A PRIMAL DUAL ACTIVE SET ALGORITHM FOR A CLASS OF NONCONVEX SPARSITY OPTIMIZATION

YULING JIAO*, BANGTI JIN†, XILIANG LU‡, AND WEINA REN§

Abstract. In this paper, we consider the problem of recovering a sparse vector from noisy measurement data. Traditionally, it is formulated as a penalized least-squares problem with an ℓ^1 penalty. Recent studies show that nonconvex penalties, e.g., ℓ^0 and bridge, allow more effective sparse recovery. We develop an algorithm of primal dual active set type for a class of nonconvex sparsity-promoting penalties, which cover ℓ^0, bridge, smoothly clipped absolute deviation, capped ℓ^1 and minimax concavity penalty. First we establish the existence of a global minimizer for the class of optimization problems. Then we derive a novel necessary optimality condition for the global minimizer using the associated thresholding operator. The solutions to the optimality system are coordinate-wise minimizers, and under minor conditions, they are also local minimizers. Upon introducing the dual variable, the active set can be determined from the primal and dual variables. This relation lends itself to an iterative algorithm of active set type which at each step involves updating the primal variable only on the active set and then updating the dual variable explicitly. When combined with a continuation strategy on the regularization parameter, the primal dual active set method has a global convergence property under the restricted isometry property. Extensive numerical experiments demonstrate its efficiency and accuracy.

Keywords: nonconvex penalty, sparsity, primal dual active set algorithm, continuation

1. Introduction.

In this paper, we develop a fast algorithm of primal dual active set (PDAS) type for a class of nonconvex optimization problems arising in sparse recovery. Sparse recovery has attracted considerable attention in signal processing, machine learning and statistics in recent years. In signal processing, especially compressive sensing, sparsity represents an important structural property that can be effectively utilized for data acquisition, signal transmission, storage and processing etc [7, 14]. In statistics, sparsity is one vital variable selection tool for constructing parsimonious models that admit easy interpretation [46]. Generally, the problem is formulated as

$$y = \Psi x + \eta, \tag{1.1}$$

where the vector $x \in \mathbb{R}^p$ denotes the signal to be recovered, the vector $\eta \in \mathbb{R}^n$ describes measurement errors, and the matrix $\Psi \in \mathbb{R}^{n\times p}$ models the system response mechanism. Throughout, we assume that the matrix Ψ has normalized column vectors $\{\psi_i\}$, i.e., $\|\psi_i\| = 1$, $i = 1, ..., p$, where $\|\cdot\|$ denotes the Euclidean norm of a vector. When $n \ll p$, problem (1.1) is underdetermined, and hence it is challenging to get a meaningful solution. The sparsity approach looks for a solution with many zero entries, and it opens a novel avenue for resolving the issue.

One popular approach to realize sparsity constraints is basis pursuit [10] or Lasso [46]. It leads to the following nonsmooth but convex optimization problem

$$\min_{x\in\mathbb{R}^p} \tfrac{1}{2}\|\Psi x - y\|^2 + \lambda\|x\|_1, \tag{1.2}$$

*School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, 430063, P.R. China. (yulingjiaomath@whu.edu.cn)

†Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK. (bangti.jin@gmail.com)

‡Corresponding author. School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China, and Computational Science Hubei Key Laboratory, Wuhan University, Wuhan, 430072, China. (xllv.math@whu.edu.cn)

§School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China. (winnerren@whu.edu.cn)

where $\|\cdot\|_1$ denotes the ℓ^1-norm of a vector, and $\lambda > 0$ is a regularization parameter. Since its introduction [10, 46], problem (1.2) has gained immense popularity, which largely can be attributed to the fact that (1.2) admits efficient numerical solution. The convexity of the problem allows designing fast and globally convergent minimization algorithms, e.g., gradient projection method and coordinate descent algorithm; see [48] for an overview. Theoretically, minimizers to (1.2) enjoy attractive statistical properties [54, 7, 39]. Under certain regularity conditions (e.g., restricted isometry property) on the matrix Ψ and the sparsity level of the true signal x, it can produce models with good estimation and prediction accuracy, and the support of the true signal can be identified correctly with a high probability [54].

However, the convex model (1.2) has several drawbacks: it requires more restrictive conditions on the matrix Ψ and more data in order to recover exactly the signal than nonconvex ones, e.g., bridge penalty [8, 18, 45]; and it tends to produce biased estimates for large coefficients [51], and hence lacks oracle property [15, 16]. To circumvent these drawbacks, nonconvex penalties have been proposed, including ℓ^0, bridge [19, 20], capped-ℓ^1 [53], smoothly clipped absolute deviation (SCAD) [15] and minimax concave penalty (MCP) [50] etc.

The nonconvex approach leads to the following optimization problem

$$\min_{x\in\mathbb{R}^p} J(x) = \tfrac{1}{2}\|\Psi x - y\|^2 + \sum_{i=1}^{p} \rho_{\lambda,\tau}(x_i), \tag{1.3}$$

where $\rho_{\lambda,\tau}$ is a nonconvex penalty, λ is a regularization parameter, and τ controls the concavity of the penalty (see Section 2.1 for details). The nonconvexity and nonsmoothess of the penalty $\rho_{\lambda,\tau}$ poses significant challenge in their analytical and numerical studies. Nonetheless, their attractive theoretical properties [52] have generated much interest in developing efficient and accurate algorithms. Below we first review existing algorithms for three popular nonconvex penalties, i.e., the ℓ^0, bridge, and SCAD penalty, and then discuss several general-purposed algorithms.

First, for the ℓ^0 penalty, iterative hard thresholding is very popular [33, 2]. The iterates generated by the algorithm are descent to the functional and converge to a local minimizer, with an asymptotic linear convergence, if Φ satisfies certain conditions [2]. Recently, Ito and Kunisch [31] developed a PDAS algorithm based on a novel necessary optimality condition and a reformulation using the complementarity function. It was derived from the augmented Lagrangian functional, and its convergence for certain diagonal-dominant operators was studied. We would also like to mention greedy methods, e.g., orthogonal matching pursuit [47] and CoSaMP [40]. These methods exploit the residual information to adaptively update the signal support, and each step involves one least-squares problem on the support.

Second, for the bridge penalty, one popular idea is to use the iteratively reweighted least-squares method together with suitable smoothing of the singularity at the origin [9, 35]. In [35, 36], the convergence to a critical point of the smoothed functional was established; see [55] for an alternative scheme and its convergence analysis. In [37], a unified convergence analysis was provided, and new variants were also developed. Each iterate of the method in [35] and [55] respectively requires solving a penalized least-squares problem and a weighted Lasso problem, and thus can be expensive. One can also employ the iterative thresholding algorithm, for which the iterates converge subsequentially, and the limit satisfies a necessary optimality condition [3].

Last, for the SCAD, Fan and Li [15] proposed to use a local quadratic approximation (LQA) to the nonconvex penalty and one single Newton step for optimizing the

resulting functional. Later a local linear approximation (LLA) was suggested [56] to replace the LQA, which leads to a one-step sparse estimator. For the closely related MCP, Zhang [50] developed an algorithm that keeps track of multiple local minima in order to select a solution with desirable statistical properties.

There are also several numerical algorithms that treat the model (1.3) in a unified manner. The first is based on majorization-minimization, where each step involves a reweighted ℓ^1 or ℓ^2 subproblem, and includes the LLA and LQA for the SCAD and multi-stage convex relaxation [53] for the smoothed bridge and the capped ℓ^1 penalty. The subproblems may be solved inexactly, e.g., with one or several gradient steps, in order to enhance the computational efficiency. Theoretically, the sequence of iterates is descent for the functional, but the convergence of the sequence itself is generally unclear. Numerically, the cost per iteration is that of the ℓ^2/ℓ^1 solver, and thus can be expensive. In [23], a general iterative shrinkage and thresholding algorithm was developed, and the convergence to a critical point was shown under a coercivity assumption on the functional. The bridge, SCAD, MCP, capped ℓ^1 and log-sum penalties were illustrated. The second is the coordinate descent algorithm, which at each step updates one component of the signal vector in either Jacobi [43] or Gauss-Seidel [4, 38] fashion for the SCAD and MCP; see also [42] for an algorithm based on smoothing and variable splitting for a class of nonconvex functionals (on the gradient). Theoretically, any cluster point of the iterates is a stationary point [49]. Numerical experiments [38] also verified its efficiency for such penalties. Third, in [21] an algorithm was proposed based on decomposing the nonconvex penalty into the difference of two convex functions and DC programming, and illustrated on the bridge, SCAD, and capped-ℓ^1 penalty. Like the first approach, each iteration involves a convex weighted LASSO problem. Last, Chen et al [12] (see also [11]) derived affine-scaled second-order necessary and sufficient conditions for local minimizers to model (1.3) in case of the bridge, SCAD and MCP, and developed a globally convergent smoothing trust-region Newton method. Meanwhile, in [27] a superlinearly convergent regularized Newton method was developed.

In this paper we develop an algorithm of PDAS type for problem (1.3). Our main contributions are as follows. First, we show the existence of a global minimizer to problem (1.3). The existence was only partially known in the general case. Second, we derive a necessary optimality condition of global minimizers to (1.3) using the associated thresholding operator, and establish that any solution to the necessary optimality condition is a coordinate-wise minimizer. Further we provide numerically verifiable sufficient conditions for a coordinate-wise to be a local minimizer. Upon introducing the dual variable, the necessary condition can be rewritten in terms of the primal and dual variables and the thresholding operator. Third, we develop a PDAS algorithm based on this relation. At each iteration, it first determines the active set from the primal and dual variables, then updates the primal variable on the active set and finally updates the dual variable explicitly. Each iteration involves only solving a standard least-squares subproblem on the active set (often of small size), which exhibits a local superlinear convergence, and thus it is very efficient when coupled with a continuation strategy. Last, we show the global convergence of the primal dual active set with continuation algorithm. Our algorithm is inspired by the interesting work [31], but different from it in several ways. In [31] only the ℓ^0 penalty was studied, while all five nonconvex penalties listed in Table 2.1 are investigated here. Further, we introduce a continuation strategy to enhance the computational efficiency of the algorithm and prove its global convergence.

The rest of the paper is organized as follows. In Section 2 we describe the nonconvex penalties and establish the existence of a global minimizer to (1.3). In Section 3 we first derive the thresholding operator for each penalty, and then use it in the necessary optimality condition, whose solutions are coordinate-wise minimizers to (1.3). Further, we give sufficient conditions for a coordinate-wise minimizer to be a local minimizer. In Section 4, by introducing a dual variable, we rewrite the necessary optimality condition and the active set using both primal and dual variables. Based on this fact, we develop a unified PDAS algorithm for all five nonconvex penalties. Further, we establish the global convergence of the algorithm when it is coupled with a continuation strategy. Finally, numerical results for several examples are presented in Section 5 to illustrate the efficiency and accuracy of the algorithm.

2. Problem formulation. Now we specify explicitly the nonconvex penalties of our interest, and discuss the existence of a global minimizer. The case of the ℓ^0 penalty was discussed in [41]. To the best of our knowledge, the existence in a general setting has not been studied for the SCAD, capped-ℓ^1 penalty and MCP earlier.

2.1. Nonconvex penalties. We focus on five commonly used nonconvex penalties for recovering sparse signals; see Table 2.1 for an overview. In Fig. 2.1, we show these penalties and their thresholding operators (cf. Section 3.1).

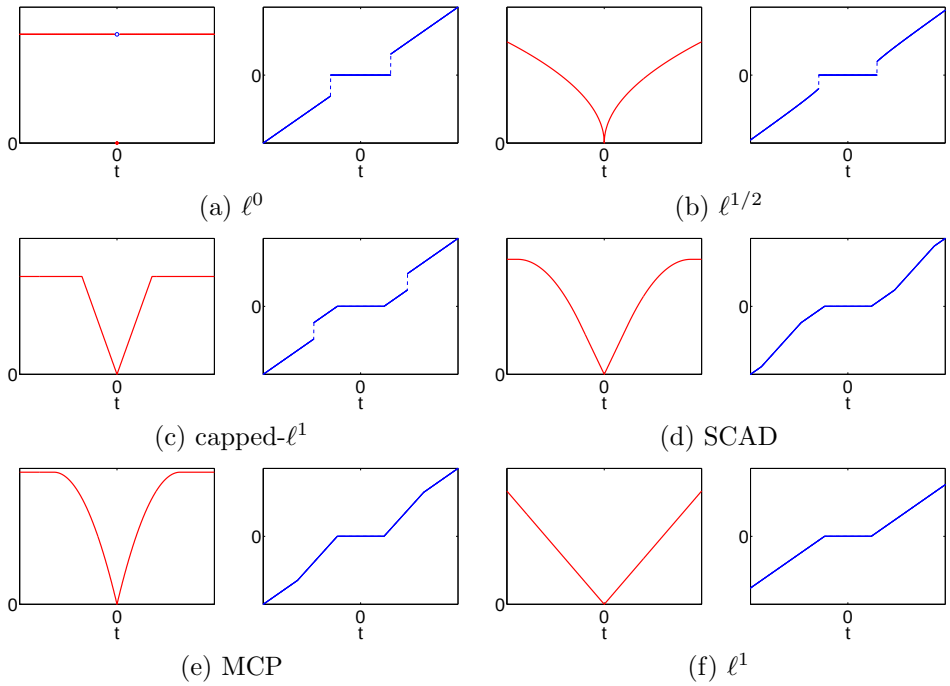


Fig. 2.1: Nonconvex penalties (left panel) and their thresholding operators (right panel). Here $\lambda = 1.2$, and $\tau = 0.5, 3.7, 2.7$ and 1.5 for the bridge, SCAD, MCP and capped-ℓ^1 penalty, respectively.

The ℓ^0-norm, denoted by $\|x\|_0$ of a vector x, is defined by $\|x\|_0 = |\{i : x_i \neq 0\}|$. It penalizes the number of nonzero components, which measures the model complexity.

Due to the discrete nature of the ℓ^0 penalty, the model (1.3) is combinatorial and hardly tractable in high-dimensional spaces. All other penalties in Table 2.1 can be regarded as approximations to the ℓ^0 penalty, and designed to alleviate its drawbacks, e.g., lack of stability [5] and computational challenges.

The bridge penalty was popularized by [19, 20]. The ℓ^τ-quasinorm $\|x\|_\tau$, $0 < \tau < 1$, of a vector x, defined by $\|x\|_\tau^\tau = \sum_i |x_i|^\tau$, is a quasi-smooth approximation of the ℓ^0 penalty as τ tends towards zero [31], and related statistical properties, e.g., variable selection and oracle property, were intensively studied [34, 28, 8, 18].

The SCAD [15, 16] was derived from the following qualitative requirements: it is singular at the origin to achieve sparsity and its derivative vanishes for large values so as to ensure unbiasedness. Specifically, for the SCAD, it is defined for $\tau > 2$ via

$$\rho_{\lambda,\tau}(t) = \lambda \int_0^{|t|} \min\left(1, \frac{\max(0, \lambda\tau - |s|)}{\lambda(\tau - 1)}\right) ds$$

and upon integration, we obtain the expression in Table 2.1. Further, variable selection consistency and asymptotic estimation efficiency were studied in [16].

The capped-ℓ^1 penalty [53] is a linear approximation of the SCAD penalty. Theoretically, it can be viewed as a variant of the two-stage optimization problem: one first solves a standard Lasso problem and then solves a Lasso problem where the large coefficients are not penalized any more, thus leading to an unbiased model. The condition $\tau > 1/2$ ensures the uniqueness of the thresholding operator [52].

Following the rationale of the SCAD, the MCP [50] is defined by

$$\rho_{\lambda,\tau}(t) = \lambda \int_0^{|t|} \max\left(0, 1 - |s|/(\lambda\tau)\right) ds.$$

The MCP minimizes the maximum concavity $\sup_{0<t_1<t_2}(\rho'_{\lambda,\tau}(t_1) - \rho'_{\lambda,\tau}(t_2))/(t_2 - t_1)$ subject to unbiasedness and feature selection constraints: $\rho'_{\lambda,\tau}(t) = 0$ for any $|t| \geq \lambda\tau$ and $\rho'_{\lambda,\tau}(0^\pm) = \pm\lambda$. Similar to the capped-$\ell^1$ penalty, the condition $\tau > 1$ ensures the wellposedness of the thresholding operator [50].

2.2. Existence of minimizers. Now we turn to the existence of a minimizer to the functional J defined in (1.3). First, we note that the ℓ^0 penalty is lower semi-continuous [31] and the rests are continuous. Hence, if the matrix Ψ is of full column rank, i.e., $\|x\| \to \infty \Rightarrow \|\Psi x\| \to \infty$, then the existence of a minimizer follows by a standard argument. However, in practice, Ψ may not have a full column rank.

First, we show one technical lemma.

LEMMA 2.1. *Let the function $\rho(t) : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ satisfy:*

(i) *ρ is even with $\rho(0) = 0$, nondecreasing for $t \geq 0$, and lower semi-continuous.*

(ii) *$\rho(t)$ is a constant when $|t| \geq t_0$ for some $t_0 > 0$.*

Then for any given subspace $\mathcal{N} \subset \mathbb{R}^p$, we have

(a) *For any $x \in \mathcal{N}^\perp$, there exists an element $z \in \mathcal{N}$ such that $\sum_{i=1}^p \rho(x_i + z_i) = \inf_{z\in\mathcal{N}} \sum_{i=1}^p \rho(x_i + z_i)$.*

(b) *Let $SE(x) = \arg\min_{z\in\mathcal{N}} \sum_{i=1}^p \rho(x_i + z_i)$. Then the function $h : \mathcal{N}^\perp \mapsto \mathbb{R}$, $h(x) = \inf_{z\in SE(x)} \|z\|$ maps a bounded set to a bounded set.*

Proof. To show part (a), let m be the dimension of $\mathcal{N}$, and $S \in \mathbb{R}^{p\times m}$ be a column orthonormal matrix whose columns form a basis of $\mathcal{N}$. We denote the rows of S by $\{\tilde{s}_i\}_{i=1}^p$. Then any $z \in \mathcal{N}$ can be written as $z = Sw$ for some $w \in \mathbb{R}^m$. Let $\{w_k\} \subset \mathbb{R}^m$ be a minimizing sequence to $\inf_{z\in\mathcal{N}} \sum_{i=1}^p \rho(x_i + z_i)$ under the representation $z = Sw$.

Table 2.1: Nonconvex penalty functions $\rho_{\lambda,\tau}(t)$ and the thresholding operators $S^{\rho}_{\lambda,\tau}(v)$. In the Table, (t^*, T^*) and $G(v)$ are given in Theorem 3.2 and the proof of Theorem 3.4, cf. Appendix B.

penalty	$\rho_{\lambda,\tau}(t)$	$S^{\rho}_{\lambda,\tau}(v)$
Lasso [46, 10]	$\lambda\|t\|$	$\operatorname{sgn}(v)\max(\|v\|-\lambda, 0)$
ℓ^0 [1]	$\begin{cases} \lambda & t \neq 0 \\ 0 & t = 0 \end{cases}$	$\begin{cases} 0 & \|v\| < \sqrt{2\lambda} \\ \{0, \operatorname{sgn}(v)\sqrt{2\lambda}\} & \|v\| = \sqrt{2\lambda} \\ v & \|v\| > \sqrt{2\lambda} \end{cases}$
bridge, $0 < \tau < 1$ [19, 20]	$\lambda\|t\|^{\tau}$	$\begin{cases} 0 & \|v\| < T^* \\ \{0, \operatorname{sgn}(v)t^*\} & \|v\| = T^* \\ \operatorname{argmin}_{u>0} G(u) & v > T^* \\ -S^{\ell^{\tau}}_{\lambda,\tau}(-v) & v < -T^* \end{cases}$
capped-ℓ^1, $\tau > \frac{1}{2}$ [53]	$\begin{cases} \lambda^2\tau & \|t\| > \lambda\tau \\ \lambda\|t\| & \|t\| \leq \lambda\tau \end{cases}$	$\begin{cases} 0 & \|v\| \leq \lambda \\ \operatorname{sgn}(v)(\|v\|-\lambda) & \lambda < \|v\| < \lambda(\tau+\frac{1}{2}) \\ \operatorname{sgn}(v)(\lambda\tau \pm \frac{\lambda}{2}) & \|v\| = \lambda(\tau+\frac{1}{2}) \\ v & \|v\| > \lambda(\tau+\frac{1}{2}) \end{cases}$
SCAD, $\tau > 2$ [15]	$\begin{cases} \frac{\lambda^2(\tau+1)}{2} & \|t\| > \lambda\tau \\ \frac{\lambda\tau\|t\| - \frac{1}{2}(t^2+\lambda^2)}{\tau-1} & \lambda < \|t\| \leq \lambda\tau \\ \lambda\|t\| & \|t\| \leq \lambda \end{cases}$	$\begin{cases} 0 & \|v\| \leq \lambda \\ \operatorname{sgn}(v)(\|v\|-\lambda) & \lambda < \|v\| \leq 2\lambda \\ \operatorname{sgn}(v)\frac{(\tau-1)\|v\|-\lambda\tau}{\tau-2} & 2\lambda < \|v\| \leq \lambda\tau \\ v & \|v\| > \lambda\tau \end{cases}$
MCP, $\tau > 1$ [50]	$\begin{cases} \lambda(\|t\| - \frac{t^2}{2\lambda\tau}) & \|t\| < \tau\lambda \\ \frac{\lambda^2\tau}{2} & \|t\| \geq \tau\lambda \end{cases}$	$\begin{cases} 0 & \|v\| \leq \lambda \\ \operatorname{sgn}(v)\frac{\tau(\|v\|-\lambda)}{\tau-1} & \lambda \leq \|v\| \leq \lambda\tau \\ v & \|v\| \geq \lambda\tau \end{cases}$

We claim that there exists a $w \in \mathbb{R}^m$ such that

$$\sum_{i=1}^{p} \rho(x_i + (Sw)_i) \leq \lim_{k\to\infty} \sum_{i=1}^{p} \rho(x_i + (Sw_k)_i).$$

First, if there is a bounded subsequence of $\{w_k\}$, the existence of a minimizer follows directly from the lower semicontinuity of ρ. Hence we assume that $\|w_k\| \to \infty$ as $k \to \infty$. Then we check the scalar sequences $\{w_k \cdot \tilde{s}_i\}_k$, $i \in \mathbb{I} \equiv \{1, ..., p\}$. For any $i \in \mathbb{I}$, if there is a bounded subsequence of $\{w_k \cdot \tilde{s}_i\}$, we may pass to a convergent subsequence and let it be the whole sequence. In this way, we divide the index set $\mathbb{I}$ into two disjoint subsets $\mathbb{I}'$ and $\mathbb{I}''$ such that for every $i \in \mathbb{I}'$, $w_k \cdot \tilde{s}_i$ converges, whereas for every $i \in \mathbb{I}''$, $|w_k \cdot \tilde{s}_i| \to \infty$. Let $\mathcal{L} = \operatorname{span}\{\tilde{s}_i\}_{i\in\mathbb{I}'}$, and decompose w_k into $w_k = w_{k,\mathcal{L}} + w_{k,\mathcal{L}^\perp}$, with $w_{k,\mathcal{L}} \in \mathcal{L}$ and $w_{k,\mathcal{L}^\perp} \in \mathcal{L}^\perp$.

If the index set $\mathbb{I}'$ is empty, then by the monotonicity of the function $\rho(t)$, one can verify directly that the zero vector 0 is a minimizer. Otherwise, by the definition of $\mathbb{I}'$ and the monotonicity of $\rho(t)$, for any $i \in \mathbb{I}'$, $w_k \cdot \tilde{s}_i = w_{k,\mathcal{L}} \cdot \tilde{s}_i$, and for any $i \in \mathbb{I}''$, $\limsup_{k\to\infty} \rho(x_i + w_{k,\mathcal{L}} \cdot \tilde{s}_i) \leq \lim_{k\to\infty} \rho(x_i + w_k \cdot \tilde{s}_i)$. Hence $\{w_{k,\mathcal{L}}\}$ is also a minimizing sequence. Next we prove that $\{w_{k,\mathcal{L}}\}$ is bounded. To this end, we let M be the submatrix of S, consisting of rows whose indices are listed in $\mathbb{I}'$. It follows from the definition of $w_{k,\mathcal{L}}$ and the convergence of $w_{k,\mathcal{L}} \cdot \tilde{s}_i$ for $i \in \mathbb{I}'$ that

$w_{k,\mathcal{L}} \in \mathcal{L} = \mathrm{Range}(M^t)$ and $Mw_{k,\mathcal{L}}$ is bounded. Now let $M = U^t \Sigma V$ be the singular value decomposition of M. Then for any $w \in \mathrm{Range}(M^t)$, i.e., $w = M^t q$, there holds

$$\|Mw\|^2 = \|\Sigma\Sigma^t Uq\|^2 = \sum_{\sigma_i>0} \sigma_i^4 (Uq)_i^2 \geq \sigma_M^2 \sum_{\sigma_i>0} \sigma_i^2 (Uq)_i^2 = \sigma_M^2 \|\Sigma^t Uq\|^2 = \sigma_M^2 \|w\|^2,$$

where σ_M is the smallest nonzero singular value of M. Hence the sequence $\{w_{k,\mathcal{L}}\}$ is bounded, from which the existence of a minimizer follows. This shows part (a).

By the construction in part (a), the set $SE(x)$ is nonempty, and by the lower semi-continuity of ρ, it is closed. Hence, there exists an element $z(x) \in SE(x)$ such that $\|z(x)\| = \inf_{z \in SE(x)} \|z\|$. We claim that the map $x \mapsto z(x)$ is bounded. To this end, let $D = \|x\|_\infty$, and recall the representation $z = Sw$ and $(x+z)_i = x_i + w \cdot \tilde{s}_i$. Denote by $\mathbb{I}' = \{i \in \mathbb{I} : |w \cdot \tilde{s}_i| \leq D + t_0\}$, and let $\mathcal{L} = \mathrm{span}\{\tilde{s}_i\}_{i \in \mathbb{I}'}$, $w = w_{\mathcal{L}} + w_{\mathcal{L}^\perp}$, with $w_{\mathcal{L}} \in \mathcal{L}$ and $w_{\mathcal{L}^\perp} \in \mathcal{L}^\perp$. Then the argument in part (a) yields

$$\left.\begin{array}{ll} \tilde{s}_i \cdot w_{\mathcal{L}} = \tilde{s}_i \cdot w & i \in \mathbb{I}' \\ \rho(x_i + \tilde{s}_i \cdot w_{\mathcal{L}}) \leq \rho(t_0) = \rho(x_i + \tilde{s}_i \cdot w) & i \in \mathbb{I} \backslash \mathbb{I}' \end{array}\right\} \Rightarrow \tilde{z}(x) = Sw_{\mathcal{L}}(x) \in SE(x),$$

and

$$\|Sw_{\mathcal{L}}(x)\| \leq C_{\mathbb{I}'}(D + t_0),$$

where the constant $C_{\mathbb{I}'}$ depends only on the smallest nonzero singular value of the submatrix whose rows are given by $\tilde{s}_i$, $i \in \mathbb{I}'$. Therefore,

$$\begin{aligned} \sup_{\|x\|_\infty \leq D} \inf_{z \in SE(x)} \|z\| &= \sup_{\|x\|_\infty \leq D} \|z(x)\| \leq \sup_{\|x\|_\infty \leq D} \|\tilde{z}(x)\| \\ &= \sup_{\|x\|_\infty \leq D} \|Sw_{\mathcal{L}}(x)\| \leq \sup_{\mathbb{I}'} C_{\mathbb{I}'}(D + t_0). \end{aligned}$$

The factor $\sup_{\mathbb{I}'} C_{\mathbb{I}'}$ is over finitely many numbers, which concludes the proof. □

Now we can show the existence of a minimizer to problem (1.3).

THEOREM 2.2. *For any of the five nonconvex penalties in Table 2.1, there exists at least one minimizer to problem* (1.3).

Proof. We discuss the cases separately.

(i) bridge. The proof is straightforward due to the coercivity of the penalty.

(ii) ℓ^0, capped-ℓ^1, SCAD and MCP. First, all these penalties satisfy the assumptions in Lemma 2.1. Let $\mathcal{N} = \mathrm{Ker}(\Psi)$, then Ψ is coercive over $\mathcal{N}^\perp$, and $\mathrm{Ker}(\Psi)^\perp = \mathrm{Range}(\Psi^t)$. Since the functional J is bounded from below by zero, the infimum $\mathrm{INF} = \inf J(x)$ exists and it is finite; further, by the very definition of the infimum INF, there exists a minimizing sequence, denoted by $\{x^k\} \subset \mathbb{R}^p$, to (1.3), i.e., $\lim_{k\to\infty} J(x^k) = \mathrm{INF}$ [22, Section 39, pp. 193]. We decomposed x^k into $x^k = P_{\mathcal{N}} x^k + P_{\mathcal{N}^\perp} x^k =: u^k + v^k$, where $P_{\mathcal{N}}$ and $P_{\mathcal{N}^\perp}$ denote the orthogonal projection into $\mathcal{N}$ and $\mathcal{N}^\perp$, respectively. By the construction of the set $SE(v^k)$ in the proof of Lemma 2.1, with the minimum-norm element $\tilde{u}^k \in SE(v^k)$ in place of u^k, the sequence $\{v^k + \tilde{u}^k\}$ is still minimizing. By the coercivity, $\{v^k\}$ is bounded, and hence $\{\tilde{u}^k\}$ is also bounded by Lemma 2.1(b). Upon passage to a convergent subsequence, the lower semi-continuity of J implies the existence of a minimizer. □

3. Necessary optimality condition for minimizers. Now we derive the necessary optimality condition for global minimizers to (1.3), which also forms the basis for the PDAS algorithm in Section 4. We shall show that the solutions to the necessary optimality condition are coordinate-wise minimizers, and provide verifiable sufficient conditions for a coordinate-wise minimizer to be a local minimizer.

3.1. Thresholding operators. First we derive thresholding operators for the penalties in Table 2.1. The thresholding operator forms the basis of many existing algorithms, e.g., coordinate descent and iterative thresholding, and thus unsurprisingly the expressions in Table 2.1 were derived earlier (see e.g. [43, 38, 4, 31, 23]), but in a different manner. We shall provide a unified derivation of the thresholding operator. To this end, for any penalty $\rho(t)$ in Table 2.1 (with the subscripts being omitted for simplicity), we define a function $g(t) : [0, \infty) \to \mathbb{R}^+ \cup \{0\}$ by

$$g(t) = \begin{cases} \frac{t}{2} + \frac{\rho(t)}{t}, & t \neq 0, \\ \liminf_{t \to 0^+} g(t), & t = 0. \end{cases}$$

LEMMA 3.1. *The value $T^* = \inf_{t>0} g(t)$ is attained at some point $t^* \geq 0$.*

Proof. By the definition of the function $g(t)$, it is continuous over $(0, +\infty)$, and approaches infinity as $t \to +\infty$. Hence any minimizing sequence $\{t_n\}$ is bounded. If it contains a positive accumulation point t^*, then $g(t^*) = T^*$ by the continuity of g. Otherwise it has only an accumulation point 0. However, by the definition of $g(0)$, $g(0) = T^*$ and hence $t^* = 0$. □

The explicit expressions of the tuple (t^*, T^*) for the penalties in Table 2.1 are given below; see Appendix A for the proof.

THEOREM 3.2. *For the five nonconvex penalties in Table 2.1, there holds*

$$(t^*, T^*) = \begin{cases} (\sqrt{2\lambda}, \sqrt{2\lambda}), & \ell^0, \\ ((2\lambda(1-\tau))^{\frac{1}{2-\tau}}, (2-\tau)\left[2(1-\tau)\right]^{\frac{\tau-1}{2-\tau}} \lambda^{\frac{1}{2-\tau}}), & \ell^\tau, \\ (0, \lambda), & \text{capped-}\ell^1,\ \text{SCAD},\ \text{MCP}. \end{cases}$$

Next we introduce the thresholding operator S^ρ defined by

$$S^\rho(v) = \operatorname*{argmin}_{u \in \mathbb{R}} \left((u - v)^2/2 + \rho(u) \right), \tag{3.1}$$

which can be set-valued. First we give a useful characterization of S^ρ.

LEMMA 3.3. *Let $u^* \in \arg\min_{u \in \mathbb{R}} \left((u - v)^2/2 + \rho(u) \right)$. Then the following three statements hold:* (a) $u^* = 0 \Rightarrow |v| \leq T^*$*;* (b) $|v| < T^* \Rightarrow u^* = 0$*; and* (c) $|v| = T^* \Rightarrow u^* = 0$ *or* $g(u^*) = \operatorname{sgn}(v) T^*$.

Proof. By the lower-semicontinuity and coercivity of the function $(u - v)^2/2 + \rho(u)$, it has at least one minimizer. Next one observes that

$$u^* \in \operatorname*{argmin}_{u \in \mathbb{R}} \left((u - v)^2/2 + \rho(u) \right) \quad \Leftrightarrow \quad u^* \in \operatorname*{argmin}_{u \in \mathbb{R}} \left(u^2/2 - uv + \rho(u) \right).$$

First, if $u^* = 0$, then for any $u \neq 0$, $u^2/2 - uv + \rho(u) = u(g(u) - v)$, which implies that u and $g(u) - v$ have the same sign. That is,

$$u > 0 \Rightarrow g(u) - v \geq 0, \quad \forall u > 0, \quad \text{then } v \leq \inf_{u>0} g(u) = T^*,$$

and

$$u < 0 \Rightarrow g(u) - v \leq 0, \quad \forall u < 0, \quad \text{then } -v \leq \inf_{u<0} -g(u) = \inf_{u<0} g(-u) = T^*.$$

From this it follows that $|v| \leq T^*$. This shows assertion (a). Second, let $G(u) = u(g(u) - v)$ for $u \neq 0$ and $G(0) = 0$. For $|v| < T^*$, since

$$u > 0 \Rightarrow g(u) \geq T^* > v \quad \text{and} \quad u < 0 \Rightarrow g(u) = -g(-u) \leq -T^* < v,$$

then $G(u) > 0$ when $u \neq 0$, which implies 0 is the only minimizer. This shows (b). Last, for $|v| = T^*$, by arguing analogously to (b) for $u > 0$ and $u < 0$, we have $G(u) \geq 0$. Then u^* satisfies that $G(u^*) = 0$, i.e., $u^* = 0$ or $g(u^*) = \text{sgn}(v)T^*$. □

REMARK 3.1. *If the minimizer t^* to $g(t)$ is unique, then assertion (c) of Lemma 3.3 can be replaced by $|v| = T^* \Rightarrow u^* = 0$ or $u^* = \text{sgn}(v)t^*$.*

Now we can derive an explicit expression of the thresholding operator S^ρ, which is summarized in Table 2.1 and given by Theorem 3.4 below. The proof is elementary but lengthy, and thus deferred to Appendix B.

THEOREM 3.4. *The thresholding operators S^ρ associated with the five nonconvex penalties (ℓ^0, bridge, capped-ℓ^1, SCAD and MCP) are as given in Table 2.1.*

REMARK 3.2. *The thresholding operator S^ρ is singled-valued, except at $v = T^*$ for the ℓ^τ, $0 \leq \tau < 1$, penalty, and at $v = \lambda(\tau + \frac{1}{2})$ for the capped-ℓ^1 penalty.*

3.2. Necessary optimality condition. Now we derive the necessary optimality condition for a global minimizer to (1.3) using the thresholding operator S^ρ. To this end, we first recall the concept of coordinate-wise minimizers. Following [49], a vector $x^* = (x_1^*, x_2^*, \ldots, x_p^*) \in \mathbb{R}^p$ is called a coordinate-wise minimizer of the functional $J(x)$ if it is the minimum along each coordinate direction, i.e.,

$$x_i^* \in \arg\min_{t \in \mathbb{R}} J(x_1^*, ..., x_{i-1}^*, t, x_{i+1}^*, ..., x_p^*). \tag{3.2}$$

Next we derive the sufficient and necessary optimality condition for a coordinate-wise minimizer x^* of problem (1.3). By the definition of x^*, there holds

$$\begin{aligned}
& x_i^* \in \operatorname*{argmin}_{t \in \mathbb{R}} J(x_1^*, ..., x_{i-1}^*, t, x_{i+1}^*, ..., x_p^*) \\
\Leftrightarrow \quad & x_i^* \in \operatorname*{argmin}_{t \in \mathbb{R}} \tfrac{1}{2}\|\Psi x^* - y + (t - x_i^*)\psi_i\|^2 + \rho_{\lambda,\tau}(t) \\
\Leftrightarrow \quad & x_i^* \in \operatorname*{argmin}_{t \in \mathbb{R}} \tfrac{1}{2}(t - x_i^*)^2 + (t - x_i^*)\psi_i^t(\Psi x^* - y) + \rho_{\lambda,\tau}(t) \\
\Leftrightarrow \quad & x_i^* \in \operatorname*{argmin}_{t \in \mathbb{R}} \tfrac{1}{2}(t - x_i^* - \psi_i^t(y - \Psi x^*))^2 + \rho_{\lambda,\tau}(t).
\end{aligned}$$

By introducing the dual variable $d_i^* = \psi_i^t(y - \Psi x^*)$ and recalling the definition of the thresholding operator $S_{\lambda,\tau}^\rho(t)$ for $\rho_{\lambda,\tau}$, we have the following characterization of x^*, which is also the necessary optimality condition of a global minimizer.

LEMMA 3.5. *An element $x^* \in \mathbb{R}^p$ is a coordinate-wise minimizer to problem (1.3) if and only if*

$$x_i^* \in S_{\lambda,\tau}^\rho(x_i^* + d_i^*) \quad \text{for } i = 1, ..., p, \tag{3.3}$$

where the dual variable d^* is defined by $d^* = \Psi^t(y - \Psi x^*)$.

REMARK 3.3. *By the expression of the thresholding operators in Table 2.1, and Remark 3.2, only in the case of $|x_i^* + d_i^*| = T^*$ for the bridge and ℓ^0 penalties, and $|x_i^* + d_i^*| = \lambda(\tau + \frac{1}{2})$ for the capped-ℓ^1 penalty, the value of x_i^* is not uniquely determined.*

The optimality condition (3.3) forms the basis of the PDAS algorithm in Section 4. Hence, the "optimal solution" by the algorithm can at best solve the necessary condition, and it is necessary to study more precisely the meaning of "optimality". First we recall a well-known result. By [49, Lemma 3.1], a coordinate-wise minimizer x^* is a stationary point in the sense that

$$\liminf_{t \to 0^+} \frac{J(x^* + td) - J(x^*)}{t} \geq 0, \quad \forall d \in \mathbb{R}^p. \tag{3.4}$$

In general, a coordinate-wise minimizer x^* is not necessarily a local minimizer, i.e., $J(x^* + \omega) \geq J(x^*)$ for all small $\omega \in \mathbb{R}^p$. Below we provide sufficient conditions for a coordinate-wise minimizer to be a local minimizer. To this end, we denote by $\mathcal{A} = \{i : \ x_i^* \neq 0\}$ and $\mathcal{I} = \mathcal{A}^c$ the active and inactive sets, respectively, of a coordinate-wise minimizer x^*. Throughout, for any $\mathcal{A} \subset \mathbb{I} = \{1, 2, ..., p\}$, we use the notation $x_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ (or $\Psi_{\mathcal{A}} \in \mathbb{R}^{n \times |\mathcal{A}|}$) for the subvector of x (or the submatrix of Ψ) consisting of entries (or columns) whose indices are listed in $\mathcal{A}$.

For any $\mathcal{A} \subset \mathbb{I}$, let $\sigma(\mathcal{A})$ be the smallest singular value of matrix $\Psi_{\mathcal{A}}^t \Psi_{\mathcal{A}}$. Then

$$\|\Psi_{\mathcal{A}} x_{\mathcal{A}}\|^2 \geq \sigma(\mathcal{A}) \|x_{\mathcal{A}}\|^2. \tag{3.5}$$

The sufficient conditions are summarized in Theorem 3.6 below. The proof is lengthy and technical, and hence deferred to Appendix C. Under the prescribed conditions, the solution generated by the PDAS algorithm, if it does converge, is a local minimizer.

THEOREM 3.6. *Let x^* be a coordinate-wise minimizer to* (1.3)*, and $\mathcal{A} = \{i : x_i^* \neq 0\}$ and $\mathcal{I} = \mathcal{A}^c$ be the active and inactive sets, respectively. Then there hold:*

(i) *ℓ^0: x^* is a local minimizer.*

(ii) *bridge: If $\sigma(\mathcal{A}) > \frac{\tau}{2}$ in* (3.5)*, then x^* is a local minimizer.*

(iii) *capped-ℓ^1: If $\{i : \ |x_i^*| = \lambda\tau\} = \emptyset$, then x^* is a local minimizer.*

(iv) *SCAD: If $\sigma(\mathcal{A}) > \frac{1}{\tau - 1}$ in* (3.5) *and $\|d_{\mathcal{I}}^*\|_\infty < \lambda$, then x^* is a local minimizer.*

(v) *MCP: If $\sigma(\mathcal{A}) > \frac{1}{\tau}$ in* (3.5) *and $\|d_{\mathcal{I}}^*\|_\infty < \lambda$, then x^* is a local minimizer.*

We have the following comments on Theorem 3.6.

REMARK 3.4. *For the ℓ^0 penalty, a coordinate-wise minimizer is always a local minimizer. For the capped-ℓ^1 penalty, the sufficient condition $\{i : \ |x_i^*| = \tau\lambda\} = \emptyset$ is related to the nondifferentiability of $\rho_{\lambda,\tau}^{c\ell^1}(t)$ at $t = \tau\lambda$. For the bridge, SCAD and MCP, the condition* (3.5) *is essential for a coordinate-wise minimizer to be a local minimizer, which requires that the size of the active set be not large. The condition $\|d_{\mathcal{I}}^*\|_\infty < \lambda$ is closely related to the uniqueness of the global minimizer. If both Ψ and η are random Gaussian, it holds except a null measure set [50].*

REMARK 3.5. *The conditions in Theorem 3.6 involve only the computed solution and the given parameters λ and τ, and can be numerically verified, which enables one to check a posteriori whether a coordinatewise minimizer is a local one.*

4. Primal dual active set algorithm. In this section, we propose an algorithm of PDAS type for the nonconvex penalties listed in Table 2.1, and discuss its efficient implementation via a continuation strategy and its global convergence.

4.1. PDAS algorithm. For the convex model (1.2), a PDAS algorithm (or equivalently, semismooth Newton method) has been applied in [24, 30]. Due to the local superlinear convergence of the semismooth Newton method, it is very efficient, when coupled with a continuation strategy [17]. For the ℓ^0 penalty (with a slightly different problem setting), a PDAS algorithm was also proposed in [31], based on an augmented Lagrangian reformulation of the necessary optimality condition. The key idea for such algorithms is to define an active set by both primal and dual variables, then update the primal variable on the active set only (and set its value to zero on the inactive set), and finally update the dual variable. Hence, there are two key ingredients in constructing a PDAS algorithm:

(i) to characterize the active set $\mathcal{A}$ by x^* and d^*;

(ii) to derive an explicit expression for the dual variable d^* on $\mathcal{A}$.

We focus on the optimality condition (3.3), i.e., a coordinate-wise minimizer x^*. Recall that the active set $\mathcal{A}$ defined in Section 3 is its support, i.e., $\mathcal{A} = \{i : x_i^* \neq 0\}$. To see

Table 4.1: Explicit expression of the dual variable $d^*_{\mathcal{A}}$ on the active set $\mathcal{A} = \{i : x^*_i \neq 0\}$, and its approximation $p_{\mathcal{A}_k}$ on $\mathcal{A}_k = \{i : |s^{k-1}_i| > T^*\}$, with $s_i = d_i + x_i$.

penalty	$d^*_{\mathcal{A}}$
ℓ^0	0
ℓ^τ	$\lambda\tau \frac{\lvert x^*_i\rvert^\tau}{x^*_i}$
capped-ℓ^1	$\begin{cases} 0 & \text{if } \lvert s^*_i\rvert > \lambda(\tau+\frac{1}{2}) \\ \operatorname{sgn}(s^*_i)\lambda & \text{if } \lambda < \lvert s^*_i\rvert < \lambda(\tau+\frac{1}{2}) \\ \{0, \operatorname{sgn}(s^*_i)\lambda\} & \text{if } \lvert s^*_i\rvert = \lambda(\tau+\frac{1}{2}) \end{cases}$
SCAD	$\begin{cases} 0 & \text{if } \lvert s^*_i\rvert \geq \lambda\tau \\ \frac{1}{\tau-1}(\operatorname{sgn}(s^*_i)\lambda\tau - x^*_i) & \text{if } \lambda\tau > \lvert s^*_i\rvert > 2\lambda \\ \operatorname{sgn}(s^*_i)\lambda & \text{if } 2\lambda \geq \lvert s^*_i\rvert > \lambda \end{cases}$
MCP	$\begin{cases} 0 & \text{if } \lvert s^*_i\rvert \geq \lambda\tau \\ \frac{1}{\tau}(\operatorname{sgn}(s^*_i)\lambda\tau - x^*_i) & \text{if } \lambda < \lvert s^*_i\rvert < \lambda\tau \end{cases}$
penalty	$p_{\mathcal{A}_k}$
ℓ^0	0
ℓ^τ	$\begin{cases} 0 & \lvert s^{k-1}_i\rvert < t^* \\ \lambda\tau \frac{\lvert x^{k-1}_i\rvert^\tau}{x^{k-1}_i} & \lvert s^{k-1}_i\rvert \geq t^* \end{cases}, \; t^* = (2\lambda(1-\tau))^{\frac{1}{2-\tau}}$
capped-ℓ^1	$\begin{cases} 0 & \text{if } \lvert s^{k-1}_i\rvert \geq \lambda(\tau+\frac{1}{2}) \\ \operatorname{sgn}(s^{k-1}_i)\lambda & \text{if } \lambda < \lvert s^{k-1}_i\rvert < \lambda(\tau+\frac{1}{2}) \end{cases}$
SCAD	$\begin{cases} \frac{1}{\tau-1}(\operatorname{sgn}(s^{k-1}_i)\lambda\tau - x^{k-1}_i) & \text{if } \lambda\tau > \lvert s^{k-1}_i\rvert > 2\lambda \text{ and } x^{k-1}_i \cdot d^{k-1}_i \geq 0 \\ \operatorname{sgn}(s^{k-1}_i)\lambda & \text{if } 2\lambda \geq \lvert s^{k-1}_i\rvert > \lambda \\ 0 & \text{otherwise} \end{cases}$
MCP	$\begin{cases} \frac{1}{\tau}(\operatorname{sgn}(s^{k-1}_i)\lambda\tau - x^{k-1}_i) & \text{if } \lambda < \lvert s^{k-1}_i\rvert < \lambda\tau \text{ and } x^{k-1}_i \cdot d^{k-1}_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$

(i), by Lemma 3.5 and the property of the operator S^ρ in Lemma 3.3, one observes

- for capped-ℓ^1, SCAD and MCP penalties, $|x^*_i + d^*_i| > T^* \Leftrightarrow x^*_i \neq 0$,
- for ℓ^τ penalty, $0 \leq \tau < 1$, $\begin{cases} |x^*_i + d^*_i| > T^* \Rightarrow x^*_i \neq 0, \\ |x^*_i + d^*_i| < T^* \Rightarrow x^*_i = 0, \\ |x^*_i + d^*_i| = T^* \Rightarrow x^*_i = 0 \text{ or } t^*. \end{cases}$

Hence, except the case $|x^*_i + d^*_i| = T^*$ for the ℓ^0 and bridge penalty, the active set $\mathcal{A}$ can be determined by the primal and dual variables together. Next we derive explicitly the dual variable d^* on $\mathcal{A}$. Straightforward computations show the formulas in Table 4.1; see Appendix D for details. We summarize this in a proposition.

PROPOSITION 4.1. *Let x^* and d^* be a coordinate-wise minimizer and the respective dual variable, $\mathcal{A}$ be the active set, and let*

$$\widetilde{\mathcal{A}} = \begin{cases} \{i : \; |x^*_i + d^*_i| = T^*\}, & \ell^0, \text{ bridge}, \\ \left\{i : |x^*_i + d^*_i| = \lambda(\tau+\frac{1}{2})\right\}, & \text{capped-}\ell^1, \\ \emptyset, & \text{SCAD, MCP}. \end{cases}$$

If the set $\widetilde{\mathcal{A}} = \emptyset$, then (i) *$\mathcal{A}$ can be characterized by $\{i : |x_i^* + d_i^*| > T^*\}$, and* (ii) *the dual variable d^* on $\mathcal{A}$ can be uniquely written as in Table 4.1.*

REMARK 4.1. *The set $\widetilde{\mathcal{A}}$ is empty for the SCAD and MCP. For the ℓ^0, bridge and capped-ℓ^1 penalty, it is likely empty, which however cannot be a priori ensured.*

Now we can derive a unified PDAS algorithm. On the active set $\mathcal{A}$, the dual variable d^* has two equivalent expressions, i.e., the defining relation

$$\Psi_{\mathcal{A}}^t(y - \Psi_{\mathcal{A}} x_{\mathcal{A}}^*) = d_{\mathcal{A}}^*,$$

and the expression $d_{\mathcal{A}}^* = d_{\mathcal{A}}(x^*, d^*)$ from Proposition 4.1(ii). This is the starting point for our algorithm. Similar to the case of convex optimization problems [26, 24], at each iteration, first we approximate the active set $\mathcal{A}$ and the inactive set $\mathcal{I}$ by $\mathcal{A}_k$ and $\mathcal{I}_k$ respectively defined by

$$\mathcal{A}_k = \{i : |x_i^{k-1} + d_i^{k-1}| > T^*\} \quad \text{and} \quad \mathcal{I}_k = \mathcal{A}_k^c.$$

Then we update the primal variable x^k on the active set $\mathcal{A}_k$ by

$$\Psi_{\mathcal{A}_k}^t(y - \Psi_{\mathcal{A}_k} x_{\mathcal{A}_k}^k) = p_{\mathcal{A}_k}, \tag{4.1}$$

where $p_{\mathcal{A}_k}$ is a suitable approximation of the dual variable d^* on the active set $\mathcal{A}_k$, and set x^k to zero on the inactive set $\mathcal{I}_k$. Finally we update the dual variable d^k by

$$d^k = \Psi^t(y - \Psi x^k).$$

REMARK 4.2. *The choice of the approximate dual variable $p_{\mathcal{A}_k}$ is related to the expression of the dual variable $d_{\mathcal{A}}^*$, cf. Proposition 4.1. For example, a natural choice of $p_{\mathcal{A}_k}$ for the bridge is given by $p_i = \lambda\tau|x_i^k|^\tau/x_i^k$ for $i \in \mathcal{A}_k$. However, it leads to a nonlinear system for updating x^k. In Algorithm 1 we choose an explicit expression for $p_{\mathcal{A}_k}$, cf. Table 4.1, which amounts to the one-step fixed-point iteration. Note that our choice of $p_{\mathcal{A}_k}$ ensures its boundedness. Namely, each component p_i satisfies*

$$|p_i| \le \begin{cases} 0 & \ell^0, \\ \lambda^{\frac{1}{2-\tau}}(2(1-\tau))^{\frac{\tau-1}{2-\tau}} & \text{bridge}, \\ \lambda & \text{capped}{-}\ell^1, \text{MCP}, \\ \frac{\tau}{\tau-1}\lambda & \text{SCAD}. \end{cases} \tag{4.2}$$

The stopping criterion at step 5 of Algorithm 1 is chosen to be either $\mathcal{A}_k = \mathcal{A}_{k+1}$ [31] or $k \ge J_{\max}$ for some fixed maximum number $J_{\max} > 0$ of iterations. Note that for nonconvex models, the stopping condition $\mathcal{A}_k = \mathcal{A}_{k+1}$ may never be reached; see [32, Example 3.1] for an example in the case of the ℓ^0 penalty. Hence, we augment it with the stopping condition $k \ge J_{\max}$.

4.2. Continuation strategy. To successfully apply Algorithm 1 to the model (1.3), there are two important practical issues, i.e., the initial guess x^0 in Algorithm 1 and the choice of the regularization parameter λ in the model (1.3), which we discuss separately below. Since the PDAS algorithm is a Newton type method, it merits fast convergence, but only in the neighborhood of a minimizer. This is also the case for the model (1.3), in light of the nonconvexity of the penalties. Hence, in order to fully exploit this feature, a good initial guess is required, which unfortunately is often

Algorithm 1 Unified primal dual active set algorithm

1: Given $J_{\max}$. Set initial guess x^0 and find $d^0 = \Psi^t(y - \Psi x^0)$.
2: **for** $k = 1, 2, ...$ **do**
3: Compute the active and inactive sets $\mathcal{A}_k$ and $\mathcal{I}_k$ respectively by

$$\mathcal{A}_k = \{i : |x_i^{k-1} + d_i^{k-1}| > T^*\} \quad \text{and} \quad \mathcal{I}_k = \mathcal{A}_k^c.$$

4: Update the primal and dual variable x^k and d^k respectively by

$$\begin{cases} x_{\mathcal{I}_k}^k = \mathbf{0}_{\mathcal{I}_k}, \\ \Psi_{\mathcal{A}_k}^t \Psi_{\mathcal{A}_k} x_{\mathcal{A}_k}^k = \Psi_{\mathcal{A}_k}^t y - p_{\mathcal{A}_k}, \\ d^k = \Psi^t(\Psi x^k - y), \end{cases}$$

where $p_{\mathcal{A}_k}$ is given in Table 4.1.
5: Check the stopping criterion.
6: **end for**

unavailable in practice. In this work, we adopt a continuation strategy to arrive at a good initial guess, which serves the role of globalizing the algorithm. Specifically, let $\lambda_s = \lambda_0 \rho^s$, $\rho \in (0, 1)$, be a decreasing sequence. Then we apply Algorithm 1 on the sequence $\{\lambda_s\}_s$, with the solution $(x(\lambda_s), d(\lambda_s))$ being the initial guess for the λ_{s+1}-problem. The overall algorithm is given in Algorithm 2. The initial guess λ_0 is chosen large enough such that 0 is the global minimizer of the model (1.3). In particular, we can choose it by

$$\lambda_0 = \begin{cases} \frac{1}{2}\|\Psi^t y\|_\infty^2 & \ell^0, \\ (\frac{\|\Psi^t y\|_\infty}{2-\tau})^{2-\tau}(2(1-\tau))^{1-\tau} & \text{bridge}, \\ \|\Psi^t y\|_\infty & \text{capped-}\ell^1, \text{SCAD, MCP}. \end{cases} \tag{4.3}$$

Algorithm 2 Unified primal dual active set with continuation algorithm

1: Given λ_0 and $\rho \in (0, 1)$. Let $x(\lambda_0) = 0$ and $d(\lambda_0) = \Psi^t y$.
2: **for** $k = 1, 2, ...$ **do**
3: Apply Algorithm 1 to problem (1.3) with $\lambda_k = \rho^k \lambda_0$, initialized with $(x^0, d^0) = (x(\lambda_{k-1}), d(\lambda_{k-1}))$.
4: Check the stopping criterion.
5: **end for**

The stopping criterion at Step 4 of Algorithm 2 determines the regularization parameter λ in the model (1.3). It compromises the tradeoff between the data fidelity and the sparsity level of the solution, and its proper choice is one notoriously difficult problem. Many useful choice rules have been proposed in the literature; see [29] for an updated overview. If the noise level ϵ of the data y is known, one may determine the regularization parameter λ by the discrepancy principle [32] (i.e., the stopping criterion for Algorithm 2):

$$\|\Psi x(\lambda_k) - y\| \le \epsilon. \tag{4.4}$$

If the noise level ϵ is unknown, one may apply the Bayesian information criterion to choose an appropriate regularization parameter λ [17]. Note that the sequence of approximate solutions $\{x(\lambda_s)\}_s$ needed in these choice rules has already been generated along with the continuation path, and this step does not incur any extra computational overhead. Hence, the continuation strategy can be seamlessly integrated with parameter choice rules.

Last we discuss the convergence of Algorithm 2. We shall focus on noise free data, and in the presence of data noise, the analysis is similar but much more involved (see [32] for the ℓ^0 case). Let $y = \Psi x^\dagger$, where $x^\dagger$ is the true sparse vector with its active set $\mathcal{A}^\dagger = \{i : x_i^\dagger \neq 0\}$ and $T = |\mathcal{A}^\dagger|$. The restricted isometry property (RIP) [6] of order k with constant δ_k of a matrix Ψ is defined as follows: Let $\delta_k \in (0,1)$ be the smallest constant such that

$$(1-\delta_k)\|x\|^2 \leq \|\Psi x\|^2 \leq (1+\delta_k)\|x\|^2$$

holds for all x with $\|x\|_0 \leq k$. Now we make the following assumption.

ASSUMPTION 4.1. *The matrix Ψ satisfies the RIP condition with an RIP constant*

$$\delta \equiv \delta_{T+1} \leq \begin{cases} \frac{1}{\sqrt{5T}+1} & \text{capped}-\ell^1, \text{MCP}, \\ \frac{1}{\sqrt{8T}+1} & \text{SCAD}, \\ \frac{2-\tau}{2-\tau+\sqrt{T[(4-2\tau)^2+1]}} & \text{bridge}. \end{cases}$$

Now we can state the convergence of Algorithm 2 under Assumption 4.1, and the proof is deferred to Appendix E.

THEOREM 4.2. *Let Assumption 4.1 hold. Then for sufficiently large λ_0 and $\rho \in (0,1)$ close to unit, Algorithm 2 is well-defined and $x(\lambda_k) \to x^\dagger$ as $k \to \infty$.*

REMARK 4.3. *In particular, Theorem 4.2 implies that for noise free data, the PDASC solution recovers the true sparse solution $x^\dagger$. Further, the proof of Theorem 4.2 in Appendix E indicates that the continuation strategy actually allows a precise control over the evolution of the active set during the iteration, cf. Lemma E.1, in addition to providing a good initial guess.*

5. Numerical experiments and discussions. In this section we showcase the performance of Algorithm 2 for the nonconvex penalties in Table 2.1 on both simulated and real data. All the experiments are done on a dual core desktop with 3.16 GHz and 4 GB RAM. The MATLAB code (package Unified-PDASC) is available at `http://www0.cs.ucl.ac.uk/staff/b.jin/software/updasc.zip`.

5.1. Experiment setup. First we describe the problem setup, i.e., data generation, parameter choice and stopping rule. In all numerical examples except example 5.5, the true signal x is given, and the vector y is generated by $y = \Psi x + \eta$, where η denotes the additive measurement noise, with the entries η_i following an independent identically distributed (i.i.d.) Gaussian distribution $N(0, \sigma^2)$ with mean zero and standard deviation σ. Unless otherwise stated, the standard deviation σ is fixed at $\sigma = 0.5$. In the examples, the matrix Ψ is constructed as follows.

(i) Random Gaussian matrix of size $n \times p$, $n \ll p$. First, a matrix $\widetilde{\Psi} \in \mathbb{R}^{n\times p}$ is generated such that its entries follow i.i.d. $N(0,1)$. Then we normalize $\widetilde{\Psi}$ to obtain Ψ such that each column is of unit length.

(ii) Random Gaussian matrix with correlation of size $n \times p$, $n \ll p$. First we generate a random Gaussian matrix $\widetilde{\Psi} \in \mathbb{R}^{n\times p}$ with its entries following i.i.d.

$N(0,1)$. Then we define a matrix $\overline{\Psi} \in \mathbb{R}^{n\times p}$ by setting $\overline{\psi}_1 = \widetilde{\psi}_1$,

$$\overline{\psi}_j = \widetilde{\psi}_j + \nu(\widetilde{\psi}_{j-1} + \widetilde{\psi}_{j+1}), \quad j = 2, ..., p-1,$$

where $\nu \in (0,1)$ is the correlation coefficient, and $\overline{\psi}_p = \widetilde{\psi}_p$. Finally, we normalize $\overline{\Psi}$ to obtain Ψ.

(iii) Heaviside matrix of size $n \times n$. It is obtained by normalizing the matrix

$$\widetilde{\Psi} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdots & 1 & 1 & 0 \\ 1 & \cdots & 1 & 1 & 1 \end{pmatrix}.$$

The true signal x is a random sparse vector, and its dynamic range (DR) is defined by $\mathrm{DR} = \max\{|x_i| : x_i \neq 0\}/\min\{|x_i| : x_i \neq 0\}$. In the model (1.3), the regularization parameter λ balances the data fidelity and the expected sparsity level of the solution. We set $\lambda_{\max}$ as in (4.3) and $\lambda_{\min} = 10^{-15}\lambda_{\max}$. The choice $\lambda_{\min}$ is quite arbitrary, since the algorithm stops at a much large λ value. The interval $[\lambda_{\min}, \lambda_{\max}]$ is divided into N ($N = 100$ in our tests) equal subintervals on a log-scale and let λ_s, $s = 0, ..., N$, be the s-th value (in descending order). Unless otherwise specified, we set $\tau = 0.5,\ 3.7,\ 2.7,$ and 1.5 for the bridge, SCAD, MCP and capped-ℓ^1 penalty, respectively.

5.2. Numerical results and discussions. Now we present numerical examples to illustrate the accuracy and efficiency of the algorithm.

Our first test illustrates the advantage of nonconvex penalties over Lasso.

EXAMPLE 5.1. *In this test we compare the exact support recovery probability at different sparsity levels for all the penalties in Table 2.1 (including Lasso). The matrix* $\Psi \in \mathbb{R}^{500\times 1000}$ *is random Gaussian, the true signal* x *has a support size of* $10 : 10 : 300$ *and a* $\mathrm{DR} = 10^3$. *The noise standard deviation* σ *is* $\sigma = 0.01$.

Lasso is solved by the proximal forward-backward splitting method [13] (with a continuation strategy), and the recovery probability is computed from 50 independent realizations of the signal x. It is observed from Fig. 5.1 that for Lasso, the recovery probability decreases rapidly as the support size exceeds 30, and almost completely vanishes when it exceeds 70. In contrast, all the nonconvex penalties perform well for a support size up to 100–200, with the range depending on the specific penalty, and hence they allow exact support recovery at much higher sparsity levels.

In the next experiment, we compare our algorithm with multi-stage convex relaxation (MSCR) (with five stages) due to Zhang [53] and a general iterative shrinkage/thresholding algorithm (GIST) due to Gong et al [23] (available online at `http://www.public.asu.edu/~jye02/Software/GIST/`, accessed on March 31, 2014). For both MSCR and GIST, we couple them with a continuation strategy, in order to obtain accurate solutions. We terminate MSCR with the Bayesian information criterion, and for GIST, we apply (4.4). Although not presented, we note that the λ values determined by the two approaches are fairly close.

EXAMPLE 5.2. *We consider the following three different problem setups*

(a) $\Psi \in \mathbb{R}^{500\times 10000}$ *is a random Gaussian matrix, and the signal* x *contains* 50 *nonzero elements with a* $\mathrm{DR} = 10^3$.

(b) $\Psi \in \mathbb{R}^{1000\times 10000}$ *is a random Gaussian matrix with a correlation coefficient* $\nu = 0.2$, *and the signal* x *contains* 100 *nonzero elements with a* $\mathrm{DR} = 10^3$.

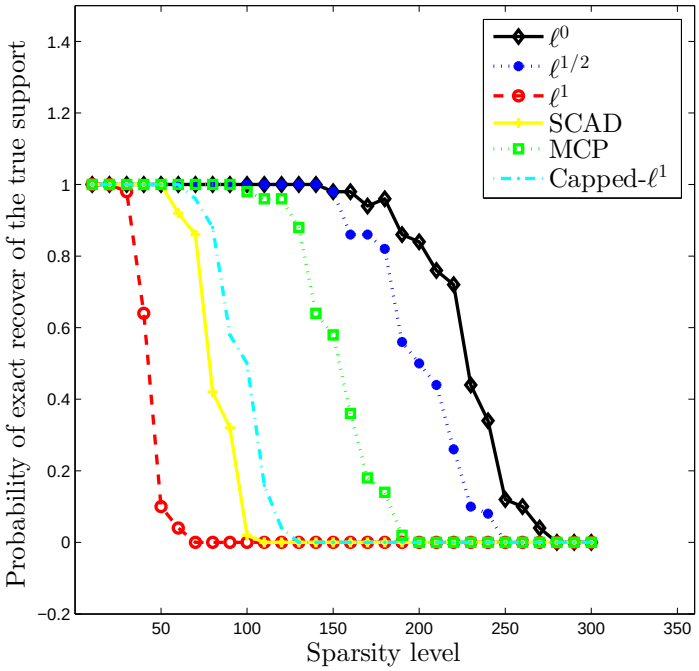


Fig. 5.1: The exact recovery probability of the support by for example 5.1.

(c) $\Psi \in \mathbb{R}^{50\times 50}$ *is a Heaviside matrix, the signal* x *contains* 5 *nonzero elements with a* DR $= 10^2$ *and* $\sigma = 1 \times 10^{-2}$.

In example 5.2, cases (a) and (b) involve random matrices and thus are easier, whereas case (c) is numerically more challenging. The performance is evaluated in terms of CPU time (in seconds) and relative error (in the ℓ^2-norm), which are computed as the average of 10 independent realizations of the problem setup. The results are summarized in Tables 5.1-5.3.

For cases (a) and (b), all three algorithms work well (since GIST does not support $\ell^\tau, \tau \in [0, 1)$ we do not show the results, indicated by $-$ in the tables.). However, our algorithm is much faster (on average at least by a factor of ten for GIST and a factor of ten - forty for MSCR), while the reconstruction errors by our algorithm are smaller. This is attributed to its local superlinear convergence, which we shall examine more closely below. For case (c), the MSCR does not work well, indicated by the large reconstruction errors. In contrast, our algorithm works well for case (c) for all nonconvex penalties, and all models yield satisfactory results. The accuracy of the GIST and PDASC algorithm is comparable.

We now examine the continuation strategy and local convergence of the algorithm.

EXAMPLE 5.3. $\Psi \in \mathbb{R}^{300\times 1000}$ *is a random Gaussian matrix, and the signal* x *contains* 50 *nonzero elements with a* DR $= 10^2$, *and* $\sigma = 0.1$.

The convergence history of Algorithm 2 for example 5.3 is shown in Fig. 5.2. Here $\mathcal{A}$ and $\mathcal{A}_s$ refer respectively to the exact active set $\mathrm{supp}(x)$ and the approximate one $\mathrm{supp}(x(\lambda_s))$, where $x(\lambda_s)$ is the solution to the λ_s-problem. It is observed that $\mathcal{A}_s \subset \mathcal{A}$ for all s, and the size $|\mathcal{A}_s|$ increases monotonically as the iteration proceeds. For each λ_s, with $(x(\lambda_{s-1}), d(\lambda_{s-1}))$ as the initial guess (with $(x(\lambda_0), d(\lambda_0)) = (0, \Psi^t y)$),

Table 5.1: Results for example 5.2(a), random Gaussian matrix.

	PDASC		GIST		MSCR	
	time	RE	time	RE	time	RE
ℓ^0	0.94	8.49e-5	-	-	3.00	2.55e-4
$\ell^{1/2}$	0.81	1.10e-4	-	-	3.75	2.61e-4
SCAD	0.61	8.49e-5	5.79	8.92e-5	6.94	6.94e-5
MCP	0.57	8.49e-5	5.65	8.80e-5	6.64	5.50e-5
capped-ℓ^1	0.55	8.49e-5	3.69	1.13e-4	7.12	4.65e-5

Table 5.2: Results for example 5.2(b), correlated random Gaussian matrix.

	PDASC		GIST		MSCR	
	time	RE	time	RE	time	RE
ℓ^0	1.89	5.24e-5	-	-	6.15	2.09e-5
$\ell^{1/2}$	1.69	7.01e-5	-	-	7.87	3.50e-5
SCAD	1.21	5.24e-5	11.5	5.62e-5	13.5	3.65e-5
MCP	1.15	5.24e-5	11.5	5.65e-5	13.7	3.16e-5
capped-ℓ^1	1.13	5.24e-5	9.25	7.13e-5	15.5	2.47e-5

Table 5.3: Results for example 5.2(c), heaviside case.

	PDASC		GIST		MSCR	
	time	RE	time	RE	time	RE
ℓ^0	6.80e-3	2.41e-4	-	-	9.71e-4	3e-1
$\ell^{1/2}$	7.10e-3	2.61e-4	-	-	8.65e-4	3e-1
SCAD	2.27e-2	1.14e-4	4.66e-2	5.14e-4	9.58e-4	3e-1
MCP	1.36e-2	1.14e-4	5.20e-2	6.89e-4	8.61e-4	3e-1
capped-ℓ^1	1.07e-2	1.14e-4	4.32e-2	1.47e-4	1.00e-3	3e-1

Algorithm 1 generally converges within one or two iterations for all five nonconvex penalties, cf., Fig. 5.2, indicating a local superlinear convergence of the active set algorithm. Hence, the coverall procedure is very efficient.

Four (bridge, capped ℓ^1, SCAD and MCP) of the penalties in Table 2.1 have a free parameter τ that controls their concavity. Our next experiment examines the sensitivity of the algorithm with respect to the concavity parameter τ.

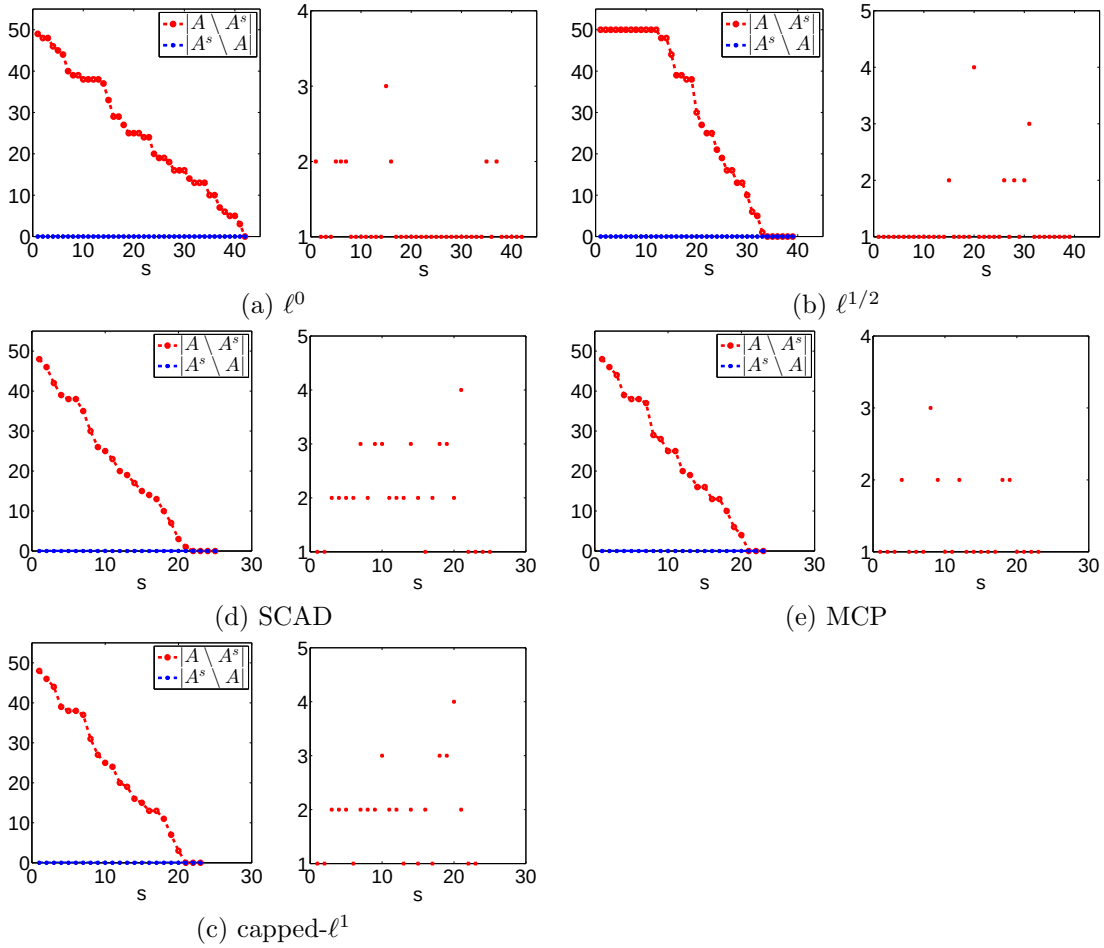
EXAMPLE 5.4. *The matrix $\Psi \in \mathbb{R}^{400\times 1000}$ is random Gaussian, and the true signal x contains* 50 *nonzero elements with a* $\mathrm{DR} = 10^3$.

Like before, we evaluate Algorithm 2 by CPU time (in seconds), relative error (in the ℓ^2-norm) and absolute ℓ^∞ error computed from ten independent realizations of the problem setup. The CPU time is fairly robust with respect to the concavity parameter τ, cf. Table 5.4. Further, the reconstruction error varies little with the parameter τ, indicating the robustness of the penalized models.

Finally, we illustrate Algorithm 2 on two real datasets.

EXAMPLE 5.5. *We consider two real datasets.*

(a) *This example is the benchmark for the Lasso solver in MATLAB 2013. It uses the 1985 auto imports database of the UCI repository, which can be download-*



(a) ℓ^0 (b) $\ell^{1/2}$

(d) SCAD (e) MCP

(c) capped-ℓ^1

Fig. 5.2: The convergence behavior of Algorithm 2 for example 5.3: change of the active sets (left panel), and the number of iterations needed for each λ_s-problem (right panel).

ed from `http://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names`. *There are 13 features: 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-size', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg', and 'highway-mpg'.*

(b) *The second is from Stamey et al. (1989) [44]. The correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy are considered. The variables are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).*

The results for example 5.5(a) are shown in Table 5.5, where the results by Lasso are computed by MATLAB built-in function `lasso`. The significant features selected

Table 5.4: Results for example 5.4: sensitivity analysis.

(a) bridge				(b) capped-ℓ^1			
τ	Time	RE	AE	τ	Time	RE	AE
0	4.58e–2	9.93e-5	7.17e-2	1.1	9.58e-2	9.93e-5	7.16e-2
0.2	2.86e-2	1.01e-4	7.21e-2	1.5	9.57e-2	9.93e-5	7.16e-2
0.4	2.65e-2	1.13e-4	7.81e-2	5	1.11e-1	1.07e-4	9.03e-2
0.6	2.81e-2	1.39e-4	1.04e-1	10	1.59e-1	1.35e-4	1.16e-2
0.8	3.22e-2	1.40e-4	1.01e-1	50	1.82e-1	8.49e-4	3.97e-1
(c) SCAD				(d) MCP			
τ	Time	RE	AE	τ	Time	RE	AE
2.1	9.67e-2	9.93e-5	7.17e-2	1.1	5.16e-2	9.93e-5	7.17e-2
3.7	8.48e-2	1.05e-4	8.62e-2	2.7	5.55e-2	1.11e-4	9.88e-2
5	8.70e-2	1.26e-4	1.26e-1	5	5.72e-2	1.30e-4	1.63e-1
10	9.70e-2	1.19e-4	1.07e-1	10	5.74e-2	1.26e-4	1.25e-1
50	1,64e-1	1.12e-3	4.71e-1	50	6.00e-2	1.31e-4	1.32e-1

by Lasso are 'length', 'width', 'curb-weight', and 'horsepower', and all the nonconvex regularized models consistently select 'width', 'curb-weight', 'compression-ratio', and 'horsepower'. Here the "exact" solution is unknown. Since the ℓ^0 model is closest to the gold standard, best subset selection, its result may be taken as a reference. All the nonconvex models can recover the "exact" solution, but the Lasso fails to recover the "exact" support, concurring with the observation in example 5.1.

Table 5.5: Results for example 5.5(a).

Features	Lasso	ℓ^0	$\ell^{1/2}$	SCAD	MCP	capped-ℓ^1
'wheel-base'						
'length'	√					
'width'	√	√	√	√	√	√
'height'						
'curb-weight'	√	√	√	√	√	√
'engine-size'						
'bore'						
'stroke'						
'compression-ratio'		√	√	√	√	√
'horsepower'	√	√	√	√	√	√
'peak-rpm'						
'city-mpg'						
'highway-mpg'						

The results for example 5.5(b) are shown in Table 5.6. The results by the best subset selection and Lasso are taken from [25]. All the nonconvex models select 'Intercept', 'lcavol', and 'lweight', which agree well with the subset selection result [25, Table 3.3], except for the SCAD penalty which selects one more feature 'lbph' than the others. Lasso selects two more features, i.e., 'lbph' and 'svi', which corroborates the empirical observations for examples 5.1 and 5.5(b).

Table 5.6: Results for example 5.5(b).

Features	best subset	Lasso	ℓ^0	$\ell^{1/2}$	SCAD	MCP	capped-ℓ^1
'Intercept'	√	√	√	√	√	√	√
'lcavol'	√	√	√	√	√	√	√
'lwight'	√	√	√	√	√	√	√
'age'							
'lbph'		√			√		
'svi'		√					
'lcp'							
'gleason'							
'pgg45'							

6. Conclusions. We have developed a primal dual active set algorithm for a class of nonconvex optimization problems arising in sparse signal recovery and high-dimensional statistics, including the ℓ^0, bridge, capped-ℓ^1, smoothly clipped absolute deviation and minimax concave penalty. Theoretically, we established the existence of a minimizer, and derived a necessary optimality condition for a global minimizer, based on the associated thresholding operator. The solutions to the necessary optimality condition are always coordinate-wise minimizers, and further, we provided numerically verifiable sufficient conditions for a coordinate-wise minimizer to be a local minimizer. Meanwhile, the necessary optimality condition and the active set can be reformulated using both primal and dual variables, which lends itself to a primal dual active set algorithm. At each iteration, it involves only solving a least-squares problem on the active set and merits a local superlinear convergence, and thus when coupled with a continuation strategy, the procedure is very efficient and accurate. The global convergence of the overall algorithm is shown. Its efficiency and accuracy is verified by extensive numerical experiments, including real data.

There are several avenues for further study. First, for the ill-conditioned sensing matrices, which are characteristic of most inverse problems, the linear systems involved in the PDAS algorithm can be challenging to solve directly, and extra regularization might be necessary. This motivates further study on the related theoretical issues. Second, in some applications, the sensing matrix Ψ is given implicitly where only matrix-vector multiplication is available. This necessitates developing iterative linear solvers, and the study of inexact inner iterations. Last, the extensions of the PDASC algorithm to structured sparsity, e.g., group sparsity penalty and the matrix analogue, are also of immense interest.

Acknowledgements. The authors are grateful to the anonymous referees and the associate editor, Prof. Wotao Yin, for their constructive comments, which have improved the quality of the paper. B. Jin is partially supported by NSF Grant DMS-1319052 and EPSRC grant EP/M025160/1, and X. Lu is supported by National Science Foundation of China No. 91230108 and No. 11471253.

Appendix A. Proof of Theorem 3.2.

Proof. We compute the tuple (t^*, T^*) for the five penalties separately.
(i) ℓ^0. $g(t) = \frac{t}{2} + \frac{\lambda}{t}$ for $t > 0$ and $g(0) = +\infty$. Hence $t^* = \sqrt{2\lambda}$ and $T^* = g(t^*) = \sqrt{2\lambda}$.
(ii) bridge. $g(t) = \frac{t}{2} + \lambda t^{\tau-1}$ for $t > 0$ and $g(0) = +\infty$. Direct computation gives $t^* = (2\lambda(1-\tau))^{\frac{1}{2-\tau}}$, and $T^* = g(t^*) = (2-\tau)\,[2(1-\tau)]^{\frac{\tau-1}{2-\tau}}\,\lambda^{\frac{1}{2-\tau}}$.

(iii) capped-ℓ^1. Then the function $g(t)$ is given by

$$g(t) = \frac{t}{2} + \begin{cases} \frac{\lambda^2\tau}{t}, & t \geq \lambda\tau, \\ \lambda, & 0 \leq t \leq \lambda\tau. \end{cases}$$

In the interval $[0, \lambda\tau]$, 0 is the minimizer of $g(t)$ with a minimum value λ, whereas in the interval $[\lambda\tau, \infty)$, the minimum value is $\lambda\sqrt{2\tau}$, which is greater than λ. Hence $t^* = 0$, and $T^* = \lambda$.
(iv) SCAD. Then the function $g(t)$ is given by

$$g(t) = \frac{t}{2} + \begin{cases} \frac{\lambda^2(\tau+1)}{2t}, & t \geq \lambda\tau, \\ \frac{\lambda\tau t - \frac{1}{2}(t^2+\lambda^2)}{(\tau-1)t}, & \lambda \leq t \leq \lambda\tau, \\ \lambda, & 0 \leq t \leq \lambda. \end{cases}$$

It can be verified directly that the minimizer of $g(t)$ in the intervals $[0, \lambda]$, $[\lambda, \lambda\tau]$, $[\lambda\tau, \infty)$ is given by 0, λ, $\lambda\sqrt{\tau+1}$, respectively. Hence $t^* = 0$, and $T^* = \lambda$.
(v) MCP. Then the function $g(t)$ is given by

$$g(t) = \frac{t}{2} + \begin{cases} \frac{\lambda^2\tau}{2t}, & t \geq \lambda\tau, \\ \lambda - \frac{t}{2\tau}, & 0 \leq t \leq \lambda\tau. \end{cases}$$

Analogous to the case of the SCAD, we can obtain $t^* = 0$, and $T^* = \lambda$. □

Appendix B. Proof of Theorem 3.4.

Proof. We discuss only the case $v > 0$, for which $u^* \geq 0$. The remaining case $v \leq 0$ can be treated similarly.
(i) ℓ^0. By Lemma 3.3, if $|v| > T^*$, then $u^* \neq 0$, which implies the minimizer u^* is v. From this the formula of $S_\lambda^{\ell^0}$ follows (see also [31]).
(ii) bridge. Let $G(u) = \frac{u^2}{2} + \lambda u^\tau - uv$ for $u \geq 0$. Its first- and second derivatives are given by

$$G'(u) = u + \lambda\tau u^{\tau-1} - v \quad \text{and} \quad G''(u) = 1 + \lambda\tau(\tau-1)u^{\tau-2}.$$

Clearly, $G'(u)$ is convex with $G'(0+) = G'(+\infty) = +\infty$. Hence, $G'(u)$ has at most two real roots, and $G(u)$ is either monotonically increasing or has three monotone intervals. This and Lemma 3.3 yield the expression $S_{\lambda,\tau}^{\ell^\tau}$. Generally there is no closed-form expression for $S_{\lambda,\tau}^{\ell^\tau}(v)$. For $|v| > T^*$, the unique minimizer to $G(u)$ is the larger root of $G'(u)$ (the other root is a local maximizer) (see also [31, 23]).
(iii) capped-ℓ^1. Let

$$G(u) = \begin{cases} \frac{u^2}{2} - uv + \lambda^2\tau, & u \geq \lambda\tau, \\ \frac{u^2}{2} - uv + \lambda u, & 0 \leq u \leq \lambda\tau. \end{cases}$$

By Lemma 3.3, for $|v| \leq \lambda$, we have $u^* = 0$. We then assume $v > \lambda$. Simple computation shows

$$S_1^* := \min_{u \geq \lambda\tau} G(u) = \lambda^2\tau - v^2/2 \text{ at } u = v,$$
$$S_2^* := \min_{u \in [0, \lambda\tau]} G(u) = -(v-\lambda)^2/2 \text{ at } u = v - \lambda.$$

Then we have

$$\begin{cases} v > \lambda(\tau+\frac{1}{2}) \Rightarrow S_1^* < S_2^*, & u^* = v > \lambda\tau, \\ v < \lambda(\tau+\frac{1}{2}) \Rightarrow S_1^* > S_2^*, & u^* = v-\lambda < \lambda\tau, \\ v = \lambda(\tau+\frac{1}{2}) \Rightarrow S_1^* = S_2^*, & u^* = \lambda\tau \pm \frac{\lambda}{2}, \end{cases}$$

whence follows the thresholding operator $S^{c\ell^1}_{\lambda,\tau}$ (see also [23]).
(iv) SCAD. We define

$$G(u) = \begin{cases} G_1(u) \triangleq \frac{u^2}{2} - uv + \frac{\lambda^2(\tau+1)}{2}, & u \geq \lambda\tau, \\ G_2(u) \triangleq \frac{u^2}{2} - uv + \frac{\lambda\tau u - \frac{1}{2}(u^2+\lambda^2)}{\tau-1}, & \lambda \leq u \leq \lambda\tau, \\ G_3(u) \triangleq \frac{u^2}{2} - uv + \lambda u, & 0 \leq u \leq \lambda. \end{cases}$$

By Lemma 3.3, $|v| \leq \lambda \Rightarrow u^* = 0$. We then assume $v > \lambda$. The three quadratic functions $G_i(u)$ achieve their minimum at $u = v$, $u = \frac{(\tau-1)v-\lambda\tau}{\tau-2}$ and $u = v-\lambda$, respectively. Next we discuss the three cases separately. First, if $v \geq \lambda\tau$, then $\frac{(\tau-1)v-\lambda\tau}{\tau-2} \geq \lambda\tau$, which implies that $G_2(u)$ is decreasing on the interval $[\lambda, \lambda\tau]$, it reaches its minimum at $\lambda\tau$. Similarly, $v-\lambda \geq \lambda$ implies that $G_3(u)$ reaches its minimum over the interval $[0,\lambda]$ at λ. Hence

$$\min_{0\leq u\leq\lambda} G_3(u) = G_3(\lambda) = G_2(\lambda) \geq \min_{\lambda\leq u\leq\lambda\tau} G_2(u) = G_2(\lambda\tau) = G_1(\lambda\tau) \geq \min_{u\geq\lambda\tau} G_1(u).$$

Second, if $\lambda\tau \geq v \geq 2\lambda$, then G_1 is increasing on $[\lambda\tau, \infty)$ and G_3 is decreasing on $[0,\lambda]$, and $\frac{(\tau-1)v-\lambda\tau}{\tau-2} \geq \lambda\tau \in [\lambda, \lambda\tau]$. Hence

$$\begin{aligned} \min_{0\leq u\leq\lambda} G_3(u) = G_3(\lambda) = G_2(\lambda) &\geq \min_{\lambda\leq u\leq\lambda\tau} G_2(u), \\ \min_{\lambda\leq u\leq\lambda\tau} G_2(u) \leq G_2(\lambda\tau) = G_1(\lambda\tau) &= \min_{u\geq\lambda\tau} G_1(u). \end{aligned}$$

Third, if $2\lambda \geq v \geq \lambda$, similar argument gives that

$$\min_{0\leq u\leq\lambda} G_3(u) \leq G_3(\lambda) = G_2(\lambda) = \min_{\lambda\leq u\leq\lambda\tau} G_2(u) \leq G_2(\lambda\tau) = G_1(\lambda\tau) = \min_{u\geq\lambda\tau} G_1(u).$$

This yields the thresholding operator $S^{\mathrm{scad}}_{\lambda,\tau}$ (see also [4, 38, 23]).
(v) MCP. Like before, we let

$$G(u) = \begin{cases} \frac{u^2}{2} - uv + \frac{1}{2}\lambda^2\tau, & u \geq \lambda\tau, \\ \frac{u^2}{2} - uv + \lambda u - \frac{u^2}{2\tau}, & 0 \leq u \leq \lambda\tau. \end{cases}$$

Similar to case (iv), we obtain the expression for $S^{\mathrm{mcp}}_{\lambda,\tau}$ (see also [4, 38, 23]). □

Appendix C. Proof of Theorem 3.6.

Proof. We prove Theorem 3.6 by establishing the inequality

$$J(x^* + \omega) \geq J(x^*) \tag{C.1}$$

for small $\omega \in \mathbb{R}^p$, using the optimality condition and thresholding operator.

(i) ℓ^0. By Lemma 3.5 and using the thresholding operator $S_\lambda^{\ell^0}$, we deduce that for $i \in \mathcal{A}$, $|x_i^*| \geq \sqrt{2\lambda}$. Further,

$$0 = d_{\mathcal{A}}^* = \Psi_{\mathcal{A}}^t(y - \Psi_{\mathcal{A}} x_{\mathcal{A}}^*) \quad \Leftrightarrow \quad x_{\mathcal{A}}^* \in \operatorname{argmin} \tfrac{1}{2}\|\Psi_{\mathcal{A}} x_{\mathcal{A}} - y\|^2. \tag{C.2}$$

Now consider a small perturbation ω, with $\|\omega\|_\infty < \sqrt{2\lambda}$, to x^*. It suffices to show (C.1) for small ω. Recall that $\omega_{\mathcal{I}}$ is the subvector of ω whose entries are listed in the index set $\mathcal{I}$. If $\omega_{\mathcal{I}} \neq 0$, then

$$J(x^* + \omega) - J(x^*) \geq \tfrac{1}{2}\|\Psi x^* - y + \Psi\omega\|^2 - \tfrac{1}{2}\|\Psi x^* - y\|^2 + \lambda \geq \lambda - |(\omega, d^*)|,$$

which is positive for small ω. Meanwhile, if $\omega_{\mathcal{I}} = 0$, by (C.2), we deduce (C.1).

(ii) bridge. We first note that on the active set $\mathcal{A}$, $|x_i^*| \geq t^* = (2\lambda(1-\tau))^{\frac{1}{2-\tau}}$. Next we claim that if the minimizer u^* of $G(u) = \frac{u^2}{2} - uv + \lambda u^\tau$ is positive, then $G(u)$ is locally strictly convex around u^*, i.e., for small t and some $\theta > 0$ such that

$$G(u^* + t) - G(u^*) = G(u^* + t) - G(u^*) - G'(u^*)t \geq \theta t^2.$$

To see this, we recall that u^* is the larger root of $u + \lambda\tau u^{\tau-1} = v$ and $v \geq T^*$. By the convexity of $u + \lambda\tau u^{\tau-1}$, $u^*(v)$ is increasing in v for $v \geq T^*$. Further, by the inequality $u^* \geq t^*$, we have

$$\begin{aligned} G''(u^*) &= 1 - \lambda\tau(1-\tau)(u^*)^{\tau-2} \\ &\geq 1 - \lambda\tau(1-\tau)(t^*)^{\tau-2} = 1 - \tfrac{\tau}{2}. \end{aligned}$$

In particular, the function $G(u)$ is locally strictly convex with $\theta = \frac{1}{2} - \frac{\tau}{4} - \epsilon$, for any $\epsilon > 0$. Hence for each $i \in \mathcal{A}$ and small t, there holds

$$J(x^* + te_i) - J(x^*) = \tfrac{1}{2}t^2 + (t\psi_i, \Psi x^* - y) + \lambda|x_i^* + t|^\tau - \lambda|x_i^*|^\tau \geq \theta t^2,$$

i.e.,

$$-td_i^* + \lambda|x_i^* + t|^\tau - \lambda|x_i^*|^\tau \geq (\theta - \tfrac{1}{2})t^2.$$

Consequently for small ω, we have

$$\begin{aligned} J(x^* + \omega) - J(x^*) &= \tfrac{1}{2}\|\Psi\omega\|^2 - (\omega, d^*) + \sum_{i\in\mathcal{A}} \lambda(|x_i^* + \omega_i|^\tau - |x_i^*|^\tau) + \lambda\sum_{i\in\mathcal{I}} |\omega_i|^\tau \\ &\geq \tfrac{1}{2}\|\Psi\omega\|^2 - (\omega_{\mathcal{I}}, d_{\mathcal{I}}^*) + \lambda\sum_{i\in\mathcal{I}} |\omega_i|^\tau + (\theta - \tfrac{1}{2})\|\omega_{\mathcal{A}}\|^2. \end{aligned}$$

Note the trivial estimates

$$-(\omega_{\mathcal{I}}, d_{\mathcal{I}}^*) \geq -\sum_{i\in\mathcal{I}} |\omega_i|\|d_{\mathcal{I}}^*\|_\infty \quad \text{and} \quad \tfrac{1}{2}\|\Psi\omega\|^2 \geq \tfrac{1}{2}\|\Psi_{\mathcal{A}}\omega_{\mathcal{A}}\|^2 + (\omega_{\mathcal{A}}, \Psi_{\mathcal{A}}^t\Psi_{\mathcal{I}}\omega_{\mathcal{I}}).$$

Further, by Young's inequality, for any $\delta > 0$

$$(\omega_{\mathcal{A}}, \Psi_{\mathcal{A}}^t\Psi_{\mathcal{I}}\omega_{\mathcal{I}}) \geq -\delta\|\omega_{\mathcal{A}}\|^2 - \tfrac{1}{4\delta}\|\Psi_{\mathcal{A}}^t\Psi_{\mathcal{I}}\omega_{\mathcal{I}}\|^2 \geq -\delta\|\omega_{\mathcal{A}}\|^2 - C_\delta\|\omega_{\mathcal{I}}\|^2.$$

Combing these four estimates together and noting $\theta = \frac{1}{2} - \frac{\tau}{4} - \epsilon$ yields

$$\begin{aligned} J(x^* + \omega) - J(x^*) \geq &\left(\tfrac{1}{2}\|\Psi_{\mathcal{A}}\omega_{\mathcal{A}}\|^2 - (\tfrac{\tau}{4} + \epsilon + \delta)\|\omega_{\mathcal{A}}\|^2\right) \\ &+ \sum_{i\in\mathcal{I}} |\omega_i|^\tau \left(\lambda - |\omega_i|^{1-\tau}\|d_{\mathcal{I}}^*\|_\infty - C_\delta|\omega_i|^{2-\tau}\right). \end{aligned}$$

The first term is nonnegative if ϵ and δ are small and Ψ satisfies (3.5) with $\sigma(\mathcal{A}) > \frac{\tau}{2}$. The sum over $\mathcal{I}$ is nonnegative for small ω, thereby showing (C.1).

The proof of the rest cases is based on the identity

$$J(x^*+\omega) - J(x^*) = \tfrac{1}{2}\|\Psi\omega\|^2 + \sum_i \underbrace{\left(\rho_{\lambda,\tau}(x_i^*+\omega_i) - \rho_{\lambda,\tau}(x_i^*) - \omega_i d_i^*\right)}_{:=s_i}. \tag{C.3}$$

(iii) capped-ℓ^1. We denote by $\mathcal{A}_1 = \{i : |x_i^*| > \lambda\tau\}$, $\mathcal{A}_2 = \{i : \lambda\tau > |x_i^*| > 0\}$. By assumption $\{i : |x_i^*| = \lambda\tau\} = \emptyset$, hence $\mathcal{I} = (\mathcal{A}_1 \cup \mathcal{A}_2)^c$. The optimality condition for x^* and the differentiability of $\rho_{\lambda,\tau}^{c\ell^1}(t)$ for $|t| \neq \lambda\tau$ yield $d_i^* = 0$ for $i \in \mathcal{A}_1$, and $d_i^* = \lambda\mathrm{sgn}(x_i^*)$ for $i \in \mathcal{A}_2$. Thus, for ω small, there holds

$$s_i = \begin{cases} 0, & i \in \mathcal{A}_1 \cup \mathcal{A}_2, \\ \lambda|\omega_i| - \omega_i d_i^*, & i \in \mathcal{I}. \end{cases}$$

Now with the fact that for $i \in \mathcal{I}$, $|d_i^*| \leq \lambda$, we deduce that for small ω, (C.1) holds.
(iv) SCAD. Let $\mathcal{A}_1 = \{i : |x_i^*| > \lambda\tau\}$, $\mathcal{A}_2 = \{i : |x_i^*| \in [\lambda, \lambda\tau]\}$, $\mathcal{A}_3 = \{i : |x_i^*| \in (0,\lambda)\}$, and $\mathcal{I} = (\cup\mathcal{A}_i)^c$. Then the optimality of x^* yields

$$d_i^* = \begin{cases} 0, & i \in \mathcal{A}_1, \\ \frac{\lambda\mathrm{sgn}(x_i^*) - x_i^*}{\tau - 1}, & i \in \mathcal{A}_2, \\ \lambda\mathrm{sgn}(x_i^*), & i \in \mathcal{A}_3, \end{cases}$$

and $|d_i^*| \leq \lambda$ on $\mathcal{I}$. Then for small ω in the sense that for

$$i \in \mathcal{A}_1 \Rightarrow |x_i^* + \omega_i| > \lambda\tau \quad \text{and} \quad i \in \mathcal{A}_3 \Rightarrow |x_i^* + \omega_i| \in (0,\lambda),$$

we obtain $s_i = 0$, $i \in \mathcal{A}_1 \cup \mathcal{A}_3$. For $i \in \mathcal{A}_2$, we have two cases:

$$s_i \begin{cases} = -\frac{w_i^2}{2(\tau-1)}, & \text{if } |x_i^* + \omega_i| \in [\lambda, \tau\lambda], \\ \geq -\frac{w_i^2}{2(\tau-1)}, & \text{otherwise.} \end{cases}$$

Finally for $i \in \mathcal{I}$, $|d_i^*| < \lambda$ by assumption, and hence

$$s_i = \lambda|w_i| - d_i^* w_i \geq |w_i|(\lambda - |d_i^*|).$$

Combining these estimates with (C.3), we arrive at

$$J(x^*+\omega) - J(x^*) \geq \tfrac{1}{2}\|\Psi\omega\|^2 - \tfrac{1}{2(\tau-1)}\|\omega_{\mathcal{A}_2}\|^2 + \sum_{i\in\mathcal{I}} |\omega_i|(\lambda - |d_i^*|),$$

Further, by Young's inequality, we bound

$$\tfrac{1}{2}\|\Psi\omega\|^2 \geq \tfrac{1}{2}\|\Psi_{\mathcal{A}}\omega_{\mathcal{A}}\|^2 + (\omega_{\mathcal{A}}, \Psi_{\mathcal{A}}^t\Psi_I\omega_I) \geq (\tfrac{1}{2} - \epsilon)\|\Psi_{\mathcal{A}}\omega_{\mathcal{A}}\|^2 - C_\epsilon\|\omega_{\mathcal{I}}\|^2.$$

Consequently, there holds

$$J(x^*+\omega) - J(x^*) \geq (\tfrac{1}{2} - \epsilon)\|\Psi_{\mathcal{A}}\omega_{\mathcal{A}}\|^2 - \tfrac{1}{2(\tau-1)}\|\omega_{\mathcal{A}_2}\|^2 - \sum_{i\in\mathcal{I}} \left(|\omega_i|(\lambda - |d_i^*| - C_\epsilon|\omega_i|\right).$$

If (3.5) with $\sigma(\mathcal{A}) > \frac{1}{\tau-1}$ and $\|d_{\mathcal{I}}^*\|_\infty < \lambda$ hold, then (C.1) follows.

(v) MCP. The proof is similar to case (iv). We let $\mathcal{A}_1 = \{i : |x_i^*| > \tau\lambda\}$, $\mathcal{A}_2 = \{i : 0 < |x_i^*| \leq \tau\lambda\}$, and $\mathcal{I} = (\cup\mathcal{A}_i)^c$. The differentiability of $\rho_{\lambda,\tau}^{\mathrm{mcp}}(t)$ yields

$$d_i^* = \begin{cases} 0, & i \in \mathcal{A}_1, \\ \lambda \mathrm{sgn}(x_i^*) - x_i^*/\tau, & i \in \mathcal{A}_2, \end{cases}$$

and on the set $\mathcal{I}$, $|d_i^*| \leq \lambda$. Note that for small ω_i, there holds $s_i = 0$ for $i \in \mathcal{A}_1$. Similarly, for $i \in \mathcal{A}_2$, there holds

$$s_i \geq \begin{cases} \frac{1}{2\tau}|\omega_i|^2, & i \in \mathcal{A}_2, \\ |\omega_i|(\lambda - |d_i^*|), & i \in \mathcal{I}, \end{cases}$$

The rest of the proof is identical with case (iv), and hence omitted. □

Appendix D. Explicit expression of $d_{\mathcal{A}}^*$. For a coordinate-wise minimizer x^*, we derive the explicit expression shown in Table 4.1 for the dual variable $d^* = \Psi^t(y - \Psi x^*)$ on the active set $\mathcal{A} = \{i : x_i^* \neq 0\}$.
(i) ℓ^0. By the expression of $S_\lambda^{\ell^0}$, we have $d_i^* x_i^* = 0$, and hence $d_i^* = 0$, for $i \in \mathcal{A}$.
(ii) bridge. Since for $i \in \mathcal{A}$, $J(x^*)$ is differentiable along the direction e_i at point x_i^*, the necessary optimality condition for x_i^* reads $d_i^* - \lambda\tau\frac{|x_i^*|^\tau}{x_i^*} = 0$.
(iii) capped-ℓ^1. We divide the active set $\mathcal{A}$ into $\mathcal{A} = \cup_i\mathcal{A}_i$, with $\mathcal{A}_1 = \{i : |x_i^* + d_i^*| > \lambda(\tau + \frac{1}{2})\}$, $\mathcal{A}_2 = \{i : \lambda < |x_i^* + d_i^*| < \lambda(\tau + \frac{1}{2})\}$, and $\mathcal{A}_3 = \{i : |x_i^* + d_i^*| = \lambda(\tau + \frac{1}{2})\}$. Then the definition of the operator $S_{\lambda,\tau}^{c\ell^1}$ gives the desired expression.
(iv) SCAD. We divide the active set $\mathcal{A}$ into $\mathcal{A} = \cup_i\mathcal{A}_i$ with $\mathcal{A}_1 = \{i : |x_i^* + d_i^*| \geq \lambda\tau\}$, $\mathcal{A}_2 = \{i : \lambda\tau > |x_i^* + d_i^*| > 2\lambda\}$, and $\mathcal{A}_3 = \{i : 2\lambda \geq |x_i^* + d_i^*| > \lambda\}$, then it follows from the necessary optimality condition for x_i^* that the desired expression holds.
(v) MCP. Similar to case (iv), we divide the active set $\mathcal{A}$ into $\mathcal{A} = \cup_i\mathcal{A}_i$ with $\mathcal{A}_1 = \{i : |x_i^* + d_i^*| \geq \lambda\tau\}$ and $\mathcal{A}_2 = \{i : \lambda < |x_i^* + d_i^*| < \lambda\tau\}$. Then the desired expression follows from the optimality condition for x_i^*.

Appendix E. Proof of Theorem 4.2. First we recall some estimates for the RIP constant δ_k (see, e.g., [40, 48]). Let $\mathcal{A} \cap \mathcal{B} = \emptyset$ and $\delta_{|\mathcal{A}|+|\mathcal{B}|}$ exists, then

$$\|\Psi_{\mathcal{A}}^t\Psi_{\mathcal{A}}x_{\mathcal{A}}\| \gtreqless (1 \mp \delta_{|\mathcal{A}|})\|x_{\mathcal{A}}\|, \quad \|(\Psi_{\mathcal{A}}^t\Psi_{\mathcal{A}})^{-1}x_{\mathcal{A}}\| \gtreqless \tfrac{1}{1\mp\delta_{|\mathcal{A}|}}\|x_{\mathcal{A}}\|,$$
$$\|\Psi_{\mathcal{A}}^t\Psi_{\mathcal{B}}\| \leq \delta_{|\mathcal{A}|+|\mathcal{B}|}, \qquad \left\|\left[I - (\Psi_{\mathcal{A}}^t\Psi_{\mathcal{A}})^{-1}\right]x_{\mathcal{A}}\right\| \leq \tfrac{\delta_{|\mathcal{A}|}}{1-\delta_{|\mathcal{A}|}}\|x_{\mathcal{A}}\|.$$

Given any index set $\mathcal{A} \subset \mathcal{A}^\dagger$, we denote $\mathcal{I} = \mathcal{A}^c$ and $\mathcal{B} = \mathcal{A}^\dagger\backslash\mathcal{A}$, and further, let

$$x_{\mathcal{A}} = (\Psi_{\mathcal{A}}^t\Psi_{\mathcal{A}})^{-1}(\Psi_{\mathcal{A}}^t y - p_{\mathcal{A}}), \quad d_{\mathcal{A}} = \Psi_{\mathcal{A}}^t(y - \Psi_{\mathcal{A}}x_{\mathcal{A}}),$$

Then we have $d_{\mathcal{A}} = p_{\mathcal{A}}$. By noting the trivial relation $y = \Psi_{\mathcal{A}}x_{\mathcal{A}}^\dagger + \Psi_{\mathcal{B}}x_{\mathcal{B}}^\dagger$, we deduce

$$\begin{aligned} \|x_{\mathcal{A}} - x_{\mathcal{A}}^\dagger\| &\leq \|(\Psi_{\mathcal{A}}^t\Psi_{\mathcal{A}})^{-1}\Psi_{\mathcal{A}}^t\Psi_{\mathcal{B}}x_{\mathcal{B}}^\dagger\| + \|(\Psi_{\mathcal{A}}^t\Psi_{\mathcal{A}})^{-1}p_{\mathcal{A}}\| \\ &\leq \tfrac{\delta}{1-\delta}\|x_{\mathcal{B}}^\dagger\| + \tfrac{1}{1-\delta}\|p_{\mathcal{A}}\|. \end{aligned} \tag{E.1}$$

Then by appealing to the identity $d_i = \Psi_i^t(\Psi_{\mathcal{B}}x_{\mathcal{B}}^\dagger - \Psi_{\mathcal{A}}(x_{\mathcal{A}} - x_{\mathcal{A}}^\dagger))$ and (E.1), we find

$$\|x_{\mathcal{A}} + d_{\mathcal{A}} - x_{\mathcal{A}}^\dagger\| \leq \tfrac{\delta}{1-\delta}\|x_{\mathcal{B}}^\dagger\| + \tfrac{\delta}{1-\delta}\|p_{\mathcal{A}}\| \triangleq h_{\mathcal{A}}, \tag{E.2}$$
$$|d_i| \geq |x_i^\dagger| - \delta\|(x_{\mathcal{A}} - x_{\mathcal{A}}^\dagger)\| - \delta\|x_{\mathcal{B}}^\dagger\| \geq |x_i^\dagger| - h_{\mathcal{A}}, \quad \forall i \in \mathcal{B}, \tag{E.3}$$
$$|d_i| \leq \delta\|(x_{\mathcal{A}} - x_{\mathcal{A}}^\dagger)\| + \delta\|x_{\mathcal{B}}^\dagger\| \leq h_{\mathcal{A}}, \quad \forall i \in \mathcal{I}^\dagger. \tag{E.4}$$

Next we define the index set $G_{\lambda,s}$ by

$$G_{\lambda,s} \triangleq \begin{cases} \left\{ i : |x_i^\dagger| \geq \lambda s \right\} & \text{capped-}\ell^1,\ \text{SCAD},\ \text{MCP}, \\ \left\{ i : |x_i^\dagger| \geq (\lambda s)^{\frac{1}{2-\tau}} \right\} & \text{bridge}. \end{cases} \tag{E.5}$$

The set $G_{\lambda,s}$ characterizes the exact sparse solution $x^\dagger$ in terms of level sets.

The general strategy of the convergence analysis is similar to that in [17, 32]. It relies crucially on the following monotonicity property on the active set. Namely, the evolution of the active set during the iteration can be precisely controlled, by suitably choosing the decreasing factor ρ and s.

LEMMA E.1. *For $\rho \in (0,1)$ close to unity and some $s > 0$, there holds*

$$G_{\lambda,s} \subset \mathcal{A}_k \subset \mathcal{A}^\dagger \Rightarrow G_{\rho\lambda,s} \subset \mathcal{A}_{k+1} \subset \mathcal{A}^\dagger. \tag{E.6}$$

Proof. Assume for some inner iteration $G_{\lambda,s} \subset \mathcal{A}_k \subset \mathcal{A}^\dagger$. Let $\mathcal{A} = \mathcal{A}_k$ and $\mathcal{B} = \mathcal{A}^\dagger \backslash \mathcal{A}$. First we derive upper bounds on the crucial term $h_{\mathcal{A}}$ in (E.2). It follows from (4.2) and the definition of $G_{\lambda,s}$ that

$$h_{\mathcal{A}} \leq \begin{cases} \frac{\delta}{1-\delta}\left(s\lambda\sqrt{|B|} + \lambda\sqrt{|A|} \right) & \text{capped-}\ell^1,\ \text{MCP}, \\ \frac{\delta}{1-\delta}\left(s\lambda\sqrt{|B|} + \lambda\frac{\tau}{\tau-1}\sqrt{|A|} \right) & \text{SCAD}, \\ \frac{\delta}{1-\delta}\left((s\lambda)^{\frac{1}{2-\tau}}\sqrt{|B|} + (\lambda c_\tau)^{\frac{1}{2-\tau}}\sqrt{|A|} \right) & \text{bridge}, \end{cases}$$

where the constant $c_\tau = [2(1-\tau)]^{\tau-1}$. Upon noting $|A| + |B| = T$ and the elementary inequality $a\sqrt{t} + b\sqrt{T-t} \leq \sqrt{a^2+b^2}\sqrt{T}$, we deduce

$$h_{\mathcal{A}} \leq \begin{cases} \frac{\delta}{1-\delta}\sqrt{s^2+1}\sqrt{T}\lambda & \text{capped-}\ell^1,\ \text{MCP}, \\ \frac{\delta}{1-\delta}\sqrt{s^2 + \frac{\tau^2}{(\tau-1)^2}}\sqrt{T}\lambda & \text{SCAD}, \\ \frac{\delta}{1-\delta}\sqrt{(\frac{s}{c_\tau})^{\frac{2}{2-\tau}} + 1}\sqrt{T}(c_\tau\lambda)^{\frac{1}{2-\tau}} & \text{bridge}. \end{cases} \tag{E.7}$$

Now we prove (E.6) for different penalties. In view of (E.2)-(E.4), it suffices to show $h_{\mathcal{A}} < T^*$ and $\rho s\lambda - h_{\mathcal{A}} > T^*$, where T^* is given in Theorem 3.2.

Capped-ℓ^1 and MCP: Since $\delta < \frac{1}{\sqrt{5T}+1}$, we have $\frac{\delta}{1-\delta}\sqrt{5T} < 1$. Then we choose $s = 2$ and $\rho \in (\frac{1+\frac{\delta}{1-\delta}\sqrt{5T}}{2}, 1)$. It follows from (E.7) that

$$\begin{aligned} &h_{\mathcal{A}} \leq \tfrac{\delta}{1-\delta}\sqrt{5T}\lambda < \lambda \Rightarrow \mathcal{A}_{k+1} \subset \mathcal{A}^\dagger, \\ &\rho s\lambda - h_{\mathcal{A}} \geq 2\rho\lambda - \tfrac{\delta}{1-\delta}\sqrt{5T}\lambda > \lambda \Rightarrow G_{\rho\lambda,s} \subset \mathcal{A}_{k+1}. \end{aligned}$$

SCAD: Like before, since $\delta < \frac{1}{\sqrt{8T}+1}$, we deduce $\frac{\delta}{1-\delta}\sqrt{8T} < 1$. We choose $s = 2$ and $\rho \in (\frac{1+\frac{\delta}{1-\delta}\sqrt{8T}}{2}, 1)$. Then by (E.7) and noting $\tau > 2 \Rightarrow \frac{\tau}{\tau-1} < 2$, we obtain

$$\begin{aligned} &h_{\mathcal{A}} \leq \tfrac{\delta}{1-\delta}\sqrt{4 + \tfrac{\tau^2}{(\tau-1)^2}}\sqrt{T}\lambda < \lambda \Rightarrow \mathcal{A}_{k+1} \subset \mathcal{A}^\dagger, \\ &\rho s\lambda - h_{\mathcal{A}} \geq 2\rho\lambda - \tfrac{\delta}{1-\delta}\sqrt{8T}\lambda > \lambda \Rightarrow G_{\rho\lambda,s} \subset \mathcal{A}_{k+1}. \end{aligned}$$

Bridge: Recall $T^* = (2-\tau)(c_\tau\lambda)^{\frac{1}{2-\tau}}$, cf. Theorem 3.2. Since $\delta < \frac{2-\tau}{2-\tau+\sqrt{T[(4-2\tau)^2+1]}}$, let $\frac{s}{c_\tau} = (4-2\tau)^{2-\tau}$ and we have $\frac{\delta}{1-\delta}\sqrt{(\frac{s}{c_\tau})^{\frac{2}{2-\tau}}+1}\sqrt{T} \le 2-\tau$. By choosing $\rho \in (\frac{2-\tau+\frac{\delta}{1-\delta}\sqrt{T[(4-2\tau)^2+1]}}{4-2\tau}, 1)$, we deduce

$$h_{\mathcal{A}} \le \tfrac{\delta}{1-\delta}\sqrt{(\tfrac{s}{c_\tau})^{\frac{2}{2-\tau}}+1}\sqrt{T}(c_\tau\lambda)^{\frac{1}{2-\tau}} < T^* \Rightarrow \mathcal{A}_{k+1} \subset \mathcal{A}^\dagger,$$
$$(\rho s\lambda)^{\frac{1}{2-\tau}} - h_{\mathcal{A}} - T^* \ge (c_\tau\lambda)^{\frac{1}{2-\tau}}\left(\rho(4-2\tau) - (2-\tau) - \tfrac{\delta}{1-\delta}\sqrt{T[(4-2\tau)^2+1]}\right) > 0 \Rightarrow G_{\rho\lambda,s} \subset \mathcal{A}_{k+1}.$$

This completes the proof of the lemma. □

Now we can give the proof of Theorem 4.2.

Proof. For each λ_k-problem, we denote by $\mathcal{A}_{k,0}$ and $\mathcal{A}_{k,\diamond}$ the active set for the initial guess and the last inner step (i.e., $\mathcal{A}(\lambda_k)$ in Algorithm 2), respectively. Since λ_0 is large enough, we deduce that $G_{\lambda_1,s} = \varnothing$ and $G_{\lambda_1,s} \subset \mathcal{A}_{1,0}$. Then by mathematical induction and Lemma E.1, we have for any k

$$G_{\lambda_k,s} \subseteq \mathcal{A}_{k,0} \subset \mathcal{A}^\dagger \quad \text{and} \quad G_{\rho\lambda_k,s} \subseteq \mathcal{A}_{k,\diamond} \subset \mathcal{A}^\dagger. \tag{E.8}$$

Therefore Algorithm 2 is well-defined and when k is large such that

$$s\lambda_k < \begin{cases} \min\left\{|x_i^\dagger| : x_i^\dagger \neq 0\right\} & \text{capped-}\ell^1\text{ , SCAD, MCP,} \\ (\min\left\{|x_i^\dagger| : x_i^\dagger \neq 0\right\})^{2-\tau} & \text{bridge,} \end{cases}$$

we have $\mathcal{A}(\lambda_k) = \mathcal{A}^\dagger$ and hence Algorithm 1 converges in one step. To show the convergence of the sequence of approximate solutions to the true solution $x^\dagger$, it suffices to check $\lim_{k\to\infty} p_{\mathcal{A}^\dagger}(\lambda_k) = 0$. The convergence of the approximate dual $p_{\mathcal{A}^\dagger}(\lambda_k)$ follows from the specific choice in Table 4.1 and its boundedness in (4.2). Hence we have

$$x(\lambda_k)_{\mathcal{A}^\dagger} = (\Psi_{\mathcal{A}^\dagger}^t\Psi_{\mathcal{A}^\dagger})^{-1}(\Psi_{\mathcal{A}^\dagger}^t y - p_{\mathcal{A}^\dagger}(\lambda_k)) \to x_{\mathcal{A}^\dagger}^\dagger.$$

This completes the proof of Theorem 4.2. □

REFERENCES

[1] Hirotugu Akaike, *A new look at the statistical model identification*, IEEE Trans. Autom. Control, 19 (1974), pp. 716–723.

[2] Thomas Blumensath and Mike E. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654.

[3] Kristian Bredies, Dirk Lorenz, and Stefan Reiterer, *Minimization of non-smooth, non-convex functionals by iterative thresholding*, J. Optim. Theory Appl., 165 (2015), pp. 78–112.

[4] Patrick Breheny and Jian Huang, *Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection*, Ann. Appl. Stat., 5 (2011), pp. 232–253.

[5] Leo Breiman, *Heuristics of instability and stabilization in model selection*, Ann. Statist., 24 (1996), pp. 2350–2383.

[6] Emmanuel J. Candés, Justin Romberg, and Terence Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[7] Emmanuel J. Candes and Terence Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[8] Rick Chartrand and Valentina Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Problems, 24 (2008), pp. 035020, 14.
[9] Rick Chartrand and Wotao Yin, *Iteratively reweighted algorithms for compressive sensing.* Proc. ICASSP 2008, 2008.
[10] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.
[11] Xiaojun Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Prog., Ser. B, 134 (2012), pp. 71–99.
[12] Xiaojun Chen, Lingfeng Niu, and Yaxiang Yuan, *Optimality conditions and a smoothing trust region Newton method for nonlipschitz optimization*, SIAM J. Optim., 23 (2013), pp. 1528–1552.
[13] Patrick L. Combettes and Valérie R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
[14] David L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[15] Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360.
[16] Jianqing Fan and Heng Peng, *Nonconcave penalized likelihood with a diverging number of parameters*, Ann. Statist., 32 (2004), pp. 928–961.
[17] Qibin Fan, Yuling Jiao, and Xiliang Lu, *A primal dual active set with continuation for compressed sensing*, IEEE Trans. Signal Proc., 62 (2014), pp. 6276–6285.
[18] Simon Foucart and Ming-Jun Lai, *Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$*, Appl. Comput. Harmon. Anal., 26 (2009), pp. 395–407.
[19] Illdiko E Frank and Jerome H Friedman, *A statistical view of some chemometrics regression tools*, Technometrics, 35 (1993), pp. 109–135.
[20] Wenjiang J. Fu, *Penalized regressions: the bridge versus the lasso*, J. Comput. Graph. Statist., 7 (1998), pp. 397–416.
[21] Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu, *Recovering sparse signals with a certain family of nonconvex penalties and DC programming*, IEEE Trans. Signal Proc., 57 (2009), pp. 4686–4698.
[22] Izrail' M. Gel'fand and Sergeĭ V. Fomin, *Calculus of Variations*, Prentice-Hall, N.J., 1963.
[23] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye, *A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems*, in Proc. 30th Int. Conf. Mach. Learn. (ICML-13), 2013, pp. 37–45.
[24] Roland Griesse and Dirk Lorenz, *A semismooth Newton method for Tikhonov functionals with sparsity constraints*, Inverse Problems, 24 (2008), pp. 035007, 19 pp.
[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, second ed., 2009.
[26] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888 (2003).
[27] Michael Hintermüller and Tao Wu, *A superlinearly convergent r-regularized newton scheme for variational models with concave sparsity-promoting priors*, Comput. Optim. Appl., 57 (2014), pp. 1–25.
[28] Jian Huang, Joel L. Horowitz, and Shuangge Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, Ann. Statist., 36 (2008), pp. 587–613.
[29] Kazufumi Ito and Bangti Jin, *Inverse Problems: Tikhonov Theory and Algorithms*, vol. 22 of Series on Applied Mathematics, World Scientific, NJ, 2014.
[30] Kazufumi Ito, Bangti Jin, and Jun Zou, *A two-stage method for inverse medium scattering*, J. Comput. Phys., 237 (2013), pp. 211–223.
[31] Kazufumi Ito and Karl Kunisch, *A variational approach to sparsity optimization based on Lagrange multiplier theory*, Inverse Problems, 30 (2014), pp. 015001, 23 pp.
[32] Yuling Jiao, Bangti Jin, and Xiliang Lu, *A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem*, Appl. Comput. Harm. Anal., (2014), p. in press.
[33] Nick Kingsbury, *Complex wavelets for shift invariant analysis and filtering of signals*, Appl. Comput. Harm. Anal., 10 (2001), pp. 234–253.
[34] Keith Knight and Wenjiang Fu, *Asymptotics for lasso-type estimators*, Ann. Statist., 28 (2000), pp. 1356–1378.
[35] Ming-Jun Lai and Jingyue Wang, *An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems*, SIAM J. Optim., 21 (2011), pp. 82–101.
[36] Ming-Jun Lai, Yangyang Xu, and Wotao Yin, *Improved iteratively reweighted least squares*

for unconstrained smoothed ℓ_q minimization, SIAM J. Numer. Anal., 51 (2013), pp. 927–957.

[37] Zhaosong Lu, *Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming*, Math. Progr., Ser. A, 147 (2014), pp. 277–307.

[38] Rahul Mazumder, Jerome H. Friedman, and Trevor Hastie, *SparseNet: coordinate descent with nonconvex penalties*, J. Amer. Statist. Assoc., 106 (2011), pp. 1125–1138.

[39] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Statist., 34 (2006), pp. 1436–1462.

[40] Deanna Needell and Joel A Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, Appl. Comput. Harm. Anal., 26 (2009), pp. 301–321.

[41] Mila Nikolova, *Description of the minimizers of least squares regularized with l^0-norm. Uniqueness of the global minimizer*, SIAM J. Imag. Sci., 6 (2013), pp. 904–937.

[42] Mila Nikolova, Michael K Ng, and Chi-Pan Tam, *Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction*, IEEE Trans. Imag. Proc., 19 (2010), pp. 3073–3088.

[43] Yiyuan She, *Thresholding-based iterative selection procedures for model selection and shrinkage*, Electron. J. Stat., 3 (2009), pp. 384–415.

[44] Thomas A Stamey, John N Kabalin, John E McNeal, Iain M Johnstone, Fuad S Freiha, Elise A Redwine, and Norman Yang, *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II: Radical prostatectomy treated patients*, J. Urol., 141 (1989), pp. 1076–1083.

[45] Qiyu Sun, *Recovery of sparsest signals via ℓ^q-minimization*, Appl. Comput. Harmon. Anal., 32 (2012), pp. 329–341.

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

[47] Joel A Tropp and Anna C Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Trans. Inf. Theory, 53 (2007), pp. 4655–4666.

[48] Joel A Tropp and Stephen J Wright, *Computational methods for sparse solution of linear inverse problems*, Proc. IEEE, 98 (2010), pp. 948–958.

[49] Paul Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl., 109 (2001), pp. 475–494.

[50] Cun-Hui Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Statist., 38 (2010), pp. 894–942.

[51] Cun-Hui Zhang and Jian Huang, *The sparsity and bias of the LASSO selection in high-dimensional linear regression*, Ann. Statist., 36 (2008), pp. 1567–1594.

[52] Cun-Hui Zhang and Tong Zhang, *A general theory of concave regularization for high-dimensional sparse estimation problems*, Statist. Sci., 27 (2012), pp. 576–593.

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

[54] Peng Zhao and Bin Yu, *On model selection consistency of Lasso*, J. Mach. Learn. Res., 7 (2006), pp. 2541–2563.

[55] Yun-Bin Zhao and Duan Li, *Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems*, SIAM J. Optim., 22 (2012), pp. 1065–1088.

[56] Hui Zou and Runze Li, *One-step sparse estimates in nonconcave penalized likelihood models*, Ann. Statist., 36 (2008), pp. 1509–1533.