Computational Optimization and Applications (2023) 86:627–667
https://doi.org/10.1007/s10589-023-00496-x

Zero-norm regularized problems: equivalent surrogates, proximal MM method and statistical error bound

Dongdong Zhang[1] · Shaohua Pan[1] · Shujun Bi[1] · Defeng Sun[2]

Received: 23 April 2022 / Accepted: 20 May 2023 / Published online: 6 June 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

For the zero-norm regularized problem, we verify that the penalty problem of its equivalent MPEC reformulation is a global exact penalty, which implies a family of equivalent surrogates. For a subfamily of these surrogates, the critical point set is demonstrated to coincide with the d-directional stationary point set and when a critical point has no too small nonzero component, it is a strongly local optimal solution of the surrogate problem and the zero-norm regularized problem. We also develop a proximal majorization-minimization (MM) method for solving the DC (difference of convex functions) surrogates, and provide its global and linear convergence analysis. For the limit of the generated sequence, the statistical error bound is established under a mild condition, which implies its good quality from a statistical respective. Numerical comparisons with ADMM for solving the DC surrogate and APG for solving its partially smoothed form indicate that our proximal MM method armed with an inexact dual PPA plus the semismooth Newton method (PMMSN for short) is remarkably superior to ADMM and APG in terms of the quality of solutions and the CPU time.

Keywords Zero-norm regularized problems · Equivalent DC surrogates · Proximal MM method · Statistical error bound

Mathematics Subject Classification 90C27 · 90C31 · 49M20

✉ Shujun Bi
bishj@scut.edu.cn

Shaohua Pan
shhpan@scut.edu.cn

Defeng Sun
defeng.sun@polyu.edu.hk

[1] School of Mathematics, South China University of Technology, Guangzhou, China

[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

1 Introduction

Let $A \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$ be the given data matrix and vector, and let $\vartheta : \mathbb{R}^n \to [0, +\infty]$ be a proper function such that $[\mathrm{Im}(A) - b] \cap \mathrm{dom}\vartheta \neq \emptyset$. We are interested in the zero-norm regularized nonsmooth loss minimization problem

$$\min_{x\in\mathbb{R}^p} \Theta_{\nu,\mu}(x) := \vartheta(Ax-b) + \nu\|x\|_0 + (\mu/2)\|x\|^2, \tag{1}$$

where $\nu > 0$ is the regularization parameter, $\|\cdot\|_0$ is the zero-norm (cardinality) of vectors, and $\mu > 0$ is a tiny constant. The term $\frac{1}{2}\mu\|x\|^2$, due to the lower boundedness of the function $x \mapsto \vartheta(Ax-b) + \nu\|x\|_0$, ensures that problem (1) has a nonempty compact global optimal solution set.

Since the zero-norm is the root to promote sparsity, problem (1) has wide applications in a host of scientific and engineering fields such as regression and variable selection in statistics (see, e.g., [21, 54]), compressed sensing [18] and source separation [9] in signal processing, feature selection and classification in machine learning [8, 57], and so on. In particular, since ϑ is not required to be differentiable, problem (1) often arises from robust models involving a piecewise linear-quadratic (PLQ) convex loss $\vartheta(Ax-b)$ with $\vartheta(z) = \frac{1}{n}\sum_{i=1}^n \theta(z_i)$ where $\theta : \mathbb{R} \to [0, +\infty]$. For example, when $\theta(t) = |t|$ for $t \in \mathbb{R}$,

$$\vartheta(Ax-b) = \frac{1}{n}\|Ax-b\|_1, \tag{2}$$

and (1) reduces to the sparsity regularized ℓ_1-loss model for robust sparse recovery [59]; and when $\theta(t) = (\tau - \mathbb{I}_{\{t\le 0\}})t$ for $t \in \mathbb{R}$ with some $\tau \in (0, 1)$,

$$\vartheta(Ax-b) = \frac{1}{n}\sum_{i=1}^n (\tau - \mathbb{I}_{\{(Ax-b)_i\le 0\}})(Ax-b)_i, \tag{3}$$

and it becomes the sparsity regularized check-loss model to monitor the heteroscedasticity of high-dimensional data [55].

1.1 Existing related works

Problem (1) is generally NP-hard due to the combination of the zero-norm, and it is impractical to seek a global optimizer via an algorithm with a polynomial-time complexity. A common way to deal with this class of problems is to adopt the convex relaxation technique to obtain a desirable solution in a statistical sense. The ℓ_1-norm convex relaxation, as a popular relaxation method, has witnessed significant progress in theory and computation since the early works [17, 54]. Although the ℓ_1-norm is the convex envelope of the zero-norm on the ℓ_∞-norm unit ball, its ability to promote sparsity is weak especially in a complicated constraint set, say, the simplex set. Inspired by this, many nonconvex surrogates have been proposed for the zero-norm function, including the non-Lipschitz ℓ_p $(0 < p < 1)$ surrogate [11, 12], smooth

concave approximation [8, 47, 57], and the folded concave functions such as SCAD and MCP [21, 64]. All of these surrogates are proposed from the primal perspective, and moreover, the surrogate problems associated to the first two classes are only an approximation to problem (1) and the approximation effect depends on whether p or the smoothing parameter is close to 0.

Soubies et al. [52] ever proposed a class of exact continuous relaxations to the ℓ_2-ℓ_0 minimization, but their proof depends on the structure of the least-square loss and it is unclear whether or not their results are applicable to model (1) with a nonsmooth loss. For the nonsmooth loss as in (1), Bian and Chen [5] recently verified that the surrogate problem associated to the capped ℓ_1-norm surrogate is an exact continuous relaxation, but their proof fully exploits the structure of the capped ℓ_1-norm and is not applicable to analyzing the exactness of other surrogates. Then, it is natural to ask whether there is a unified mechanism to analyze the exactness or equivalence of nonconvex surrogates. This work provides such a unified analysis technique by leveraging the metric subregularity, a Lipschitz-like continuity, of the sparsity constrained system.

For zero-norm regularized smooth loss optimization problems, a large number of algorithms have been developed by using the sufficient smoothness of the loss function and/or the closed form of the proximal operator of the zero-norm (see, e.g., [28, 35, 60, 65]). By contrast, the algorithms for zero-norm regularized nonsmooth loss optimization problems receive less attention except [5, 61], in which the nonsmooth difficulty of the loss function is overcome by the smoothing technique. As will be shown by the numerical results in Sect. 5.2, an appropriate smoothing parameter is hard to choose because it is very sensitive to the data, especially for those with a highly-relevant covariance and a heavily-tailed noise. Another way to overcome the nonsmooth difficulty of the loss function is to adopt the alternating direction of multiplier method (ADMM), but for zero-norm or its nonconvex surrogate regularized nonsmooth loss problems, there is lack of convergence certificate for the ADMM, and the existing convergence analysis in [7, 56] is inapplicable to it. Thus, to develop an algorithm, which is efficient in practice and has a theoretical certificate, even for problem (1) with a nonsmooth convex loss is imperative.

For optimization models involving a smooth loss and a nonconvex surrogate of the zero-norm, there are some works to focus on the error bounds of their stationary points to the true vector (see, e.g., [10, 36]) or the oracle property of a local optimizer yielded by a specific algorithm [22]. However, for the models involving a nonsmooth loss and a DC surrogate of zero-norm, to the best of our knowledge, there are no work to investigate the statistical error bound of the stationary point yielded by an algorithm. The model in [53] involves a square-root loss and a DC surrogate of zero-norm, but the statistical error bound of the stationary point was not derived. For the zero-norm regularized nonsmooth convex loss model, Bian and Chen [5] proposed a smoothing proximal gradient algorithm to seek a lifted stationary point of the capped ℓ_1-surrogate, but they did not provide a statistical error bound for the stationary point.

1.2 The main contributions

The first contribution of this work is to provide a unified mechanism to capture equivalent surrogates for problem (1) from a primal-dual viewpoint. We achieve this goal by leveraging its mathematical program with equilibrium constraint (MPEC) reformulation and showing that the penalty problem induced by the equilibrium constraint is a global exact penalty of the MPEC. The SCAD, MCP and capped ℓ_1 functions are illustrated to be a member of this family. In particular, for a subfamily of these surrogates, we demonstrate the strong local optimality of their critical points to the surrogate problems and the zero-norm regularized problem. To reformulate the ℓ_0-norm regularized problem as an MPEC is a common practice, just as in [4, 23] do for the zero-norm minimization problem, but to verify that its penalty problem induced by the equilibrium constraint is a global exact penalty is more troublesome than to do for the latter, because the former involves the growth of an additional function. For the zero-norm constrained problems, Gotoh et al. [25] recently presented an equivalent DC surrogate by proving the penalty problem induced by the zero-norm constraint to be exact in a global sense, but their exact penalty analysis technique is inapplicable to our MPEC reformulation. Le Thi et al. [31] got an equivalent DC surrogate for the zero-norm regularized problem from a primal-dual viewpoint, while their surrogate is different from ours.

The second contribution is to develop a proximal MM method for solving the equivalent DC surrogate models. Different from the work [53], our proximal MM method is not a special case of DC algorithms [32, 43] because it is based on a tighter majorization of the DC surrogate, and moreover, its linear convergence rate could be achieved without any conditions provided that ϑ and ϕ (to induce the DC surrogate) are PLQ functions definable in an o-minimal structure. From [2, Section 4], such PLQ functions are extensive. In addition, there are huge literature on MM methods and DC algorithms for nonconvex and nonsmooth problems, but few of them discuss the local optimality of the limit to the iterate sequence generated. It is shown that the limit of the iterate sequence of our algorithm is a strongly local optimal solution to problem (1) and its DC surrogate whenever its smallest nonzero component is not too small.

The last contribution, for the scenario where the data (b, A) comes from a noisy linear observation model with respect to a true but unknown $x^* \in \mathbb{R}^p$, is to achieve a non-asymptotic statistical error bound for the limit of the generated sequence to the true x^*. This error bound not only implies the good quality of the obtained limit from a statistical perspective, but also clarifies the relation between the sample size and the sparsity of the true x^*.

For the proposed proximal MM armed with an inexact dual proximal point algorithm (PPA) plus the powerful semismooth Newton to solve the subproblems, termed as PMMSN, we conduct numerical experiments on synthetic and real data examples, and compare its performance with that of ADMM for the DC surrogate problem and the accelerated proximal gradient (APG) method for its partially smoothed form. The results indicate that PMMSN has a significant superiority in the quality of solutions and the CPU time, and is very robust for the data with a highly-relevant covariance and a heavily-tailed noise.

2 Notation and preliminaries

Throughout this paper, I and e denote an identity matrix and a vector of all ones with dimension known from the context, and $\|\cdot\|$, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the ℓ_2-norm, ℓ_1-norm and ℓ_∞-norm of vectors, respectively. For a matrix $X \in \mathbb{R}^{n\times p}$, $\|X\|$, $\|X\|_\infty$ and $|||X|||_1$ respectively denote the spectral norm, the elementwise maximum norm and the column sum norm of X; and for the given index sets $\mathcal{I} \subseteq \{1,\ldots,n\}$ and $\mathcal{J} \subseteq \{1,\ldots,p\}$, $X_{\mathcal{I}\cdot}$ and $X_{\cdot\mathcal{J}}$ are the submatrix consisting of those rows $X_{i\cdot}$ for $i \in \mathcal{I}$ and those columns $X_{\cdot j}$ for $j \in \mathcal{J}$, respectively. For a set S, $\mathrm{conv}(S)$ means the convex hull of S, $\mathbb{I}_S$ means the characteristic function of S that takes 1 if $x \in S$ and 0 otherwise, and δ_S denotes the indicator function of S that takes 0 if $x \in S$ and ∞ otherwise. For a vector $x \in \mathbb{R}^p$, $|x|_{\mathrm{nz}}$ represents the smallest nonzero entry of the vector $|x|{:=}(|x_1|,\ldots,|x_p|)^{\mathbb{T}}$ (also called the smallest nonzero entry of x though not accurately), and $|x|^{\downarrow}$ is the vector obtained by arranging the entries of $|x|$ in a nonincreasing order. For given vector $x \in \mathbb{R}^p$, $\mathrm{dist}(x,S)$ means the Euclidean distance of x from a closed set $S \subseteq \mathbb{R}^p$, and $\mathbb{B}_\infty(x,\delta)$ and $\mathbb{B}(x,\delta)$ for a $\delta > 0$ denote the closed ball on the ℓ_∞-norm and the ℓ_2-norm centered at x with radius δ, which are respectively simplified as $\mathbb{B}_\infty(\delta)$ and $\mathbb{B}(\delta)$ if $x = 0$. For a given $p \times p$ positive definite matrix Q, $\|x\|_Q{:=}\sqrt{\langle x, Qx\rangle}$ represents the norm associated to Q. For an extended real-valued $h : \mathbb{R}^p \to [-\infty,+\infty]$, h is said to be proper if its domain $\mathrm{dom}\, h{:=}\{x \in \mathbb{R}^p \mid h(x) < +\infty\}$ is nonempty, and its conjugate is defined by $h^*(z){:=}\sup_{x\in\mathbb{R}^p}\{\langle x,z\rangle - h(x)\}$. In the sequel, we often use

$$f(x){:=}\vartheta(Ax-b) \ \text{ and } \ f_\mu(x){:=}f(x)+(\mu/2)\|x\|^2 \ \text{ for } x \in \mathbb{R}^p. \tag{4}$$

2.1 Partial calmness of optimization problems

Let $\varphi : \mathbb{R}^p \to (-\infty,+\infty]$ be a proper lower semicontinuous (lsc) function, $h : \mathbb{R}^p \to \mathbb{R}$ be a continuous function, and Δ be a nonempty closed set of $\mathbb{R}^p$. In order to introduce the partial calmness of the optimization problem

$$\text{(MP)} \qquad \min_{z\in\mathbb{R}^p}\{\varphi(z) \ \text{s.t.} \ h(z)=0,\ z\in\Delta\},$$

we consider its following perturbation:

$$(\text{MP}_\epsilon) \qquad \min_{z\in\mathbb{R}^p}\{\varphi(z) \ \text{s.t.} \ h(z)=\epsilon,\ z\in\Delta\},$$

where $\epsilon \in \mathbb{R}$ is a parameter, and denote by $\mathcal{F}_\epsilon$ the feasible set of (MP_ϵ).

Definition 1 (see [62, Definition 3.1] or [63, Definition 2.1]) Problem (MP) is said to be partially calm at a solution point z^* if there exist $\varepsilon > 0$ and $\mu > 0$ such that for all $\epsilon \in [-\varepsilon,\varepsilon]$ and all $z \in (z^*+\varepsilon\mathbb{B}) \cap \mathcal{F}_\epsilon$, one has $\varphi(z)-\varphi(z^*)+\mu|h(z)| \geq 0$, where $\mathbb{B}$ denotes the unit ball of $\mathbb{R}^p$ cetered at the origin, and (MP) is said to be partially calm over the global optimal solution set $\mathcal{F}^*$ if it is partially calm at each $z^* \in \mathcal{F}^*$.

2.2 Calmness of a multifunction

Definition 2 (see [19, Section 3 H]) Given a multifunction $\Phi : \mathbb{R}^l \rightrightarrows \mathbb{R}^p$, we say that Φ is calm at a point $\overline{x}$ for $\overline{z} \in \Phi(\overline{x})$ if there exist a constant $\kappa \geq 0$ and neighborhoods $\mathcal{N}_{\overline{x}}$ of $\overline{x}$ and $\mathcal{N}_{\overline{z}}$ of $\overline{z}$ such that for all $x \in \mathcal{N}_{\overline{x}}$,

$$\Phi(x) \cap \mathcal{N}_{\overline{z}} \subseteq \Phi(\overline{x}) + \kappa \|x - \overline{x}\|\mathbb{B}.$$

Note that the neighborhood $\mathcal{N}_{\overline{x}}$ of $\overline{x}$ in Definition 2 can be removed. By [19, Exercise 3 H.4], the calmness of Φ at $\overline{x}$ for $\overline{z} \in \Phi(\overline{x})$ is equivalent to the metric subregularity of its inverse mapping Φ^{-1} at $\overline{z}$ for $\overline{x} \in \Phi^{-1}(\overline{z})$, i.e., there exist a constant $\kappa \geq 0$ and a neighborhood $\mathcal{N}_{\overline{z}}$ of $\overline{z}$ such that for all $z \in \mathcal{N}_{\overline{z}}$,

$$\mathrm{dist}(z, \Phi(\overline{x})) \leq \kappa \mathrm{dist}(\overline{x}, \Phi^{-1}(z)).$$

2.3 Proximal mapping and Moreau envelope

For a proper lower semicontinuous (lsc) function $h : \mathbb{R}^p \to (-\infty, \infty]$ and a parameter $\gamma > 0$, we denote $\mathcal{P}_\gamma h$ and $e_\gamma h$ by the proximal mapping and Moreau envelope of h associated to γ, which are respectively defined as

$$\mathcal{P}_\gamma h(x) := \arg\min_{z \in \mathbb{R}^p} \left\{ \frac{1}{2\gamma}\|z - x\|^2 + h(z) \right\}, \; e_\gamma h(x) := \min_{z \in \mathbb{R}^p} \left\{ \frac{1}{2\gamma}\|z - x\|^2 + h(z) \right\}.$$

When h is convex, $\mathcal{P}_\gamma h$ is a Lipschitz mapping from $\mathbb{R}^p$ to $\mathbb{R}^p$ with Lipschitz constant 1, and $e_\gamma h$ is a smooth convex function with $\nabla e_\gamma h(x) = \frac{1}{\gamma}(x - \mathcal{P}_\gamma h(x))$. When $\gamma = 1$, we replace $\mathcal{P}_\gamma h$ with $\mathcal{P}h$. The following lemma presents the expression of the proximal operator of the weighted ℓ_1-norm.

Lemma 1 *For any given $\omega \in \mathbb{R}^p_+$ and $\mu \geq 0$, let $h_{\omega,\mu}(x) := \|\omega \circ x\|_1 + \frac{1}{2}\mu\|x\|^2$ for $x \in \mathbb{R}^p$. The proximal operator of $h_{\omega,\mu}$ associated to a parameter $\gamma > 0$ is given by*

$$\mathcal{P}_{\gamma^{-1}} h_{\omega,\mu}(z) = \frac{\gamma}{\gamma + \mu} \mathrm{sign}(z) \circ \max(|z| - \gamma^{-1}\omega, 0) \quad \forall z \in \mathbb{R}^p.$$

2.4 Subderivatives and generalized subdifferentials

The second subserivative of an extended real-valued function plays a key role in verifying the local optimality of a stationary point of a nonsmooth optimization problem, though its characterization is not an easy task. Here we introduce the notion of subderivative and second subderivative.

Definition 3 ([50, Definition 8.1 & 13.3]) For a function $h : \mathbb{R}^p \to (-\infty, \infty]$, a point $x \in \mathrm{dom}\, h$ and any $v \in \mathbb{R}^p$, the subderivative function $dh(x) : \mathbb{R}^p \to [-\infty, \infty]$ is

defined by

$$dh(x)(w) := \liminf_{\tau \downarrow 0, w' \to w} \frac{h(x+\tau w') - h(x)}{\tau},$$

while the second subderivative of h at x for v and w is defined by

$$d^2h(x|v)(w) := \liminf_{\tau \downarrow 0, w' \to w} \frac{h(x+\tau w') - h(x) - \tau\langle v, w'\rangle}{\frac{1}{2}\tau^2}.$$

Definition 4 ([50, Definition 8.3]) Consider a function $h : \mathbb{R}^p \to (-\infty, \infty]$ and a point $x \in \operatorname{dom} h$. The regular subdifferential of h at x, denoted by $\widehat{\partial} h(x)$, is defined as

$$\widehat{\partial} h(x) := \left\{ v \in \mathbb{R}^p \mid \liminf_{x \neq x' \to x} \frac{h(x') - h(x) - \langle v, x' - x\rangle}{\|x' - x\|} \geq 0 \right\};$$

and the (limiting) subdifferential of h at x, denoted by $\partial h(x)$, is defined as

$$\partial h(x) := \left\{ v \in \mathbb{R}^p \mid \exists\, x^k \to x \text{ with } h(x^k) \to h(x) \text{ and } v^k \in \widehat{\partial} h(x^k) \text{ with } v^k \to v \right\}.$$

Remark 1 At any $x \in \operatorname{dom} h$, the inclusion $\widehat{\partial} h(x) \subseteq \partial h(x)$ holds, and $\widehat{\partial} h(x)$ is a closed convex set, while $\partial h(x)$ is closed but generally nonconvex. When h is convex, they reduce to the subdifferential of h at x in the sense of [49]. A point x at which $0 \in \partial h(x)$ (respectively, $0 \in \widehat{\partial} h(x)$) is called a limiting (respectively, regular) critical point of h, denoted by $\operatorname{crit} h$ (respectively, $\widehat{\operatorname{crit}}\, h$). By Definition 4, obviously, a local minimizer of h is a regular critical point.

2.5 Semismoothness of local Lipschitz mappings

Semismoothness was originally introduced by Mifflin [38] for functionals, and Qi and Sun [44] later developed the class of vector-valued semismooth functions. To introduce the concept of semismoothness, for a locally Lipschitz mapping $F : \mathcal{O} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ where $\mathcal{O}$ is an open set, we denote by $\partial_C F(x)$ the Clarke Jacobian of F at $x \in \mathcal{O}$ (see [13] for its detailed discussion).

Definition 5 (see [44]) Let $F : \mathcal{O} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be a local Lipschitz mapping on an open set $\mathcal{O}$. The mapping F is said to be semismooth at a point $x \in \mathcal{O}$ if F is directionally differentiable at x and for any $\Delta x \to 0$ and any $V \in \partial_C F(x + \Delta x)$,

$$F(x + \Delta x) - F(x) - V\Delta x = o(\|\Delta x\|);$$

and F is said to be strongly semismooth at x if F is semismooth at x and for any $\Delta x \to 0$,

$$F(x + \Delta x) - F(x) - V\Delta x = O(\|\Delta x\|^2).$$

The mapping F is said to be a semismooth (respectively, strongly semismooth) on $\mathcal{O}$ if it is semismooth (respectively, strongly semismooth) everywhere in $\mathcal{O}$.

The proximal mapping of the function $h_{\omega,\mu}$ in Lemma 1 is piecewise affine, so it is strongly semismooth by [20, Proposition 7.4.7]. The following lemma provides the characterization for its Clarke Jacobian. Since its proof is direct by an elementary calculation, we omit the detail.

Lemma 2 *For any given $\omega \in \mathbb{R}^p_+$ and $\mu \geq 0$, let $h_{\omega,\mu}$ be the function in Lemma 1. Then, at any given $z \in \mathbb{R}^p$, the Clarke Jacobian of its proximal mapping associated to $\gamma > 0$ has the form*

$$\partial_C(\mathcal{P}_{\gamma^{-1}h_{\omega,\mu}})(z) = \left\{ \begin{array}{rl} \xi_i = \frac{\gamma}{\gamma+\mu} & \text{if } |z_i| > \gamma^{-1}\omega_i, \\ \mathrm{diag}(\xi_1,\ldots,\xi_p) \mid \xi_i \in \big[0, \frac{\gamma}{\gamma+\mu}\big] & \text{if } |z_i| = \gamma^{-1}\omega_i, \\ \xi_i = 0 & \text{if } |z_i| < \gamma^{-1}\omega_i \end{array} \right\}.$$

3 Equivalent surrogates

In this section, we derive a family of equivalent surrogates for problem (1) by leveraging the global exact penalty for its MPEC reformulation, and for a subfamily of these surrogates, investigate the local optimality of their critical points. In order to introduce the MPEC reformulation of (1), let $\mathscr{L}$ denote the family of proper lsc functions ϕ that is convex on $[0, 1]$ and satisfies

$$\mathrm{int}(\mathrm{dom}\,\phi) \supseteq [0,1],\ t^* := \mathop{\arg\min}_{0\leq t\leq 1} \phi(t),\ \phi(t^*) = 0 \ \text{ and } \ \phi(1) = 1. \tag{5}$$

As will be illustrated later, the condition (5) is rather weak and $\mathscr{L}$ contains many proper lsc functions ϕ. For each $\phi \in \mathscr{L}$, denote its convex truncation by

$$\psi(t) := \left\{ \begin{array}{ll} \phi(t) & \text{if } t \in [0,1], \\ +\infty & \text{otherwise.} \end{array} \right. \tag{6}$$

Pick an arbitrary $\phi \in \mathscr{L}$. By equation (5), t^* is the unique minimizer of ϕ on $[0, 1]$ with $\phi(t^*) = 0$ and $\phi(1) = 1$, which implies that for any $z \in \mathbb{R}^p$,

$$\|z\|_0 = \min_{w\in\mathbb{R}^p} \Big\{ \textstyle\sum_{i=1}^p \phi(w_i) \ \ \text{s.t. } \big\langle e - w, |z| \big\rangle = 0,\ 0 \leq w \leq e \Big\}.$$

Such a characterization for $z \in \mathbb{R}^p_+$ with $\phi(t) \equiv t$ first appeared in [37], and it immediately implies that problem (1) is equivalent to the following MPEC

$$\min_{x\in\mathbb{R}^p, w\in[0,e]} \Big\{ f_\mu(x) + \nu \textstyle\sum_{i=1}^p \phi(w_i) \ \ \text{s.t. } \langle e - w, |x| \rangle = 0 \Big\}, \tag{7}$$

in the sense that if $(\widehat{x}, \widehat{w})$ is a global (or local) optimal solution of (7), then $\widehat{x}$ is globally (or locally) optimal to (1); and if $\widehat{x}$ is globally (or locally) optimal to (1),

$(\widehat{x}, \max(\mathrm{sign}(|\widehat{x}|), t^*e)$ is a global (or local) optimal solution of (7). The equivalence between (7) and (1) discloses that the difficulty of the zero-norm regularized problem comes from the implicit constraint $\langle e-w, |x|\rangle = 0$.

As well known, the handling of nonconvex constraints is numerically more troublesome than that of nonconvex cost functions. Hence, we are interested in the following penalty problem of the MPEC (7):

$$\min_{x,w\in\mathbb{R}^p} \left\{ f_\mu(x) + \nu\textstyle\sum_{i=1}^p \psi(w_i) + \rho\nu\langle e-w, |x|\rangle \right\}, \tag{8}$$

where $\rho > 0$ is the penalty parameter. In order to establish that the problem (8) is a global exact penalty for (7), i.e., there exists $\overline{\rho} > 0$ such that the problem (8) associated to every $\rho > \overline{\rho}$ has the same global optimal solution set as the MPEC (7) does, we need the following technical lemma, and the coerciveness of f_μ and the discreteness of the zero-norm accounts for this.

Lemma 3 *Let $\mathcal{X}^*$ denote the global optimal solution set of the problem (1). Then, there exists a constant $\alpha > 0$ such that for all $x \in \mathcal{X}^*$, $|x|_{\mathrm{nz}} > \alpha$.*

Proof Suppose on the contradiction that the conclusion does not hold. There exists a sequence $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathcal{X}^*$ such that for each $k \in \mathbb{N}$, $|x^k|_{\mathrm{nz}} \le 1/k$. Clearly, there exist an index set $K \subseteq \mathbb{N}$ and an index $j \in \{1, 2, \ldots, p\}$ such that $0 < |x_j^k| \le 1/k$ for all $k \in K$. Recall that $\mathcal{X}^*$ is compact due to the coerciveness of f_μ. By taking a subsequence if necessary, we may assume that $\{x^k\}_{k\in K}$ is convergent, say, $\lim_{K\ni k\to+\infty} x^k = x^* \in \mathcal{X}^*$. Together with $0 < |x_j^k| \le 1/k$ for all $k \in K$, we obtain $|x_j^*| = 0$. Thus, for all sufficiently large $k \in K$, $\|x^k\|_0 \ge \|x^*\|_0 + 1$. Since f_μ is lsc, for all sufficiently large $k \in K$, $f_\mu(x^k) \ge f_\mu(x^*) - \nu/2$. From $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathcal{X}^*$ and $x^* \in \mathcal{X}^*$, we obtain

$$\begin{aligned} f_\mu(x^*) + \nu\|x^*\|_0 &= f_\mu(x^k) + \nu\|x^k\|_0 \\ &\ge f_\mu(x^*) - \nu/2 + \nu\|x^*\|_0 + \nu = f_\mu(x^*) + \nu\|x^*\|_0 + \nu/2, \end{aligned}$$

which is impossible. This shows that the desired conclusion holds. □

Next under a rather weak assumption on the function ϑ, we establish the partial calmness of the MPEC (7) on its global optimal solution set, which by [34, Proposition 2.1] implies that the penalty problem (8) is a global exact penalty. For the notion of partial calmness, please refer to [34, 63].

Proposition 1 *Let $\mathcal{F}^*$ denote the global optimal solution set of the MPEC (7). Suppose that the function ϑ has a polyhedral domain* $\mathrm{dom}\vartheta$ *and is strictly continuous relative to* $\mathrm{dom}\vartheta$. *Then, for each $(\widehat{x}, \widehat{w}) \in \mathcal{F}^*$, there exist $\delta > 0$ and $\widehat{\rho} > 0$ such that for all $(x, w) \in \mathbb{B}((\widehat{x}, \widehat{w}), \delta) \cap (\mathrm{dom} f \times [0, e])$,*

$$\left[f_\mu(x) + \nu\textstyle\sum_{i=1}^p \psi(w_i)\right] - \left[f_\mu(\widehat{x}) + \nu\textstyle\sum_{i=1}^p \psi(\widehat{w}_i)\right] + \widehat{\rho}\nu\langle e - w, |x|\rangle \ge 0, \tag{9}$$

and consequently, there exists $\overline{\rho} \ge \widehat{\rho}$ such that the problem (8) associated to every $\rho \ge \overline{\rho}$ has the same global optimal solution set as the MPEC (7) does.

Proof By the given assumption and the definition of f in (4), clearly, f is strictly continuous relative to its domain. Also, since $\mathrm{dom}\vartheta$ is polyhedral, the function f also has a polyhedral domain $\mathrm{dom} f$. Let $\Pi(x,w):=x$ for $(x,w)\in\mathbb{R}^p\times\mathbb{R}^p$, and let $R>0$ be a constant such that $\mathbb{B}_\infty(R)\supseteq\bigcup_{x\in\Pi(\mathcal{F}^*)}\mathbb{B}(x,1/2)$. Such R exists by recalling that $\mathcal{X}^*=\Pi(\mathcal{F}^*)$ is nonempty and compact. For each $k\in\{1,2,\ldots,p\}$, define

$$\mathcal{F}_k(\tau):=\big\{x\in\mathrm{dom} f\cap\mathbb{B}_\infty(R)\mid\|x\|_1-\|x\|_{(k)}=\tau\big\}\ \text{ for }\tau\geq 0,$$

where $\|\cdot\|_{(k)}$ denotes the Ky Fan k-norm of vectors. Since each $\mathcal{F}_k$ is a polyhedral multifunction, i.e., its graph is the union of finitely many polyhedral convex sets, by [51, Proposition 1] each $\mathcal{F}_k$ is calm at the origin for all $z\in\mathcal{F}_k(0)$. From the compactness of $\mathrm{dom} f\cap\mathbb{B}_\infty(R)$ and [45, Theorem 3.1], for each $k\in\{1,2,\ldots,p\}$, there exists $\gamma_k>0$ such that for all $z\in\mathrm{dom} f\cap\mathbb{B}_\infty(R)$,

$$\mathrm{dist}(z,\mathcal{F}_k(0))\leq\gamma_k\big[\|z\|_1-\|z\|_{(k)}\big]. \tag{10}$$

Fix any $(\widehat{x},\widehat{w})\in\mathcal{F}^*$. Since f is strictly continuous relative to $\mathrm{dom} f$, the function f_μ is strictly continuous relative to $\mathrm{dom} f$. Then, there exists $\delta\in(0,\frac{1}{2})$ such that

$$|\ f_\mu(x')-f_\mu(x'')\ |\leq L_{f_\mu}\|x'-x''\|\quad\text{for all } x',x''\in\mathbb{B}(\widehat{x},\delta)\cap\mathrm{dom} f. \tag{11}$$

By Lemma 3, $|\widehat{x}|_{\mathrm{nz}}>\alpha$. Let $\kappa=\|\widehat{x}\|_0$. From the continuity, for all $x\in\mathbb{B}(\widehat{x},\delta)$ we have $|x|^\downarrow_\kappa>\alpha$ (if necessary by reducing δ). Let $\widehat{\rho}:=\max\big(\frac{\phi'_-(1)}{\alpha},\frac{\gamma\phi'_-(1)(1-t^*)L_{f_\mu}}{\nu(1-t_0)}\big)$ for $\gamma=\max\{\gamma_1,\ldots,\gamma_p\}$, where $t_0\in[0,1)$ is such that $\frac{1}{1-t^*}\in\partial\phi(t_0)$ and its existence is by [34, Lemma 1]. Pick any $(x,w)\in\mathbb{B}((\widehat{x},\widehat{w}),\delta/2)\cap(\mathrm{dom} f\times[0,e])$. Let $J:=\big\{j\in\{1,\ldots,p\}\mid\widehat{\rho}|\widehat{x}|^\downarrow_j>\phi'_-(1)\big\}$ and $r:=|J|$. Recall that $\widehat{x}\in\Pi(\mathcal{F}^*)$. By the definition of $\mathbb{B}_\infty(R)$, $x\in\mathrm{dom} f\cap\mathbb{B}_\infty(R)$. By invoking (10) with $k=r$ and $z=x$, there exists $x^{\widehat{\rho}}\in\mathcal{F}_r(0)$ with $\|x-x^{\widehat{\rho}}\|=\mathrm{dist}(x,\mathcal{F}_r(0))$ such that

$$\|x-x^{\widehat{\rho}}\|\leq\gamma\big[\|x\|_1-\|x\|_{(r)}\big]=\gamma\textstyle\sum_{j=r+1}^p|x|^\downarrow_j. \tag{12}$$

Since $\widehat{\rho}\geq\frac{\phi'_-(1)}{\alpha}$, clearly, $\kappa\leq r$. Together with $\kappa=\|\widehat{x}\|_0$ and $(\widehat{x},\widehat{w})\in\mathcal{F}^*$, we have $\widehat{x}\in\mathcal{F}_r(0)$, and consequently, $\|x^{\widehat{\rho}}-\widehat{x}\|\leq\|x^{\widehat{\rho}}-x\|+\|x-\widehat{x}\|\leq 2\|x-\widehat{x}\|\leq\delta$. By (11),

$$|\ f_\mu(x)-f_\mu(x^{\widehat{\rho}})\ |\leq L_{f_\mu}\|x-x^{\widehat{\rho}}\|. \tag{13}$$

Note that $w\in[0,e]$ implied by $(x,w)\in\mathrm{dom} f\times[0,e]$. It is immediate to obtain that

$$\textstyle\sum_{i=1}^p\phi(w_i^\downarrow)+\widehat{\rho}\big(\|x\|_1-\langle w^\downarrow,|x|^\downarrow\rangle\big)\geq\sum_{i=1}^p\min_{t\in[0,1]}\big\{\phi(t)+\widehat{\rho}|x|^\downarrow_i(1-t)\big\}.$$

Write $J_1:=\big\{j\mid\frac{1}{1-t^*}\leq\widehat{\rho}|x|^\downarrow_j\leq\phi'_-(1)\big\}$ and $J_2:=\big\{j\mid 0\leq\widehat{\rho}|x|^\downarrow_j<\frac{1}{1-t^*}\big\}$. For the minimization problem on the right hand side of the last equation, by invoking [34,

Lemma 1] with $\omega = |x|^{\downarrow}_j$ for each j, it is immediate to obtain that

$$\begin{aligned}
&\textstyle\sum_{i=1}^p \phi(w_i^{\downarrow}) + \widehat{\rho}\big(\|x\|_1 - \langle w^{\downarrow}, |x|^{\downarrow}\rangle\big)\\
&\geq \|x^{\widehat{\rho}}\|_0 + \frac{\widehat{\rho}(1-t_0)}{\phi'_-(1)(1-t^*)} \sum_{j\in J_1} |x|^{\downarrow}_j + \widehat{\rho}(1-t_0) \sum_{j\in J_2} |x|^{\downarrow}_j\\
&\geq \|x^{\widehat{\rho}}\|_0 + \frac{\widehat{\rho}(1-t_0)}{\phi'_-(1)(1-t^*)} \sum_{j\in J_1\cup J_2} |x|^{\downarrow}_j = \|x^{\widehat{\rho}}\|_0 + \frac{\widehat{\rho}(1-t_0)}{\phi'_-(1)(1-t^*)} \sum_{j=r+1}^p |x|^{\downarrow}_j\\
&\geq \|x^{\widehat{\rho}}\|_0 + \frac{\widehat{\rho}(1-t_0)}{\gamma\phi'_-(1)(1-t^*)}\|x - x^{\widehat{\rho}}\| \geq \|x^{\widehat{\rho}}\|_0 + \nu^{-1}\big[f_\mu(x^{\widehat{\rho}}) - f_\mu(x)\big]
\end{aligned}$$

where the first inequality is also using $\|x^{\widehat{\rho}}\|_0 = r$ implied by $x^{\widehat{\rho}} \in \mathcal{F}_r(0)$, the second one is due to $-1 = \phi(t^*) - \phi(1) \geq \phi'_-(1)(t^*-1)$, the third one is due to (12), and the last one is using the definition of $\widehat{\rho}$ and (13). Since $\langle w^{\downarrow}, |x|^{\downarrow}\rangle \geq \langle w, |x|\rangle$, we have

$$f_\mu(x) + \nu\textstyle\sum_{i=1}^p \phi(w_i) + \widehat{\rho}\nu\big(\|x\|_1 - \langle w, |x|\rangle\big) \geq f_\mu(x^{\widehat{\rho}}) + \nu\|x^{\widehat{\rho}}\|_0.$$

Now take $w_i^{\widehat{\rho}} = 1$ for $i \in \mathrm{supp}(x^{\widehat{\rho}})$ and $w_i^{\widehat{\rho}} = t^*$ for $i \notin \mathrm{supp}(x^{\widehat{\rho}})$. It is easy to check that $(x^{\widehat{\rho}}, w^{\widehat{\rho}})$ is a feasible point of (7) with $\sum_{i=1}^p \phi(w_i^{\widehat{\rho}}) = \|x^{\widehat{\rho}}\|_0$. Then, we have $f_\mu(x^{\widehat{\rho}}) + \sum_{i=1}^p \phi(w_i^{\widehat{\rho}}) \geq f_\mu(\widehat{x}) + \sum_{i=1}^p \phi(\widehat{w}_i)$. Together with the last inequality,

$$f_\mu(x) + \nu\textstyle\sum_{i=1}^p \phi(w_i) + \widehat{\rho}\nu\big(\|x\|_1 - \langle w, |x|\rangle\big) \geq f_\mu(\widehat{x}) + \sum_{i=1}^p \phi(\widehat{w}_i).$$

By the arbitrariness of $(\widehat{x}, \widehat{w})$ in $\mathcal{F}^*$ and the expression of ψ, we obtain the first part. The second part holds by combining the first part with [34, Proposition 2.1]. □

Remark 2 When replacing the function $x \mapsto \vartheta(Ax - b)$ in (1) with a general proper function $\widetilde{f} : \mathbb{R}^p \to (-\infty, +\infty]$, which has a polyhedral domain and is strictly continuous relative to its domain, the conclusion of Proposition 1 still holds. For example, $\widetilde{f} = h + \delta_\Omega$ where $\Omega \subseteq \mathbb{R}^p$ is a polyhedral set and $h : \mathbb{R}^p \to \mathbb{R}$ is strictly continuous relative to Ω, by comparing Proposition 1 with [34, Theorem 3.2], we see that the former not only weakens the Lipschitz continuity of $\widetilde{f}$ on Ω to be its strict continuity relative to Ω, but also removes the restricted assumption on the structure of Ω.

The penalty problem (8), compared with (7), is easier to operate because its nonconvexity is caused by the coupled term $\langle e-w, |x|\rangle$ rather than the equilibrium constraint, and moreover, by the conjugate of ψ, the optimal value of the minimization problem in (8) with respect to w is $-\nu\sum_{i=1}^p \psi^*(\rho|x_i|)$. Consequently, the penalty problem (8) can be compactly written as

$$\min_{x\in\mathbb{R}^p} \Theta_{\rho,\nu,\mu}(x) := f_\mu(x) + \rho\nu\big(\|x\|_1 - g_\rho(x)\big) \tag{14}$$

with $g_\rho(x) := \rho^{-1}\sum_{i=1}^p \psi^*(\rho|x_i|)$ for $x \in \mathbb{R}^p$. Combining Proposition 1 with the equivalence between (7) and (1), we immediately have the following result.

Theorem 2 *If ϑ has a polyhedral domain* $\mathrm{dom}\vartheta$ *and is strictly continuous relative to* $\mathrm{dom}\vartheta$*, then the problem (14) associated to any $\rho \geq \overline{\rho}$ has the same global optimal solution set as (1) does, i.e., $\Theta_{\rho,\nu,\mu}$ for each $\rho \geq \overline{\rho}$ is an equivalent surrogate of $\Theta_{\nu,\mu}$.*

Notice that ψ^* is a nondecreasing finite convex function on $\mathbb{R}$ because by [49, Section 24] the range of $\partial\psi^*$ is contained in $\mathrm{dom}\partial\psi = [0,1]$. Hence, the function $t \mapsto \psi^*(\rho|t|)$ associated to every $\rho > 0$ is a finite convex function. Next we illustrate that the function $x \mapsto \rho\nu(\|x\|_1 - g_\rho(x))$ covers some common surrogates of the zero-norm such as the capped ℓ_1-norm, SCAD and MCP.

Example 1 Let $\phi(t) = t$ for $t \in \mathbb{R}$. One can check that $\phi \in \mathscr{L}$ with $t^* = 0$ and $\psi^*(s) = \max(s-1, 0)$ for $s \in \mathbb{R}$. Now the function $x \mapsto \|x\|_1 - g_\rho(x)$ is precisely the capped ℓ_1-norm, and when ϑ satisfies the assumption of Theorem 2, the function $\Theta_{\rho,\nu,\mu}$ associated to each $\rho > \overline{\rho}$ is an equivalent surrogate of the function $\Theta_{\nu,\mu}$. Compared with [5, Theorem 2.4], Theorem 2 removes the Lipschitz continuity of f on its domain but requires the strict continuity of f and the coerciveness of f_μ.

Example 2 Let $\phi(t) = \frac{a^2}{4}t^2 - \frac{a^2}{2}t + at + \frac{(a-2)^2}{4}$ $(a > 2)$ for $t \in \mathbb{R}$. One can check that $\phi \in \mathscr{L}$ with $t^* = \frac{a-2}{a}$ and the conjugate ψ^* of ψ takes the following form

$$\psi^*(s) = \begin{cases} -\frac{(a-2)^2}{4} & \text{if } s \leq a - \frac{a^2}{2}, \\ \frac{1}{a^2}(\frac{a(a-2)}{2} + s)^2 - \frac{(a-2)^2}{4} & \text{if } a - \frac{a^2}{2} < s \leq a, \\ s - 1 & \text{if } s > a \end{cases} \quad \text{for } s \in \mathbb{R}.$$

Now $\frac{a}{2}\lambda[|t| - \rho^{-1}\psi^*(\rho|t|)]$ with $\rho = \frac{1}{\lambda}$ is the MCP function $p(t;\lambda)$ in [64], and when ϑ satisfies the assumption of Theorem 2, the function $\Theta_{\rho,\nu,\mu}$ associated to each $\rho > \overline{\rho}$ is an equivalent surrogate of the function $\Theta_{\nu,\mu}$.

Example 3 Let $\phi(t) = \frac{a-1}{a+1}t^2 + \frac{2}{a+1}t$ $(a > 1)$ for $t \in \mathbb{R}$. Then $t^* = 0$, $t_0 = \frac{1}{2}$ and

$$\psi^*(s) = \begin{cases} 0 & \text{if } s \leq \frac{2}{a+1}, \\ \frac{((a+1)s-2)^2}{4(a^2-1)} & \text{if } \frac{2}{a+1} < s \leq \frac{2a}{a+1}, \\ s - 1 & \text{if } s > \frac{2a}{a+1} \end{cases} \quad \text{for } s \in \mathbb{R}. \tag{15}$$

Now $\lambda[|t| - \rho^{-1}\psi^*(\rho|t|)]$ with $\rho = \frac{2}{(a+1)\lambda}$ is exactly the SCAD function $p_\lambda(t)$ in [21], and when ϑ satisfies the assumption of Theorem 2, the function $\Theta_{\rho,\nu,\mu}$ associated to every $\rho \geq \overline{\rho}$ is an equivalent surrogate of $\Theta_{\nu,\mu}$.

In the sequel, we denote by $\mathscr{L}_{\sigma,\gamma}$ the set consisting of those $\phi \in \mathscr{L}$ that are strongly convex on $[0,1]$ with modulus σ and satisfies $\phi(0) = 0$ and $\phi'_-(0) \geq \gamma$ for some $\gamma > 0$. Clearly, the function ϕ in Example 2 and 3 belongs to $\mathscr{L}_{\sigma,\gamma}$. By Lemma 7 in "Appendix A", the associated g_ρ for every $\rho > 0$ is continuously differentiable on $\mathbb{R}^p$ and at any $x \in \mathbb{R}^p$, $\nabla g_\rho(x) = w_\rho(x) \circ \mathrm{sign}(x)$ with

$$w_\rho(x) := \big((\psi^*)'(\rho|x_1|), \ldots, (\psi^*)'(\rho|x_p|)\big)^{\mathbb{T}}, \tag{16}$$

where w_ρ is Lipschitz continuous with modulus ρ/σ. The following proposition summarizes some desirable properties of $\Theta_{\rho,\nu,\mu}$; see "Appendix A" for its proof.

Proposition 3 *Suppose that ϑ is a convex function with a polyhedral domain $\mathrm{dom}\,\vartheta$ and is strictly continuous relative to $\mathrm{dom}\,\vartheta$. Then, for any given $\phi \in \mathscr{L}_{\sigma,\gamma}$ and $\rho > 0$, the following statements hold.*

(i) $\Theta_{\rho,\nu,\mu}$ is a nonnegative and coercive DC function, and at any $x \in \mathrm{dom}\, f$,

$$\widehat{\partial}\Theta_{\rho,\nu,\mu}(x) = \partial\Theta_{\rho,\nu,\mu}(x) = A^{\mathbb{T}}\partial\vartheta(Ax-b) + \rho\nu\big[\partial\|x\|_1 - \nabla g_\rho(x)\big] + \mu x.$$

(ii) The set $\mathrm{crit}\,\Theta_{\rho,\nu,\mu}$ coincides with the d-stationary point set of (14).

(iii) Every $\overline{x} \in \mathrm{crit}\,\Theta_{\rho,\nu,\mu}$ with $|\overline{x}|_{\mathrm{nz}} \geq \phi'_+(1)/\rho$ is a strongly local optimal solution of (14), i.e., there exist $\varepsilon > 0$ and $c_0 > 0$ such that for all $x \in \mathbb{B}(\overline{x},\varepsilon)$,

$$\Theta_{\rho,\nu,\mu}(x) \geq \Theta_{\rho,\nu,\mu}(\overline{x}) + c_0\|x-\overline{x}\|^2.$$

(iv) Every $\overline{x} \in \mathrm{crit}\,\Theta_{\rho,\nu,\mu}$ with $|\overline{x}|_{\mathrm{nz}} \geq \phi'_+(1)/\rho$ belongs to $\widehat{\mathrm{crit}}\,\Theta_{\nu,\mu} = \mathrm{crit}\,\Theta_{\nu,\mu}$, which is also a strongly local optimal solution set of the problem (1).

When $\overline{x} \in \mathrm{crit}\,\Theta_{\rho,\nu,\mu}$ is sparse enough and ρ is chosen to be suitably large, it is highly possible for $|\overline{x}|_{\mathrm{nz}} \geq \phi'_+(1)/\rho$ to hold, and now every stationary point $\overline{x}$ of (14) is a strongly local optimal solution to problems (14) and (1). It is worth pointing out that when $\phi \in \mathscr{L}\backslash\mathscr{L}_{\sigma,\gamma}$, for example, the function ϕ in Example 1, the stationary point of (14) may be a strongly local optimal solution to problems (14) and (1) by the recent work [14, 15].

4 Proximal MM method

In this section, we develop a tailored proximal MM method for seeking a critical point of the surrogate problem (14) under the following assumption.

Assumption 1 ϑ is a convex function with a polyhedral domain $\mathrm{dom}\,\vartheta$ and is strictly continuous relative to $\mathrm{dom}\,\vartheta$.

Assumption 1 implies that problem (14) is a DC program. Fix any $x' \in \mathbb{R}^p$. For any $x \in \mathbb{R}^p$, the convexity and smoothness of the function ψ^* implies that

$$\rho^{-1}\textstyle\sum_{i=1}^p \psi^*(\rho|x_i|) \geq \rho^{-1}\sum_{i=1}^p \psi^*(\rho|x'_i|) + \langle w_\rho(x'), |x| - |x'|\rangle, \tag{17}$$

where $w_\rho : \mathbb{R}^p \to \mathbb{R}^p$ is the mapping in (16). By the expression of $\Theta_{\rho,\nu,\mu}$,

$$\Theta_{\rho,\nu,\mu}(x) \leq \Xi_{\rho,\nu,\mu}(x,x') := f_\mu(x) + \rho\nu\big[\|x\|_1 - \langle w_\rho(x'), |x|\rangle\big] + R_{\rho,\nu,\mu}(x')$$

where $R_{\rho,\nu,\mu}(x') = \rho\nu[\langle w_\rho(x'), |x'|\rangle - g_\rho(x')]$. This, together with $\Xi_{\rho,\nu,\mu}(x',x') = \Theta_{\rho,\nu,\mu}(x')$, means that $\Xi_{\rho,\nu,\mu}(\cdot,x')$ is a majorization of the function $\Theta_{\rho,\nu,\mu}$ at x'. This majorization is tighter than the one in [53] which was obtained by the convexity

of g_ρ. Indeed, since range$(\partial\psi^*) \subseteq [0,1]$, we have $w_\rho(x') \geq 0$, which along with (16) implies that $\langle w_\rho(x'), |x|\rangle \geq \langle \nabla g_\rho(x'), x\rangle$, and consequently,

$$\Xi_{\rho,\nu,\mu}(x,x') \leq \widetilde{\Xi}_{\rho,\nu,\mu}(x,x') := f_\mu(x) + \rho\nu[\|x\|_1 - \langle \nabla g_\rho(x'), x\rangle] + R_{\rho,\nu,\mu}(x').$$

The majorization $\widetilde{\Xi}_{\rho,\nu,\mu}(\cdot,x')$ is precisely the one used in [53]. Our proximal MM method is designed by minimizing a proximal version of $\Xi_{\rho,\nu,\mu}(\cdot,x')$.

Algorithm 1 (Proximal MM method for solving (14))

Require: $\widetilde{\lambda} > 0, \widetilde{\gamma}_{1,0} > 0, \widetilde{\gamma}_{2,0} > 0, x^{-1} \in \mathrm{dom}\, f$. To seek a starting point

$$x^0 \approx \mathop{\arg\min}_{x\in\mathbb{R}^p} \Big\{ f(x) + \widetilde{\lambda}\|x\|_1 + \frac{\widetilde{\gamma}_{1,0}}{2}\|x\|^2 + \frac{\widetilde{\gamma}_{2,0}}{2}\|Ax\|^2 \Big\}. \tag{18}$$

1: Choose $\phi \in \mathscr{L}_{\sigma,\gamma}$, $\rho \geq 1$, $\underline{\gamma_1} > 0$, $\underline{\gamma_2} > 0$, $\varrho \in (0,1]$, $0 < \gamma_{1,0} \leq \widetilde{\gamma}_{1,0}$ and $0 < \gamma_{2,0} \leq \widetilde{\gamma}_{2,0}$. Set $\lambda = \rho\nu$ and $B = \underline{\gamma_1} I + \underline{\gamma_2} A^{\mathbb{T}} A$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Compute $v^k = e - w_\rho(x^k)$.
4: Choose an error vector $\delta^k \in \mathbb{R}^p$ with $\|\delta^k\| \leq \frac{\|B^{1/2}(x^k - x^{k-1})\|}{\sqrt{2}\|B^{-1/2}\|}$.
5: Set $B_k = \gamma_{1,k} I + \gamma_{2,k} A^{\mathbb{T}} A$ and compute the optimal solution x^{k+1} to

$$\min_{x\in\mathbb{R}^p} \Big\{ f(x) + \lambda\langle v^k, |x|\rangle + \frac{\mu}{2}\|x\|^2 + \frac{1}{2}\|x - x^k\|^2_{B_k} - \langle \delta^k, x - x^k\rangle \Big\}. \tag{19}$$

6: Let $\gamma_{1,k+1} = \max(\underline{\gamma_1}, \varrho\gamma_{1,k})$ and $\gamma_{2,k+1} = \max(\underline{\gamma_2}, \varrho\gamma_{2,k})$.
7: **end for**

Remark 3 **(a)** As well known, for nonconvex optimization problems, the choice of the starting point is crucial to the quality of the limit of the sequence generated from this point. Inspired by the good performance of the ℓ_1-norm regularized minimization problem, we choose a starting point x^0 by solving the problem (18) inexactly. Such an initial point, as will be shown in Sect. 4.3, is good in a statistical sense when appropriate $\widetilde{\gamma}_{1,0}$ and $\widetilde{\gamma}_{2,0}$ are used. The inexactness of x^0 means that there exists an error vector $\widetilde{\delta}^0 \in \mathbb{R}^p$ with $\|\widetilde{\delta}^0\|_\infty \leq \widetilde{\epsilon}_0$ for some $\widetilde{\epsilon}_0 > 0$ such that

$$x^0 = \mathop{\arg\min}_{x\in\mathbb{R}^p} \Big\{ f(x) + \widetilde{\lambda}\|x\|_1 + \frac{\widetilde{\gamma}_{1,0}}{2}\|x\|^2 + \frac{\widetilde{\gamma}_{2,0}}{2}\|Ax\|^2 - \langle \widetilde{\delta}^0, x\rangle \Big\}. \tag{20}$$

(b) From (16), obviously, $w_\rho(x^k) \neq \nabla g_\rho(x^k)$, which implies that Algorithm 1 does not belong to the DCA framework proposed in [32, 43] even if ϑ is convex. By contrast, the proximal MM in [53] is a DC algorithm. The proximal term $\frac{1}{2}\|x - x^k\|^2_{B_k}$ in the subproblem plays a twofold role: one is to ensure that the subproblem (19) is solvable and the other, as will be shown in the sequel, is to guarantee the decreasing of the objective value sequence of (14) and then its global convergence. The subproblem

(19) is seeking an inexact solution to the following strongly convex program

$$\min_{x\in\mathbb{R}^p}\Big\{f(x)+\lambda\langle v^k,|x|\rangle+\frac{\mu}{2}\|x\|^2+\frac{1}{2}\|x-x^k\|_{B_k}^2\Big\}.$$

The inexactness means $\delta^k\in\partial f(x^{k+1})+\lambda\partial\langle v^k,|\cdot|\rangle(x^{k+1})+B_k(x^{k+1}-x^k)+\mu x^{k+1}$. Since the error δ^k only depends on the past two iterates, such an inexact solution x^{k+1} is available from a specific algorithm. In Sect. 4.3, we develop an efficient inexact dual PPA plus the powerful semismooth Newton method for computing x^{k+1}.

(c) We suggest that the parameter ρ is chosen to be $\frac{\alpha_0}{\|x^0\|_\infty}$ for a suitable $\alpha_0>0$. Indeed, by the expression of w_ρ in (16) and the range of $\partial\psi^*\subseteq[0,1]$, we have

$$v^k=\Big(1-(\psi^*)'\Big(\frac{\alpha_0|x_1^k|}{\|x^0\|_\infty}\Big),\ldots,1-(\psi^*)'\Big(\frac{\alpha_0|x_p^k|}{\|x^0\|_\infty}\Big)\Big)^{\mathbb{T}}\in[0,e].$$

Note that ψ^* is a nondecreasing finite convex function on $\mathbb{R}$. Such a choice of ρ ensures that those very small x_i^k (likely corresponding to zero components) become zero quickly because a weight close to 1 is imposed on $|x_i|$, while those very large x_i^k (likely corresponding to nonzero components) continue to keep large because a weight close to 0 is imposed on $|x_i|$.

4.1 Convergence analysis of Algorithm 1

By the steps of Algorithm 1 and Remark 3 (b), the sequence $\{x^k\}_{k\in\mathbb{N}}$ is well defined. To analyze its convergence, we define the potential function

$$\Psi_{\rho,\nu,\mu}(x,y):=\Theta_{\rho,\nu,\mu}(x)+\frac{1}{4}\|x-y\|_B^2\quad\forall x,y\in\mathbb{R}^p.$$

The following lemma provides the properties of $\Psi_{\rho,\nu,\mu}$ on the sequence $\{x^k\}_{k\in\mathbb{N}}$.

Lemma 4 *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then,*

(i) for each $k\in\mathbb{N}$, $\Psi_{\rho,\nu,\mu}(x^{k+1},x^k)\le\Psi_{\rho,\nu,\mu}(x^k,x^{k-1})-\frac{1}{4}\|x^{k+1}-x^k\|_B^2$;

(ii) for each $k\in\mathbb{N}$, there exists $\zeta^k\in\partial\Psi_{\rho,\nu,\mu}(x^k,x^{k-1})$ with $\|\zeta^k\|\le b_1\|x^k-x^{k-1}\|+b_2\|x^{k-1}-x^{k-2}\|$, where b_1 and b_2 are positive constants independent of k.

Lemma 4 (i) implies that the sequence $\{\Psi_{\rho,\nu,\mu}(x^k,x^{k-1})\}_{k\in\mathbb{N}}$ is nonincreasing, while by Proposition 3 the function $\Psi_{\rho,\nu,\mu}$ is lower bounded. Thus, the sequence $\{\Psi_{\rho,\nu,\mu}(x^k,x^{k-1})\}_{k\in\mathbb{N}}$ is convergent, and we denote by ϖ^* its limit.

Proof **(i)** By Step 2 of Algorithm 1, $\{x^k\}_{k\in\mathbb{N}}\subseteq\operatorname{dom}f$. By invoking (17) with $x=x^{k+1}$ and $x'=x^k$ yields that

$$g_\rho(x^{k+1})-\langle w_\rho(x^k),|x^{k+1}|\rangle\ge g_\rho(x^k)-\langle w_\rho(x^k),|x^k|\rangle.$$

Note that the objective function of (19) is a sum of a convex function and a strongly convex quadratic function. From the definition of x^{k+1} and x^k, $x^{k+1} \in \operatorname{dom} f$,

$$f(x^k) + \lambda\langle v^k, |x^k|\rangle + \frac{\mu}{2}\|x^k\|^2 \geq f(x^{k+1}) + \lambda\langle v^k, |x^{k+1}|\rangle + \frac{\mu}{2}\|x^{k+1}\|^2 + \|x^{k+1}-x^k\|^2_{B_k} + \langle \delta^k, x^k - x^{k+1}\rangle.$$

Combining the last two inequalities with the expression of $\Theta_{\rho,\nu,\mu}$ and $\lambda = \rho\nu$ yields

$$\begin{aligned}\Theta_{\rho,\nu,\mu}(x^k) &\geq \Theta_{\rho,\nu,\mu}(x^{k+1}) + \langle \delta^k, x^k - x^{k+1}\rangle + \|x^{k+1}-x^k\|^2_{B_k}\\ &\geq \Theta_{\rho,\nu,\mu}(x^{k+1}) + \langle \delta^k, x^k - x^{k+1}\rangle + \|x^{k+1}-x^k\|^2_{B}\\ &\geq \Theta_{\rho,\nu,\mu}(x^{k+1}) - \frac{1}{2}\|B^{-1/2}\delta^k\|^2 - \frac{1}{2}\|B^{1/2}(x^{k+1}-x^k)\|^2 + \|x^{k+1}-x^k\|^2_{B}\\ &\geq \Theta_{\rho,\nu,\mu}(x^{k+1}) - \frac{1}{4}\|x^k - x^{k-1}\|^2_B + \frac{1}{2}\|x^{k+1}-x^k\|^2_B\end{aligned}$$

where the second inequality is due to the positive semidefiniteness of $B_k - B$, the third one is using the Cauchy-Schwartz inequality, and the last one is due to the upper bound for $\|\delta^k\|$. Along with the definition of $\Psi_{\rho,\nu,\mu}$, we get the desired inequality.
(ii) By the definition of x^k in (19) and [49, Theorem 23.8], for each $k \in \mathbb{N}$,

$$0 \in \partial f(x^k) + \mu x^k + \lambda\big[v_1^{k-1}\partial|x_1^k| \times \cdots \times v_p^{k-1}\partial|x_p^k|\big] + B_{k-1}(x^k - x^{k-1}) - \delta^{k-1}.$$

In addition, from equation (16) and Proposition 3 (i), it follows that

$$\partial\Theta_{\rho,\nu,\mu}(x^k) = \partial f(x^k) + \mu x^k + \lambda\big[\partial|x_1^k| \times \cdots \times \partial|x_p^k|\big] - \lambda w_\rho(x^k) \circ \operatorname{sign}(x^k).$$

Note that $\partial|x_i^k| = \{\operatorname{sign}(x_i^k)\}$ if $x_i^k \neq 0$ and $\partial|x_i^k| = [-1, 1]$ if $x_i^k = 0$. Then, we have

$$\delta^{k-1} - B_{k-1}(x^k - x^{k-1}) + \lambda[w_\rho(x^{k-1}) - w_\rho(x^k)] \circ \operatorname{sign}(x^k) \in \partial\Theta_{\rho,\nu,\mu}(x^k).$$

Let $u^k := \delta^{k-1} - B_{k-1}(x^k - x^{k-1}) + \lambda[w_\rho(x^{k-1}) - w_\rho(x^k)] \circ \operatorname{sign}(x^k)$. Then, by combining $u^k \in \partial\Theta_{\rho,\nu,\mu}(x^k)$ with the definition of $\Psi_{\rho,\nu,\mu}$, it is not hard to verify that $\zeta^k := (u^k + B(x^k - x^{k-1}); B(x^k - x^{k-1})) \in \partial\Psi_{\rho,\nu,\mu}(x^k, x^{k-1})$ with

$$\begin{aligned}\|\zeta^k\| \leq 2\|B\|\|x^k - x^{k-1}\| + \|\delta^{k-1}\| + \|B_{k-1}\|\|x^k - x^{k-1}\|\\ + \lambda\|[w_\rho(x^{k-1}) - w_\rho(x^k)] \circ \operatorname{sign}(x^k)\|.\end{aligned}$$

Recall that w_ρ is Lipschitz continuous with modulus ρ/σ. From (16), it follows that

$$\|[w_\rho(x^{k-1}) - w_\rho(x^k)] \circ \operatorname{sign}(x^k)\| \leq \|w_\rho(x^{k-1}) - w_\rho(x^k)\| \leq (\rho/\sigma)\|x^{k-1} - x^k\|.$$

Note that $\|B_{k-1}\| \le \|B_0\|$ with $B_0 := \gamma_{1,0} I + \gamma_{2,0} A^{\mathbb{T}} A$. From $\|\delta^{k-1}\| \le \frac{\|x^{k-1}-x^{k-2}\|}{\sqrt{2}\|B^{-1/2}\|}$,

$$\|\zeta^k\| \le (2\|B\| + \|B_0\| + \lambda\rho/\sigma)\|x^k - x^{k-1}\| + \frac{1}{\sqrt{2}\|B^{-1/2}\|}\|x^{k-1} - x^{k-2}\|.$$

The desired result follows with $b_1 = 2\|B\| + \|B_0\| + \frac{\lambda\rho}{\sigma}$ and $b_2 = \frac{1}{\sqrt{2}\|B^{-1/2}\|}$. □

Lemma 5 *Let $\varpi(x^0)$ denote the cluster point set of the sequence $\{x^k\}_{k\in\mathbb{N}}$ and define the set $\Omega := \{(x, x) \mid x \in \varpi(x^0)\}$. Then, the following assertions hold.*

(i) $\varpi(x^0)$ is a nonempty compact set and $\varpi(x^0) \subseteq \mathrm{crit}\Theta_{\rho,\nu,\mu}$;
(ii) $\Omega \subseteq \mathrm{crit}\Psi_{\rho,\nu,\mu}$ and $\lim_{k\to\infty} \mathrm{dist}((x^k, x^{k-1}), \Omega) = 0$;
(iii) the function $\Psi_{\rho,\nu,\mu}$ is finite and keeps the constant on the set Ω.

Proof **(i)** Since $\{(x^k, x^{k-1})\} \subseteq \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^p \mid \Psi_{\rho,\nu,\mu}(x, y) \le \Psi_{\rho,\nu,\mu}(x^0, x^{-1})\}$ by Lemma 4 (i), the boundedness of $\{x^k\}_{k\in\mathbb{N}}$ is implied by the coerciveness of $\Psi_{\rho,\nu,\mu}$. This means that $\varpi(x^0)$ is a nonempty compact set. Pick any $\overline{x} \in \varpi(x^0)$. Then there exists a subsequence $\{x^{k_q}\}_{q\in\mathbb{N}}$ such that $\lim_{q\to\infty} x^{k_q} = \overline{x}$. By Lemma 4 (i), it is easy to obtain $\lim_{q\to\infty} x^{k_q-1} = \overline{x}$. We next argue that $\lim_{q\to\infty} \Theta_{\rho,\nu,\mu}(x^{k_q}) = \Theta_{\rho,\nu,\mu}(\overline{x})$. From the lower semicontinuity of $\Theta_{\rho,\nu,\mu}$, clearly, $\liminf_{q\to\infty} \Theta_{\rho,\nu,\mu}(x^{k_q}) \ge \Theta_{\rho,\nu,\mu}(\overline{x})$. In addition, from the definition of x^k in Step 2 and $\overline{x} \in \mathrm{dom}\, f$, it follows that

$$\begin{aligned} &f_\mu(x^{k_q}) + \lambda\langle v^{k_q-1}, |x^{k_q}|\rangle + \frac{1}{2}\|x^{k_q} - x^{k_q-1}\|^2_{B_{k_q-1}} - \langle \delta^{k_q-1}, x^{k_q} - x^{k_q-1}\rangle \\ &\le f_\mu(\overline{x}) + \lambda\langle v^{k_q-1}, |\overline{x}|\rangle + \frac{1}{2}\|\overline{x} - x^{k_q-1}\|^2_{B_{k_q-1}} - \langle \delta^{k_q-1}, \overline{x} - x^{k_q-1}\rangle \end{aligned}$$

which, by the definition of $\Theta_{\rho,\nu,\mu}$ and $\lambda = \rho\nu$, is equivalent to saying that

$$\begin{aligned} &\Theta_{\rho,\nu,\mu}(x^{k_q}) - \lambda\langle w_\rho(x^{k_q-1}), |x^{k_q}|\rangle + \lambda g_\rho(x^{k_q}) + \frac{1}{2}\|x^{k_q} - x^{k_q-1}\|^2_{B_{k_q-1}} \\ &\le \Theta_{\rho,\nu,\mu}(\overline{x}) - \lambda\langle w_\rho(x^{k_q-1}), |\overline{x}|\rangle + \lambda g_\rho(\overline{x}) + \frac{1}{2}\|\overline{x} - x^{k_q-1}\|^2_{B_{k_q-1}} + \langle \delta^{k_q-1}, x^{k_q} - \overline{x}\rangle. \end{aligned}$$

Passing the limit $q \to \infty$ to this inequality and using the continuity of w_ρ and g_ρ and the boundedness of $\{B_k\}_{k\in\mathbb{N}}$, we obtain $\limsup_{q\to\infty} \Theta_{\rho,\nu,\mu}(x^{k_q}) \le \Theta_{\rho,\nu,\mu}(\overline{x})$. Consequently, the stated limit holds. By the proof of Lemma 4 (ii), $u^{k_q} \in \partial\Theta_{\rho,\nu,\mu}(x^{k_q})$, and moreover, from the definition of u^{k_q}, it is easy to verify that $\lim_{q\to\infty} u^{k_q} = 0$. Together with $\lim_{q\to\infty} x^{k_q} = \overline{x}$ and $\lim_{q\to\infty} \Theta_{\rho,\nu,\mu}(x^{k_q}) = \Theta_{\rho,\nu,\mu}(\overline{x})$, we obtain $0 \in \partial\Theta_{\rho,\nu,\mu}(\overline{x})$, i.e. $\overline{x} \in \mathrm{crit}\Theta_{\rho,\nu,\mu}$. Consequently, $\varpi(x^0) \subseteq \mathrm{crit}\Theta_{\rho,\nu,\mu}$.
(ii)-(iii) Part (ii) is immediate by part (i) and the expression of $\Psi_{\rho,\nu,\mu}$. For part (iii), by picking any $(\overline{x}, \overline{x}) \in \Omega$, there exists a subsequence $\{x^{k_q}\}_{q\in\mathbb{N}}$ such that $\lim_{q\to\infty} x^{k_q} = \overline{x}$. By Lemma 4 (i), $\lim_{q\to\infty} \Theta_{\rho,\nu,\mu}(x^{k_q}) = \varpi^*$ for some $\varpi^* \ge 0$. In addition, from the expression of $\Psi_{\rho,\nu,\mu}$ and $\lim_{q\to\infty} x^{k_q-1} = \overline{x}$, it follows that $\Psi_{\rho,\nu,\mu}(\overline{x}, \overline{x}) = \lim_{q\to\infty} \Theta_{\rho,\nu,\mu}(x^{k_q})$. Thus, $\Psi_{\rho,\nu,\mu}(\overline{x}, \overline{x}) = \varpi^*$. □

When $\Theta_{\rho,\nu,\mu}$ is a KL function, $\Psi_{\rho,\nu,\mu}$ is also a KL function, and if $\Theta_{\rho,\nu,\mu}$ has the KL property of exponent $1/2$ at $\overline{x} \in \operatorname{dom} \Theta_{\rho,\nu,\mu}$, then by the proof of [33, Theorem 3.6], it is easy to verify that $\Psi_{\rho,\nu,\mu}$ has the KL property of exponent $1/2$ at $(\overline{x}, \overline{x}) \in \operatorname{dom} \Psi_{\rho,\nu,\mu}$. Thus, by using Lemma 4-5 and following the similar arguments to those for [6, Theorem 1] and [1, Theorem 2], we can achieve the following convergence results although the subgradient lower bound in Lemma 4 (ii) has a different form the one used in [6, Theorem 1] and [1, Theorem 2].

Theorem 4 *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. The following assertions hold:*

(i) If $\Theta_{\rho,\nu,\mu}$ is a KL function, then $\{x^k\}_{k\in\mathbb{N}}$ is convergent, and its limit, say $\overline{x}$, is a strongly local optimal solution of (14) and (1) whenever $|\overline{x}|_{\rm nz} \geq \phi'_+(1)/\rho$.

(ii) If $\Theta_{\rho,\nu,\mu}$ has the KL property of exponent $1/2$ at $\overline{x}$, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to $\overline{x}$ in a R-linear rate.

Remark 4 By [2, Section 4.3], there are many types of ϑ and $\phi \in \mathscr{L}_{\sigma,\gamma}$ such that the associated $\Theta_{\rho,\nu,\mu}$ is a KL function. For example, if ϑ and $\phi \in \mathscr{L}_{\sigma,\gamma}$ are definable in an o-minimal structure over $\mathbb{R}$, then so is $\Theta_{\rho,\nu,\mu}$ and hence is a KL function.

Suppose that ϑ and $\phi \in \mathscr{L}_{\sigma,\gamma}$ are PLQ functions definable in an o-minimal structure over $\mathbb{R}$. Then the associated $\Theta_{\rho,\nu,\mu}$ is a PLQ KL function, which means that its subdifferential mapping is a polyhedral multifunction and hence is metrically subregular by [51, Proposition 1], and moreover, it is not hard to verify that Assumption 3.1 of [41] holds when $\Theta_{\rho,\nu,\mu}$ is a KL function. By [41, Theorem 3.2 (ii)], the associated $\Theta_{\rho,\nu,\mu}$ is a KL function of exponent $1/2$. Clearly, the functions ϑ defined in (2) or (3), and the ϕ from Example 2 and 3 are such convex functions.

4.2 Statistical error bound of the limit

This section focuses on the scenario where each row $a_i^{\mathbb{T}}$ of A follows the normal distribution $N(0, \Sigma)$ and $b = (b_1, \ldots, b_n)^{\mathbb{T}}$ is from a linear observation model

$$b_i = a_i^{\mathbb{T}} x^* + \varpi_i, \quad i = 1, 2, \ldots, n, \tag{21}$$

where $x^* \in \operatorname{dom} f$ is the true but unknown vector with sparsity $s^* \ll p$, and $\varpi = (\varpi_1, \ldots, \varpi_n)^{\mathbb{T}} \neq 0$ is the noise vector. We shall verify that the limit $\overline{x}$ of the sequence $\{x^k\}_{k\in\mathbb{N}}$ is good from the statistical perspective by establishing its error bound to the true x^*. This requires the following assumption on ϑ, which is satisfied by the functions ϑ defined in (2) and (3).

Assumption 2 $\vartheta(z) \equiv \frac{1}{n}\sum_{i=1}^n \theta(z_i)$ for a PLQ convex function $\theta : \mathbb{R} \to \mathbb{R}_+$, where $\theta(0) = 0$, θ^2 is strongly convex with modulus $\tau > 0$, and there exists $\widetilde{\tau} > 0$ such that

$$|\eta| \leq \widetilde{\tau} \quad \text{for all } \eta \in \partial\theta(t) \text{ and all } t \in \mathbb{R}. \tag{22}$$

Since we are interested in the high-dimensional case, i.e. $n < p$, the sample covariance matrix $\frac{1}{n}A^{\mathbb{T}}A$ is not positive definite, but it may be positive definite

on a subset $\mathcal{C}$ of $\mathbb{R}^p$. For a given index set $S \subset \{1,\ldots,p\}$, the sample covariance matrix $\frac{1}{n}A^{\mathbb{T}}A$ is said to satisfy the restricted eigenvalue (RE) condition over $\mathcal{C}=\{x\in\mathbb{R}^p \mid \|x_{S^c}\|_1 \le 3\|x_S\|_1\}$ with parameter $\kappa>0$ and $S^c:=\{1,\ldots,p\}\backslash S$ if

$$\frac{1}{2n}\|Ax\|^2 \ge \kappa\|x\|^2 \quad \text{for all } x\in\mathcal{C}.$$

Let S^* represent the support of the true vector x^* in model (21). In the sequel, we need a RE condition of $\frac{1}{n}A^{\mathbb{T}}A$ over a set $\mathcal{C}(S^*)$ with parameter $\kappa>0$, where

$$\mathcal{C}(S^*):=\bigcup_{S^*\subset S,|S|\le 1.5s^*}\left\{x\in\mathbb{R}^p:\ \|x_{S^c}\|_1\le 3\|x_S\|_1\right\}$$

comprises those vector x with small components x_j for $j\notin S^*$. By [46, Corollary 1], if Σ satisfies the RE condition over $\mathcal{C}(S^*)$ with parameter $\kappa>0$ (for example, Σ is positive definite), then for $n>\alpha\frac{\max_{1\le j\le p}\Sigma_{jj}}{\kappa}s^*\log p$, the matrix $\frac{1}{n}A^{\mathbb{T}}A$ satisfies the RE condition over $\mathcal{C}(S^*)$ of parameter $\sqrt{2\kappa}/8$ with probability at least $1-\alpha'\exp(-\alpha'' n)$, where α, α' and α'' are the universal positive constants. This means that, for those A with $a_i^{\mathbb{T}}\sim N(0,\Sigma)$ for a positive definite Σ, the matrx $\frac{1}{n}A^{\mathbb{T}}A$ for an appropriately large n satisfies the RE condition over $\mathcal{C}(S^*)$ with a high probability.

by Lemma 9 in "Appendix B", we achieve the following error bound result.

Theorem 5 *Suppose that $\frac{1}{n}A^{\mathbb{T}}A$ satisfies the RE condition of parameter $\kappa>0$ over $\mathcal{C}(S^*)$, and that the sample size n satisfies $n>\frac{72\widetilde{\tau}^2 s^*\|A\|_\infty\|\|A_{\mathcal{I}\cdot}\|\|_1}{\tau\kappa-2\mu\widetilde{\tau}(27\,s^*\|A\|_\infty\|x^*\|_\infty-\|\varpi\|_\infty)}$ for $\|\varpi\|_\infty<27s^*\|A\|_\infty\|x^*\|_\infty$ with $\mathcal{I}=\mathrm{supp}(\varpi)$. If the chosen λ belongs to the interval*

$$\Big[\frac{16\widetilde{\tau}}{n}\|\|A_{\mathcal{I}\cdot}\|\|_1+12\mu\|x^*\|_\infty,\ \frac{2\mu\widetilde{\tau}\|\varpi\|_\infty+\tau\kappa-4\widetilde{\tau}\|A\|_\infty(2\widetilde{\tau}n^{-1}\|\|A_{\mathcal{I}\cdot}\|\|_1+1.5\mu\|x^*\|_\infty)s^*}{4\widetilde{\tau}\|A\|_\infty s^*}\Big],$$

then the limit $\overline{x}$ of the sequence $\{x^k\}_{k\in\mathbb{N}}$ with $|\overline{x}_i|\le\frac{\gamma}{2}$ for $i\notin S^*$ satisfies

$$\|\overline{x}-x^*\|\le\frac{2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda+2\widetilde{\tau}n^{-1}\|\|A_{\mathcal{I}\cdot}\|\|_1+\mu\|x^*\|_\infty\big)\sqrt{s^*}}{2\mu\widetilde{\tau}\|\varpi\|_\infty+\tau\kappa-4\widetilde{\tau}\|A\|_\infty\big(\lambda+2\widetilde{\tau}n^{-1}\|\|A_{\mathcal{I}\cdot}\|\|_1+\mu\|x^*\|_\infty\big)s^*}\quad(23)$$

Theorem 5 states that, when solving the surrogate problem (14) for an appropriate $\lambda=\rho\nu$ with Algorithm 1, the limit $\overline{x}$ of the generated sequence with $|\overline{x}_i|\le\gamma/2$ for $i\notin S^*$ has the error bound in (23). The requirement on $\overline{x}$ is rather mild and is more reasonable than the one used in [10, Assumption 3.7] for smooth loss functions because it allows $\overline{x}$ to have more small nonzero entries than the true x^*. From (23), we see that as the sample size n increases, the error bound of $\overline{x}$ to x^* becomes better, and when the data (b,A) from (21) for ϖ with a higher sparsity or a smaller $\|\varpi\|_\infty$, the error bound is also better.

Proof Since $x^k\to\overline{x}$ as $k\to\infty$ and $\gamma_{1,k}\in[\underline{\gamma_1},\gamma_{1,0}]$ and $\gamma_{2,k}\in[\underline{\gamma_2},\gamma_{2,0}]$, by the definition of ξ^k in (B8) and $\|\delta^{k-1}\|\le\frac{\|B^{1/2}(x^{k-1}-x^{k-2})\|}{\sqrt{2}\|B^{-1/2}\|}$, we have $\xi^k\to-\mu x^*$. Then

there exists $\widehat{k} \in \mathbb{N}$ such that for all $k \geq \widehat{k}$, $\frac{\mu}{2}\|x\|_\infty^* < \|\xi^k\|_\infty < \frac{3\mu}{2}\|x^*\|_\infty$. From $x^k \to \overline{x}$,there exists $\widetilde{k} \in \mathbb{N}$ such that for all $k \geq \widetilde{k}$ and each $i \in \{1, 2, \ldots, n\}$,

$$|x_i^k| - |\overline{x}_i| \leq |x_i^k - \overline{x}_i| \leq \gamma/3.$$

This, by the assumption on $\overline{x}$, implies that $|x_i^k| \leq \frac{5\gamma}{6}$ for $i \notin S^*$. By the proof of Lemma 7 and the expression of $w_\rho(x^k)$, we have $v_i^k = 1$ for each $i \in (S^*)^c$ when $k \geq \widetilde{k}$. Set $\overline{k} := \max(\widehat{k}, \widetilde{k})$. Then, for all $k \geq \overline{k}+1$, $S^{k-1} \equiv S^*$ satisfies the assumption of Lemma 9. Since $\|v_{S^*}^{k-1}\|_\infty \leq 1$ and $\frac{1}{2}\mu\|x\|_\infty^* < \|\xi^k\|_\infty < \frac{3}{2}\mu\|x^*\|_\infty$ for all $k \geq \overline{k}+1$, the choice of the parameter λ implies that for all $k \geq \overline{k}+1$,

$$\frac{16\widetilde{\tau}}{n}|||A_{\mathcal{I}\cdot}|||_1 + 8\|\xi^k\|_\infty \leq \lambda < \frac{2\mu\widetilde{\tau}\|\varpi\|_\infty + \tau\kappa - 4\widetilde{\tau}\|A\|_\infty(2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1 + \|\xi^k\|_\infty)|S^{k-1}|}{4\widetilde{\tau}\|A\|_\infty\|v_{S^*}^{k-1}\|_\infty|S^{k-1}|}.$$

From Lemma 9, we immediately obtain the following inequality

$$\begin{aligned}\|x^k - x^*\| &\leq \frac{2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda\|v_{S^*}^{k-1}\|_\infty + 2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1 + \|\xi^k\|_\infty\big)\sqrt{|S^{k-1}|}}{2\mu\widetilde{\tau}\|\varpi\|_\infty + \tau\kappa - 4\widetilde{\tau}\|A\|_\infty\big(\lambda\|v_{S^*}^{k-1}\|_\infty + 2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1 + \|\xi^k\|_\infty\big)|S^{k-1}|} \\ &\leq \frac{2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda + 2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1 + \|\xi^k\|_\infty\big)\sqrt{s^*}}{2\mu\widetilde{\tau}\|\varpi\|_\infty + \tau\kappa - 4\widetilde{\tau}\|A\|_\infty\big(\lambda + 2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1 + \|\xi^k\|_\infty\big)s^*}.\end{aligned}$$

Taking the limit $k \to \infty$ to the both sides and using $\xi^k \to \mu x^*$ yields the result. □

Remark 5 Let $\overline{n} = \frac{72\widetilde{\tau}^2 s^*\|A\|_\infty|||A_{\mathcal{I}\cdot}|||_1}{\tau\kappa - 2\mu\widetilde{\tau}(27\,s^*\|A\|_\infty\|x^*\|_\infty - \|\varpi\|_\infty)}$. When $n > \overline{n}$, by recalling that $\mu > 0$ is a tiny constant, the choice interval of λ is nonempty and will become larger as n increases. Obviously, the threshold $\overline{n}$ of the sample size increases as the sparsity of ϖ becomes worse. Similar to [36] for the smooth loss, our choice interval of λ involves $\|\varpi\|_\infty$ and a restriction on the sparsity of ϖ. As the sparsity of ϖ increases, the value of λ becomes smaller and the error bound of $\overline{x}$ to the true x^* becomes better. Similar to the ℓ_1-regularized squared-root loss in [3], the parameter λ is required to lie in an interval depending on the sparsity s^*, and a large s^* implies a small interval of λ.

The limit $\overline{x}$ of the sequence $\{x^k\}_{k\in\mathbb{N}}$ depends on the starting point. For the starting point x^0 used in Algorithm 1, the following proposition states that it satisfies the error bound $\|x^0 - x^*\| \leq \overline{\alpha}\sqrt{s^*}$ for a certain constant $\overline{\alpha} > 0$ related to $\widetilde{\gamma}_{1,0}$ and $\widetilde{\gamma}_{2,0}$. Its proof is included in "Appendix B".

Proposition 6 *Suppose that the inexactness of x^0 is defined in the sense of (20), and that $\widetilde{\lambda} \geq 2(\frac{\widetilde{\tau}}{n}|||A|||_1 + \widetilde{\gamma}_{1,0}\|x^*\|_\infty + \widetilde{\gamma}_{2,0}\|A^{\mathbb{T}}Ax^*\|_\infty + \widetilde{\epsilon}_0)$. Then, $\|x^0 - x^*\| \leq \frac{3\widetilde{\lambda}\sqrt{s^*}}{2\widetilde{\gamma}_{1,0}}$.*

4.3 Dual PPA plus semismooth Newton method

The practical efficiency of Algorithm 1 depends on whether the strongly convex subproblem (19) is well solved. Since it involves two nonsmooth terms, ADMM becomes a conventional solver for it. However, it is extremely time-consuming for ADMM to

solve those subproblems with smaller $\gamma_{1,k}$ and $\gamma_{2,k}$, especially under the scenarios where $A^{\mathbb{T}}A$ has a large spectral norm, because the strong convexity of (19) becomes worse and worse as the proximal parameters $\gamma_{1,k}$ and $\gamma_{2,k}$ decrease. Inspired by this, we develop a powerful dual PPA armed with the semismooth Newton method to solve the subproblem (19).

By introducing a variable $z \in \mathbb{R}^p$, the subproblem (19) can be rewritten as

$$\begin{aligned}&\min_{x\in\mathbb{R}^p, z\in\mathbb{R}^p} \vartheta(z) + h_k(x) + \frac{\gamma_{1,k}}{2}\|x-x^k\|^2 + \frac{\gamma_{2,k}}{2}\|z-(Ax^k-b)\|^2 - \langle \delta^k, x-x^k\rangle\\ &\quad \text{s.t.}\quad Ax-b-z=0 \quad \text{with}\ \ h_k(x):=\lambda\|v^k\circ x\|_1 + \frac{\mu}{2}\|x\|^2,\end{aligned} \tag{24}$$

whose dual problem, after an elementary calculation, takes the following form

$$\min_{u\in\mathbb{R}^n} \frac{\|u\|^2}{2\gamma_{2,k}} - e_{\gamma_{2,k}^{-1}\vartheta}\Big(Ax^k-b+\frac{u}{\gamma_{2,k}}\Big) - e_{\gamma_{1,k}^{-1}h_k}\Big(x^k - \frac{A^{\mathbb{T}}u-\delta^k}{\gamma_{1,k}}\Big) + \frac{\|A^{\mathbb{T}}u-\delta^k\|^2}{2\gamma_{1,k}} \tag{25}$$

Clearly, the strong duality holds for problems (24) and (25). It is worth emphasizing that a direct application of the semismooth Newton method to (25) will fail since its generalized Hessian may be singular. Hence, we propose the following inexact PPA armed with the semismooth Newton method, where Φ_k represents the objective function of the dual problem (25).

Algorithm A Inexact PPA with semismooth Newton (dPPASN)

Require: Fix a $k\in\mathbb{N}$. Choose $\underline{\tau}\geq 0$, $\tau_0>0$ and an initial $u^0\in\mathbb{R}^n$. Set $j=0$.

1: **while** the stopping conditions are not satisfied **do**

2: Compute the following u^{j+1} with the semismooth Newton method

$$u^{j+1} \approx \min_{u\in\mathbb{R}^n} \Upsilon_{k,j}(u) := \Phi_k(u) + \frac{\tau_j}{2}\|u-u^j\|^2. \tag{26}$$

3: Update $\tau_{j+1}\downarrow\underline{\tau}$. Let $j\leftarrow j+1$, and then go to Step 1.

4: **end while**

By [48, Section 3] we use the following criterion for the inexactness in (26):

$$\|\nabla\Upsilon_{k,j}(u^{j+1})\| \leq \alpha_j\tau_j\|u^{j+1}-u^j\| \ \text{ with } \ \textstyle\sum_{j=0}^{\infty}\alpha_j<\infty.$$

For the global and linear convergence analysis of Algorithm A under this criterion, the reader can refer to the papers [16, 48].

Solving the subproblem (26) is equivalent to seeking the root of the system $\nabla\Upsilon_{k,j}=0$. Since the mapping $\nabla\Upsilon_{k,j}:\mathbb{R}^n\to\mathbb{R}^n$ is Lipschitz continuous, we define the generalized Hessian of $\Upsilon_{k,j}$ at u by $\partial^2\Upsilon_{k,j}(u):=\partial_C\nabla\Upsilon_{k,j}(u)$. By [27, Theorem 2.2],

$\partial^2\Upsilon_{k,j}(u)d = \widehat{\partial}^2\Upsilon_{k,j}(u)d$ for all $d \in \mathbb{R}^n$ with

$$\widehat{\partial}^2\Upsilon_{k,j}(u) = \tau_j I + \frac{\partial_C \mathcal{P}_{\gamma_{2,k}^{-1}\vartheta}\big(Ax^k - b + \frac{u}{\gamma_{2,k}}\big)}{\gamma_{2,k}} + \frac{A\partial_C \mathcal{P}_{\gamma_{1,k}^{-1}h_k}\big(x^k - \frac{A^{\mathbb{T}}u-\delta^k}{\gamma_{1,k}}\big)A^{\mathbb{T}}}{\gamma_{1,k}}.$$

By mimicking the proof in [40, Section 3.3.4], every $U \in \partial_C \mathcal{P}_{\gamma_{2,k}^{-1}\vartheta}\big(Ax^k - b + \frac{u}{\gamma_{2,k}}\big)$ is symmetric and positive semidefinite, while by Lemma 2 every $V \in \partial_C \mathcal{P}_{\gamma_{1,k}^{-1}h_k}\big(x^k - \frac{A^{\mathbb{T}}u-\delta^k}{\gamma_{1,k}}\big)$ is a positive semidefinite diagonal matrix. Along with $\tau_j > 0$ for each $j \in \mathbb{N}$, the matrix $\tau_j I + \gamma_{2,k}^{-1}U + \gamma_{1,k}^{-1}AVA^{\mathbb{T}}$ is positive definite, so every element in $\widehat{\partial}^2\Upsilon_{k,j}(u)$ is nonsingular. Thus, the following semismooth Newton method can be employed to seek the inexact u^{j+1} in (26), whose global and local convergence analysis can be found in [66, Theorem 3.3−3.4].

Algorithm A.1 Semismooth Newton method

Require: Fix $k, j \in \mathbb{N}$. Choose $\underline{\eta}, \beta \in (0,1)$, $\varsigma \in (0,1]$, $0 < c_1 < c_2 < \frac{1}{2}$. Let $u^0 = u^j$ and set $l = 0$.

1: **while** the stopping conditions are not satisfied **do**

2: Choose $U^l \in \partial_C \mathcal{P}_{\gamma_{2,k}^{-1}\vartheta}\big(Ax^k - b + \frac{u^l}{\gamma_{2,k}}\big)$ and $V^l \in \partial_C \mathcal{P}_{\gamma_{1,k}^{-1}h_k}\big(x^k - \frac{A^{\mathbb{T}}u^l-\delta^k}{\gamma_{1,k}}\big)$.

3: Solve the following linear system exactly or by an iterative method

$$W^l d = -\nabla\Upsilon_{k,j}(u^l) \ \text{ with } \ W^l = \tau_j I + \gamma_{2,k}^{-1}U^l + \gamma_{1,k}^{-1}AV^lA^{\mathbb{T}}$$

to find d^l such that $\|W^l d^l + \nabla\Upsilon_{k,j}(u^l)\| \le \min(\underline{\eta}, \|\nabla\Upsilon_{k,j}(u^l)\|^{1+\varsigma})$.

4: Set $\alpha_l = \beta^{m_l}$, where m_l is the smallest nonnegative integer m satisfying

$$\Upsilon_{k,j}(u^l + \beta^m d^l) \le \Upsilon_{k,j}(u^l) + c_1\beta^m\langle\nabla\Upsilon_{k,j}(u^l), d^l\rangle$$
$$|\langle\nabla\Upsilon_{k,j}(u^l + \beta^m d^l), d^l\rangle| \le c_2|\langle\nabla\Upsilon_{k,j}(u^l), d^l\rangle|.$$

5: Set $u^{l+1} = u^l + \alpha_l d^l$. Let $l \leftarrow l+1$, and then go to Step 1.

6: **end while**

In the sequel, Algorithm 1 armed with dPPASN is termed as PMMSN. To close this subsection, we take a look at the stopping criterion for PMMSN. Let $h_{\lambda,\mu}(x) := \lambda\|x\|_1 + \frac{\mu}{2}\|x\|^2$ with $\lambda = \rho\nu$ for $x \in \mathbb{R}^p$. By Proposition 3 (i), a vector $\overline{x} \in \mathbb{R}^p$ is a critical point of $\Theta_{\rho,\nu,\mu}$ if and only if there exists $\overline{u} \in \mathbb{R}^n$ such that the triple $(\overline{x}, \overline{z}, \overline{u})$ with $\overline{z} = A\overline{x} - b$ satisfies the following system

$$0 \in \partial\vartheta(z) - u, \ 0 \in A^{\mathbb{T}}u - \lambda w_\rho(x) \circ \mathrm{sign}(x) + \partial h_{\lambda,\mu}(x) \ \text{ and } \ Ax - b - z = 0.$$

which by the proximal mappings of ϑ and $h_{\lambda,\mu}$ can be equivalently written as

$$z - \mathcal{P}\vartheta(z+u) = 0, \ x - \mathcal{P}h_{\lambda,\mu}(x - A^{\mathbb{T}}u + \lambda w_\rho(x) \circ \mathrm{sign}(x)) = 0 \ \text{ and } \ Ax - b - z = 0.$$

In view of this, we terminate Algorithm 1 at (x^k, z^k, u^k) when $k \geq k_{\max}$ or

$$E_k := \frac{\|[R_1^k;\, R_2^k;\, Ax^k - b - z^k]\|}{1 + \|b\|} \leq \text{tol} \tag{27}$$

with $R_1^k := z^k - \mathcal{P}\vartheta(z^k + u^k)$ and $R_2^k := x^k - \mathcal{P}h_{\lambda,\mu}(x^k - A^{\mathbb{T}}u^k + \lambda w_\rho(x^k) \circ \text{sign}(x^k))$. Here, $z^k = Ax^k - b$ and u^k is the inexact solution to (25) given by dPPASN.

5 Numerical experiments

We test the performance of PMMSN and compare its performance with that of ADMM and accelerated proximal gradient (APG) for solving the DC surrogate problem (14). Among others, APG is applied to the partially smoothed version of problem (14). Such two methods are very common to solve some structured nonconvex and nonsmooth problems. All numerical tests are performed in MATLAB on a laptop computer running on 64-bit Windows Operating System with an Intel(R) Core(TM) i7-8565U CPU 1.80GHz and 8 GB memory.

5.1 ADMM and APG for surrogate problem

For any given $\nu > 0$, $\mu > 0$, $\rho > 0$, let $g_{\rho,\lambda,\mu}(x) := \lambda(\|x\|_1 - g_\rho(x)) + \frac{\mu}{2}\|x\|^2$ with $\lambda = \rho\nu$ for $x \in \mathbb{R}^p$. Observe that problem (14) is equivalent to

$$\min_{x,s\in\mathbb{R}^p, z\in\mathbb{R}^n} \big\{\vartheta(z) + g_{\rho,\lambda,\mu}(s) \;\text{ s.t. }\; Ax - z - b = 0,\; x - s = 0\big\}, \tag{28}$$

and the proximal mapping of $g_{\rho,\lambda,\mu}$ has a closed form. It is natural to apply the classical ADMM [24] to solving the problem (28). For a given $\beta > 0$, the augmented Lagrangian function of problem (28) is defined by

$$\begin{aligned} L_\beta(x, z, s; u, v) := \vartheta(z) + g_{\rho,\lambda,\mu}(s) + \langle u, Ax - b - z\rangle + \langle v, x - s\rangle \\ + \frac{\beta}{2}\big[\|Ax - b - z\|^2 + \|x - s\|^2\big]. \end{aligned}$$

The iteration steps of the ADMM for solving (14) are described as follows.

Algorithm 2 (ADMM for surrogate problem (14))

Require: Choose $\rho \geq 1$, $\beta > 0$ and an initial (z^0, s^0, u^0, v^0). Let $\lambda = \rho\nu$.
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Compute the optimal solution of the following problems

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathbb{R}^p} L_\beta(x, z^k, s^k; u^k, v^k); & \text{(29a)} \\ (z^{k+1}, s^{k+1}) = \arg\min_{z\in\mathbb{R}^n, s\in\mathbb{R}^p} L_\beta(x^{k+1}, z, s; u^k, v^k) & \text{(29b)}. \end{cases}$$

3: Let $u^{k+1} = u^k + \beta(Ax^{k+1} - z^{k+1} - b)$ and $v^{k+1} = v^k + \beta(x^{k+1} - s^{k+1})$.
4: **end for**

Remark 6 **(a)** By the expression of L_β, an elementary calculation yields that

$$\begin{aligned} x^{k+1} &= (I + A^{\mathbb{T}}A)^{-1}\big[s^k + A^{\mathbb{T}}(b + z^k - u^k/\beta) - v^k/\beta\big], \\ z^{k+1} &= \mathcal{P}_{\beta^{-1}\vartheta}(Ax^{k+1} - b + \beta^{-1}u^k) \ \text{ and } \ s^{k+1} \in \mathcal{P}_{\beta^{-1}g_{\rho,\lambda,\mu}}\big(x^{k+1} + v^k/\beta\big). \end{aligned} \tag{30}$$

Since the proximal mapping of $g_{\rho,\lambda,\mu}$ has a closed form, the main cost of ADMM in each step is to solve system (30). When n is small or medium-scale, one can use the direct method which, by the Sherman-Morrison-Woodbury formula, is equivalent to requiring a Cholesky decomposition $I + AA^{\mathbb{T}}$; when n is large-scale, one may use an iterative method such as the conjugate gradient to get x^{k+1}. For Algorithm 2, we terminate it at the iterate $(x^k, z^k, s^k, u^k, v^k)$ when $k \geq k_{\max}$, or the primal infeasibility $\text{pinf}^k \leq \epsilon^1_{\text{admm}}$ and the relative KKT residual $\text{Err}^k \leq \epsilon^2_{\text{admm}}$ where

$$\begin{aligned} \text{pinf}^k &:= \sqrt{\|Ax^k - z^k - b\|^2 + \|x^k - s^k\|^2}, \\ \text{Err}^k &:= \frac{\sqrt{\|s^k - \mathcal{P}g_{\rho,\lambda,\mu}(s^k + v^k)\|^2 + \|z^k - \mathcal{P}\vartheta(z^k + u^k)\|^2 + \|Ax^k - z^k - b\|^2}}{1 + \|b\|}. \end{aligned}$$

(b) To the best of our knowledge, there is no convergence certificate on ADMM for the DC problem (28) due to the nonsmoothness of ϑ, and the convergence results in [56] are inapplicable to it. Nevertheless, as will be shown in Sect. 5.2, this method still works for our test examples.

The APG for the problem (14) is developed by its partially smoothed form

$$\min_{x\in\mathbb{R}^p} \Big\{ e_\varepsilon\vartheta(Ax - b) + \lambda(\|x\|_1 - \varphi_{\rho,\lambda}(x)) + \frac{\mu}{2}\|x\|^2 \Big\}, \tag{31}$$

where $e_\varepsilon\vartheta$ is the Moreau envelope of ϑ w.r.t. parameter $\varepsilon > 0$, and $\varphi_{\rho,\lambda}$ is same as in the proof of Proposition 3 (iii). Since $e_\varepsilon\vartheta$ is smooth with Lipschitz gradient and the proximal mapping of $\varphi_{\rho,\lambda}$ has a closed form, one may apply Nesterov's APG [39] for solving (31). The iterates of APG are as follows, where $f_{\varepsilon,\mu}(x) := e_\varepsilon\vartheta(Ax - b) + \frac{\mu}{2}\|x\|^2$ and L is the Lipschitz modulus of $\nabla f_{\varepsilon,\mu}$.

Algorithm 3 (APG for DC surrogate problem (14))

Require: Choose $\rho \geq 1$, $\varepsilon > 0$ and an initial x^0. Let $\lambda = \rho\nu$. Set $t_{-1} = t_0 = 1$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Let $y^k = x^k + \frac{t_{k-1}-1}{t_k}(x^k - x^{k-1})$ and compute

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^p} \Big\{ \langle \nabla f_{\varepsilon,\mu}(y^k), x - y^k\rangle + (L/2)\|x - y^k\|^2 + \varphi_{\rho,\lambda}(x) \Big\}.$$

3: Set $t_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_k^2})$.

4: **end for**

For the convergence analysis of APG under the above nonconvex and nonsmooth setting, the reader may refer to [58]. For Algorithm 3, we terminate at x^k when

$$\frac{\|x^k - \mathcal{P}\varphi_{\rho,\lambda}(x^k - A^{\mathbb{T}}\nabla e_\varepsilon \vartheta(Ax^k - b) - \mu x^k)\|}{1 + \|b\|} \leq \epsilon_{\text{apg}}. \tag{32}$$

5.2 Implementation details of three solvers

We choose $\vartheta(z) = \frac{1}{n}\sum_{i=1}^n \theta(z_i)$ with $\theta(t) = |t|$ to test the performance of PMMSN, ADMM and APG for solving the surrogate problem (14) associated to ϕ in Example 3 for $a = 4.0$. For such ϑ and ϕ, the assumptions of Theorems 2 and 4 are satisfied and Assumption 2 in Sect. 4.2 also holds. We measure the performance of a solver in terms of the approximate sparsity and relative ℓ_2-error of its output x^{out}, the objective value of (14) at x^{out}, and the CPU time, where the approximate sparsity and relative ℓ_2-error of x^{out} is defined as

$$N_{\text{nz}}(x^{\text{out}}) := \sum_{i=1}^p \mathbb{I}\big\{|x_i^{\text{out}}| > 10^{-8}\|x^{\text{out}}\|_\infty\big\} \quad \text{and} \quad \textbf{L2err} := \frac{\|x^{\text{out}} - x^*\|}{\|x^*\|}.$$

We terminate Algorithm 1 at x^k when condition (27) is satisfied for $\text{tol} = 10^{-6}$ and $N_{\text{nz}}(x^k) = N_{\text{nz}}(x^{k-1}) = N_{\text{nz}}(x^{k-2})$; terminate ADMM at x^k when $\text{Err}^k \leq 10^{-4}$, $\text{pinf}^k \leq 10^{-4}$ and $N_{\text{nz}}(x^k) = \cdots = N_{\text{nz}}(x^{k-49})$; and terminate APG at x^k when the criterion (32) holds with $\epsilon_{\text{apg}} = 10^{-5}$ and $N_{\text{nz}}(x^k) = \cdots = N_{\text{nz}}(x^{k-49})$. Since the approximate sparsity yielded by ADMM and APG has a worse stability, we terminate the two solvers when the approximate sparsity keeps unchanged in the past 50 steps instead of 2 steps as for Algorithm 1. The three solvers are also terminated when $k > k_{\max}$, where $k_{\max} = 100, 25000$ and 20000 are respectively used for Algorithm 1, ADMM and APG.

During the testing, we choose $\mu = 10^{-8}$, $\rho = 2$ and the parameter ν (or $\lambda = \rho\nu$) is specified in the examples. The parameters of Algorithm 1 are choose as follows: $\widetilde{\gamma}_{1,0} = \gamma_{1,0} = 10^{-3}$, $\widetilde{\gamma}_{2,0} = \gamma_{2,0} = 10^{-4}$, and $\underline{\gamma_1} = \underline{\gamma_2} = 10^{-6}$. For ADMM, we choose $\beta = 1$. The starting point of ADMM is chosen to be $(z^0, s^0, u^0, v^0) = (-b, 0, 0, 0)$, and that of APG is chosen to be $x^0 = 0$.

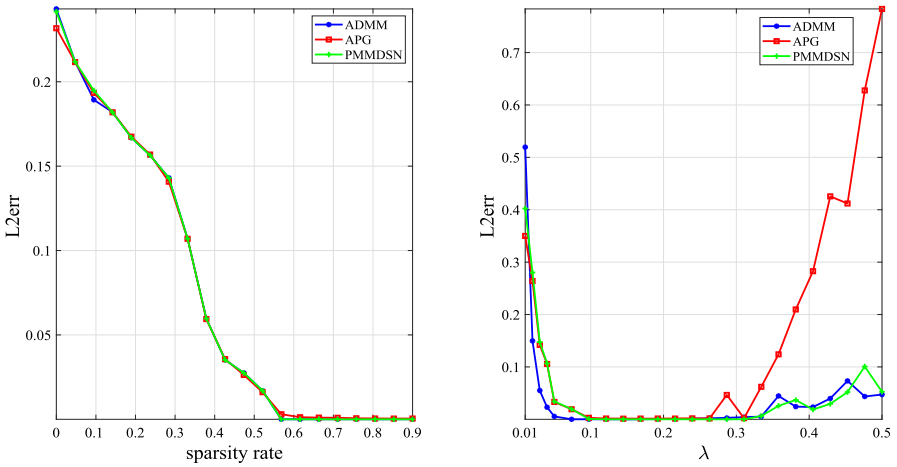

Fig. 1 The relative ℓ_2-error of three solvers under different sparsity rate $(n - |\mathcal{I}|)/n$ and λ

For the smoothing parameter ε in (31), preliminary tests indicate that, when ε increases, the sparsity and relative ℓ_2-error of the output $\overline{x}^\varepsilon$ for APG become better and keep unchanged once ε is over a threshold; while the loss $\frac{1}{n}\|A\overline{x}^\varepsilon - b\|_1$ first increases and then decreases a certain level and keeps it unchanged. In view of this, we regard the smallest one among those ε for which $\overline{x}^\varepsilon$ has the sparsity closest to the sparsity of the true x^* for synthetic data (respectively, the output of PMMSN for real data) as the best, by noting that such $\overline{x}^\varepsilon$ usually has a favorable ℓ_2-error and model (31) with a smaller ε is closer to (14). We search such a suboptimal ε^* from an appropriate interval by comparing the sparsity of $\overline{x}^\varepsilon$ corresponding to 20 ε's. The search of such ε^*, albeit impracticality, is just for numerical comparisons.

5.3 Numerical comparisons on synthetic data

We test the performance of three solvers on synthetic data (b, A), where b is obtained from the linear observation model (21) with each $a_i^{\mathbb{T}} \sim N(0, \Sigma)$ and a noise vector ϖ with i.i.d. nonzero entries. The covariance matrix Σ, the true x^* and the distribution of the nonzero entries of ϖ are specified below.

5.3.1 Performance under different sparsity of ϖ

We use the example in [26] for testing, which is generated randomly as follows.

Example 4 Take $(n, p) = (200, 1000)$, $\Sigma_{ij} = 0.8^{|i-j|}$ and $\varpi_{\mathcal{I}} \sim N(0, 2I)$. The true x^* has the form $(2,\ 0,\ 1.5,\ 0,\ 0.8,\ 0,\ 0,\ 1,\ 0,\ 1.75,\ 0,\ 0,\ 0.75,\ 0,\ 0,\ 0.3,\ 0, \ldots, 0)^{\mathbb{T}}$.

The left subfigure of Fig. 1plots the average relative ℓ_2-error of three solvers for solving 10 test problems under different sparsity rate of the noise vector ϖ. PMMSN is solving the problem (14) with $\lambda = \max(10^{-4}, \frac{1}{2p}|||A|||_1)$, ADMM is solving (28) with the same λ, and APG is solving the partially smoothed form (31) with the same λ and

$\varepsilon^* = 1.5$. We see that the average relative ℓ_2-error of three solvers decreases as the sparsity rate increases, which is consistent with the conclusion of Theorem 5, and the average ℓ_2-errors of three solvers are comparable under different sparsity rate for this example.

5.3.2 Performance under different λ

We test the performance of three solvers under different λ by using the following example from [26], which involves a heavily-tailed noise.

Example 5 Be same as Example 4 except that $\varpi_{\mathcal{I}} = \frac{\|Ax^*\|}{3\|\xi_{\mathcal{I}}\|}\xi_{\mathcal{I}}$ with $|\mathcal{I}| = \lfloor 0.5n \rfloor$, where all entries of $\xi_{\mathcal{I}}$ obey the Cauchy distribution of density $d(u) = \frac{1}{\pi(1+u^2)}$.

The right subfigure of Fig. 1 plots the average relative ℓ_2-error curve of three solvers under different λ, where ADMM is solving (28) with the same λ, and APG is solving the partially smoothed form (31) with the same λ and $\varepsilon^* = 1.5$. We see that the average relative ℓ_2-error of three solvers have an arc curve as λ increases in $[0.01, 0.5]$, and the relative ℓ_2-error of PMMSN and ADMM is lower than that of APG. This means that if the sparsity of ϖ is well controlled, there exists an interval of λ in which the error bound of the outputs of three solvers has a small variation, and the interval for PMMSN and ADMM is larger than the one for APG.

5.3.3 Performance on other sparse noises

We test the performance of three solvers for other types of sparse noises via 25 examples, generated with $p = 5000$, $s^* = \lfloor \sqrt{p}/2 \rfloor$ and $n = \lfloor 2\, s^* \ln p \rfloor$. The sparsity of ϖ is set to be $|\mathcal{I}| = \lfloor 0.3n \rfloor$ and the nonzero entries of x^* follow $N(0, 4)$. The noise ϖ comes from the distributions used in [26], including **(1)** the normal distribution $N(0, 10^2)$ **(2)** the scaled Student's t-distribution with 4 degrees of freedom $\sqrt{2} \times t_4$ **(3)** the Cauchy distribution with density $d(u) = \frac{1}{\pi(1+u^2)}$ **(4)** the mixture normal distribution $N(0, \sigma^2)$ with $\sigma \sim \mathrm{Unif}(1, 5)$, denoted by MN_2 **(5)** the Laplace distribution with density $d(u) = 0.5\exp(-|u|)$. The covariance matrix Σ includes **(1)** $\Sigma = I$ **(2)** $\Sigma = (0.5^{|i-j|})$ **(3)** $\Sigma = (0.8^{|i-j|})$ **(4)** $\Sigma = (\alpha + (1-\alpha)\mathbb{I}_{\{i=j\}})$ for $\alpha = 0.5$ **(5)** $\Sigma = (\alpha + (1-\alpha)\mathbb{I}_{\{i=j\}})$ for $\alpha = 0.8$.

Table 1 reports the average result of **10** test problems for each case with $\lambda = \lambda_c \max(10^{-4}, \|\|A\|\|_1/p)$, where a=PMMSN, b=ADMM and c=APG, **Nz** denotes the average approximate sparsity of x^{out}, **Obj** means the average objective value of (14), **FP** and **FZ** represent the average number of false positives and false zeros of x^{out}, respectively, and ε^* columns list the suboptimal ε used for APG. We see that PMMSN yields the lowest objective value and relative ℓ_2-error and the best (**FP**, **FZ**) for all examples, and the average sparsity of its output equals that of x^* for all examples except $(\Sigma, \varpi) = (3, 1)$; while ADMM has a comparable performance with APG in terms of the relative ℓ_2-error and the objective value. We find that APG yields the worst sparsity and (**FP**, **FZ**) for those examples with the covariance matrix Σ from **(4)**–**(5)** or noise ϖ from the Cauchy distribution because the parameter ε is very sensitive to the data and the suboptimal ε^* is not suitable for all 10 test problems. This shows that

Table 1 The performance of three solvers for the sparse noise examples

Prob	λ_c	ε^*_{apg}	Nz			Obj			L2err			FP			FZ			Time(s)		
(Σ, ϖ)			a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
(1,1)	1.2	15	35.0	38.0	35.1	4.403	4.882	4.405	3.07e−10	1.42e−01	1.30e−03	0.0	6.9	0.2	0.0	3.9	0.1	1.8	189.7	8.6
(1,2)	1.2	15	35.0	35.0	35.0	2.418	2.418	2.420	1.08e−12	9.63e−05	1.26e−03	0.0	0.0	0.1	0.0	0.0	0.1	1.4	196.5	7.8
(1,3)	1.2	15	35.0	35.7	34.9	4.970	4.976	4.973	9.52e−12	1.51e−03	1.25e−03	0.0	0.7	0.0	0.0	0.0	0.1	1.5	189.3	7.1
(1,4)	1.2	15	35.0	36.4	35.0	2.693	2.708	2.696	1.46e−12	4.53e−03	1.29e−03	0.0	1.4	0.1	0.0	0.0	0.1	1.3	190.6	7.6
(1,5)	1.2	15	35.0	35.0	34.9	2.266	2.265	2.268	2.14e−11	1.05e−04	1.23e−03	0.0	0.0	0.0	0.0	0.0	0.1	1.6	185.1	7.2
(2,1)	1.2	15	35.0	38.0	35.1	4.410	5.055	4.413	2.44e−12	2.09e−01	1.44e−03	0.0	9.1	0.2	0.0	6.1	0.1	2.0	177.0	9.3
(2,2)	1.2	15	35.0	35.4	35.0	2.425	2.434	2.428	2.37e−11	2.27e−03	1.42e−03	0.0	0.4	0.1	0.0	0.0	0.1	1.7	182.4	7.6
(2,3)	1.2	15	35.0	36.1	34.9	4.977	5.036	4.980	6.58e−11	1.21e−02	1.41e−03	0.0	1.4	0.0	0.0	0.3	0.1	1.7	181.0	7.2
(2,4)	1.2	15	35.0	35.9	35.0	2.701	2.757	2.703	5.36e−12	1.33e−02	1.44e−03	0.0	1.1	0.1	0.0	0.2	0.1	1.6	181.4	7.6
(2,5)	1.2	15	35.0	35.6	34.9	2.273	2.277	2.276	5.12e−12	1.27e−03	1.39e−03	0.0	0.6	0.0	0.0	0.0	0.1	1.7	187.5	7.7
(3,1)	1.2	40	35.7	39.0	35.9	4.428	5.096	4.435	2.60e−02	2.85e−01	3.00e−02	1.2	12.1	1.6	0.5	8.1	0.7	2.3	180.2	8.7
(3,2)	1.2	40	35.0	35.0	35.3	2.415	2.415	2.423	5.89e−12	5.32e−05	3.99e−03	0.0	0.0	0.5	0.0	0.0	0.2	1.8	175.6	6.5
(3,3)	1.2	40	35.0	35.6	35.8	4.968	4.969	4.975	7.58e−11	1.16e−03	4.01e−03	0.0	0.6	1.0	0.0	0.0	0.2	1.9	174.9	6.4
(3,4)	1.2	40	35.0	39.9	35.4	2.691	2.732	2.698	8.60e−12	2.07e−02	4.02e−03	0.0	5.0	0.6	0.0	0.1	0.2	1.9	175.1	6.9
(3,5)	1.2	40	35.0	35.1	35.3	2.263	2.263	2.270	1.77e−12	4.69e−05	3.90e−03	0.0	0.1	0.5	0.0	0.0	0.2	2.2	173.0	6.3
(4,1)	0.8	500	35.0	39.5	31.6	3.731	4.221	3.965	1.23e−11	2.14e−01	1.40e−01	0.0	9.7	1.9	0.0	5.2	5.3	2.9	193.1	53.1
(4,2)	0.7	500	35.0	36.1	31.9	1.583	1.606	1.664	8.22e−11	7.61e−03	4.58e−02	0.0	1.5	0.0	0.0	0.4	3.1	2.5	173.6	34.8
(4,3)	0.7	500	35.0	35.2	32.4	4.135	4.172	4.225	8.04e−11	1.04e−02	4.97e−02	0.0	0.8	0.0	0.0	0.6	2.6	2.3	174.4	32.7
(4,4)	0.7	500	35.0	36.1	31.8	1.858	1.893	1.943	4.17e−12	1.05e−02	5.00e−02	0.0	1.6	0.0	0.0	0.5	3.2	2.3	174.3	38.5
(4,5)	0.7	500	35.0	35.9	32.4	1.431	1.442	1.513	2.12e−12	4.63e−03	4.35e−02	0.0	1.1	0.0	0.0	0.2	2.6	2.8	171.7	32.8
(5,1)	0.5	500	35.0	38.6	35.7	3.229	3.436	3.753	7.28e−12	9.49e−02	3.45e−01	0.0	5.7	11.2	0.0	2.1	10.5	2.3	149.4	43.1
(5,2)	0.4	500	35.0	35.0	35.3	1.084	1.084	1.192	3.52e−12	2.51e−05	6.73e−02	0.0	0.0	3.6	0.0	0.0	3.3	2.0	138.9	48.2
(5,3)	0.5	500	35.0	34.9	32.7	3.797	3.864	4.028	2.86e−11	2.49e−02	1.33e−01	0.0	0.5	1.9	0.0	0.6	4.2	1.7	132.1	44.7
(5,4)	0.5	500	35.0	34.5	35.8	1.520	1.563	1.772	6.30e−12	1.79e−02	1.66e−01	0.0	0.2	5.9	0.0	0.7	5.1	1.8	136.6	45.4
(5,5)	0.4	500	35.0	35.0	32.1	0.932	0.932	1.027	1.12e−11	2.64e−05	5.77e−02	0.0	0.0	0.2	0.0	0.0	3.1	2.4	137.9	48.1

replacing ϑ with its smooth approximation $e_\varepsilon\vartheta$ is not effective for highly-relevant Σ and heavily-tailed sparse noises, though ε is elaborately selected. In addition, PMMSN requires the least CPU time that is at most a fiftieth of the CPU time for ADMM and a fifth of the CPU time for APG.

5.4 Numerical comparisons on real data

This part uses the LIBSVM datasets from https://www.csie.ntu.edu.tw to test the efficiency of PMMSN for large scale problems. For those data sets with a few features, such as **pyrim, abalone, bodyfat, housing, space ga**, we follow the same technique as in [53] to expand their original features by using polynomial basis functions over those features. For example, the last digit in **pyrim5** indicates that a polynomial of order 5 is used to generate the basis function. Such a naming convention is also applicable to the other expanded data sets. These data sets are quite difficult in terms of the dimension and the largest eigenvalues of $A^{\mathbb{T}}A$. Table 2reports the results of three solvers with λ specified in the third column. We see that the output of PMMSN has the lowest objective value for ten examples, and for **abalone7**, **housing7** and **mpg7** its objective value is much less than that of the output yielded by ADMM and APG; while the output of ADMM has much worse objective value than that of APG since the maximum number of iterates 25000 is not enough for these examples. The CPU time taken by PMMSN is less than **one-third** of the CPU time taken by APG and **one-fiftieth** of the CPU time taken by ADMM for all examples except E2006.train and E2006.test.with ADMM, APG has a little better performance and takes much less CPU time, but its performance depends much on the smoothing parameter ε that is very sensitive to those data with a highly-relevant covariance matrix or a heavily-tailed sparse noise. From the numerical results on synthetic examples, we see that when the sparsity of ϖ attains a certain level, say, $|\mathcal{I}| \le 0.6n$ for the examples in Table 1, the relative ℓ_2-error has an order about 10^{-10}, close to the exact recovery. Then, it is natural to ask for which kind of covariances and noises, the exact recovery of the limit $\overline{x}$ can be achieved by controlling the sparsity of ϖ. We leave this question for a future topic.

6 Conclusions

For the zero-norm regularized problem, we verified that the penalty problem of its equivalent MPEC reformulation is a global exact penalty, which implies a family of equivalent DC surrogates. For a subfamily of these DC surrogates, we showed that the critical point set coincides with the d-directional stationary point set, and when a critical point has no too small nonzero component, it is a strongly local optimal solution of the surrogate problem and the zero-norm regularized problem. We also developed a proximal MM method for solving these DC surrogates, and provided its theoretical certificates by establishing the global convergence and the local R-linear rate of the generated iterate sequence. In particular, for the data (b, A) from a linear observation model, the statistical error bound was also achieved for the limit to the true

Table 2 The performance of three solvers for the eight data from LIBSVM datasets

Name of data	$\|A\|^2$	λ	ε^*_{apg}	NZ			Obj			Time(s)		
				a	b	c	a	b	c	a	b	c
log1p.E2006.train 16087;4272227	5.86e+7	1.14e−2	2000	197	99	228	0.252	2.424	0.255	749.3	20746.3	3321.6
log1p.E2006.test 3308;4272226	1.46e+7	2.48e−2	1000	94	163	89	0.242	0.249	0.247	513.5	9148.1	1314.5
E2006.train 16087;150360	1.91e+5	5.00e−5	1500	17	10	84	0.265	3.509	0.266	219.2	302.3	6.9
E2006.test 3308;150358	4.79e+4	5.00e−5	1000	30	5	17	0.235	3.801	0.236	26.2	13.7	4.9
abalone7 4177;6435	5.23e+5	3.25e−2	1500	8	5	20	1.817	5.769	1.910	15.8	1305.7	35.9
bodyfat7 252;116280	5.30e+4	1.08e−3	50	109	99	50	0.001	0.001	0.008	21.8	208.6	107.4
housing7 506;77520	3.28e+5	1.63e−2	3000	69	112	97	2.253	3.065	2.641	19.1	1537.2	124.9
mpg7 392;3432	1.30e+4	5.71e−3	500	63	88	77	1.562	1.695	1.689	0.9	49.7	2.4
pyrim5 74;201376	1.22e+6	1.47e−4	2500	95	132	118	0.003	0.003	0.093	6.9	175.4	36.8
space ga9 3107;5005	4.10e+3	3.10e−2	8	5	5	7	0.140	0.143	0.140	2.5	150.5	9.9

vector. Numerical comparisons with ADMM and APG show that the proximal MM method armed with dPPASN has a remarkable superiority in the quality of solutions and the CPU time, and is very robust for A with a large spectral norm and b corrupted by the heavily-tailed noise. It is worth pointing out that our global exact penalty result and the proposed proximal MM method are applicable to the function $\vartheta = h + \delta_\Omega$, where $h : \mathbb{R}^p \to \mathbb{R}$ is a finite convex function and Ω is a general polyhedral set $\{x \in \mathbb{R}^p \mid Cx - d \geq 0\}$.

Acknowledgements The first two authors would like to express their sincere thanks to Prof. Kim-Chuan Toh from National University of Singapore for helpful suggestions on the implementation of Algorithm A.1 when visiting SCUT, and give thanks to Prof. Liping Zhu from RenMin University of China for helpful discussion on Theorem 5.

Funding The funding was provided by the National Natural Science Foundation of China under projects No. 11971177 and the Hong Kong Research Grant Council under grant No. 15304019

Data availability The data used to form the test problems in Subsection 5.4 are freely available in https://www.csie.ntu.edu.tw.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest to this work.

Appendix A: Proof of Proposition 3

The following two technical lemmas are need for the proof of Proposition 3.

Lemma 6 *Fix any $\nu > 0$ and $\mu > 0$. Let $h_\nu(x) := \nu\|x\|_0$ for $x \in \mathbb{R}^p$. If ϑ is regular and strictly continuous relative to* $\mathrm{dom}\,\vartheta$*, then for any $x \in \mathrm{dom}\, f$ and $\zeta \in \mathbb{R}^p$,*

$$\widehat{\partial}\Theta_{\nu,\mu}(x) = \partial\Theta_{\nu,\mu}(x) = \partial f_\mu(x) + \partial h_\nu(x), \tag{A1}$$

$$\widehat{d}\Theta_{\nu,\mu}(x)(\zeta) = d\Theta_{\nu,\mu}(x)(\zeta) = df_\mu(x)(\zeta) + dh_\nu(x)(\zeta). \tag{A2}$$

Proof Fix any $x \in \mathrm{dom}\, f$ and $\zeta \in \mathbb{R}^p$. Since ϑ is strictly continuous relative to $\mathrm{dom}\,\vartheta$, by the expression of f_μ in (4), the function f_μ can be rewritten as $\widetilde{f}_\mu + \delta_{\mathrm{dom}\, f}$ where $\widetilde{f}_\mu$ is a finite strictly continuous function on $\mathbb{R}^p$. Clearly, $\widetilde{f}_\mu$ is regular by the regularity of ϑ, and $\delta_{\mathrm{dom}\, f}$ is also regular by the polyhedrality of $\mathrm{dom}\,\vartheta$. By invoking [50, Exercise 10.10] and the first inclusion of [50, Corollary 10.9], it is not hard to obtain that

$$\partial\widetilde{f}_\mu(x) + \widehat{\partial}(\delta_{\mathrm{dom}\, f} + h_\nu)(x) \subseteq \widehat{\partial}\Theta_{\nu,\mu}(x) \subseteq \partial\Theta_{\nu,\mu}(x) \subseteq \partial\widetilde{f}_\mu(x) + \partial(\delta_{\mathrm{dom}\, f} + h_\nu)(x).$$

Since $\mathrm{epi}\, h_\nu$ is a union of finitely many polyhedral sets and $\mathrm{dom}\, f$ is polyhedral, from [29, Page 213] it follows that $\partial(\delta_{\mathrm{dom}\, f} + h_\nu)(x) \subseteq \mathcal{N}_{\mathrm{dom}\, f}(x) + \partial h_\nu(x)$ and $\partial^\infty(\delta_{\mathrm{dom}\, f} + h_\nu)(x) \subseteq \mathcal{N}_{\mathrm{dom}\, f}(x) + \partial^\infty h_\nu(x)$. The first inclusion, along with the first inclusion of [50, Corollary 10.9] and the regularity of $\delta_{\mathrm{dom}\, f}$ and h_ν, implies that

$$\mathcal{N}_{\mathrm{dom}\, f}(x) + \partial h_\nu(x) \subseteq \widehat{\partial}(\delta_{\mathrm{dom}\, f} + h_\nu)(x) \subseteq \partial(\delta_{\mathrm{dom}\, f} + h_\nu)(x) \subseteq \mathcal{N}_{\mathrm{dom}\, f}(x) + \partial h_\nu(x).$$

The regularity of h_ν is implied by [30, Theorem 1]. The last two equations imply the first equality in (A1). By the strict continuity of $\widetilde{f}_\mu$ and [50, Corollary 10.9],

$$\begin{aligned} d\Theta_{\nu,\mu}(x)(\zeta) &\geq d\widetilde{f}_\mu(x)(\zeta)+d\delta_{\mathrm{dom}\, f}(x)(\zeta)+dh_\nu(x)(\zeta) = df_\mu(x)(\zeta)+dh_\nu(x)(\zeta),\\ \widehat{d}\Theta_{\nu,\mu}(x)(\zeta) &\leq \widehat{d}\widetilde{f}_\mu(x)(\zeta)+\widehat{d}(\delta_{\mathrm{dom}\, f}+h_\nu)(x)(\zeta)\\ &\leq \widehat{d}\widetilde{f}_\mu(x)(\zeta)+\widehat{d}\delta_{\mathrm{dom}\, f}(x)(\zeta)+\widehat{d}h_\nu(x)(\zeta)\\ &= d\widetilde{f}_\mu(x)(\zeta)+d\delta_{\mathrm{dom}\, f}(x)(\zeta)+dh_\nu(x)(\zeta)\\ &= df_\mu(x)(\zeta)+dh_\nu(x)(\zeta) \end{aligned} \tag{A3}$$

where the second inequality in (A3) is due to $\partial^\infty(\delta_{\mathrm{dom}\, f}+h_\nu)(\overline{x}) \subseteq \mathcal{N}_{\mathrm{dom}\, f}(\overline{x}) + \partial^\infty h_\nu(\overline{x})$ and [50, Exercise 8.23], and the first equality in (A3) is due to the regularity of $\widetilde{f}_\mu$, h_ν and $\mathrm{dom}\, f$. Note that $\widehat{d}\Theta_{\nu,\mu}(\overline{x})(\zeta) \geq d\Theta_{\nu,\mu}(\overline{x})(\zeta)$. From the last two inequality, we obtain the second equality in (A1). The proof is completed. □

Lemma 7 *Pick any $\phi \in \mathscr{L}_{\sigma,\gamma}$. The associated function g_ρ for any $\rho > 0$ is continuously differentiable on $\mathbb{R}^p$.*

Proof Recall that ψ^* is a finite nondecreasing convex function on $\mathbb{R}$. If in addition ϕ is strongly convex on [0, 1] with modulus σ, then by [49, Theorem 26.3] and [50, Proposition 12.60], ψ^* is smooth on $\mathbb{R}$ and $(\psi^*)'$ is Lipschitz continuous with constant $1/\sigma$. Thus, by the expression of g_ρ, it suffices to argue that $h(t):=\rho^{-1}\psi^*(\rho|t|)$ for $t \in \mathbb{R}$ is continuously differentiable at $t = 0$. Indeed, by the assumption on ϕ, it is easy to verify that $\psi^*(s) = 0$ for all $s \in [0, \gamma]$. Then, for all $|t| \leq \gamma$, $h(t) = 0$. Together with $h(0) = 0$, h is differentiable at $t = 0$ with $h'(0) = 0$. □

Proof **(i)** Since the range of $\partial\psi^*$ is contained in $\mathrm{dom}\partial\psi = [0, 1]$, for any $x \in \mathbb{R}^p$ it holds that $\|x\|_1 - g_\rho(x) \geq 0$. Together with the nonnegativity and coerciveness of f_μ, it follows that $\Theta_{\rho,\nu,\mu}$ is nonnegative and coercive. Fix any $x \in \mathrm{dom}\, f$. From Lemma 7 and [50, Exericse 8.8], $\widehat{\partial}\Theta_{\rho,\nu,\mu}(x) = \partial\Theta_{\rho,\nu,\mu}(x) = \partial(f_\mu+\rho\nu\|\cdot\|_1)(x)-\rho\nu\nabla g_\rho(x)$. Recall that $[\mathrm{Im}(A)-b]\cap\mathrm{dom}\vartheta \neq \emptyset$. By the convexity of ϑ and [49, Theorem 23.9],

$$\partial(f_\mu+\rho\nu\|\cdot\|_1)(x) = A^{\mathbb{T}}\partial\vartheta(Ax-b)+\mu x+\rho\nu\partial\|x\|_1.$$

The characterization on the regular and limiting subdifferentials of $\Theta_{\rho,\nu,\mu}$ then holds.
(ii) By the definition of d-stationary point for DC program (see [42, Section 3]), a point $x \in \mathrm{dom}\, f$ is a d-stationary point of (14) iff $\rho\nu\nabla g_\rho(x) \in \partial(f_\mu+\rho\nu\|\cdot\|_1)(x)$, which by part (i) is equivalent to saying that $x \in \mathrm{dom}\, f$ is a limiting critical point of $\Theta_{\rho,\nu,\mu}$.
(iii) By [50, Theorem 13.24 (c)], it suffices to argue that $d^2\Theta_{\rho,\nu,\mu}(\overline{x}|0)(\zeta) > 0$ for all $\zeta \neq 0$. Fix any $\zeta \in \mathbb{R}^p\backslash\{0\}$. Let $\varphi_{\rho,\lambda}(x):=\lambda[\|x\|_1-g_\rho(x)]$ with $\lambda = \rho\nu$ for $x \in \mathbb{R}^p$. Clearly, $\varphi_{\rho,\lambda}$ is Lipschitz continuous and regular by the smoothness of g_ρ. Note that $\Theta_{\rho,\nu,\mu} = f_\mu + \varphi_{\rho,\lambda}$. By invoking [50, Proposition 13.19], it follows that

$$d^2\Theta_{\rho,\nu,\mu}(\overline{x}|0)(\zeta) \geq \sup_{u\in\partial f_\mu(\overline{x}),v\in\partial\varphi_{\rho,\lambda}(\overline{x})}\Big\{d^2f_\mu(\overline{x}\,|\,u)(\zeta)+d^2\varphi_{\rho,\lambda}(\overline{x}|v)(\zeta)\ \text{ s.t. }\ u+v=0\Big\}. \tag{A4}$$

Recall that f_μ is strongly convex with modulus μ. By Definition 3, we have that

$$d^2 f_\mu(\overline{x}\,|\,u)(\zeta) \geq \mu\|\zeta\|^2 > 0 \quad \forall u \in \partial f_\mu(\overline{x}). \tag{A5}$$

Fix any $v \in \partial\varphi_{\rho,\lambda}(\overline{x})$. Since $\varphi_{\rho,\lambda}$ is Lipschitz and directionally differentiable,

$$\langle v, \zeta\rangle \leq d\varphi_{\rho,\lambda}(\overline{x})(\zeta) = \varphi'_{\rho,\lambda}(\overline{x}, \zeta) = \lambda(\|\cdot\|_1)'(\overline{x}, \zeta) - \lambda\langle\nabla g_\rho(\overline{x}), \zeta\rangle.$$

By [50, Proposition 13.5], $d^2\varphi_{\rho,\lambda}(\overline{x}|v)(\zeta) = +\infty$ when $d\varphi_{\rho,\lambda}(\overline{x})(\zeta) > \langle v, \zeta\rangle$. This, together with (A4)-(A5), implies that $d^2\Theta_{\rho,\nu,\mu}(\overline{x}|0)(\zeta) > 0$, so it suffices to consider that $\varphi'_{\rho,\lambda}(\overline{x}; \zeta) = \langle v, \zeta\rangle$. In this case, from Definition 3 it follows that

$$\begin{aligned} d^2\varphi_{\rho,\lambda}(\overline{x}|v)(\zeta) &= \liminf_{\substack{\tau\downarrow 0\\ \zeta'\to\zeta}} \frac{\varphi_{\rho,\lambda}(\overline{x}+\tau\zeta') - \varphi_{\rho,\lambda}(\overline{x}) - \tau\varphi'_{\rho,\lambda}(\overline{x}, \zeta')}{\frac{1}{2}\tau^2} \\ &= \lambda \liminf_{\substack{\tau\downarrow 0\\ \zeta'\to\zeta}} \frac{-g_\rho(\overline{x}+\tau\zeta') + g_\rho(\overline{x}) + \tau\langle\nabla g_\rho(\overline{x}), \zeta'\rangle}{\frac{1}{2}\tau^2}, \end{aligned} \tag{A6}$$

where the second equality is because $\|\overline{x}+\tau\zeta'\|_1 - \|\overline{x}\|_1 - \tau(\|\cdot\|_1)'(\overline{x}, \zeta') = 0$ for any $\tau > 0$ small enough. Let $h(t) := \rho^{-1}\psi^*(\rho|t|)$ for $t \in \mathbb{R}$. Clearly, $g_\rho(z) = \sum_{i=1}^p h(z_i)$ for $z \in \mathbb{R}^p$. When $i \notin \mathrm{supp}(\overline{x})$, from the proof of Lemma 7, for all $\tau > 0$ small enough, we have $h(\overline{x}_i+\tau\zeta'_i) - h(\overline{x}_i) - \tau h'(\overline{x}_i)\zeta'_i = 0$. When $i \in \mathrm{supp}(\overline{x})$, by noting that $\psi^*(s) = s - \phi(1)$ for all $s \geq \phi'_+(1)$ and using the assumption $|\overline{x}|_{\mathrm{nz}} \geq \phi'_+(1)/\rho$,

$$h(\overline{x}_i+\tau\zeta'_i) - h(\overline{x}_i) - \tau h'(\overline{x}_i)\zeta'_i = 0 = |\overline{x}_i+\tau\zeta'_i| - \overline{x}_i - \tau\mathrm{sign}(\overline{x}_i)\zeta'_i = 0$$

for all sufficiently $\tau > 0$. This means that, for all $\tau > 0$ small enough,

$$-g_\rho(\overline{x}+\tau\zeta') + g_\rho(\overline{x}) + \tau\langle\nabla g_\rho(\overline{x}), \zeta'\rangle = 0.$$

By combining this with (A6), we obtain from (A4)-(A5) that $d^2\Theta_{\rho,\nu,\mu}(\overline{x}|0)(\zeta) > 0$.
(iv) By Lemma 6, $\widehat{\mathrm{crit}}\,\Theta_{\nu,\mu} = \mathrm{crit}\,\Theta_{\nu,\mu}$. We next argue that $\overline{x} \in \widehat{\mathrm{crit}}\,\Theta_{\nu,\mu}$. Since $\overline{x} \in \mathrm{crit}\,\Theta_{\rho,\nu,\mu}$, from part (i) it follows that

$$0 \in A^{\mathbb{T}}\partial\vartheta(A\overline{x}-b) + \mu\overline{x} + \rho\nu\big[(1-(w_\rho(\overline{x}))_1)\partial|\overline{x}_1| \times \cdots \times (1-(w_\rho(\overline{x}))_p)\partial|\overline{x}_p|\big]$$

where $[w_\rho(\overline{x})]_i = (\psi^*)'(\rho|\overline{x}_i|)$ for $i = 1, 2, \ldots, p$. By the definition of ψ^*, it is easy to deduce that $\psi^*(s) = s - \phi(1)$ for all $s \geq \phi'_+(1)$. Together with $|\overline{x}|_{\mathrm{nz}} \geq \phi'_+(1)/\rho$, it holds that $[w_\rho(\overline{x})]_i = (\psi^*)'(\rho|\overline{x}_i|) = 1$ for all $i \in \mathrm{supp}(\overline{x})$. From [30, Theorem 1], we know that $\widehat{\partial}\|\overline{x}\|_0 = \{v \in \mathbb{R}^p \mid v_i = 0 \text{ for } i \in \mathrm{supp}(\overline{x})\}$. This means that

$$\rho\nu\big[(1-(w_\rho(\overline{x}))_1)\partial|\overline{x}_1| \times \cdots \times (1-(w_\rho(\overline{x}))_p)\partial|\overline{x}_p|\big] \subseteq \nu\widehat{\partial}\|\overline{x}\|_0.$$

From the last two equations, $0 \in A^{\mathbb{T}}\partial\vartheta(A\overline{x}-b) + \mu\overline{x} + \widehat{\partial}\|\overline{x}\|_0 = \widehat{\partial}\Theta_{\nu,\mu}(\overline{x})$, where the equality is by Lemma 6. This means that $\overline{x} \in \widehat{\mathrm{crit}}\,\Theta_{\nu,\mu}$. For the rest, it suffices to

argue that every point in crit $\Theta_{\nu,\mu}$ is a strongly local optimal solution of (1). Pick any $\overline{x} \in \text{crit}\,\Theta_{\nu,\mu}$. We only need to argue that $d^2\Theta_{\nu,\mu}(\overline{x}|0)(\zeta) > 0$ for all $\zeta \neq 0$. Fix any $0 \neq \zeta \in \mathbb{R}^p$. By combining Lemma 6 with [50, Proposition 13.19], it holds that

$$d^2\Theta_{\nu,\mu}(\overline{x}|0)(\zeta) \geq \sup_{u\in\partial f_\mu(\overline{x}),v\in\partial h_\nu(\overline{x})}\Big\{d^2 f_\mu(\overline{x}|u)(\zeta)+d^2 h_\nu(\overline{x}|v)(\zeta)\ \text{ s.t. } u+v=0\Big\} \tag{A7}$$

where h_ν is same as in Lemma 6. Fix any $v \in \partial h_\nu(\overline{x})$. Let $\overline{J}:=\{1,\ldots,p\}\backslash\text{supp}(\overline{x})$. Then, $\langle v,\zeta\rangle = \langle v_{\overline{J}},\zeta_{\overline{J}}\rangle$. A simple calculation yields $dh_\nu(\overline{x})(\zeta) = \sum_{i\in\overline{J}}\delta_{\{0\}}(\zeta_i)$. This means that $dh_\nu(\overline{x})(\zeta) \geq \langle v,\zeta\rangle$. When $dh_\nu(\overline{x})(\zeta) > \langle v,\zeta\rangle$, by [50, Proposition 13.5] we have $d^2h_\nu(\overline{x}|\xi)(\zeta) = +\infty$. This along with (A5) and (A7) means that $d^2\Theta_{\nu,\mu}(\overline{x}|0)(\zeta) > 0$, so it suffices to consider the case $dh_\nu(\overline{x})(\zeta) = \langle v,\zeta\rangle$. For this case, from $dh_\nu(\overline{x})(\zeta) = \sum_{i\in\overline{J}}\delta_{\{0\}}(\zeta_i)$, we have $\zeta_{\overline{J}} = 0$. Consequently,

$$\begin{aligned} d^2h_\nu(\overline{x}|v)(\zeta) &= \liminf_{\tau\downarrow 0,\zeta'\to\zeta}\frac{h_\nu(\overline{x}+\tau\zeta')-h_\nu(\overline{x})-\tau\langle v_{\overline{J}},\zeta'_{\overline{J}}\rangle}{\frac{1}{2}\tau^2} \\ &= \liminf_{\tau\downarrow 0,\zeta'_{\overline{J}}\to\zeta_{\overline{J}}}\frac{\sum_{i\in\overline{J}}[\text{sign}(\tau|\zeta'_i|)-\tau v_i\zeta'_i]}{\frac{1}{2}\tau^2} \geq 0. \end{aligned}$$

This along with (A5) and (A7) implies that $d^2\Theta_{\nu,\mu}(\overline{x}|0)(\zeta) > 0$. □

Appendix B: Proof of results in Sect. 4.2

In this section, let x^* be the true vector in model (21), and for each $k \in \mathbb{N}$ write

$$y^k{:=}Ax^k-b,\ \ \Delta x^k{:=}x^k-x^*\ \text{and}\ \xi^k{:=}B_{k-1}(x^{k-1}-x^k)+\delta^{k-1}-\mu x^*. \tag{B8}$$

By Assumption 2 and [50, Theorem 10.49], for any $\bar{t} \in \mathbb{R}$ we have $\partial(\theta^2)(\bar{t}) = 2D^*\theta(\bar{t})(\theta(\bar{t}))$ where $D^*\theta(\bar{t}):\mathbb{R}\rightrightarrows\mathbb{R}$ is the coderivative of θ at $\bar{t}$. Together with [50, Proposition 9.24(b)], $D^*\theta(\bar{t})(\theta(\bar{t})) = \partial(\theta(\bar{t})\theta)(\bar{t})$. Thus,

$$\partial(\theta^2)(\bar{t}) = \begin{cases}\{0\} & \text{if } \theta(\bar{t})=0;\\ 2\theta(\bar{t})\partial\theta(\bar{t}) & \text{otherwise}\end{cases}\qquad \text{for any } \bar{t}\in\mathbb{R}. \tag{B9}$$

By using (B9) and the above notation, we can establish the following lemma.

Lemma 8 *Suppose that for a certain $k \geq 1$ there exists an index set $S^{k-1} \supseteq S^*$ satisfying $\min_{i\in(S^{k-1})^c} v_i^{k-1} \geq 1/2$. Let $\mathcal{I}:=\{i \in \{1,\ldots,n\} \mid \varpi_i \neq 0\}$. Then, when $\lambda \geq 16\widetilde{\tau}n^{-1}\|A_{\mathcal{I}\cdot}\|_1 + 8\|\xi^k\|_\infty$, it holds that $\|\Delta x^k_{(S^{k-1})^c}\|_1 \leq 3\|\Delta x^k_{S^{k-1}}\|_1$.*

Proof From $x^* \in \text{dom}\,f$ and the definition of x^k in Step 2, it is not difficult to obtain

$$f(x^*)+\frac{\mu}{2}\|x^*\|^2+\lambda\langle v^{k-1},|x^*|\rangle+\frac{1}{2}\|x^*-x^{k-1}\|^2_{B_{k-1}}$$

$$\geq f(x^k)+\frac{\mu}{2}\|x^k\|^2+\lambda\langle v^{k-1},|x^k|\rangle+\frac{1}{2}\|x^k-x^{k-1}\|^2_{B_{k-1}}$$
$$+\frac{1}{2}\langle x^*-x^k,(\mu I+B_{k-1})(x^*-x^k)\rangle+\langle\delta^{k-1},x^*-x^k\rangle,$$

where the strong convexity of the objective function of (19) is used. After a suitable rearrangement for the last inequality, we obtain

$$f(x^k)-f(x^*)+\mu\|\Delta x^k\|^2\leq\lambda\langle v^{k-1},|x^*|-|x^k|\rangle+\langle\xi^k,x^k-x^*\rangle. \quad \text{(B10)}$$

For each $k\in\mathbb{N}$, let $\mathcal{J}_k:=\{i\notin\mathcal{I}\,|\,y_i^k\neq 0\}$. By the expression of ϑ and $\varpi=b-Ax^*$,

$$\begin{aligned}&\vartheta(Ax^k-b)-\vartheta(Ax^*-b)\\&=\frac{1}{n}\Big[\sum_{i\in\mathcal{J}_k}\frac{\theta^2(y_i^k)-\theta^2(\varpi_i)}{\theta(y_i^k)+\theta(\varpi_i)}+\sum_{i\in\mathcal{I}}\frac{\theta^2(y_i^k)-\theta^2(\varpi_i)}{\theta(y_i^k)+\theta(\varpi_i)}\Big]\\&\geq\frac{1}{n}\Big[\sum_{i\in\mathcal{J}_k}\frac{\theta^2(y_i^k)-\theta^2(\varpi_i)}{\widetilde{\tau}\|y^k\|_\infty}+\sum_{i\in\mathcal{I}}\frac{\theta^2(y_i^k)-\theta^2(\varpi_i)}{\theta(y_i^k)+\theta(\varpi_i)}\Big].\end{aligned} \quad \text{(B11)}$$

where the inequality is since $\theta(y_i)\leq\widetilde{\tau}\|y\|_\infty$ for $i=1,\ldots,n$, implied by $\theta(0)=0$ and (22). Fix any $\eta_i\in\partial(\theta^2)(\varpi_i)$. Since θ^2 is strongly convex with modulus τ, we have

$$\theta^2(y_i^k)-\theta^2(\varpi_i)\geq\eta_i(y_i^k-\varpi_i)+0.5\tau(y_i^k-\varpi_i)^2\ \text{ for }\ i=1,\ldots,n. \quad \text{(B12)}$$

Along with (B9), for each $i\in\mathcal{J}_k$, $\eta_i=0$ and $\theta^2(y_i^k)-\theta^2(\varpi_i)\geq\frac{\tau}{2}(y_i^k-\varpi_i)^2$, and consequently,

$$\sum_{i\in\mathcal{J}_k}\frac{\theta^2(y_i^k)-\theta^2(\varpi_i)}{\widetilde{\tau}\|y^k\|_\infty}\geq\frac{\tau}{2\widetilde{\tau}}\sum_{i\in\mathcal{J}_k}\frac{(y_i^k-\varpi_i)^2}{\|y^k\|_\infty}.$$

For each $i\in\mathcal{I}$, write $\widetilde{y}_i^k:=\frac{\eta_i}{\theta(y_i^k)+\theta(\varpi_i)}$. From (B9) and (22), it is not hard to obtain $|\widetilde{y}_i^k|\leq 2\widetilde{\tau}$ for all $i\in\mathcal{I}$. Together with (B12), $\varpi=b-Ax^*$ and $\theta(y_i^k)\leq\widetilde{\tau}\|y^k\|_\infty$,

$$\begin{aligned}\sum_{i\in\mathcal{I}}\frac{\theta^2(y_i^k)-\theta^2(\varpi_i)}{\theta(y_i^k)+\theta(\varpi_i)}&\geq\sum_{i\in\mathcal{I}}\widetilde{y}_i^k(y_i^k-\varpi_i)+\frac{\tau}{2}\sum_{i\in\mathcal{I}}\frac{(y_i^k-\varpi_i)^2}{\theta(y_i^k)+\theta(\varpi_i)}\\&\geq-2\widetilde{\tau}\|[A(x^k-x^*)]_{\mathcal{I}}\|_1+\frac{\tau}{2}\sum_{i\in\mathcal{I}}\frac{(y_i^k-\varpi_i)^2}{\widetilde{\tau}(\|y^k\|_\infty+\|\varpi\|_\infty)}.\end{aligned}$$

Substituting the last two inequalities into (B11) and using the definition of f yields

$$f(x^k)-f(x^*)=\vartheta(Ax^k-b)-\vartheta(Ax^*-b)$$

$$\geq -\frac{2\widetilde{\tau}}{n}\|[A(x^k-x^*)]_{\mathcal{I}}\|_1+\frac{\tau\|A(x^k-x^*)\|^2}{2n\widetilde{\tau}(\|y^k\|_\infty+\|\varpi\|_\infty)}.$$

Write $\Upsilon^k:=\frac{\tau\|A(x^k-x^*)\|^2}{2n\widetilde{\tau}(\|y^k\|_\infty+\|\varpi\|_\infty)}$. By combining this inequality and (B10), we get

$$\begin{aligned}\mu\|\Delta x^k\|^2+\Upsilon^k &\leq \lambda\langle v^{k-1},|x^*|-|x^k|\rangle+2\widetilde{\tau}n^{-1}\big\|[A(x^k-x^*)]_{\mathcal{I}}\big\|_1+\langle\xi^k,x^k-x^*\rangle\\ &\leq \lambda\Big(\textstyle\sum_{i\in S^*}v_i^{k-1}|\Delta x_i^k|-\sum_{i\in(S^{k-1})^c}v_i^{k-1}|\Delta x_i^k|\Big)\\ &\quad+\big(2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty\big)\big(\|\Delta x^k_{S^{k-1}}\|_1+\|\Delta x^k_{(S^{k-1})^c}\|_1\big). \quad \text{(B13)}\end{aligned}$$

Since $S^{k-1}\supseteq S^*$ and $v_i^{k-1}\in[0.5,1]$ for $i\in(S^{k-1})^c$, from the last inequality we have

$$\begin{aligned}\mu\|\Delta x^k\|^2+\Upsilon^k &\leq \textstyle\sum_{i\in S^{k-1}}\big(\lambda v_i^{k-1}+2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty\big)|\Delta x_i^k|\\ &\quad+\textstyle\sum_{i\in(S^{k-1})^c}\big(n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty-\lambda/2\big)|\Delta x_i^k|\\ &=\big(\lambda+2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty\big)\big\|\Delta x^k_{S^{k-1}}\big\|_1\\ &\quad+\big(2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty-0.5\lambda\big)\big\|\Delta x^k_{(S^{k-1})^c}\big\|_1.\end{aligned}$$

From the nonnegativity of the left hand side and the given assumption on λ, we have

$$\big\|\Delta x^k_{(S^{k-1})^c}\big\|_1\leq\frac{\lambda+2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty}{0.5\lambda-2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1-\|\xi^k\|_\infty}\big\|\Delta x^k_{S^{k-1}}\big\|_1\leq 3\big\|\Delta x^k_{S^{k-1}}\big\|_1.$$

This implies that the desired result holds. The proof is completed. □

By invoking (B13) and Lemma 8, we can obtain the following conclusion.

Lemma 9 *Suppose that $A^{\mathbb{T}}A/n$ satisfies the RE condition of parameter $\kappa>0$ on $\mathcal{C}(S^*)$, and that for some $k\geq 1$ there is an index set $S^{k-1}\supseteq S^*$ and $|S^{k-1}|\leq 1.5s^*$ such that $\min_{i\in(S^{k-1})^c}v_i^{k-1}\geq\frac{1}{2}$. Let $\mathcal{I}:=\{i\mid\varpi_i\neq 0\}$. If λ is chosen such that $16\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+8\|\xi^k\|_\infty\leq\lambda<\frac{2\mu\widetilde{\tau}\|\varpi\|_\infty+\tau\kappa-4\widetilde{\tau}\|A\|_\infty(2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty)|S^{k-1}|}{4\widetilde{\tau}\|A\|_\infty\|v^{k-1}_{S^*}\|_\infty|S^{k-1}|}$,*

$$\|\Delta x^k\|\leq\frac{2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty\big)\sqrt{|S^{k-1}|}}{2\mu\widetilde{\tau}\|\varpi\|_\infty+\tau\kappa-4\widetilde{\tau}\|A\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}|||A_{\mathcal{I}\cdot}|||_1+\|\xi^k\|_\infty\big)|S^{k-1}|}.$$

Proof Note that $\|y^k\|_\infty+\|\varpi\|_\infty=\|\varpi-A\Delta x^k\|_\infty+\|\varpi\|_\infty\leq\|A\Delta x^k\|_\infty+2\|\varpi\|_\infty$. Then

$$\frac{\tau\|A(x^k-x^*)\|^2}{2n\widetilde{\tau}(\|z^k\|_\infty+\|\varpi\|_\infty)}\geq\frac{\tau\|A\Delta x^k\|^2}{2n\widetilde{\tau}(\|A\Delta x^k\|_\infty+2\|\varpi\|_\infty)}:=\widetilde{\Upsilon}^k.$$

Together with inequality (B13) and $v_i^{k-1}\in[0.5,1]$ for $i\in(S^{k-1})^c$, it follows that

$$\mu\|\Delta x^k\|^2+\widetilde{\Upsilon}^k\leq\lambda\textstyle\sum_{i\in S^*}v_i^{k-1}|\Delta x_i^k|-(\lambda/2)\sum_{i\in(S^{k-1})^c}|\Delta x_i^k|$$

$$
\begin{aligned}
&+\big(2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)\big(\|\Delta x^k_{S^{k-1}}\|_1+\|\Delta x^k_{(S^{k-1})^c}\|_1\big)\\
&\le\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)\|\Delta x^k_{S^{k-1}}\|_1
\end{aligned}
$$

where the second inequality is due to $\lambda\ge 16\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+8\|\xi^k\|_\infty$. By Lemma 8, $\|\Delta x^k_{(S^{k-1})^c}\|_1\le 3\|\Delta x^k_{S^{k-1}}\|_1$, which means that $\Delta x^k\in\mathcal{C}(S^*)$. From the assumption on $\frac{1}{n}A^{\mathbb{T}}A$, we have $\|A\Delta x^k\|^2\ge 2n\kappa\|\Delta x^k\|^2$. Then, it holds that

$$
\begin{aligned}
&\mu\|\Delta x^k\|^2+\frac{\tau\kappa\|\Delta x^k\|^2}{\widetilde{\tau}(\|A\Delta x^k\|_\infty+2\|\varpi\|_\infty)}\\
&\le\Big(\lambda\|v^{k-1}_{S^*}\|_\infty+\frac{2\widetilde{\tau}}{n}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\Big)\big\|\Delta x^k_{S^{k-1}}\big\|_1.
\end{aligned}
$$

Multiplying the both sides of this inequality with $\widetilde{\tau}(\|A\Delta x^k\|_\infty+2\|\varpi\|_\infty)$ yields that

$$
\begin{aligned}
&\big[\mu\widetilde{\tau}(\|A\Delta x^k\|_\infty+2\|\varpi\|_\infty)+\tau\kappa\big]\|\Delta x^k\|^2\\
&\le\widetilde{\tau}\|A\Delta x^k\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)\big\|\Delta x^k_{S^{k-1}}\big\|_1\\
&\quad+2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)\big\|\Delta x^k_{S^{k-1}}\big\|_1.
\end{aligned}
$$

Note that $\|A\Delta x^k\|_\infty\le\|A\|_\infty\|\Delta x^k\|_1$. Together with $\|\Delta x^k_{(S^{k-1})^c}\|_1\le 3\|\Delta x^k_{S^{k-1}}\|_1$, $\|A\Delta x^k\|_\infty\le 4\|A\|_\infty\|\Delta x^k_{S^{k-1}}\|_1$, so the right hand side of the last inequality satisfies

$$
\begin{aligned}
\mathrm{RHS}&\le 4\widetilde{\tau}\|A\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)|S^{k-1}|\big\|\Delta x^k_{S^{k-1}}\big\|^2\\
&\quad+2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)\sqrt{|S^{k-1}|}\big\|\Delta x^k_{S^{k-1}}\big\|.
\end{aligned}
$$

From the last two equations, a suitable rearrangement yields that

$$
\begin{aligned}
&\Big[2\mu\widetilde{\tau}\|\varpi\|_\infty+\tau\kappa-4\widetilde{\tau}\|A\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)|S^{k-1}|\Big]\|\Delta x^k\|^2\\
&\le 2\widetilde{\tau}\|\varpi\|_\infty\big(\lambda\|v^{k-1}_{S^*}\|_\infty+2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty\big)\sqrt{|S^{k-1}|}\big\|\Delta x^k_{S^{k-1}}\big\|,
\end{aligned}
$$

which along with $\lambda<\frac{2\mu\widetilde{\tau}\|\varpi\|_\infty+\tau\kappa-4\widetilde{\tau}\|A\|_\infty(2\widetilde{\tau}n^{-1}\|\!|A_{\mathcal{I}\cdot}|\!\|_1+\|\xi^k\|_\infty)|S^{k-1}|}{4\widetilde{\tau}\|A\|_\infty\|v^{k-1}_{S^*}\|_\infty|S^{k-1}|}$ implies the desired result. The proof is then completed. □

B.1 Proof of Proposition 6:

Let $\Delta x^0{:=}x^0-x^*$. From $x^*\in\mathrm{dom}\,f$ and the strong convexity of (20),

$$
\begin{aligned}
&f(x^*)+\widetilde{\lambda}\|x^*\|_1+\frac{\widetilde{\gamma}_{1,0}}{2}\|x^*\|^2+\frac{\widetilde{\gamma}_{2,0}}{2}\|Ax^*\|^2\\
&\ge f(x^0)+\widetilde{\lambda}\|x^0\|_1+\frac{\widetilde{\gamma}_{1,0}}{2}\|x^0\|^2+\frac{\widetilde{\gamma}_{2,0}}{2}\|Bx^0\|^2
\end{aligned}
$$

$$+\langle\widetilde{\delta}^0, x^*-x^0\rangle+\frac{1}{2}\langle(x^*-x^0),(\widetilde{\gamma}_{1,0}I+\widetilde{\gamma}_{20}A^{\mathbb{T}}A)(x^*-x^0)\rangle.$$

From $\vartheta(z)=\frac{1}{n}\sum_{i=1}^{n}\theta(z_i)$ and Assumption 2, $f(x^*)-f(x^0)\le\frac{\widetilde{\tau}}{n}\|A(x^*-x^0)\|_1$. Notice that $\|x^0\|^2-\|x^*\|^2=\|x^0-x^*\|^2+2\langle x^0-x^*,x^*\rangle$. Together with the last inequality and $\|\widetilde{\delta}^0\|_\infty\le\widetilde{\epsilon}_0$, it follows that

$$\begin{aligned}\|\Delta x^0\|^2_{\widetilde{\gamma}_{1,0}I+\widetilde{\gamma}_{2,0}A^{\mathbb{T}}A}&\le\widetilde{\lambda}(\|x^*\|_1-\|x^0\|_1)+n^{-1}\widetilde{\tau}\|A(x^*-x^0)\|_1\\&\quad+\langle x^0-x^*,\widetilde{\delta}^0+\widetilde{\gamma}_{2,0}A^{\mathbb{T}}Ax^*-\widetilde{\gamma}_{1,0}x^*\rangle\\&\le\big(\widetilde{\lambda}+\widetilde{\tau}n^{-1}|||A|||_1+\widetilde{\gamma}_{1,0}\|x^*\|_\infty+\widetilde{\gamma}_{2,0}\|A^{\mathbb{T}}Ax^*\|_\infty+\widetilde{\epsilon}_0\big)\|\Delta x^0_{S^*}\|_1\\&\quad+\big(\widetilde{\tau}n^{-1}|||A|||_1+\widetilde{\gamma}_{1,0}\|x^*\|_\infty+\widetilde{\gamma}_{2,0}\|A^{\mathbb{T}}Ax^*\|_\infty+\widetilde{\epsilon}_0-\widetilde{\lambda}\big)\|\Delta x^0_{(S^*)^c}\|_1.\end{aligned}$$

By the assumption on $\widetilde{\lambda}$ and the nonnegativity of $\|\Delta x^0\|^2_{\widetilde{\gamma}_{1,0}I+\widetilde{\gamma}_{2,0}A^{\mathbb{T}}A}$, we get $\|\Delta x^0_{(S^*)^c}\|_1\le 3\|\Delta x^0_{S^*}\|_1$. Substituting this into the last inequality yields

$$\begin{aligned}&\|\Delta x^0\|^2_{\widetilde{\gamma}_{1,0}I+\widetilde{\gamma}_{2,0}A^{\mathbb{T}}A}\\&\le\big(\widetilde{\lambda}+\widetilde{\tau}n^{-1}|||A|||_1+\widetilde{\gamma}_{1,0}\|x^*\|_\infty+\widetilde{\gamma}_{2,0}\|A^{\mathbb{T}}Ax^*\|_\infty+\widetilde{\epsilon}_0\big)\big\|\Delta x^0_{S^*}\big\|_1\\&\le\frac{3\widetilde{\lambda}\sqrt{s^*}}{2}\big\|\Delta x^0\big\|\end{aligned}$$

which implies that the desired conclusion holds. The proof is completed. □

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**, 5–16 (2009)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**, 438–457 (2010)
3. Belloni, A., Chernozhukov, V.: Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika **4**, 791–806 (2010)
4. Bi, S.J., Liu, X.L., Pan, S.H.: Exact penalty decomposition method for zero-norm minimization based on MPEC formulation. SIAM J. Sci. Comput. **36**, A1451–A1477 (2014)
5. Bian, W., Chen, X.J.: A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. SIAM J. Numer. Anal. **58**, 858–883 (2020)
6. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)
7. Bot, R.I., Nguyen, D.K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Oper. Res. **45**, 1–31 (2018)
8. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Proceeding of ICML (1998)
9. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**, 34–81 (2009)
10. Cao, S.S., Huo, X.M., Pang, J.S.: *A unifying framework of high-dimensional sparse estimation with difference-of-convex (DC) regularizations*, arXiv:1812.07130 (2018)
11. Chartrand, R.: Exact reconstruction of sparse signals via nonconvex minimization. IEEE Signal Process. Lett. **14**, 707–710 (2007)

12. Chen, X.J., Xu, F.M., Ye, Y.Y.: Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. SIAM J. Sci. Comput. **32**, 2832–2852 (2010)
13. Clarke, F.H.: Optimization and Nonsmooth Analysis. John Wiley and Sons, New York (1983)
14. Cui, Y., Chang, T.H., Hong, M., Pang, J.S.: A study of piecewise linear-quadratic programs. J. Optim. Theory Appl. **186**, 523–553 (2020)
15. Cui, Y., Pang, J.S.: Modern Nonconvex Nondifferentiable Optimization. Society for Industrial and Applied Mathematics, Philadelphia (2022)
16. Cui, Y., Sun, D.F., Toh, K.C.: On the R-superlinear convergence of the KKT residuals generated by the augmented Lagrangian method for convex composite conic programming. Math. Program. **178**, 381–415 (2019)
17. Donoho, D.L., Stark, B.F.: Uncertainty principles and signal recovery. SIAM J. Appl. Math. **49**, 906–931 (1989)
18. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
19. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings-a View from Variational Analysis. Springer Monographs in Mathematics, LLC, New York (2009)
20. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer, New York (2003)
21. Fan, J.Q., Li, R.Z.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
22. Fan, J.Q., Xue, L.Z., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**, 819–849 (2014)
23. Feng, M.B., Mitchell, J.E., Pang, J.S., Shen, X., Wächter, A.: Complementarity formulations of ℓ_0-norm optimization problems. Pac. J. Optim. **14**, 273–305 (2018)
24. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**, 17–40 (1976)
25. Gotoh, J.Y., Takeda, A., Tono, K.: DC formulations and algorithms for sparse optimization problems. Math. Program. **169**, 141–176 (2018)
26. Gu, Y.W., Fan, J., Kong, L.C., Ma, S.Q., Zou, H.: ADMM for high-dimensional sparse penalized quantile regression. Technometrics **60**, 319–331 (2018)
27. Hiriart-Urruty, J.B., Strodiot, J.J., Nguyen, V.H.: Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data. Appl. Math. Optim. **11**, 43–56 (1984)
28. Huang, J., Hom, H., Jiao, Y., Liu, Y., Lu, X.: A constructive approach to L_0 penalized regression. J. Mach. Learn. Res. **19**, 1–37 (2018)
29. Ioffe, A.D., Outrata, J.V.: On metric and calmness qualification conditions in subdifferential calculus. Set-Valued Anal. **16**, 199–227 (2008)
30. Le, H.Y.: Generalized subdifferentials of the rank function. Optim. Lett. **7**, 731–743 (2013)
31. Le Thi, H.A., Le, H.M., Pham Dinh, T.: Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm. Mach. Learn. **101**, 163–186 (2015)
32. Le Thi, H.A., Pham Dinh, T.: DC programming and DCA: thirty years of developments. Mathematical Programming B, Special Issue dedicated to: DC Programming-Theory, Algorithms and Applications **169**, 5-68 (2018)
33. Li, G.Y., Pong, T.K.: Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. **18**, 1199–1232 (2018)
34. Liu, Y.L., Bi, S.J., Pan, S.H.: Equivalent Lipschitz surrogates for zero-norm and rank optimization problems. J. Global Optim. **72**, 679–704 (2018)
35. Lu, Z.: Iterative hard thresholding methods for ℓ_0 regularized convex cone programming. Math. Program. **147**, 125–154 (2014)
36. Loh, P.L., Wainwright, M.J.: Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. J. Mach. Learn. Res. **16**, 559–616 (2015)
37. Mangasarian, O.L.: Machine learning via polyhedral concave minimization. In: Fischer, H., Riedmueller, B., Schaeffler, S. (eds.) Applied Mathematics and Parallel Computing-Festschrift for Klaus Ritter, pp. 175–188. Physica-Verlag, Heidelberg (1996)
38. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM J. Control. Optim. **15**, 959–972 (1977)
39. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Dokl. **27**, 372–376 (1983)

40. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York (1970)
41. Pan, S.H., Liu, Y.L.: *Subregularity of subdifferential mappings relative to the critical set and KL property of exponent 1/2*, arXiv:1812.00558v3
42. Pang, J.S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. Math. Oper. Res. **42**, 95–118 (2017)
43. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to DC programming: theory, algorithms and applications. Acta Math. Vietnamica **22**, 289–355 (1997)
44. Qi, L.Q., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**, 353–367 (1993)
45. Qian, Y.T., Pan, S.H., Liu, Y.L.: *Calmness of partial perturbation to composite rank constraint systems and its applications*, arXiv:2102.10373v2, October 8 (2021)
46. Raskutti, G., Wainwright, M.J.: Restricted eigenvalue properties for correlated Gaussian designs. J. Mach. Learn. Res. **11**, 2241–2259 (2010)
47. Rinaldi, F., Schoen, F., Sciandrone, M.: Concave programming for minimizing the zero-norm over polyhedral sets. Comput. Optim. Appl. **46**, 467–486 (2010)
48. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**, 97–116 (1976)
49. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
50. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Cham (1998)
51. Robinson, S.M.: Some continuity properties of polyhedral multifunctions. Math. Program. Study **14**, 206–214 (1981)
52. Soubies, E., Blang-Fraud, L., Aubert, G.: A unified view of exact continuous penalities for ℓ_2-ℓ_0 minimization. SIAM J. Optim. **8**, 1067–1639 (2017)
53. Tang, P.P., Wang, C.J., Sun, D.F., Toh, K.C.: A sparse semismooth Newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problems. J. Mach. Learn. Res. **21**, 1–38 (2020)
54. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B **58**, 267–288 (1996)
55. Wang, L., Wu, Y.C., Li, R.Z.: Quantile regression for analyzing heterogeneity in ultra high dimension. J. Am. Stat. Assoc. **107**, 214–222 (2012)
56. Wang, Y., Yin, W.T., Zeng, J.S.: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Sci. Comput. **78**, 29–63 (2019)
57. Weston, J., Elisseef, A., Schölkopf, B., Tipping, M.: Use of the zero norm with linear models and kernel methods. J. Mach. Learn. Res. **3**, 1439–1461 (2003)
58. Wen, B., Chen, X.J., Pong, T.K.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM J. Optim. **27**, 124–145 (2017)
59. Wright, J., Ma, Y.: Dense error correction via ℓ_1-mininization. IEEE Trans. Inf. Theory **56**, 3540–3560 (2010)
60. Wu, F., Bian, W.: Accelerated iterative hard thresholding algorithm for ℓ_0 regularized regression problem. J. Global Optim. **76**, 819–840 (2020)
61. Wu, F., Bian, W., Xue, X.P.: Smoothing fast iterative hard thresholding algorithm for ℓ_0 regularized nonsmooth convex regression problem, arXiv:2104.13107v1
62. Ye, J.J., Zhu, D.L.: Optimality conditions for bilevel programming problems. Optimization **33**, 9–27 (1995)
63. Ye, J.J., Zhu, D.L., Zhu, Q.J.: Exact penalization and necessary optimality conditions for generalized bilevel programming problems. SIAM J. Optim. **7**, 481–507 (1997)
64. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894–942 (2010)
65. Zhang, X., Zhang, X.Q.: A new proximal iterative hard thresholding method with extrapolation for ℓ_0 minimization. J. Sci. Comput. **79**, 809–826 (2019)
66. Zhao, X.Y., Sun, D.F., Toh, K.C.: A Newton-CG augmented Lagrangian method for semidefinite programming. SIAM J. Optim. **20**, 1737–1765 (2010)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.