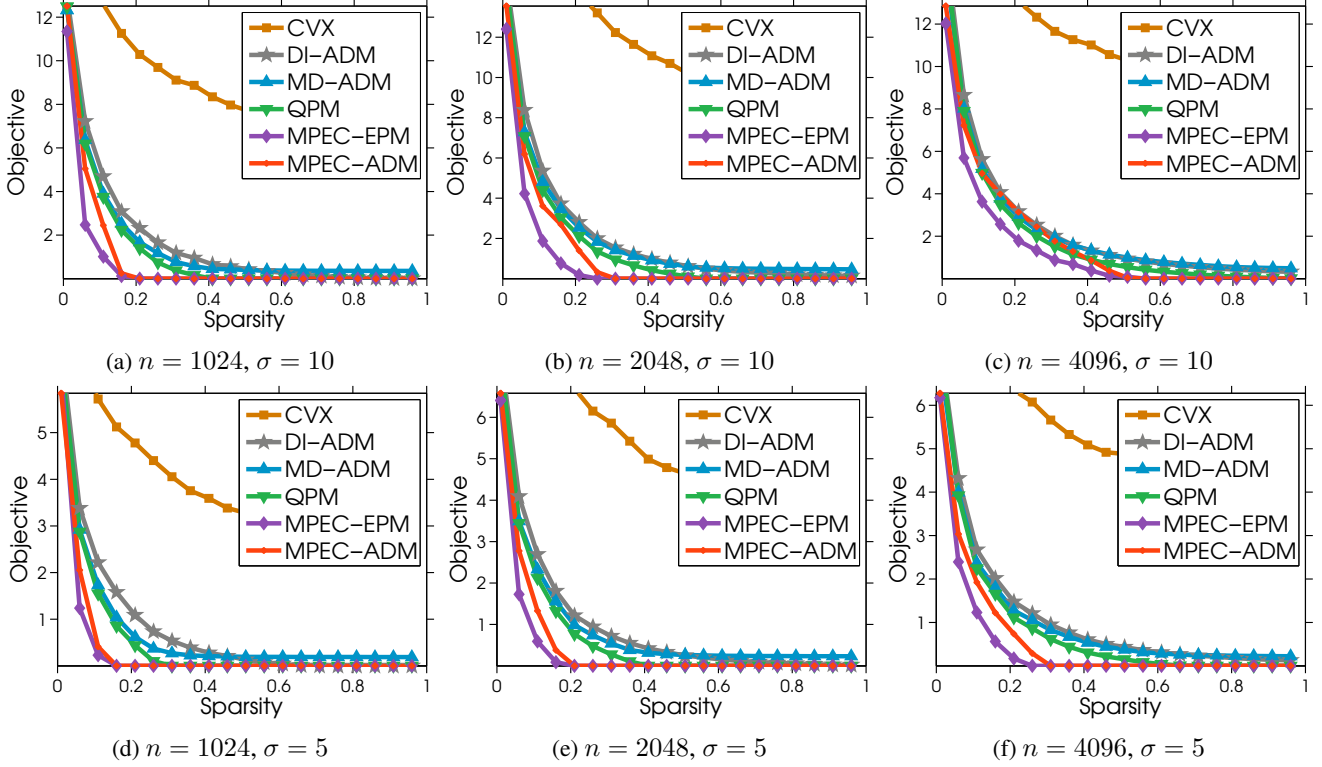(d) $n = 1024$, $\sigma = 5$ (e) $n = 2048$, $\sigma = 5$ (f) $n = 4096$, $\sigma = 5$

Figure 4: Segmented regression. Results for unit-column-norm $\mathbf{A}$ for different parameter settings.

6.1. Feature Selection

Given a set of labeled patterns $\{\mathbf{s}_i, \mathbf{y}_i\}_{i=1}^p$, where $\mathbf{s}_i$ is an instance with n features and $\mathbf{y}_i \in \{\pm 1\}$ is the label. Feature selection considers the following optimization problem:

$$\min_{\mathbf{x}} \; \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \sum_i^p \ell\left(\langle \mathbf{s}_i, \mathbf{x}\rangle, \mathbf{y}_i\right), \; s.t. \; \|\mathbf{x}\|_0 \le k \quad (17)$$

where the loss function $\ell(\cdot)$ is chosen to be the logistic loss: $\ell(r, t) = \log(1 + \exp(r)) - rt$ or the hinge loss: $\ell(r, t) = \max(0, 1 - rt)$.

In our experiments, we test on 12 well-known benchmark datasets[11] which contain high dimensional ($n \ge 7000$) data vectors and thousands of data instances. Since we are solving an optimization problem, we only consider measuring the quality of the solutions by comparing the objective values.

For logistic loss feature selection, since the objective is smooth, we also include a comparisons with the state-of-the-art solver Gradient Support Pursuit (GSP)[12] and greedy hard thresholding (GREEDY). Based on our experimental results in Figure 2, we make the following observations. (i) Convex methods generally gives good results in this task, but they fail on 'duke' and 'madelon' datasets.

(ii) The quadratic penalization techniques (NI-ADM, MD-ADM and QPM) generally demonstrate bad performance in this task. (iii) GREEDY and GSP often achieve good performance, but they are not stable in some cases. (iv) Both our MPEC-EPM and MPEC-ADM often achieve better performance than existing solutions. For hinge loss feature selection, we demonstrate our experimental results in Figure 3. We make the following observations. (i) Convex methods generally fail in this situation, as they usually produce much larger objectives. (ii) NI-ADM and MD-ADM outperform QPM in all test cases. (iii) MPEC-EPM and MPEC-ADM generally achieve lower objective values.

6.2. Segmented Regression

Given a design matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and an observation vector $\mathbf{b} \in \mathbb{R}^m$, segmented regression involves solving the following optimization problem:

$$\min_{\mathbf{x}} \; \|\mathbf{A}^T(\mathbf{A}x - \mathbf{b})\|_\infty, \; s.t. \; \|\mathbf{x}\|_0 \le k \quad (18)$$

This regression model is closely related to Dantzig selector [11] in the literature, where ℓ_0 norm is replaced by ℓ_1 norm in Eq (18). In our experiments, we consider design matrices with unit column norms. Similar to [11], we first generate an $\mathbf{A}$ with independent Gaussian entries and then normalize each column to have norm 1. We then select a support set

---

[11] www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
[12] users.ece.gatech.edu/sbahmani7/GraSP.html

(a) by Graph Cut
$f(\mathbf{x}) = 95741.1$

(b) by NI-ADM
$f(\mathbf{x}) = 95852.6$

(c) by MD-ADM
$f(\mathbf{x}) = 95825.2$

(d) by QPM
$f(\mathbf{x}) = 95827.3$

(e) by MPEC-EPM
$f(\mathbf{x}) = 95798.8$
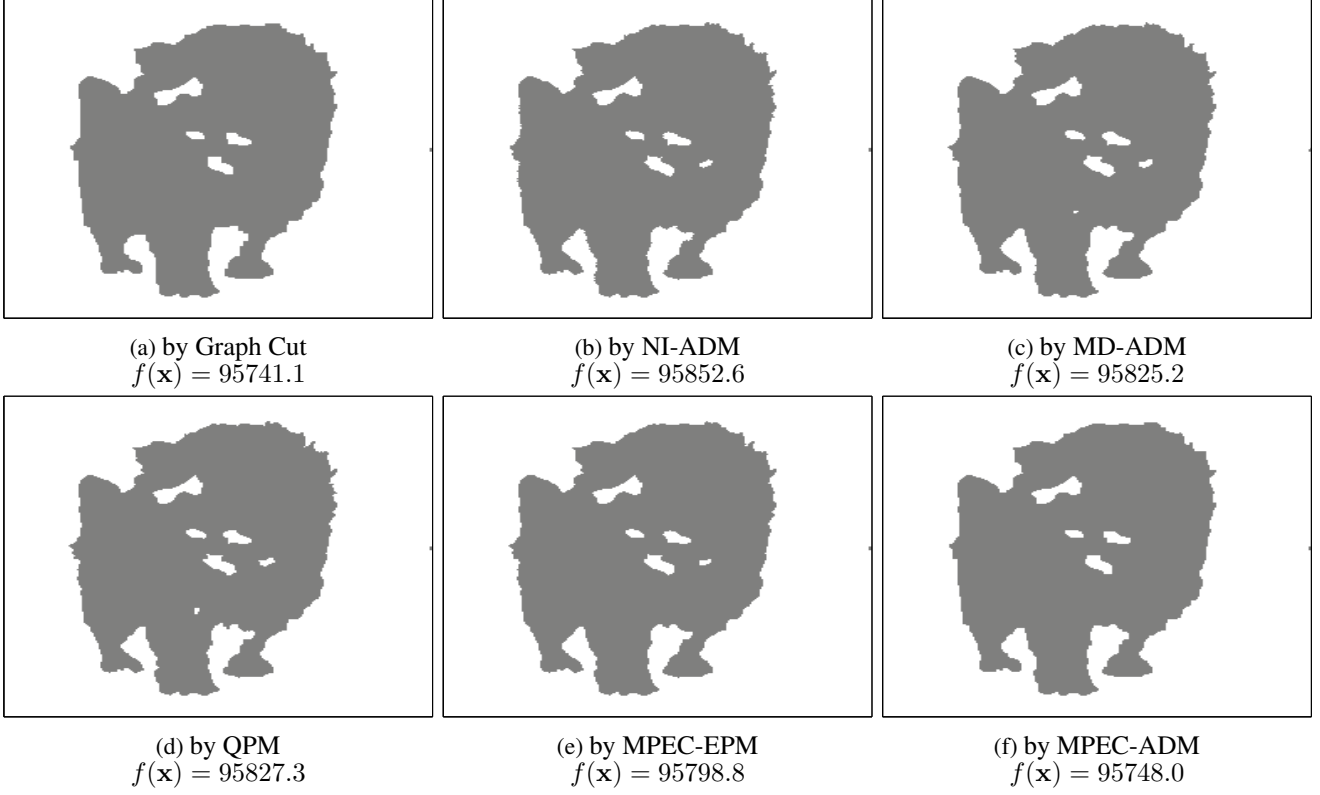
(f) by MPEC-ADM
$f(\mathbf{x}) = 95748.0$

Figure 5: MRF optimization on 'cat' image. The objective value for the unary image $\mathbf{b}$ is $f(\mathbf{b}) = 96439.8$.
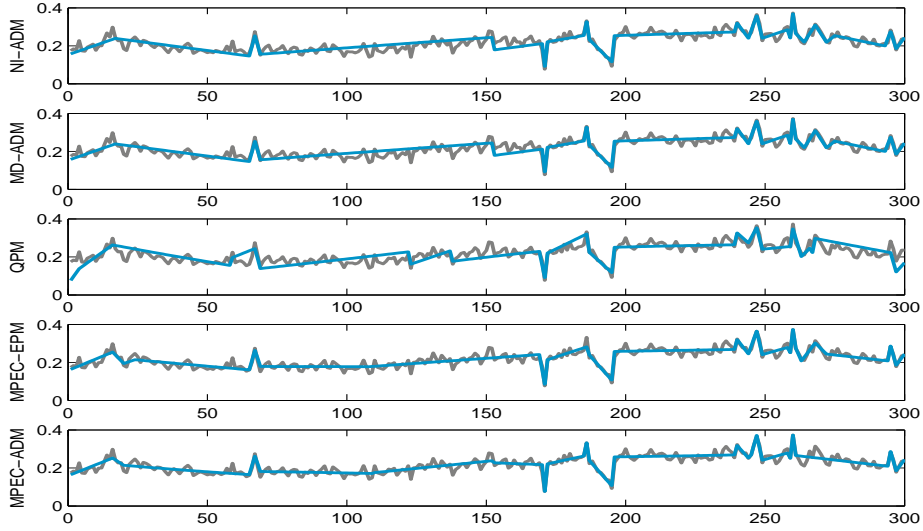

Figure 6: Trend filtering result on 'snp500.dat' data [38]. The objective values achieved by the five methods are $0.38, 0.38, 0.53, 0.32, 0.33$, respectively.

of size $0.5 \times \min(m,n)$ uniformly at random. We finally set $\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$. We vary n from $\{1024, 2048, 4096\}$ and set $m = n/8$. We demonstrate our experimental results in Figure 4. The proposed penalization techniques (MPEC-EPM and MPEC-ADM) outperform the quadratic penalization techniques (NI-ADM, MD-

10

Table 2: Random-value impulse noise removal problems. The results separated by '/' are SNR_0, SNR_1 and SNR_2, respectively.

Alg. / Img.	CVX (i.e. $\ell_1 TV$)	DI-ADM	MD-ADM	QPM	MPEC-EPM	MPEC-ADM
mandrill+10%	0.83/3.06/5.00	0.78/3.68/7.06	0.84/3.71/7.10	0.85/3.76/7.16	0.85/3.71/6.90	**0.86/3.87/7.64**
mandrill+30%	0.69/2.38/4.42	0.67/2.96/**5.13**	0.74/2.91/4.96	0.75/2.94/5.02	0.75/2.89/4.87	**0.76/3.02/5.13**
mandrill+50%	0.55/1.80/3.33	0.58/2.32/3.86	0.65/2.23/3.64	0.66/2.30/3.79	0.66/2.22/3.58	**0.67/2.39/3.88**
mandrill+70%	0.41/0.91/1.64	0.47/1.49/2.42	0.52/1.39/2.21	0.52/1.40/2.22	0.53/1.39/2.12	**0.55/1.56/2.43**
mandrill+90%	0.31/0.02/0.09	0.33/0.14/**0.13**	0.35/0.04/-0.08	0.35/0.04/-0.03	0.36/0.07/-0.14	**0.37/0.19/**0.05
lenna+10%	0.97/8.08/13.48	0.97/8.05/13.51	0.97/8.13/13.59	0.97/8.12/13.87	0.97/8.10/13.72	**0.98/8.31/14.55**
lenna+30%	0.92/7.40/13.60	0.92/6.71/10.53	0.91/6.62/10.07	0.94/7.38/12.27	0.93/7.19/11.86	**0.96/7.90/13.68**
lenna+50%	0.84/5.50/9.46	0.84/5.32/8.13	0.83/5.18/7.53	0.88/6.28/10.10	0.86/5.75/8.69	**0.91/6.65/10.89**
lenna+70%	0.57/2.48/4.17	0.67/3.34/5.11	0.66/3.17/4.40	0.71/3.66/5.34	0.69/3.34/4.53	**0.75/4.09/5.81**
lenna+90%	0.35/0.53/0.89	0.41/0.76/**0.91**	0.40/0.54/0.48	0.41/0.66/0.80	0.43/0.54/0.38	**0.48/0.89/**0.80
lake+10%	0.94/8.03/14.66	0.95/8.21/13.64	**0.96**/8.31/13.81	**0.96**/8.40/14.52	**0.96**/8.33/14.14	**0.96/8.45/14.72**
lake+30%	0.88/7.29/12.83	0.86/6.86/10.87	0.87/6.81/10.02	0.91/7.57/12.45	0.89/7.26/11.33	**0.92/7.87/13.34**
lake+50%	0.61/4.89/8.31	0.76/5.69/8.52	0.77/5.65/8.02	0.83/6.57/10.36	0.80/5.98/8.78	**0.85/6.75/10.56**
lake+70%	0.40/2.07/3.61	0.51/3.68/5.49	0.56/3.72/5.14	0.60/4.01/5.60	0.61/3.94/5.28	**0.62/4.08/5.61**
lake+90%	0.27/0.54/0.87	0.25/**0.94/1.06**	0.27/0.87/0.65	0.26/0.75/0.65	**0.29**/0.81/0.44	0.28/0.90/0.71
jetplane+10%	**0.42/3.33/7.98**	0.37/3.16/7.10	0.39/3.16/7.15	0.39/3.20/7.43	0.39/3.16/7.28	0.39/3.22/7.52
jetplane+30%	0.36/2.94/7.00	0.35/2.70/5.69	0.38/2.68/5.60	0.39/2.98/6.68	0.39/2.87/6.24	**0.40/3.14/7.23**
jetplane+50%	0.26/1.75/4.24	0.31/2.02/4.21	0.37/2.08/4.00	0.38/2.56/5.45	0.37/2.26/4.48	**0.40/2.72/5.75**
jetplane+70%	0.24/-0.55/-0.17	0.24/0.62/1.46	0.31/0.83/1.49	0.30/0.81/1.44	**0.34/0.98/1.55**	0.33/0.92/1.38
jetplane+90%	0.19/-1.81/-3.38	0.17/-1.98/-3.23	0.18/-1.96/-3.47	0.18/-2.18/-3.90	**0.20/-1.66/-2.13**	0.19/-2.05/-3.80
blonde+10%	0.26/**0.92**/2.83	0.25/0.87/2.69	0.25/0.86/2.68	0.22/0.90/2.80	**0.27**/0.88/2.72	0.25/**0.92/2.87**
blonde+30%	0.25/0.82/2.57	0.26/0.76/2.21	0.26/0.70/2.14	0.25/0.80/2.47	**0.27**/0.76/2.35	**0.27/0.85/2.59**
blonde+50%	0.25/0.60/1.89	0.25/0.61/1.80	0.24/0.58/1.77	0.25/0.67/2.13	0.25/0.64/1.96	**0.26/0.75/2.26**
blonde+70%	0.24/0.02/0.22	0.22/0.35/1.02	0.21/0.35/1.05	0.22/0.33/1.11	0.20/0.38/1.19	**0.26/0.49/1.47**
blonde+90%	0.20/-0.81/-1.39	**0.21**/-0.62/-1.15	0.20/-0.65/-1.26	0.18/-0.75/-1.39	0.20/-0.52/-1.03	0.19/**-0.45/-0.88**

ADM and QPM) consistently.

6.3. MRF Optimization

Markov Random Field (MRF) optimization involves solving the following problem [39],

$$\min_{\mathbf{x}} \ \tfrac{1}{2}\mathbf{x}^T\mathbf{L}\mathbf{x} + \mathbf{b}^T\mathbf{x}, \ s.t. \ \mathbf{x} \in \{0,1\}^n \qquad (19)$$

where $\mathbf{b}$ is determined by the unary term defined for the graph and $\mathbf{L}$ is the Laplacian matrix, which is based on the binary term relating pairs of graph nodes together. The quadratic term is usually considered a smoothness prior on the node labels. The MRF formulation is widely used in many labeling applications, including image segmentation.

In our experiments, we include the comparison with the graph cut method [39], which is known to achieve the global optimal solution for this specific class of binary problem. Figure 5 demonstrates a qualitative result for image segmentation. MPEC-ADM produces a solution that is very close to the globally optimal one. Moreover, both our methods achieve lower objectives than the other compared methods.

6.4. Trend Filtering

Given a time series $\mathbf{y} \in \mathbb{R}^n$, the goal of trend filtering is to find another nearest time series $\mathbf{x}$ such that $\mathbf{x}$ is sparse after a gradient mapping [38]. Trend filtering involves solving the following optimization, where $\mathbf{D} \in \mathbb{R}^{(n-2)\times n}$ is a 2nd difference matrix.

$$\min_{\mathbf{x}} \ \tfrac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2, \ s.t. \ \|\mathbf{D}\mathbf{x}\|_0 \leq k \qquad (20)$$

In our experiments, we perform trend filtering on the 'snp500.dat' data set [38]. For better visualization, we only use the first 300 time instances of the signal $\mathbf{y}$, i.e. $n = 300$. The sparsity parameter k is set to 30. We stop all the algorithms when the same strict stopping criterion is met, in order to ensure that the hard constraint $\|\mathbf{A}\mathbf{x}\|_0 \leq k$ is fully satisfied. As shown in Figure 6, the proposed methods (MPEC-EPM and MPEC-ADM) achieve more natural trend filtering results (refer to position 120-170), since they achieve lower objective values (0.32 and 0.33).

6.5. Impulse Noise Removal

Given a noisy image $\mathbf{b}$ which was contaminated by random value impulsive noise, we consider a denoised image of $\mathbf{x}$ as a minimizer of the following constrained $\ell_0 TV$ model [65]:

$$\min_{\mathbf{x}} \ TV(\mathbf{x}), \ s.t. \ \|\mathbf{x} - \mathbf{b}\|_0 \leq k \qquad (21)$$

where $TV(\mathbf{x}) \triangleq \|\boldsymbol{\nabla}\mathbf{x}\|_{p,1}$ denotes the total variation norm function, $\|\mathbf{x}\|_{p,1} \triangleq \sum_{i=1}^{n}(|\mathbf{x}_i|^p + |\mathbf{x}_{n+i}|^p)^{\frac{1}{p}}$; $\boldsymbol{\nabla} \triangleq$
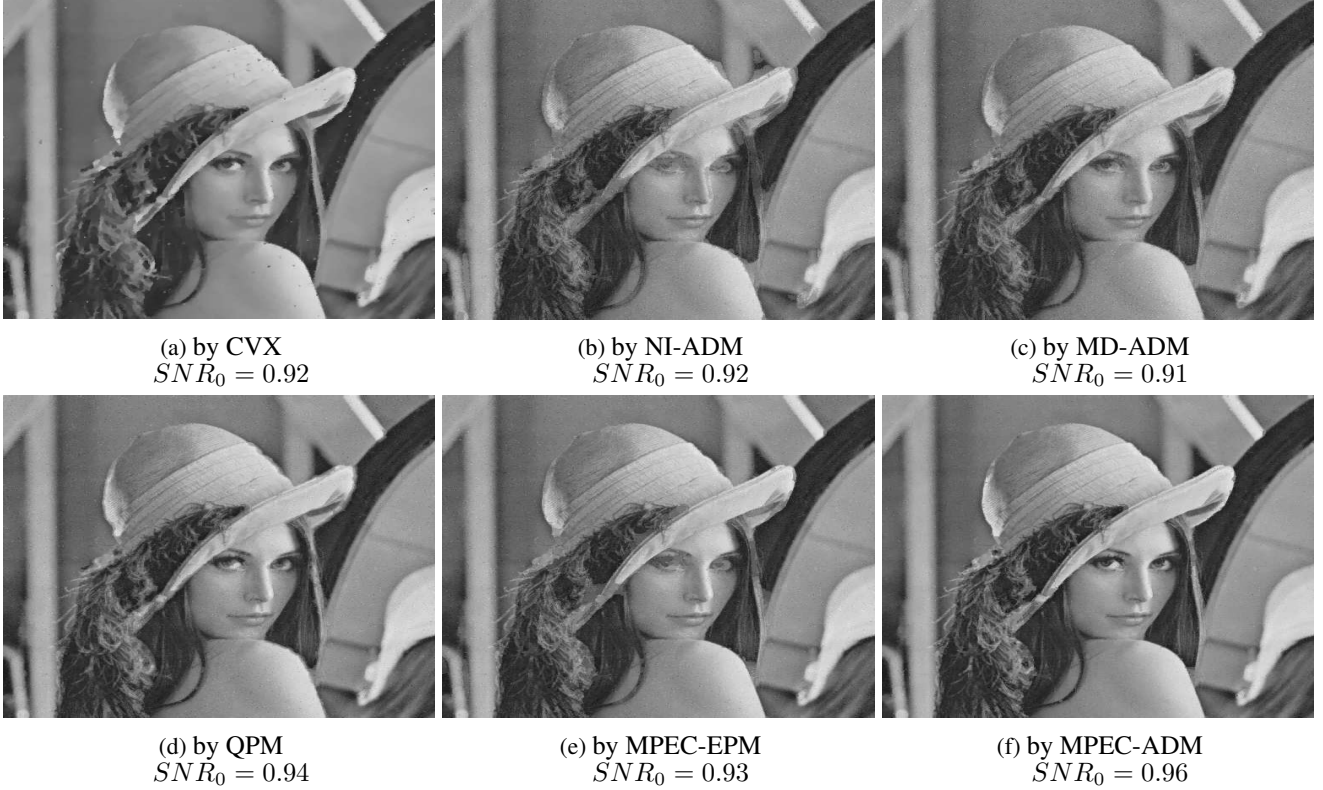
(a) by CVX	(b) by NI-ADM	(c) by MD-ADM
$SNR_0 = 0.92$	$SNR_0 = 0.92$	$SNR_0 = 0.91$
(d) by QPM	(e) by MPEC-EPM	(f) by MPEC-ADM
$SNR_0 = 0.94$	$SNR_0 = 0.93$	$SNR_0 = 0.96$

Figure 7: Random value impulse noise removal on 'lenna' image.

$[\boldsymbol{\nabla}_x | \boldsymbol{\nabla}_y]^T \in \mathbb{R}^{2n \times n}$, and k specifies the noise level. Furthermore, $\boldsymbol{\nabla}_x \in \mathbb{R}^{n \times n}$ and $\boldsymbol{\nabla}_y \in \mathbb{R}^{n \times n}$ are two suitable linear transformation matrices that computes their discrete gradients. Following the experimental settings in [65], we use three ways to measure SNR, i.e. SNR_0, SNR_1, SNR_2. We show image recovery results in Table 2 when random-value impulse noise is added to the clean images {'mandrill', 'lenna', 'lake', 'jetplane', 'blonde'}. We also demonstrate a visualized example of random value impulse noise removal on 'lenna' image in Figure 7. A conclusion can be drawn that both MPEC-EPM and MPEC-ADM achieve state-of-the-art performance while MPEC-ADM is generally better than MPEC-EPM in this impulse noise removal problem.

7. Conclusions and Future Work

This paper presents two methods (exact penalty and alternating direction) to solve the general sparsity constrained minimization problem. Although it is non-convex, we design effective penalization/regularization schemes to solve its equivalent MPEC. We also prove that both of our methods are convergent to first-order KKT points. Experimental results on a variety of sparse optimization applications demonstrate that our methods achieve state-of-the-art performance.

Our future work focuses on several directions. **(i)** We will investigate the optimality qualification of our multi-stage convex relaxation method for some specific objective functions (e.g. squared or logistic function). **(ii)** It is interesting to extend the present results to deal with rank minimization problem and its applications. **(iii)** We are also interested in applying the proposed algorithms to other sparse optimization and binary optimization applications in machine learning and signal processing.

References

[1] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *Neural Information Processing Systems (NIPS)*, pages 1466–1474, 2012. 2, 3

[2] Hedy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for non-convex problems: An approach based on the kurdyka-lojasiewicz inequality. *Mathematics*

of Operations Research (MOR), 35(2):438–457, 2010. 3, 5, 6, 21

[3] Sohail Bahmani, Bhiksha Raj, and Petros T. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research (JMLR)*, 14(1):807–841, 2013. 2

[4] Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. Dictionary learning for sparse coding: Algorithms and convergence analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(7):1356–1369, 2016. 1

[5] Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization (SIOPT)*, 23(3):1480–1509, 2013. 3

[6] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *arXiv preprint:1507.03133*, 2015. 3

[7] Shujun Bi, Xiaolan Liu, and Shaohua Pan. Exact penalty decomposition method for zero-norm minimization based on MPEC formulation. *SIAM Journal on Scientific Computing (SISC)*, 36(4), 2014. 3

[8] Daniel Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical programming*, 74(2):121–140, 1996. 2, 3

[9] Thomas Blumensath and Mike E Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008. 2

[10] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014. 3, 5, 6, 21

[11] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007. 1, 9

[12] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. 1, 2

[13] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008. 1, 2, 3

[14] Antoni B. Chan, Nuno Vasconcelos, and Gert R. G. Lanckriet. Direct convex relaxations of sparse svm. In *International Conference on Machine Learning (ICML)*, pages 145–153, 2007. 2, 3

[15] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on Scientific Computing (SISC)*, 20(1):33–61, 1998. 2

[16] Jon Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing USA, 2011. 3, 18

[17] Victor DeMiguel, Michael P Friedlander, Francisco J Nogales, and Stefan Scholtes. A two-sided relaxation scheme for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization (SIOPT)*, 16(2):587–609, 2005. 5

[18] Bin Dong and Yong Zhang. An efficient algorithm for l0 minimization in wavelet frame based image restoration. *Journal of Scientific Computing*, 54(2-3):350–368, 2013. 1, 3, 6

[19] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 1

[20] John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1 ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, pages 272–279, 2008. 4

[21] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2797, 2009. 1

[22] Francisco Facchinei, Houyuan Jiang, and Liqun Qi. A smoothing method for mathematical programs with equilibrium constraints. *Mathematical programming*, 85(1):107–134, 1999. 5

[23] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association (JASA)*, 96(456):1348–1360, 2001. 3

[24] Mingbin Feng, John E Mitchell, Jong-Shi Pang, Xin Shen, and Andreas Wachter. Complementarity formulations of $\ell 0$-norm optimization problems. 2013. 3

[25] Fajwel Fogel, Rodolphe Jenatton, Francis R. Bach, and Alexandre d'Aspremont. Convex relaxations for permutation problems. *SIAM Journal on Matrix Analysis and Applications (SIMAX)*, 36(4):1465–1488, 2015. 1

[26] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. 1

[27] Jerome H Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722–738, 2012. 3

[28] Cuixia Gao, Naiyan Wang, Qi Yu, and Zhihua Zhang. A feasible nonconvex relaxation approach to feature selection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 356–361, 2011. 3

[29] Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of ℓ_p minimization. *Mathematical Programming*, 129(2):285–299, 2011. 2, 3

[30] Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing (TIP)*, 4(7):932–946, 1995. 3

[31] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning (ICML)*, pages 37–45, 2013. 3

[32] Bingsheng He and Xiaoming Yuan. On the $\mathcal{O}(1/n)$ convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis (SINUM)*, 50(2):700–709, 2012. 4, 17

[33] Richard V. Helgason, Jeffery L. Kennington, and H. S. Lall. A polynomially bounded algorithm for a singly constrained quadratic program. *Mathematical Programming*, 18(1):338–343, 1980. 4, 5, 22

[34] XM Hu and Daniel Ralph. Convergence of a penalty method for mathematical programming with complementarity constraints. *Journal of Optimization Theory and Applications*, 123(2):365–390, 2004. 3, 5

[35] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Neural Information Processing Systems (NIPS)*, pages 685–693, 2014. 3

[36] Bo Jiang, Ya-Feng Liu, and Zaiwen Wen. l_p-norm regularization algorithms for optimization over permutation matrices. Technical report, Optimization online, November 2015. 1

[37] Abdeslam Kadrani, Jean-Pierre Dussault, and Abdelhamid Benchakroun. A new regularization scheme for mathematical programs with complementarity constraints. *SIAM Journal on Optimization (SIOPT)*, 20(1):78–103, 2009. 3

[38] Seung-Jean Kim, Kwangmoo Koh, Stephen P. Boyd, and Dimitry M. Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009. 1, 10, 11

[39] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph

cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(2):147–159, 2004. 10, 11

[40] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *Neural Information Processing Systems (NIPS)*, pages 801–808, 2006. 1

[41] Yuanqing Li, Andrzej Cichocki, and Shun-ichi Amari. Analysis of sparse representation and blind source separation. *Neural computation*, 16(6):1193–1234, 2004. 1

[42] Ya-Feng Liu, Yu-Hong Dai, and Shiqian Ma. Joint power and admission control: Non-convex l_q approximation and an effective polynomial time deflation approach. *IEEE Transactions on Signal Processing*, 63(14):3641–3656, 2015. 1

[43] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Generalized nonconvex nonsmooth low-rank minimization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4130–4137, 2014. 3

[44] Zhaosong Lu and Yong Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization (SIOPT)*, 23(4):2448–2478, 2013. 2, 3, 4, 6

[45] Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996. 3

[46] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. 2

[47] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association (JASA)*, 2012. 3

[48] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. Spectral k-support norm regularization. In *Neural Information Processing Systems (NIPS)*, pages 3644–3652, 2014. 3

[49] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009. 2

[50] Deanna Needell and Roman Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of computational mathematics*, 9(3):317–334, 2009. 2

[51] Andrew Y Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *International Conference on Machine Learning (ICML)*, page 78, 2004. 1

[52] Mert Pilanci, Martin J. Wainwright, and Laurent El Ghaoui. Sparse learning via boolean relaxations. *Mathematical Programming*, 151(1):63–87, 2015. 3

[53] Nikhil S. Rao, Parikshit Shah, and Stephen J. Wright. Forward-backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Signal Processing*, 63(21):5798–5811, 2015. 2

[54] Joel Tropp, Anna C Gilbert, et al. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. 2

[55] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. 2

[56] Joshua Trzasko and Armando Manduca. Highly undersampled magnetic resonance image reconstruction via homotopic-minimization. *IEEE Transactions on Medical Imaging*, 28(1):106–121, 2009. 3

[57] Peng Wang, Chunhua Shen, and Anton van den Hengel. A fast semidefinite approach to solving binary quadratic problems. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1312–1319, 2013. 1

[58] Andreas Weinmann, Martin Storath, and Laurent Demaret. The l^1-potts functional for robust jump-sparse reconstruction. *SIAM Journal on Numerical Analysis (SINUM)*, 53(1):644–673, 2015. 1

[59] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Michael E. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research (JMLR)*, 3:1439–1461, 2003. 1

[60] Bin Wu, Chao Ding, Defeng Sun, and Kim-Chuan Toh. On the moreau-yosida regularization of the vector k-norm related functions. *SIAM Journal on Optimization (SIOPT)*, 24(2):766–794, 2014. 18

[61] Zhen James Xiang, Yongxin Taylor Xi, Uri Hasson, and Peter J. Ramadge. Boosting with spatial regularization. In *Neural Information Processing Systems (NIPS)*, pages 2107–2115, 2009. 1

[62] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via ℓ_0 gradient minimization. *ACM Transactions on Graphics (TOG)*, 30(6):174, 2011. 1

[63] Penghang Yin, Yifei Lou, Qi He, and Jack Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM Journal on Scientific Computing (SISC)*, 37(1), 2015. 2, 3

[64] Jin Yu, Anders Eriksson, Tat-Jun Chin, and David Suter. An adversarial optimization approach to efficient outlier removal. In *International Conference on Computer Vision (ICCV)*, pages 399–406, 2011. 2, 3

[65] Ganzhao Yuan and Bernard Ghanem. ℓ_0tv: A new method for image restoration in the presence of impulse noise. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5369–5377, 2015. 1, 2, 3, 4, 11

[66] Ganzhao Yuan and Bernard Ghanem. A proximal alternating direction method for semidefinite rank minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016. 4

[67] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research (JMLR)*, 11:2261–2286, 2010. 1

[68] Xiaotong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning (ICML)*, pages 127–135, 2014. 3

[69] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010. 3

[70] Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research (JMLR)*, 11:1081–1107, 2010. 1, 3

[71] Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011. 2

Appendix

The appendix is organized as follows. Section A presents the proof of Lemma 1. Section B presents the convergence analysis of the exact penalty method. Section C presents the convergence analysis of the alternating direction method. Section D presents a Matlab code of the breakpoint search algorithm.

A. Proof of Lemma 1

Here we provide the proof of our new non-separable MPEC of the ℓ_0-norm function.

Proof. It is not hard to validate that the ℓ_0 norm function of $\mathbf{x} \in \mathbb{R}^m$ can be repressed as the following minimization problem over $\mathbf{u} \in \mathbb{R}^m$:

$$\|\mathbf{x}\|_0 = \min_{-\mathbf{1} \leq \mathbf{u} \leq \mathbf{1}} \|\mathbf{u}\|_1, \ s.t. \ |\mathbf{x}| = \mathbf{u} \odot \mathbf{x} \quad (22)$$

Note that the minimization problem in Eq (22) can be decomposed into m subproblems. When $\mathbf{x}_i = 0$, the optimal solution in position i will be achieved at $\mathbf{u}_i = 0$ by minimization; when $\mathbf{x}_i > 0$ ($\mathbf{x}_i < 0$, respectively), the optimal solution in position i will be achieved at $\mathbf{u}_i = 1$ ($\mathbf{u}_i = -1$, respectively) by constraint. In other words,

$\mathbf{u}^* = sign(\mathbf{x})$ will be achieved for Eq (22), leading to $\|\mathbf{x}\|_0 = \|sign(\mathbf{x})\|_1$.

We now focus on Eq (22). Due to the box constraints $-\mathbf{1} \le \mathbf{u} \le \mathbf{1}$, it always holds that $|\mathbf{x}| - \mathbf{u} \odot \mathbf{x} \ge \mathbf{0}$. Therefore, the error generated by the difference of $|\mathbf{x}|$ and $\mathbf{u} \odot \mathbf{x}$ is summarizable. We naturally have the following results:

$$|\mathbf{x}| - \mathbf{u} \odot \mathbf{x} = \mathbf{0} \quad \Leftrightarrow \quad \langle \mathbf{1}, |\mathbf{x}| - \mathbf{u} \odot \mathbf{x} \rangle = 0$$
$$\Leftrightarrow \quad \|\mathbf{x}\|_1 = \langle \mathbf{u}, \mathbf{x} \rangle$$

$\square$

B. Convergence Analysis of Algorithm 1

This section presents the proof of Theorem 1 and Theorem 2. Our main novelties are establishing the inversely proportional relationship between the weight of ℓ_1 norm and the sparsity of $\mathbf{Ax}$ (in the proof of exactness) and qualifying the sparsity upper bound for $\langle \mathbf{w}, |\mathbf{Ax}| \rangle$ (in the proof of convergence rate). Our results make use of the specified structure of the ℓ_1 norm problem.

The following lemma is useful and it establishes a lower bound on the norm of matrix vector product.

Lemma 2. *Assume that $\mathbf{A} \in \mathbb{R}^{m \times n}$ has right inverse. It always holds that $\|\mathbf{A}^T \mathbf{c}\|_2 \ge \sigma(\mathbf{A})\|\mathbf{c}\|_2$ for all $\mathbf{c} \in \mathbb{R}^m$, where $\sigma(\mathbf{A})$ is the smallest singular value of $\mathbf{A}$.*

Proof. We denote $\mathbf{Z} = \mathbf{AA}^T - (\sigma(\mathbf{A}))^2 \mathbf{I} \in \mathbb{R}^{m \times m}$. Since $\mathbf{A}$ has right inverse, we have $\sigma(\mathbf{A}) > 0$ and $\mathbf{Z} \succeq \mathbf{0}$. We let $\mathbf{Z} = \mathbf{LL}^T$ and have the following inequalities:

$$\begin{aligned} \|\mathbf{A}^T \mathbf{c}\|_2^2 &= \langle \mathbf{AA}^T, \mathbf{cc}^T \rangle \\ &= \langle \mathbf{Z} + (\sigma(\mathbf{A}))^2 \mathbf{I}, \mathbf{cc}^T \rangle \\ &= \|\mathbf{L}^T \mathbf{c}\|_2^2 + \langle (\sigma(\mathbf{A}))^2 \mathbf{I}, \mathbf{cc}^T \rangle \\ &\ge 0 + \langle (\sigma(\mathbf{A}))^2 \|\mathbf{c}\|^2 \rangle \end{aligned}$$

Taking the square root of both sides, we naturally complete the proof of this lemma.

$\square$

The following lemma is useful in our convergence analysis.

Lemma 3. *Assume that $h(\cdot)$ is convex and L-Lipschitz continuous and $\mathbf{w}$ is a non-negative vector. We have the following results: (i) The optimal solution of the following optimization problem:*

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \ h(\mathbf{x}) + \langle \mathbf{w}, |\mathbf{Ax}| \rangle \tag{23}$$

will be achieved with $|\mathbf{Ax}^*|_i = 0$ when $\mathbf{w}_i > L/\sigma(\mathbf{A})$ for all $i \in [m]$. (ii) Moreover, if $\|\mathbf{x}^*\| \le \delta$, it always holds that $\langle \mathbf{w}, |\mathbf{Ax}^*| \rangle \le \delta L$.

Proof. (i) We now prove the first part of this lemma. The convex optimization problem in Eq (23) is equivalent to the following minimax saddle point problem:

$$(\mathbf{x}^*, \mathbf{y}^*) = \arg\min_{\mathbf{x}} \ \arg\max_{-\mathbf{1} \le \mathbf{y} \le \mathbf{1}} \ h(\mathbf{x}) + \langle \mathbf{y} \odot \mathbf{w}, \mathbf{Ax} \rangle$$

According to the optimality with respect to $\mathbf{y}$, we have the following result:

$$\begin{aligned} \mathbf{y}_i^* = 1 &\Rightarrow (\mathbf{Ax}^*)_i > 0, \\ |\mathbf{y}_i^*| < 1 &\Rightarrow (\mathbf{Ax}^*)_i = 0, \\ \mathbf{y}_i^* = -1 &\Rightarrow (\mathbf{Ax}^*)_i < 0 \end{aligned} \tag{24}$$

According to the optimality with respect to $\mathbf{x}$, we have the following result:

$$h'(\mathbf{x}) + \mathbf{A}^T(\mathbf{y} \odot \mathbf{w}) = \mathbf{0} \tag{25}$$

where $h'(\mathbf{x})$ denotes the subgradient of $h(\cdot)$ with respect to $\mathbf{x}$.

We now prove that $|\mathbf{y}_i|$ is strictly less than 1 when $\mathbf{w}_i > L/\sigma(A)$, then we obtain $|\mathbf{Ax}^*|_i = 0$ due to the optimality condition in Eq (24). Our proof is as follows. We assume that $\mathbf{y}_i \ne 0$, since otherwise $|\mathbf{y}_i| < 0$, the conclusion holds. Then we derive the following inequalities:

$$\begin{aligned} |\mathbf{y}_i| &< |\mathbf{y}_i| \cdot \frac{\mathbf{w}_i}{L} \cdot \sigma(\mathbf{A}) \\ &\le |\mathbf{y}_i| \cdot \frac{\mathbf{w}_i}{L} \cdot \frac{\|\mathbf{A}^T(\mathbf{y} \odot \mathbf{w})\|}{\|\mathbf{y} \odot \mathbf{w}\|} \\ &\le |\mathbf{y}_i| \cdot \frac{\mathbf{w}_i}{L} \cdot \frac{L}{\|\mathbf{y} \odot \mathbf{w}\|} \\ &\le \frac{|\mathbf{y}_i| \cdot \mathbf{w}_i}{\|\mathbf{y} \odot \mathbf{w}\|_\infty} \\ &\le 1 \end{aligned} \tag{26}$$

where the first step uses the choice that $\mathbf{w}_i > L/\sigma(A)$ and the assumption that $\mathbf{y}_i \ne 0$, the second step uses Lemma 2, the third step uses the fact that $\|\mathbf{A}^T(\mathbf{y} \odot \mathbf{w})\| = \|h'(\mathbf{x})\| \le L$ which can be derived from Eq (25) and the L-Lipschitz continuity of $h(\cdot)$, the fourth step uses the norm inequality that $\|\mathbf{c}\|_\infty \le \|\mathbf{c}\|_2 \ \forall \mathbf{c}$, the last step uses the nonnegativity of $\mathbf{w}$.

We remark that such results are natural and commonplace, since it is well-known that ℓ_1 norm induces sparsity.

(ii) We now prove the second part of this lemma. Recall that if $\mathbf{x}^*$ solves Eq (23), then it must satisfy the following variational inequality [32]:

$$\langle \mathbf{w}, |\mathbf{Ax}| - |\mathbf{Ax}^*| \rangle + \langle \mathbf{x} - \mathbf{x}^*, h'(\mathbf{x}^*) \rangle \ge 0, \ \forall \mathbf{x} \tag{27}$$

Letting $\mathbf{x} = \mathbf{0}$ in Eq (27), using the nonnegativity of $\mathbf{w}$, Cauchy-Schwarz inequality, and the Lipschitz continuity of

17

$h(\cdot)$, we have:

$$
\begin{aligned}
\langle \mathbf{w}, |\mathbf{A}\mathbf{x}^*| \rangle &\leq -\langle \mathbf{x}^*, h'(\mathbf{x}^*) \rangle \\
&\leq \|\mathbf{x}^*\| \cdot \|h'(\mathbf{x}^*)\| \\
&\leq \delta L
\end{aligned}
$$

$\square$

The following lemma shows that the biconvex minimization problem will lead to basic feasibility for the complementarity constraint when the penalty parameter ρ is larger than a threshold.

Lemma 4. *When $\rho > L/\sigma(\mathbf{A})$, any local optimal solution of the minimization problem: $\min_{\mathbf{x},\mathbf{u}} \ \mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$ will be achieved with $\|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle = 0$.*

Proof. We let $\{\mathbf{x}, \mathbf{u}\}$ be any local optimal solution of the biconvex minimization problem $\min_{\mathbf{x},\mathbf{u}} \ \mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$. We denote O as the index of the largest-k value of $|\mathbf{A}\mathbf{x}|$, $G \triangleq \{i | (\mathbf{A}\mathbf{x})_i > 0\}$, and $S \triangleq \{j | (\mathbf{A}\mathbf{x})_j < 0\}$. Moreover, we define $I \triangleq O \cap G$, $J \triangleq O \cap S$ and $K \triangleq \{1, 2, ..., m\} \setminus \{I \cup J\}$.

(i) First of all, we consider the minimization problem of the penalty function with respect to $\mathbf{u}$ (i.e. $\min_{\mathbf{u}} \ \mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$). It reduces to the following minimization problem:

$$
\min_{-\mathbf{1} \leq \mathbf{u} \leq \mathbf{1}} \ \langle -\mathbf{A}\mathbf{x}, \mathbf{u} \rangle, \ s.t. \ \|\mathbf{u}\|_1 \leq k \tag{28}
$$

The objective of Eq (28) essentially computes the k-largest norm function of $\mathbf{A}\mathbf{x}$, see [60, 16]. It is not hard to validate that the optimal solution of Eq (28) can be computed as follows:

$$
\mathbf{u}_i = \begin{cases} +1, & i \in I \\ -1, & i \in J \\ 0, & u \in K \end{cases}
$$

(ii) We now consider the minimization problem of the penalty function with respect to $\mathbf{x}$ (i.e. $\min_{\mathbf{x}} \ \mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$). Clearly, we have that: $\langle (\mathbf{A}\mathbf{x})_I, \mathbf{u}_I \rangle = \|(\mathbf{A}\mathbf{x})_I\|_1$ and $\langle (\mathbf{A}\mathbf{x})_J, \mathbf{u}_J \rangle = \|(\mathbf{A}\mathbf{x})_J\|_1$. It reduces to the following optimization problem for the $\mathbf{x}$-subproblem:

$$
\min_{\mathbf{x}} \ f(\mathbf{x}) + \rho(\|(\mathbf{A}\mathbf{x})_K\|_1 - \langle (\mathbf{A}\mathbf{x})_K, \mathbf{u}_K \rangle).
$$

Since $\mathbf{u}_K = 0$, it can be further simplified as:

$$
\min_{\mathbf{x}} \ f(\mathbf{x}) + \langle \mathbf{w}, |\mathbf{A}\mathbf{x}| \rangle, \ \text{where } \mathbf{w}_i = \begin{cases} \rho, & i \in K \\ 0 & i \notin K \end{cases} \tag{29}
$$

Applying Lemma 3 with $h(\mathbf{x}) = f(\mathbf{x})$, we conclude that $|\mathbf{A}\mathbf{x}|_K = \mathbf{0}$ will be achieved when $\rho \geq L/\sigma(A)$.

Since $|\mathbf{A}\mathbf{x}|_K = \mathbf{0}$, $\langle (\mathbf{A}\mathbf{x})_I, \mathbf{u}_I \rangle = \|(\mathbf{A}\mathbf{x})_I\|_1$ and $\langle (\mathbf{A}\mathbf{x})_J, \mathbf{u}_J \rangle = \|(\mathbf{A}\mathbf{x})_J\|_1$, we achieve that the complementarity constraint $\|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle = 0$ is fully satisfied. This finishes the proof for claim in the lemma.

$\square$

We now show that when the penalty parameter ρ is larger than a threshold, the biconvex objective function $\mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$ is equivalent to the original constrained MPEC problem.

Theorem 1. *Exactness of the Penalty Function. The penalty problem $\min_{\mathbf{x},\mathbf{v}} \ \mathcal{J}_\rho(\mathbf{x}, \mathbf{v})$ admits the same local and global optimal solutions as the original MPEC problem when $\rho > L/\sigma(\mathbf{A})$.*

Proof. First of all, based on the non-separable MPEC reformulation, we have the following Lagrangian function $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$:

$$
\mathcal{H}(\mathbf{x}, \mathbf{u}, \chi) = f(\mathbf{x}) + I(\mathbf{u}) + \chi(\|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle)
$$

Based on the Lagrangian function[13], we have the following KKT conditions for any KKT solution $(\mathbf{x}^*, \mathbf{u}^*, \chi^*)$:

$$
\begin{aligned}
\mathbf{0} &\in \partial f(\mathbf{x}^*) + \chi^* \mathbf{A}^T \partial |\mathbf{A}\mathbf{x}^*| - \chi^* \mathbf{A}^T \mathbf{u}^* \\
\mathbf{0} &\in \partial I(\mathbf{u}^*) - \chi^* \mathbf{A}\mathbf{x}^* \\
0 &= \|\mathbf{A}\mathbf{x}^*\|_1 - \langle \mathbf{A}\mathbf{x}^*, \mathbf{u}^* \rangle
\end{aligned} \tag{30}
$$

The KKT solution is defined as the first-order minimizer (respectively maximizer) for the primal (respectively dual) variables of the Lagrange function. It can be simply derived from setting the (sub-)gradient of the Lagrange function to zero with respect to each block of variables.

Secondly, we focus on the penalty function $\mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$. When $\rho > L/\sigma(\mathbf{A})$, by Lemma 4 we have:

$$
0 = \|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle \tag{31}
$$

By the local optimality of the penalty function $\mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$, we have the following equalities:

$$
\begin{aligned}
\mathbf{0} &\in \partial f(\mathbf{x}) + \rho \mathbf{A}^T \partial |\mathbf{A}\mathbf{x}| - \rho \mathbf{A}^T \mathbf{u} \\
\mathbf{0} &\in \partial I(\mathbf{u}) - \rho \mathbf{A}\mathbf{x}
\end{aligned} \tag{32}
$$

Since Eq (32) and Eq (31) coincide with Eq (30), we conclude that the solution of $\mathcal{J}_\rho(\mathbf{x}, \mathbf{u})$ admits the same local and global optimal solutions as the original non-separable MPEC reformulation, when $\rho > L/\sigma(\mathbf{A})$.

$\square$

Theorem 2. *Convergence rate of Algorithm MPEC-EPM. Assume that $\|\mathbf{x}^t\| \leq \delta$ for all t. Algorithm 1 will converge to the first-order KKT point in at most $\lceil (\ln(L\delta) - \ln(\epsilon \rho^0))/\ln 2 \rceil$ outer iterations with the accuracy at least $\|\mathbf{A}\mathbf{x}\|_1 - \langle \mathbf{A}\mathbf{x}, \mathbf{v} \rangle \leq \epsilon$.*

---

[13]Note that the Lagrangian function $\mathcal{H}$ has the same form as the penalty function $\mathcal{J}$ and the multiplier χ also plays a role of penalty parameter. It always holds at the optimal solution that $\chi^* > 0$ since $\langle \mathbf{A}\mathbf{x}, \mathbf{u} \rangle \leq \|\mathbf{A}\mathbf{x}\|_1 \cdot \|\mathbf{u}\|_\infty \leq \|\mathbf{A}\mathbf{x}\|_1$ and the equality holds at the optimal solution.

Proof. Assume Algorithm 1 takes s outer iterations to converge. We have the following inequalities:

$$
\begin{aligned}
& \|\mathbf{Ax}^{s+1}\|_1 - \langle \mathbf{Ax}^{s+1}, \mathbf{v}^{s+1} \rangle \\
= & \quad \|(\mathbf{Ax}^{s+1})_K\|_1 \\
\leq & \quad \frac{\delta L}{\rho^s}
\end{aligned}
$$

where the first step uses the notations and the results (See Eq (29)) in Lemma 4, the second step uses the second part of Lemma 3 with $h(\mathbf{x}) = f(\mathbf{x})$. The above inequality implies that when $\rho^s \geq \frac{\delta L}{\epsilon}$, Algorithm MPEC-EPM achieves accuracy at least $\|\mathbf{Ax}\|_1 - \langle \mathbf{Ax}, \mathbf{v} \rangle \leq \epsilon$. Noticing that $\rho^s = 2^s \rho^0$, we have that ϵ accuracy will be achieved when

$$
\begin{aligned}
& 2^s \rho^0 \geq \frac{\delta L}{\epsilon} \\
\Rightarrow \quad & 2^s \geq \frac{\delta L}{\epsilon \rho^0} \\
\Rightarrow \quad & s \geq \big(\ln(\delta L) - \ln(\epsilon \rho^0) \big) / \ln 2
\end{aligned}
$$

Thus, we finish the proof of this theorem. $\qquad \square$

C. Convergence Analysis of Algorithm 2

This section presents the proof of Theorem 3 and Theorem 4. Note that our proof is also built upon Lemma 3.

Lemma 5. *When $\bar{\pi} > L/\sigma(\mathbf{A})$, any local optimal solution of the minimization problem:* $\min_{\mathbf{x}, \mathbf{v}} \mathcal{L}(\mathbf{x}, \mathbf{v}, \bar{\pi})$ *will be achieved with* $|\mathbf{Ax}| \odot \mathbf{v} = \mathbf{0}$.

Proof. We assume that $\mathbf{x}$ and $\mathbf{v}$ are arbitrary local optimal solutions of the augmented Lagrangian function for a given $\bar{\pi}$.
(i) Firstly, we now focus on the **v**-subproblem, we have:

$$
\mathbf{v} = \arg \min_{\mathbf{v}} I(\mathbf{v}) + \langle \bar{\pi}, |\mathbf{Ax}| \odot \mathbf{v} \rangle + \frac{\alpha}{2} \| |\mathbf{Ax}| \odot \mathbf{v} \|_2^2
$$

Then there exists a constant $\theta \geq 0$ (that depends on the local optimal solutions $\mathbf{x}$ and $\mathbf{v}$) such that it solves the following minimax saddle point problem:

$$
\begin{aligned}
\max_{\theta \geq 0} \min_{\mathbf{0} \leq \mathbf{v} \leq \mathbf{1}} & \quad \langle \bar{\pi}, |\mathbf{Ax}| \odot \mathbf{v} \rangle + \frac{\alpha}{2} \| |\mathbf{Ax}| \odot \mathbf{v} \|_2^2 \\
& + \theta(n - k - \langle \mathbf{v}, \mathbf{1} \rangle)
\end{aligned}
$$

Clearly, $\mathbf{v}$ can be computed as:

$$
\mathbf{v} = \max(0, \min(1, \frac{\theta - |\mathbf{Ax}| \odot \bar{\pi}}{\alpha |\mathbf{Ax}| \odot |\mathbf{Ax}|})) \tag{33}
$$

We now define $I \triangleq \{i | \theta - |\mathbf{Ax}|_i \cdot \bar{\pi}_i \leq 0\}$, $J \triangleq \{i | \theta - |\mathbf{Ax}|_i \cdot \bar{\pi}_i > 0\}$. Clearly, we have

$$
\mathbf{v}_I = 0, \mathbf{v}_J \neq 0
$$

(ii) Secondly, we focus on the **x**-subproblem:

$$
\min_{\mathbf{x}} \quad f(\mathbf{x}) + \langle |\mathbf{Ax}| \odot \mathbf{v}, \pi \rangle + \frac{\alpha}{2} \| |\mathbf{Ax}| \odot \mathbf{v} \|^2
$$

Noticing $\mathbf{v}_J \neq 0$, we apply Lemma 3 with $\mathbf{w} = \bar{\pi} \odot \mathbf{v}$ and $h(\mathbf{x}) = f(\mathbf{x}) + \frac{\alpha}{2} \| |\mathbf{Ax}| \odot \mathbf{v} \|^2$, then $|\mathbf{Ax}|_J = 0$ will be achieved whenever $\bar{\pi}_J \odot \mathbf{v}_J > L_g / \sigma(\mathbf{A})$, where L_h is the Lipschitz constant of $h(\cdot)$. Since $\mathbf{v}_I = 0$, $(\mathbf{Ax})_J = 0$ and $I \cup J = \{1, 2, ..., m\}$, we have $\frac{\alpha}{2} \| |\mathbf{Ax}| \odot \mathbf{v} \|^2 = 0$, thus, $L_h = L$. Moreover, incorporating $(\mathbf{Ax})_J = 0$ into Eq(33), we have $\mathbf{v}_J = \mathbf{1}$.
(iii) Finally, since $\bar{\pi} > 0$ and $\theta \geq 0$, by the definition of I, we have $|\mathbf{Ax}|_I \geq 0$. In summery, we have the following results:

$$
\mathbf{v}_I = 0, \ \mathbf{v}_J = 1, \ (\mathbf{Ax})_I \geq 0, \ (\mathbf{Ax})_J = 0 \tag{34}
$$

We conclude that the complementarity constraint will hold automatically whenever $\bar{\pi}_J > (L/\sigma(\mathbf{A})) \odot \frac{1}{\mathbf{v}_J} = L/\sigma(\mathbf{A})$, where J is the index that $|\mathbf{Ax}|_J = 0$. However, we do not have any prior information of the index set J (i.e. the value of $\mathbf{Ax}^*$) and J can be any subset of $\{1, 2, ..., m\}$, the condition $\bar{\pi}_J > L/\sigma(\mathbf{A})$ needs to be further restricted to $\bar{\pi} > L/\sigma(\mathbf{A})$. Thus, we complete the proof of this Lemma. $\qquad \square$

Theorem 3. *Boundedness of multiplier and exactness of the Augmented Lagrangian Function. The augmented Lagrangian problem* $\min_{\mathbf{x}, \mathbf{v}} \mathcal{L}(\mathbf{x}, \mathbf{v}, \bar{\pi})$ *admits the same local and global optimal solutions as the original MPEC problem when* $\bar{\pi} \geq L/\sigma(\mathbf{A})$.

Proof. First of all, based on the Lagrangian function $\mathcal{L}(\cdot)$, we have the following KKT conditions for any KKT solution $(\mathbf{x}^*, \mathbf{v}^*, \pi^*)$:

$$
\begin{aligned}
\mathbf{0} & \in \partial f(\mathbf{x}^*) + \mathbf{A}^T (\partial |\mathbf{Ax}^*| \odot \mathbf{v}^* \odot \pi^*) \\
\mathbf{0} & \in \partial I(\mathbf{v}^*) + \mathbf{Ax}^* \odot \pi^* \\
\mathbf{0} & = |\mathbf{Ax}^*| \odot \mathbf{v}^*
\end{aligned} \tag{35}
$$

The KKT solution is defined as the first-order minimizer (respectively maximizer) for the primal (respectively dual) variables of the Lagrange function. It can be simply derived from setting the (sub-)gradient of the Lagrange function to zero with respect to each block of variables.

Secondly, we focus on the penalty function $\mathcal{L}(\mathbf{x}, \mathbf{u}, \bar{\pi})$. When $\bar{\pi} > L/\sigma(\mathbf{A})$, by Lemma 5 we have:

$$
\mathbf{0} = |\mathbf{Ax}| \odot \mathbf{v} \tag{36}
$$

By the local optimality of the penalty function $\mathcal{L}(\mathbf{x}, \mathbf{u}, \bar{\pi})$, we have the following equalities:

$$
\begin{aligned}
\mathbf{0} & \in \partial f(\mathbf{x}) + \mathbf{A}^T (\partial |\mathbf{Ax}| \odot \mathbf{v} \odot \pi) \\
\mathbf{0} & \in \partial I(\mathbf{v}) + \mathbf{Ax} \odot \pi
\end{aligned} \tag{37}
$$

19

Since Eq (36) and Eq (37) coincide with Eq (35), we conclude that the solution of $\mathcal{L}(\mathbf{x}, \mathbf{u}, \bar{\boldsymbol{\pi}})$ admits the same local and global optimal solutions as the original non-separable MPEC reformulation, when $\bar{\boldsymbol{\pi}} > L/\sigma(\mathbf{A})$. $\qquad\square$

In the following, we present the proof of Theorem 4. For the ease of discussions, we define:

$$\mathbf{s} \triangleq \{\mathbf{x}, \mathbf{v}, \boldsymbol{\pi}\}, \ \mathbf{w} \triangleq \{\mathbf{x}, \mathbf{v}\}.$$

First of all, we prove the subgradient lower bound for the iterates gap by the following lemma.

Lemma 3. *Assume that $\mathbf{x}^t$ are bounded for all t, then there exists a constant $\varpi > 0$ such that the following inequality holds:*

$$\|\partial\mathcal{L}(\mathbf{s}^{t+1})\| \leq \varpi\|\mathbf{s}^{t+1} - \mathbf{s}^t\| \qquad (38)$$

Proof. For notation simplicity, we denote $\mathbf{p} = \partial|\mathbf{A}\mathbf{x}^{t+1}|$ and $\mathbf{z} = |\mathbf{A}\mathbf{x}^{t+1}|$. By the optimal condition of the $\mathbf{x}$-subproblem and $\mathbf{v}$-subproblem, we have:

$$
\begin{aligned}
\mathbf{0} \ \in \ & \mathbf{A}^T \left(\mathbf{p} \odot \mathbf{v}^t \odot (\alpha\mathbf{z} \odot \mathbf{v}^t + \boldsymbol{\pi}^t)\right) + f'(\mathbf{x}^{t+1}) \\
& + \mu(\mathbf{x}^{t+1} - \mathbf{x}^t) \\
\mathbf{0} \ \in \ & \left(\alpha\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^t\right) \odot \mathbf{z} + \partial I(\mathbf{v}^{t+1}) + \\
& \mu(\mathbf{v}^{t+1} - \mathbf{v}^t) \qquad\qquad\qquad\qquad (39)
\end{aligned}
$$

By the definition of $\mathcal{L}$ we have that

$$
\begin{aligned}
& \partial\mathcal{L}_{\mathbf{x}}(\mathbf{s}^{t+1}) \\
= \ & \mathbf{A}^T(\mathbf{p} \odot \mathbf{v}^{t+1} \odot \alpha(\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^{t+1}/\alpha)) + f'(\mathbf{x}^{t+1}) \\
= \ & \mathbf{A}^T\mathbf{p} \odot (\mathbf{v}^{t+1} \odot \alpha(\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^{t+1}/\alpha) \\
& -\mathbf{v}^t \odot \alpha(\mathbf{z} \odot \mathbf{v}^t + \boldsymbol{\pi}^t/\alpha)) + \mu(\mathbf{x}^t - \mathbf{x}^{t+1}) \\
= \ & \mathbf{A}^T\mathbf{p} \odot (\mathbf{v}^{t+1} \odot \alpha(\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^{t+1}/\alpha) \\
& -\mathbf{v}^t \odot \alpha(\mathbf{z} \odot \mathbf{v}^t + \boldsymbol{\pi}^t/\alpha)) + \mu(\mathbf{x}^t - \mathbf{x}^{t+1}) \\
= \ & \mathbf{A}^T\mathbf{p} \odot (\mathbf{v}^{t+1} \odot \alpha(\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^{t+1}/\alpha \\
& -\mathbf{v}^t \odot \alpha(\mathbf{z} \odot \mathbf{v}^t + (\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t+1} + \boldsymbol{\pi}^t)/\alpha)) \\
& + \mu(\mathbf{x}^t - \mathbf{x}^{t+1}) \\
= \ & \mathbf{A}^T\mathbf{p} \odot (\mathbf{v}^{t+1} \odot \alpha(\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^{t+1}/\alpha \\
& -\mathbf{v}^t \odot \alpha(\mathbf{z} \odot \mathbf{v}^t + (\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t+1} + \boldsymbol{\pi}^t)/\alpha)) \\
& + \mu(\mathbf{x}^t - \mathbf{x}^{t+1}) \\
= \ & \mathbf{A}^T\mathbf{p} \odot \mathbf{v}^t \odot (\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t+1}) + \mu(\mathbf{x}^t - \mathbf{x}^{t+1}) + \\
& \mathbf{A}^T\mathbf{p} \odot (\mathbf{v}^{t+1} \odot (\alpha\mathbf{z} \odot \mathbf{v}^{t+1} + \boldsymbol{\pi}^{t+1}) - \\
& \mathbf{v}^t \odot (\alpha\mathbf{z} \odot \mathbf{v}^t + \boldsymbol{\pi}^{t+1})) \\
= \ & \mathbf{A}^T\mathbf{p} \odot \mathbf{v}^t \odot (\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t+1}) + \mu(\mathbf{x}^t - \mathbf{x}^{t+1}) + \\
& \mathbf{A}^T\mathbf{p} \odot \boldsymbol{\pi}^{t+1} \odot (\mathbf{v}^{t+1} - \mathbf{v}^t) + \\
& \mathbf{A}^T\mathbf{p} \odot \alpha\mathbf{z} \odot (\mathbf{v}^{t+1} \odot \mathbf{v}^{t+1} - \mathbf{v}^t \odot \mathbf{v}^t) \qquad (40)
\end{aligned}
$$

The first step uses the definition of $\mathcal{L}_{\mathbf{x}}(\mathbf{s}^{t+1})$, the second step uses Eq (39), the third step uses $\mathbf{v}^t + \mathbf{v}^{t+1} - \mathbf{v}^{t+1} = \mathbf{v}^t$, the fourth step uses the multiplier update rule for $\boldsymbol{\pi}$.

We assume that $\mathbf{x}^{t+1}$ is bounded by δ for all t, i.e. $\|\mathbf{x}^{t+1}\| \leq \delta$. By Theorem 3, $\boldsymbol{\pi}^{t+1}$ is also bounded. We assume it is bounded by a constant κ, i.e. $\|\boldsymbol{\pi}^{t+1}\| \leq \kappa$. Using the fact that $\|\partial\|\mathbf{A}\mathbf{x}\|_1\|_\infty \leq 1$, $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|\|\mathbf{x}\|$, $\|\mathbf{x} \odot \mathbf{y}\|_2 \leq \|\mathbf{x}\|_\infty\|\mathbf{y}\|$ and Eq (40), we have the following inequalities:

$$
\begin{aligned}
& \|\partial\mathcal{L}_{\mathbf{x}}(\mathbf{s}^{t+1})\| \\
\leq \ & \|\mathbf{A}\| \cdot \|\mathbf{p}\|_\infty \cdot \|\mathbf{v}\|_\infty \cdot \|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t+1}\| + \mu\|\mathbf{x}^t - \mathbf{x}^{t+1}\| + \\
& \|\mathbf{A}\| \cdot \|\mathbf{p}\|_\infty \cdot \|\boldsymbol{\pi}^{t+1}\|_\infty \cdot \|\mathbf{v}^{t+1} - \mathbf{v}^t\| + \\
& \|\mathbf{A}\| \cdot \|\mathbf{p}\|_\infty \cdot \alpha\|\mathbf{z}\| \cdot \|\mathbf{v}^{t+1} + \mathbf{v}^t\|_\infty \cdot \|\mathbf{v}^{t+1} - \mathbf{v}^t\| \\
\leq \ & \|\mathbf{A}\|\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t+1}\| + \mu\|\mathbf{x}^{t+1} - \mathbf{x}^t\| + \\
& (\kappa + 2\delta\alpha\|\mathbf{A}\|) \cdot \|\mathbf{A}\| \cdot \|\mathbf{v}^{t+1} - \mathbf{v}^t\| \qquad (41)
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\partial\mathcal{L}_{\mathbf{v}}(\mathbf{s}^{t+1}) \ = \ & \partial I(\mathbf{v}^{t+1}) + \alpha\mathbf{z} \odot (\boldsymbol{\pi}^{t+1}/\alpha + \mathbf{z} \odot \mathbf{v}^{t+1}) \\
= \ & \mathbf{z} \odot (\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^t) - \mu(\mathbf{v}^{t+1} - \mathbf{v}^t)
\end{aligned}
$$

$$\partial\mathcal{L}_{\boldsymbol{\pi}}(\mathbf{s}^{t+1}) = |\mathbf{A}\mathbf{x}^{t+1} \odot \mathbf{v}^{t+1}| = \tfrac{1}{\alpha}(\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^t)$$

Then we derive the following inequalities:

$$\|\partial\mathcal{L}_{\mathbf{v}}(\mathbf{s}^{t+1})\| \leq \delta\|\mathbf{A}\|\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t+1}\| + \mu\|\mathbf{v}^{t+1} - \mathbf{v}^t\| \quad (42)$$

$$\|\partial\mathcal{L}_{\boldsymbol{\pi}}(\mathbf{s}^{t+1})\| \leq \tfrac{1}{\alpha}\|\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^t\| \qquad\qquad (43)$$

Combining Eqs (41-43), we conclude that there exists $\varpi > 0$ such that the following inequality holds

$$\|\partial\mathcal{L}(s^{t+1})\| \leq \varpi\|\mathbf{s}^{t+1}\|$$

We thus complete the proof of this lemma. $\qquad\square$

The following lemma is useful in our convergence analysis.

Proposition 1. *Assume that $\mathbf{x}^t$ are bounded for all t, then we have the following inequality:*

$$\sum_{t=0}^{\infty} \|\mathbf{s}^t - \mathbf{s}^{t+1}\|^2 < +\infty$$

In particular the sequence $\|\mathbf{s}^t - \mathbf{s}^{t+1}\|$ is asymptotic regular, namely $\|\mathbf{s}^t - \mathbf{s}^{t+1}\| \to 0$ as $t \to \infty$. Moreover any cluster point of $\mathbf{s}^t$ is a stationary point of $\mathcal{L}$.

Proof. Due to the initialization and the update rule of $\boldsymbol{\pi}$, we conclude that $\boldsymbol{\pi}^t$ is nonnegative and monotone nondecreasing. Moreover, using the result of Theorem 3, as $t \to \infty$, we have: $|\mathbf{A}\mathbf{x}^{t+1}| \odot \mathbf{v}^{t+1} = \mathbf{0}$. Therefore, we conclude that as $t \to +\infty$ it must hold that

$$
\begin{aligned}
& |\mathbf{A}\mathbf{x}^{t+1}| \odot \mathbf{v}^{t+1} = \mathbf{0}, \\
& \sum_{i=1}^t \|\boldsymbol{\pi}^{i+1} - \boldsymbol{\pi}^i\| < +\infty, \\
& \sum_{i=1}^t \|\boldsymbol{\pi}^{i+1} - \boldsymbol{\pi}^i\|^2 < +\infty
\end{aligned}
$$

On the other hand, we naturally derive the following inequalities:

$$\mathcal{L}(\mathbf{x}^{t+1}, \mathbf{v}^{t+1}; \boldsymbol{\pi}^{t+1})$$
$$= \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{v}^{t+1}; \boldsymbol{\pi}^{t}) + \langle \boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t}, |\mathbf{A}\mathbf{x}^{t+1}| \odot \mathbf{v}^{t+1} \rangle$$
$$= \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{v}^{t+1}; \boldsymbol{\pi}^{t}) + \frac{1}{\alpha} \|\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t}\|^2$$
$$\leq \mathcal{L}(\mathbf{x}^{t}, \mathbf{v}^{t+1}; \boldsymbol{\pi}^{t}) - \frac{\mu}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|^2 + \frac{1}{\alpha}\|\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t}\|^2$$
$$\leq \mathcal{L}(\mathbf{x}^{t}, \mathbf{v}^{t}; \boldsymbol{\pi}^{t}) - \frac{\mu}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|^2 - \frac{\mu}{2}\|\mathbf{v}^{t+1} - \mathbf{v}^{t}\|^2$$
$$+ \frac{1}{\alpha}\|\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t}\|^2 \qquad (44)$$

The first step uses the definition of $\mathcal{L}$; the second step uses update rule of the Lagrangian multiplier $\boldsymbol{\pi}$; the third and fourth step use the μ-strongly convexity of $\mathcal{L}$ with respect to $\mathbf{x}$ and $\mathbf{v}$, respectively.

We define $C = \frac{1}{\alpha}\sum_{i=1}^{t} \|\boldsymbol{\pi}^{i+1} - \boldsymbol{\pi}^{i}\|^2 + \mathcal{L}(\mathbf{x}^0, \mathbf{v}^0; \boldsymbol{\pi}^0) - \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{v}^{t+1}; \boldsymbol{\pi}^{t+1})$. Clearly, by the boundedness of $\{\mathbf{x}^t, \mathbf{v}^t, \boldsymbol{\pi}^t\}$, both C and $\mathcal{L}(\mathbf{x}^{t+1}, \mathbf{v}^{t+1}; \boldsymbol{\pi}^{t+1})$ are bounded. Summing Eq (44) over $i = 1, 2..., t$, we have:

$$\frac{\mu}{2}\sum_{i=1}^{t} \|\mathbf{w}^{i+1} - \mathbf{w}^{i}\|^2 \leq C \qquad (45)$$

Therefore, combining Eq(44) and Eq (45), we have $\sum_{t=1}^{+\infty} \|\mathbf{s}^{t+1} - \mathbf{s}^{t}\|^2 < +\infty$; in particular $\|\mathbf{s}^{t+1} - \mathbf{s}^{t}\| \to 0$. By Eq (38), we have that:

$$\|\partial \mathcal{L}(\mathbf{s}^{t+1})\| \leq \varpi \|\mathbf{s}^{t+1} - \mathbf{s}^{t}\| \to 0$$

which implies that any cluster point of $\mathbf{s}^t$ is a stationary point of $\mathcal{L}$. We complete the proof of this lemma.

□

Remarks: Lemma 1 states that any cluster point is the KKT point. Strictly speaking, this result does not imply the convergence of the algorithm. This is because the boundedness of $\sum_{t=0}^{\infty} \|\mathbf{s}^t - \mathbf{s}^{t+1}\|^2$ does not imply that the sequence $\mathbf{s}^t$ is convergent [14]. In what follows, we aim to prove stronger result in Theorem 4.

Our analysis is mainly based on a recent non-convex analysis tool called Kurdyka-Łojasiewicz inequality [2, 10]. One key condition of our proof requires that the Lagrangian function $\mathcal{L}(\mathbf{s})$ satisfies the so-call (KL) property in its effective domain. It is so-called the semi-algebraic function satisfy the Kurdyka-Łojasiewicz property. We note that semi-algebraic functions include (i) real polynomial functions, (ii) finite sums and products of semi-algebraic functions , and (iii) indicator functions of semi-algebraic sets [10]. Using these definitions repeatedly, the graph of $\mathcal{L}(\mathbf{s}) : \{(\mathbf{s}, z)| z = \mathcal{L}(\mathbf{s})\}$ can be proved to be a semi-algebraic set. Therefore, the Lagrangian function $\mathcal{L}(\mathbf{s})$ is a semi-algebraic function. This is not surprising since semi-algebraic function is ubiquitous in applications [10]. We now present the following proposition established in [2].

---

[14]One typical counter-example is $\mathbf{s}^t = \sum_{i=1}^{t} \frac{1}{i}$. Clearly, $\sum_{t=0}^{\infty} \|\mathbf{s}^t - \mathbf{s}^{t+1}\|^2 = \sum_{t=1}^{\infty} (\frac{1}{t})^2$ is bounded by $\frac{\pi^2}{6}$; however, $\mathbf{s}^t$ is divergent since $s^t = \ln(k) + C_e$, where C_e is the well-known Euler's constant.

Proposition 2. *For a given semi-algebraic function $\mathcal{L}(s)$, for all $s \in dom\mathcal{L}$, there exists $\theta \in [0, 1)$, $\eta \in (0, +\infty]$ a neighborhood $\mathcal{S}$ of s and a concave and continuous function $\varphi(\varsigma) = \varsigma^{1-\theta}, \varsigma \in [0, \eta)$ such that for all $\bar{s} \in \mathcal{S}$ and satisfies $\mathcal{L}(\bar{s}) \in (\mathcal{L}(s), \mathcal{L}(s) + \eta)$, the following inequality holds:*

$$dist(0, \partial \mathcal{L}(\bar{\mathbf{s}}))\varphi'(\mathcal{L}(\mathbf{s}) - \mathcal{L}(\bar{\mathbf{s}})) \geq 1, \ \forall \mathbf{s}$$

where $dist(0, \partial \mathcal{L}(\bar{\mathbf{s}})) = \min\{\|w^*\| : w^* \in \partial \mathcal{L}(\bar{\mathbf{s}})\}$.

We now ready to prove the convergence of our alternating direction method.

Theorem 4. *Assume that $\mathbf{x}^t$ are bounded for all t. Then we have the following inequality:*

$$\sum_{t=0}^{+\infty} \|\mathbf{s}^t - \mathbf{s}^{t+1}\| < \infty \qquad (46)$$

Moreover, as $t \to +\infty$, Algorithm MPEC-ADM converges to a first order KKT point of the reformulated MPEC problem.

Proof. For simplicity, we define $r^t = \varphi(\mathcal{L}(\mathbf{s}^t) - \mathcal{L}(\mathbf{s}^*)) - \varphi(\mathcal{L}(\mathbf{s}^{t+1}) - \mathcal{L}(\mathbf{s}^*))$. We naturally derive the following inequalities:

$$\frac{\mu}{2}\|\mathbf{w}^{t+1} - \mathbf{w}^{t}\|^2 - \frac{1}{\alpha}\|\boldsymbol{\pi}^{t+1} - \boldsymbol{\pi}^{t}\|^2$$
$$\leq \mathcal{L}(\mathbf{s}^t) - \mathcal{L}(\mathbf{s}^{t+1})$$
$$= (\mathcal{L}(\mathbf{s}^t) - \mathcal{L}(\mathbf{s}^*)) - (\mathcal{L}(\mathbf{s}^{t+1}) - \mathcal{L}(\mathbf{s}^*))$$
$$\leq r^t / \varphi'(\mathcal{L}(\mathbf{s}^t) - \mathcal{L}(\mathbf{s}^*))$$
$$\leq r^t \cdot dist(0, \partial \mathcal{L}(\mathbf{s}^t))$$
$$\leq r^t \cdot \varpi \left(\|\mathbf{x}^t - \mathbf{x}^{k-1}\| + \|\mathbf{v}^t - \mathbf{v}^{k-1}\| + \|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{k-1}\| \right)$$
$$\leq r^t \cdot \varpi \left(\sqrt{2}\|\mathbf{w}^t - \mathbf{w}^{k-1}\| + \|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{k-1}\| \right)$$
$$= r^t \varpi \sqrt{\frac{16}{\mu}} \left(\sqrt{\frac{\mu}{8}}\|\mathbf{w}^t - \mathbf{w}^{k-1}\| + \sqrt{\frac{\mu}{16}}\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{k-1}\| \right)$$
$$\leq \left(\sqrt{\frac{\mu}{8}}\|\mathbf{w}^t - \mathbf{w}^{k-1}\| + \sqrt{\frac{\mu}{16}}\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{k-1}\| \right)^2 + (r^t \varpi \sqrt{\frac{4}{\mu}})^2$$

The first step uses Eq (44); the third step uses the concavity of φ such that $\varphi(a) - \varphi(b) \geq \varphi'(a)(a - b)$ for all $a, b \in \mathbb{R}$; the fourth step uses the KL property such that $dist(0, \partial \mathcal{L}(\mathbf{s}^t))\varphi'(\mathcal{L}(\mathbf{s}^t) - \mathcal{L}(\mathbf{s}^*)) \geq 1$ as in Proposition 2; the fifth step uses Eq (38); the sixth step uses the inequality that $\|x; y\| \leq \|x\| + \|y\| \leq \sqrt{2}\|x; y\|$, where ';' in $[\cdot]$ denotes the row-wise partitioning indicator as in Matlab; the last step uses that fact that $2ab \leq a^2 + b^2$ for all $a, b \in \mathbb{R}$. As $k \to \infty$, we have:

$$\frac{\mu}{2}\|\mathbf{w}^{t+1} - \mathbf{w}^{t}\|^2 \leq \frac{4(r^t \kappa)^2}{\mu} + \left(\sqrt{\frac{\mu}{8}}\|\mathbf{w}^t - \mathbf{w}^{k-1}\| \right)^2$$

Taking the squared root of both side and using the inequality that $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \in \mathbb{R}_+$, we have

$$2\sqrt{\frac{\mu}{8}}\|\mathbf{w}^{t+1} - \mathbf{w}^{t}\| \leq \frac{2r^t \kappa}{\sqrt{\mu}} + \sqrt{\frac{\mu}{8}}\|\mathbf{w}^t - \mathbf{w}^{k-1}\|$$

Then we have

$$\sqrt{\frac{\mu}{8}}\|\mathbf{w}^{t+1} - \mathbf{w}^t\|$$

$$\leq \quad \frac{2r^t\kappa}{\sqrt{\mu}} + \sqrt{\frac{\mu}{8}}(\|\mathbf{w}^t - \mathbf{w}^{k-1}\| - \|\mathbf{w}^{t+1} - \mathbf{w}^t\|) \quad (47)$$

Summing Eq (47) over $i = 1, 2..., t$, we have:

$$\sum_{i=1}^{t}\sqrt{\frac{\mu}{8}}\|\mathbf{w}^{i+1} - \mathbf{w}^i\|$$

$$\leq \quad \frac{\kappa}{\sqrt{\mu}}\sum_{i=1}^{t} r^i + \left(\|\mathbf{w}^1 - \mathbf{w}^0\| + \|\mathbf{w}^{t+1} - \mathbf{w}^t\|\right) \quad (48)$$

The first term in the right-hand side of Eq(48) is bounded since $\sum_{i=1}^{t} r^i = \varphi(\mathcal{L}(\mathbf{s}^0) - \mathcal{L}(\mathbf{s}^*)) - \varphi(\mathcal{L}(\mathbf{s}^{t+1}) - \mathcal{L}(\mathbf{s}^*))$ is bounded. Therefore, we conclude that as $k \rightarrow \infty$, we obtain:

$$\sum_{i=1}^{t}\|\mathbf{w}^{i+1} - \mathbf{w}^i\| < +\infty$$

By the boundedness of $\boldsymbol{\pi}$ in Theorem 3, we have $\sum_{i=1}^{t}\|\boldsymbol{\pi}^{i+1} - \boldsymbol{\pi}^i\| < +\infty$. Therefore, we obtain:

$$\sum_{i=1}^{t}\|\mathbf{s}^{i+1} - \mathbf{s}^i\| < +\infty$$

Finally, by Eq (38) in Lemma 3 we have $\partial\mathcal{L}(\mathbf{s}^{t+1}) = 0$. In other words, we have the following results:

$$\mathbf{0} \quad \in \quad f'(\mathbf{x}^{t+1}) + \mathbf{A}^T(\partial|\mathbf{A}\mathbf{x}^{t+1}| \odot \mathbf{v}^{t+1} \odot \boldsymbol{\pi}^{t+1})$$
$$\mathbf{0} \quad \in \quad \boldsymbol{\pi}^{t+1} \odot |\mathbf{A}\mathbf{x}^{t+1}| + \partial I(\mathbf{v}^{t+1})$$
$$\mathbf{0} \quad = \quad |\mathbf{A}\mathbf{x}^{t+1}| \odot \mathbf{v}^{t+1}$$

which coincide with the KKT conditions in Eq (35). Therefore, we conclude that the solution $\{\mathbf{x}^{t+1}, \mathbf{v}^{t+1}, \boldsymbol{\pi}^{t+1}\}$ converges to a first-order KKT point of the reformulated MPEC problem.

□

D. Matlab Code of Breakpoint Search Algorithm [33]

```matlab
function [x] = BreakPointSearch(d,a,s,oper)
% This program solves the following QP:
% min_x 0.5*x'*diag(d)*x + a'*x
% s.t. sum(x) oper s,  0 <= x <= 1
% oper can be '==', '>=', or '<='
% We assume d > 0.

n = length(d); assert( s>=0 && s<=n );
y = sort([-a-d;-a],'ascend');
l = 1; r = 2*n; L = n; R = 0;
while 1
    if(r-l==1)
        S = (s-L)/(R-L);
        rho = y(l)+(y(r) - y(l))*S;
        break;
    else
        m = floor(0.5*(l+r));
        t = (-a-y(m))./d;
        C = sum(max(0,min(t,1)));
        if(C==s),
            rho = y(m); break;
        elseif(C>s)
            l = m;L=C;
        elseif(C<s)
            r = m;R=C;
        end
    end
end
if (oper(1)=='>')
rho = min(rho,0);
elseif(oper(1)=='<')
rho = max(rho,0);
end
t = (-a-rho)./d;
x = max(0,min(t,1));
```