

A General Theory of Concave Regularization for High Dimensional Sparse Estimation Problems

Cun-Hui Zhang*

Department of Statistics and Biostatistics
Rutgers University, NJ

Tong Zhang[†]

Department of Statistics and Biostatistics
Rutgers University, NJ

Abstract. Concave regularization methods provide natural procedures for sparse recovery. However, they are difficult to analyze in the high dimensional setting. Only recently a few sparse recovery results have been established for some specific local solutions obtained via specialized numerical procedures. Still, the fundamental relationship between these solutions such as whether they are identical or their relationship to the global minimizer of the underlying nonconvex formulation is unknown. The current paper fills this conceptual gap by presenting a general theoretical framework showing that under appropriate conditions, the global solution of nonconvex regularization leads to desirable recovery performance; moreover, under suitable conditions, the global solution corresponds to the unique sparse local solution, which can be obtained via different numerical procedures. Under this unified framework, we present an overview of existing results and discuss their connections. The unified view of this work leads to a more satisfactory treatment of concave high dimensional sparse estimation procedures, and serves as guideline for developing further numerical procedures for concave regularization.

*Research partially supported by the NSF Grants DMS 0906420, DMS-11-06753 and NSA Grant H98230-11-1-0205

[†]Research partially supported by the following grants: AFOSR-10097389, NSA -AMS 081024, NSF DMS-1007527, and NSF IIS-1016061

1. INTRODUCTION

Let \mathbf{X} be an $n \times p$ design matrix and $\mathbf{y} \in \mathbb{R}^n$ a response vector satisfying

$$(1) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a target vector of regression coefficients and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a noise vector. This paper concerns the estimation of the value of $\mathbf{X}\boldsymbol{\beta}$, that of $\boldsymbol{\beta}$, or its support set $\text{supp}(\boldsymbol{\beta})$, where $\text{supp}(\mathbf{b}) := \{j : b_j \neq 0\}$ for any vector $\mathbf{b} = (b_1, \dots, b_p)^\top \in \mathbb{R}^p$.

We are interested in the high-dimensional case where n and p are both allowed to diverge, including the case of $p \gg n$. We assume that the target vector $\boldsymbol{\beta}$ is sparse in some sense; such as the ℓ_0 sparsity $|\text{supp}(\boldsymbol{\beta})| \leq s$ or the capped- ℓ_1 sparsity $\sum_{j=1}^p \min(1, |\beta_j|/\lambda) \leq s$ for some positive number $s > 0$. Usually, in the context of high dimensional sparsity analysis, we can allow s as large as $c_0 n / \ln p$ for a fixed small constant c_0 , and $\lambda = \sigma \sqrt{2 \ln p / n}$, where σ is a certain noise level. While we are mainly interested in the Gaussian noise $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_{n \times n})$ or zero-mean sub-Gaussian noise, the specific noise properties required in our analysis will be provided later.

We consider the following class of penalized least squares estimators

$$(2) \quad \hat{\boldsymbol{\beta}} := \arg \min_{\mathbf{b} \in \mathbb{R}^p} L_\lambda(\mathbf{b}), \quad L_\lambda(\mathbf{b}) := \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \sum_{j=1}^p \rho(b_j; \lambda),$$

where $\mathbf{b} = (b_1, \dots, b_p)^\top$ and $\rho(t; \lambda)$ is a scalar regularization function with a certain regularization parameter $\lambda > 0$. As an example, we may let $\rho(t; \lambda) = (\lambda^2/2)I(t \neq 0)$, which corresponds to the ℓ_0 regularization problem. Here $I(\cdot)$ denotes $\{0, 1\}$ valued indicator function. Since $I(t \neq 0)$ is a discontinuous function at $t = 0$, the corresponding ℓ_0 optimization problem may be difficult to solve. In practice, one also looks at continuous regularizers that approximate ℓ_0 regularization, such as those described in Table 1 and plotted in Figure 1. The quantities λ^* and γ^* in Table 1 will be introduced later in our analysis. As we will show in the paper, sparse local solutions of such regularizers can be obtained using standard numerical procedures (such as gradient descent), and they are closely related to the global solution of (2).

2. SURVEY OF EXISTING CONCAVE REGULARIZATION RESULTS

While this survey is not intended to be comprehensive, it presents a high-level view of some important contributions to the area of concave regularization. We will discuss both methodological and analytical contributions.

2.1 Terminologies

The following notation is used throughout the paper. For any dimension d , bold face letters denote vectors and normal face their elements, e.g. $\mathbf{v} = (v_1, \dots, v_d)^\top$, with $\text{supp}(\mathbf{v})$ being its support $\{j : v_j \neq 0\} \cap \{0, \dots, d\}$. Capital bold face letters denote matrices, e.g.

TABLE 1
Examples of concave penalties of form $\rho(t; \lambda) = \lambda^2 \rho(t/\lambda)$

$\lambda = \lambda^*$ in all examples, $C_\alpha = \{2(1 - \alpha)\}^{1-\alpha}/(2 - \alpha)^{2-\alpha}$, $\gamma \geq 1$
 Threshold level of the penalty $\lambda^* := \inf_{t>0} \{t/2 + \rho(t; \lambda)/t\}$
 Scaled maximum penalty $\gamma^* := \max_t \rho(t; \lambda)/\lambda^2$

Penalty	$\rho(t)$	$\text{sgn}(t)(d/dt)\rho(t)$	γ^*
ℓ_0	$(1/2)I\{ t > 0\}$		1/2
Bridge ($0 < \alpha < 1$)	$C_\alpha t ^\alpha$	$\alpha C_\alpha t ^{\alpha-1}$	∞
ℓ_1	$ t $	1	∞
Capped- ℓ_1	$\min(\gamma/2, t)$	$I(t \leq \gamma/2)$	$\gamma/2$
MCP	$\int_0^{ t } (1 - x/\gamma)_+ dx$	$(1 - t /\gamma)_+$	$\gamma/2$
SCAD	$\int_0^{ t } 1 \wedge \left(1 - \frac{x-1}{\gamma-1}\right)_+ dx$	$1 \wedge \left(1 - \frac{ t -1}{\gamma-1}\right)_+$	$(\gamma + 1)/2$

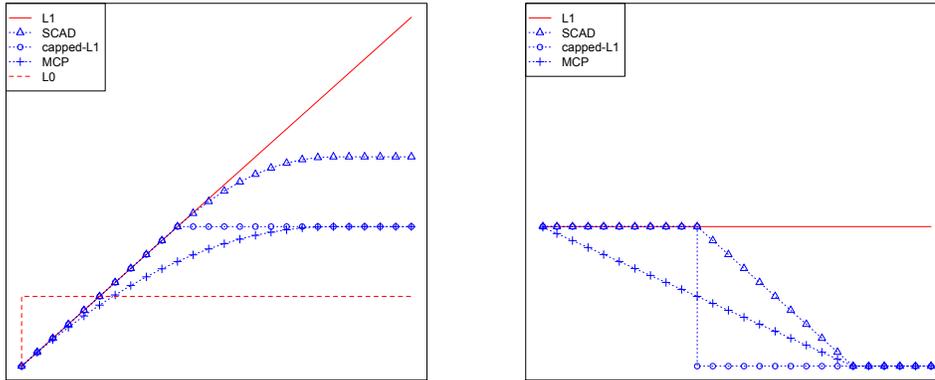


FIG 1. *Examples of concave penalty functions (left) and their derivatives (right)*

\mathbf{X} and Σ . The ℓ_q “norm” of \mathbf{v} is $\|\mathbf{v}\|_q := (\sum_{j=1}^d |v_j|^q)^{1/q}$ for $0 < q < \infty$, with the usual extension $\|\mathbf{v}\|_0 := |\text{supp}(\mathbf{v})|$ and $\|\mathbf{v}\|_\infty := \max_{j \leq d} |v_j|$. Design vectors, or columns of \mathbf{X} , are denoted by \mathbf{x}_j . For simplicity, we assume throughout the paper that the columns \mathbf{X} are normalized to

$$\|\mathbf{x}_j\|_2 = \sqrt{n}.$$

This condition is not essential but it simplifies some notations. For variable sets $A \subseteq \{1, \dots, p\}$, $\mathbf{X}_A = (\mathbf{x}_j, j \in A)$ denotes the restriction of columns of \mathbf{X} to A , and $\mathbf{b}_A = (b_j, j \in A)^\top$ the restriction of vector $\mathbf{b} \in \mathbb{R}^p$ to A . The maximum and minimum eigenvalues of matrix Σ are denoted by $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$.

DEFINITION 1. *The following terminologies will be used to simplify discussion.*

- (a) *The ℓ_0 sparsity of β means the ℓ_0 norm of β is small: β is s^* ℓ_0 -sparse if $\|\beta\|_0 \leq s^*$. To allow β with many more components near zero, a weaker notion of capped- ℓ_1 sparsity is defined as: β is s^* capped- ℓ_1 -sparse if $\sum_j \min(1, |\beta_j|/\lambda_{\text{univ}}) \leq s^*$, where $\lambda_{\text{univ}} = \sigma\sqrt{(2/n)\ln p}$ is the universal threshold level for a certain noise level σ .*
- (b) *A regularity condition on \mathbf{X} is a class \mathcal{X} of (column-normalized) matrices that match a sparsity condition on β to guarantee a desired result. Such a regularity condition can be stated as $\mathbf{X} \in \mathcal{X}_{s^*}^{n \times p}$, with matrix classes $\mathcal{X}_{s^*}^{n \times p} \subseteq \mathbb{R}^{n \times p}$ indexed by (n, p, s^*) , where s^* is the sparsity level of the matching regularity condition on β . Such a condition on \mathbf{X} is called an ℓ_2 regularity condition (or simply ℓ_2 regular) if the matrix classes $\mathcal{X}_{s^*}^{n \times p}$ are sufficiently large to satisfy the following condition:*
- *Given any $u_0 \geq 1$, there exists a constant $c_0 > 0$ such that for all $0 < \delta \leq 1/e$*

$$\inf_{\mu, n, p, s^*} \left\{ \mu(Q^{-1}(\mathcal{X}_{s^*}^{n \times p})) : \mu \in \mathcal{M}_{u_0}^{n \times p}, \frac{s^*}{n} \ln \left(\frac{p}{\delta} \right) \leq c_0, \min(n, p, s^*) \geq 1 \right\} \geq 1 - \delta,$$

where $\mathcal{M}_{u_0}^{n \times p}$ is the set of probability measures in $\mathbb{R}^{n \times p}$ under which the rows of $\mathbb{R}^{n \times p}$ are iid $N(0, \Sigma)$ for some Σ with $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \leq u_0$ and identical diagonal elements, and Q is the column normalization mapping given by $Q(\mathbf{X}) = (\mathbf{x}_j n^{1/2} / \|\mathbf{x}_j\|_2, j \leq p)$.

- (c) *An estimator $\hat{\beta}$ is selection consistent if $\text{supp}(\hat{\beta}) = \text{supp}(\beta)$, and sign-consistent if $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta)$, with the convention $\text{sgn}(0) = 0$ for the sign function.*
- (d) *An estimator has the oracle property if*

$$(3) \quad \hat{\beta} = \hat{\beta}^o, \quad \hat{\beta}_S^o = (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} \mathbf{X}_S^\top \mathbf{y}, \quad \text{supp}(\hat{\beta}^o) \subseteq S,$$

where $S = \text{supp}(\beta)$. The estimator $\hat{\beta}^o$ is called the oracle LSE.

For $0 < r \leq 1$, the capped- ℓ_1 sparsity condition holds for all vectors with $\|\beta\|_r \leq R$ as long as $(R/\lambda_{\text{univ}})^r \leq s^*$.

The standard regularity condition for the classical low-dimensional statistical scenario of $p \leq n$ is that the rank of \mathbf{X} is p . The purpose of Definition 1 (b) is to generalize this classical regularity condition to allow $p \gg n$. For example, $\inf_{|A| \leq 3s^*} \{\text{rank}(\mathbf{X}_A)/|A|\} = 1$ is ℓ_2 regular. The ℓ_2 notion allows an assessment of the strength of assumptions on \mathbf{X} by random matrix theory without repeating technical statements of more specialized conditions. Although the definition of ℓ_2 regularity is abstract, the underlying intuition is that columns of \mathbf{X} should not be highly correlated and for Gaussian random matrices the condition should be satisfied with large probability when n is larger than the order $s^* \ln p$ with s^* being the sparsity of β . We may explicitly include the classical situation into the definition of ℓ_2 regularity (that is, require $\mathcal{X}_{s^*}^{n \times p}$ to contain all column-normalized $n \times p$ matrices of rank p) if we confine our discussion to fixed sample conditions. See Remark 2 and the last paragraph of this subsection for more discussion. In this regard, the definition of ℓ_2 regularity condition excludes some of the conditions used in high dimensional sparsity analysis (such as irrepresentable condition which we will discuss later) because they do not generalize the classical rank-based regularity condition.

Throughout the paper, \mathbf{X} and β in (1) are treated as deterministic. Since the ℓ_2 criterion is about the size of $\mathcal{X}_{s^*}^{n \times p}$, it does not imply randomness of \mathbf{X} . In fact, since the ℓ_2 criterion is required to hold simultaneously for all $\mu \in \mathcal{M}_{u_0}^{n \times p}$ with the same $\mathcal{X}_{s^*}^{n \times p}$ in $\mathbb{R}^{n \times p}$, an ℓ_2 regularity condition is weaker than the condition of a random \mathbf{X} with distribution $\mu(Q^{-1}(\cdot))$ for a fixed $\mu \in \mathcal{M}_{u_0}^{n \times p}$ and typically requires a more explicit specification of the matrix class $\mathcal{X}_{s^*}^{n \times p}$. We call the criterion ℓ_2 , since it depends only on the range of the spectrum (the smallest and largest eigenvalues) of Σ .

If we consider a sequence of models in (1) with $n \rightarrow \infty$, then asymptotically an estimator has the oracle property (allowing statistical inference for all linear functionals of β) if

$$\sup_{\mathbf{a}} P\{|\mathbf{a}^\top(\hat{\beta} - \hat{\beta}^o)|^2 > \epsilon \text{Var}(\mathbf{a}^\top \hat{\beta}^o)\} = o(1) \quad \forall \epsilon > 0,$$

and this is a weaker requirement than (3) because it allows $\hat{\beta}$ to converge only asymptotically to $\hat{\beta}^o$. While this work focuses on the stronger requirement (3) that is easier to interpret in the finite sample situation, the weaker definition has been used in some previous asymptotic analysis.

The rest of the subsection discusses different forms of ℓ_2 conditions. Since the meaning of sparsity level is always clear in its proper context, for simplicity we may discuss design matrix conditions without explicitly referring to their sparsity levels.

In what follows, we will briefly explain some ℓ_2 -regularity conditions appeared in the literature. Related conditions have been introduced first in the compressive sensing literature to analyze ℓ_1 -regularized recovery of a sparse β from its random projection $\mathbf{X}\beta$ with iid $N(0, 1)$ entries in \mathbf{X} . The most well-known of such conditions is the restricted isometry condition (RIP) introduced in [11]. In order to explain RIP, we first define the lower and upper sparse eigenvalues as

$$(4) \quad \kappa_-(m) := \min_{\|\mathbf{u}\|_0 \leq m; \|\mathbf{u}\|_2 = 1} \|\mathbf{X}\mathbf{u}\|_2^2/n, \quad \kappa_+(m) := \max_{\|\mathbf{u}\|_0 \leq m; \|\mathbf{u}\|_2 = 1} \|\mathbf{X}\mathbf{u}\|_2^2/n.$$

RIP requires $\delta_k + \delta_{2k} + \delta_{3k} < 1$ with $k = \|\boldsymbol{\beta}\|_0$ and $\delta_m = \max\{\kappa_+(m) - 1, 1 - \kappa_-(m)\}$. A related condition is the uniform uncertainty principle (UUP) $\delta_{2k} + \theta_{2k,k} < 1$ in [10], where

$$(5) \quad \theta_{k,\ell} = \max_{A \cap B = \emptyset, |A|=k, |B|=\ell} \max_{\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1} (\mathbf{X}_A \mathbf{v}_A)^\top (\mathbf{X}_B \mathbf{u}_B) / n.$$

For ℓ_1 regularized estimators, bounds of the optimal order for the ℓ_2 -norm estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ can be obtained under RIP, UUP, as well as their improvement $\delta_{1.25k} + \theta_{1.25k,k} < 1$ in [8]. While the conditions for RIP and UUP are specialized to hold for random designs with covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_{p \times p}$, related conditions using sparse eigenvalues can be defined to fulfill the ℓ_2 criterion in Definition 1(b); for example the sparse Riesz condition (SRC) $\|\boldsymbol{\beta}\|_0 < \max_m 2m / \{1 + \kappa_+(m) / \kappa_-(m)\}$ in [44, 43], and some other extensions in [46, 42]. These more general conditions are ℓ_2 regularity conditions by our definition, and they lead to ℓ_2 -norm estimation error bounds of the optimal order for ℓ_1 regularized estimators.

While a suitable lower bound for a lower sparse eigenvalue is required for the identification of sparse $\boldsymbol{\beta}$ (even for $p < n$), it is not completely clear that an upper bound for an upper sparse eigenvalue or the cross-product in (5) is required for prediction and estimation error bounds of optimal order for a computationally manageable estimator. Refinements of RIP/UUP and SRC were introduced in the literature, such as the restricted eigenvalue of [4, 23],

$$\text{RE}_2 = \text{RE}_2(\xi, S) := \inf_{\mathbf{u}} \left\{ \|\mathbf{X}\mathbf{u}\|_2 / (\|\mathbf{u}\|_2 n^{1/2}) : \|\mathbf{u}_{S^c}\|_1 < \xi \|\mathbf{u}_S\|_1 \right\}$$

where $S = \text{supp}(\boldsymbol{\beta})$, and the compatibility factor of [38, 40],

$$\text{RE}_1 = \text{RE}_1(\xi, S) := \inf_{\mathbf{u}} \left\{ |S|^{1/2} \|\mathbf{X}\mathbf{u}\|_2 / (\|\mathbf{u}_S\|_1 n^{1/2}) : \|\mathbf{u}_{S^c}\|_1 < \xi \|\mathbf{u}_S\|_1 \right\}.$$

These quantities are directly connected to the Lasso and Dantzig selector since their estimation error $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ lives in the cone $\{\mathbf{u} : \|\mathbf{u}_{S^c}\|_1 < \xi \|\mathbf{u}_S\|_1\}$ under mild conditions on the noise. Thus, without requiring an additional condition on the upper sparse eigenvalue, $\text{RE}_1 > 0$ provides ℓ_1 -norm estimation and ℓ_2 -norm prediction error bounds of optimal order for these estimators, and $\text{RE}_2 > 0$ provides ℓ_2 -norm estimation bounds of optimal order. It can be shown that $\text{RE}_1 \geq \text{RE}_2$ and appropriate sparse eigenvalues imply $\text{RE}_2 > 0$. Therefore both $\text{RE}_2 > 0$ and $\text{RE}_1 > 0$ are ℓ_2 regularity conditions.

This paper employs an even weaker condition involving a restricted invertibility factor RIF_q in (15) which is related to the cone invertibility factor CIF_q ($q \geq 1$) defined below:

$$(6) \quad \text{CIF}_q = \text{CIF}_q(\xi, S) := \inf \left\{ \frac{|S|^{1/q} \|\mathbf{X}^\top \mathbf{X}\mathbf{u}\|_\infty}{n \|\mathbf{u}\|_q} : \|\mathbf{u}_{S^c}\|_1 < \xi \|\mathbf{u}_S\|_1 \right\}.$$

The quantity CIF_q and its sign-restricted version have appeared in [42], where invertibility factor-based ℓ_q error bounds of the form (20) below have been proven to sharpen earlier results for the Lasso and Dantzig selector [10, 44, 4, 46, 40] when $q \in [1, 2]$. Such error

bounds are of optimal order [42, 31]. Compared with RE_q , CIF_q weakens the lower bound condition with the stronger ℓ_∞ norm in the numerator of its definition. Of special interests are $q \in [1, 2]$ for which

$$(7) \quad \text{CIF}_1(\xi, S) \geq \frac{\text{RE}_1^2(\xi, S)}{(1 + \xi)^2}, \quad \text{CIF}_2(\xi, S) \geq \frac{\text{RE}_1(\xi, S)\text{RE}_2(\xi, S)}{(1 + \xi)} \geq \frac{\text{RE}_2^2(\xi, S)}{(1 + \xi)}.$$

Thus, $\text{CIF}_q > 0$ is an ℓ_2 regularity condition for $q \in [1, 2]$.

A main advantage of using invertibility factor is that for $q > 2$, invertibility factors still yield ℓ_q error bounds of optimal order which match results in [46, 42]. However, the sparse and restricted eigenvalues do not yield error bounds of optimal order for $q > 2$ due to the unboundedness of $\max_{\|\mathbf{u}\|_2=1} \|\mathbf{u}_S\|_q \|\mathbf{u}_S\|_1 / |S|^{1/q}$ in $|S|$.

We shall point out that different ℓ_2 regularity conditions are typically not equivalent since different norms are involved in the definitions of different quantities. For instance, in a specific example given in [4, 40], RE_1 and CIF_2 , uniformly bounded from away from zero, respectively yield ℓ_1 and ℓ_2 error bounds of optimal order, but RE_2 does not.

In the above discussion, we focus on fixed sample conditions like $\text{RE}_2 > 0$ and $\text{CIF}_2 > 0$, which hold when $\text{rank}(\mathbf{X}) = p$. These conditions can be directly seen as ℓ_2 regular from their existing lower bounds for $p > n$ such as those in [4, 42]. The optimality of the order of the error bounds based on such quantities can be also stated as ℓ_2 regularity conditions by comparing them with sparse eigenvalues. See Remark 2 for more discussion.

2.2 Previous Results

The Lasso (ℓ_1 regularization) is a special case of (2) with $\rho(t; \lambda) = \lambda|t|$ [36, 12]:

$$(8) \quad \hat{\boldsymbol{\beta}}^{(\ell_1)} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \left[\frac{1}{2n} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right].$$

As a function of λ , the Lasso path $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\ell_1)}(\lambda)$ matches that of ℓ_1 constrained quadratic programming. One may use the homotopy/Lars algorithm to compute the complete Lasso path for $\lambda \in [0, \infty)$ [29, 30, 14] or simply use a standard convex optimization algorithm to compute the Lasso solution for a finite set of λ . The Dantzig selector, proposed in [10], is an ℓ_1 -minimization method related to the Lasso, which solves

$$\hat{\boldsymbol{\beta}} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1 \quad \text{subject to } \|\mathbf{X}^\top(\mathbf{X}\mathbf{b} - \mathbf{y})/n\|_\infty \leq \lambda.$$

It has analytical properties similar to that of the Lasso, but can be computed by linear programming rather than quadratic programming as in the Lasso case. Analytic properties of the Lasso or Dantzig selector have been studied in [22, 18, 27, 37, 50, 41, 10, 7, 39, 44, 28, 4, 23, 46, 40, 8, 42]. A basic story is described in the following three paragraphs.

Under various ℓ_2 regularity conditions on \mathbf{X} and the ℓ_0 sparsity condition on $\boldsymbol{\beta}$, the Lasso and Dantzig selector control the estimation errors and the dimension of the selected

model in the sense

$$(9) \quad \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{M_{\text{pred}}\sigma^2 \ln p} + \frac{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q}{M_{\text{est}}\{(\sigma^2/n) \ln p\}^{q/2}} + \frac{\|\hat{\boldsymbol{\beta}}\|_0}{M_{\text{dim}}} \leq s^*, \quad 1 \leq q \leq 2,$$

with large probability [10, 39, 44, 4, 23, 46, 42, 6]. More precisely, for iid $N(0, \sigma^2)$ errors and $\lambda = A\sqrt{(2/n) \ln p}$, $A > 1$ for the the Lasso and $A = 1$ for the Dantzig selector, (9) holds with $M_{\text{pred}} = M_0/F_2$, $M_{\text{est}} = M_0/F_q$, and $M_{\text{dim}} = M_{\text{dim}}(m) = 1 + \{\kappa_+(m)/\kappa_-(m) - 1\}/(2 - 2\alpha_0)$ for the Lasso when $k \leq m/M_{\text{dim}}(m)$, where M_0 denotes a numerical constant, $k = \|\boldsymbol{\beta}\|_0$, $\alpha_0 \in (0, 1)$, and F_q are allowed to be $F_2 = \kappa_-(1.25k) - \theta_{1.25k, k}$ for the Dantzig selector and $q = 2$, $F_q = \text{RE}_q$ for $q \in \{1, 2\}$, or $F_q = \text{CIF}_q$ for $q \in [1, \infty]$.

Compared with the oracle $\hat{\boldsymbol{\beta}}^o$ in (3), the estimation loss of $\hat{\boldsymbol{\beta}}$ is inflated by a factor of no greater order than $\sqrt{\ln p}$, and the size of the selected model is of the same order as the true one. This inflation factor can be viewed as the cost of not knowing $\text{supp}(\boldsymbol{\beta})$. When $\ln(p/n) \asymp \ln p$, it has been proved in [42, 31] that (9) matches the order of the risk of a Bayes estimator for a class of (weak) signals close to zero, so that the order of this loss inflation factor $\sqrt{\ln p}$ is the smallest possible without further assumption on the strength of the signal $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ is strong (in the sense that its minimum nonzero coefficient is not close to zero), it is possible to achieve the oracle property, which removes the inflation factor. However, even in such cases, the logarithmic inflation is still present for the Lasso solution, and it is generally referred to as the *Lasso bias*; it means that the Lasso does not have the oracle property even when the signal is strong [15, 16]. Nonconvex penalty can be used to remedy this issue. For the Lasso and Dantzig selector, extensions of (9) have been established for capped- ℓ_1 sparse $\boldsymbol{\beta}$ and for $2 < q \leq \infty$ [44, 46, 42], under certain ℓ_q regularity conditions on \mathbf{X} [46, 42]. Error bounds of type (9) have been used in the analysis of the joint estimation of the noise level $\sigma^* := \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$ and $\boldsymbol{\beta}$ [33, 2, 34, 35]. For example, the scaled Lasso

$$\{\hat{\boldsymbol{\beta}}, \hat{\sigma}\} = \arg \min_{\{\mathbf{b}, \sigma\}} \{ \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 / (2n\sigma) + 2\sqrt{(\ln p)/n} \|\mathbf{b}\|_1 \}$$

provides $|\hat{\sigma}/\sigma^* - 1| = O_P(|S|(\ln p)/n)$ along with (9) under ℓ_2 regularity conditions [35]. If the penalty level is set at $\lambda = \hat{\sigma}2\sqrt{(\ln p)/n}$, the Lasso estimator becomes the square-root Lasso [3].

For variable selection, the Lasso is sign consistent in the event

$$(10) \quad \text{sgn}(\hat{\boldsymbol{\beta}}^o) = \text{sgn}(\boldsymbol{\beta}), \quad \min_{j \in S} |\hat{\beta}_j^o| \geq \theta_1^* \lambda, \quad \lambda \geq \frac{\sigma \sqrt{(2/n) \ln(p - |S|)}}{(1 - \theta_2^*)_+},$$

where $\theta_1^* = \|(\mathbf{X}_S^\top \mathbf{X}_S/n)^{-1} \text{sgn}(\boldsymbol{\beta}_S)\|_\infty$, $\theta_2^* = \|\mathbf{X}_{S^c}^\top \mathbf{X}_S (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} \text{sgn}(\boldsymbol{\beta}_S)\|_\infty$, $S = \text{supp}(\boldsymbol{\beta})$, and $\hat{\boldsymbol{\beta}}^o$ is the oracle estimator in (3) [27, 37, 50, 41]. Since $\|\hat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}\|_\infty = O_P(1)\sqrt{(\ln \|\boldsymbol{\beta}\|_0)/n} = o_P(\lambda)$ under mild conditions, θ_1^* and θ_2^* are key quantities in (10). For fixed $\kappa_0 < 1$, $\theta_2^* \leq \kappa_0$ is called the neighborhood stability/strong irrepresentable condition [27, 50]. For \mathbf{X} with

iid $N(0, \Sigma)$ rows and given S , θ_1^* and θ_2^* are within a small fraction of their population versions with Σ in place of $\mathbf{X}^\top \mathbf{X}/n$ [41]. For random β with $\|\beta\|_0 \lesssim n/\{\|\mathbf{X}\mathbf{X}^\top/p\|_2 \ln p\}$ and uniformly distributed $\text{sgn}(\beta)$ given $\|\beta\|_0$, $\theta_1^* \leq 2$ and $\theta_2^* \leq 1 - 1/\sqrt{2}$ with large probability under the incoherence condition $\max_{j \neq k} |\mathbf{x}_j^\top \mathbf{x}_k/n| \lesssim 1/(\ln p)$ [9]. It is worth mentioning that neither the incoherence condition nor the strong irrepresentable condition is ℓ_2 regular: in fact they may both fail with $\theta_2^* \asymp |S|^{1/2}$ and $\min_{j \in S} |\hat{\beta}_j^o| \geq \theta_1^* \lambda$ even in the classical low-dimensional setting of \mathbf{X} being rank p . Since $\theta_2^* \leq 1$ is necessary for the selection consistency of the Lasso under the first two conditions of (10) [37, 41], this means that Lasso is not model selection consistent under ℓ_2 regularity conditions. In order to achieve model selection consistency under ℓ_2 regularity, we have to employ a nonconvex penalty in (2).

For sparse estimation, ℓ_0 penalized LSE corresponds to the choice of $\rho(t; \lambda) = \lambda^2/2I(t \neq 0)$ in (2), and it was introduced in the literature [1, 25, 32] before Lasso. Formally,

$$(11) \quad \hat{\beta}^{(\ell_0)} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \left[\frac{1}{2n} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \frac{\lambda^2}{2} \|\mathbf{b}\|_0 \right].$$

This method is important for sparse recovery because with the Gaussian noise model $\varepsilon \sim N(0, \sigma^2 I)$, uniform distribution on support set, and appropriate flat distribution of β within support, it may be considered as a Bayesian procedure for support set recovery. However, this penalty is not easy to work with numerically because it is discontinuous at zero. The Lasso can be viewed as a convex surrogate of (11), but it does not achieve model selection consistency under ℓ_2 regularity, nor does it have the oracle property when the signal is uniformly strong.

Continuous concave penalties other than Lasso have been introduced to remedy these problems. These concave functions approximate ℓ_0 penalty better than the Lasso, and thus can remove the Lasso bias problem. Most concave penalties are interpolations between the Lasso and the ℓ_0 penalty. The earliest example in the literature is the ℓ_α (bridge) penalty [17] with $0 < \alpha < 1$. While the bridge penalty is continuous, its derivative is ∞ at $t = 0$, which may still cause numerical problems. In fact, the ∞ derivative value means that $\hat{\beta} = 0$ is always a local solution of (2) for bridge penalty, which prevents any possibility for the uniqueness of a reasonable local solution among sparse local solutions— a topic which we will investigate in this paper. In order to address this issue, additional penalty functions $\rho(t; \lambda)$ with finite derivatives at $t = 0$ have been suggested in the literature, such as the SCAD penalty [15], and the MCP [43]. These penalties can be written in a more general form as $\rho(t; \lambda) = \lambda^2 \rho(t/\lambda)$ with $\rho(0) = 0$ and $1 - t \leq (d/dt)\rho(t) \leq 1$ for $t > 0$. It can be verified that the ℓ_α penalty for $0 \leq \alpha \leq 1$, the SCAD, and the MCP are all concave in $[0, \infty)$. Another simple concave penalty is the capped- ℓ_1 penalty introduced in [47]. The penalties discussed in this paragraph are all described in Table 1 and plotted in Figure 1.

The above mentioned nonconvex interpolations of the ℓ_0 and ℓ_1 penalties typically gain smoothness over the ℓ_0 penalty and thus allow more computational options. Meanwhile, they may improve variable selection accuracy and gain oracle properties by reducing the

bias of the Lasso. A more direct way to reduce the bias of the Lasso is via the adaptive Lasso procedure [51], which solves the following weighted ℓ_1 regularization problem for some $\alpha \in (0, 1)$:

$$\min_{\mathbf{b} \in \mathbb{R}^p} \left[\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \sum_{j=1}^p |\hat{w}_j|^{-\alpha} |b_j| \right],$$

where \hat{w} is an estimator of $\boldsymbol{\beta}$ (for example, the solution of the standard unweighted Lasso with regularization parameter λ). A low-dimensional analysis in [51] showed that the adaptive Lasso solution can achieve the oracle property asymptotically. A high dimensional analysis of this procedure was given in [19]. For variable selection consistency and oracle properties to hold, the adaptive Lasso requires stronger conditions in terms of the minimum signal strength $\min_{j \in \text{supp}(\boldsymbol{\beta})} |\beta_j|$ than what is optimal. Specifically, the optimal requirement is $\min_{j \in \text{supp}(\boldsymbol{\beta})} |\beta_j| \geq \gamma \lambda_{univ}$ with $\lambda_{univ} = \sigma \sqrt{(2/n) \ln p}$ for some constant γ that may depend on an ℓ_2 regularity condition (also see Eq (12) below), which can be achieved by other procedures [43, 49]; however, the adaptive Lasso requires $\min_{j \in \text{supp}(\boldsymbol{\beta})} |\beta_j|$ to be significantly larger than the optimal order of λ_{univ} . This means the adaptive Lasso is sub-optimal for sparse estimation problems. We also observe that the adaptive Lasso does not directly minimize a concave loss function, and hence it is not an instance of (2). It was later noted that this procedure is only one iteration of using the so-called MM (majorization-minimization) principle to solve (2) with the bridge penalty [52]. The corresponding MM procedure is referred to as multi-stage convex relaxation in [47, 49]. For sparse estimation problem (2) with a penalty $\rho(t; \lambda)$ that is concave in $|t|$, this method iteratively invokes the solution of the following reweighted ℓ_1 regularization problem for stage $\ell = 1, 2, \dots$, starting with the initial value of $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$:

$$\hat{\boldsymbol{\beta}}^{(\ell)} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \left[\frac{1}{2n} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \sum_{j=1}^p \lambda_j^{(\ell)} |b_j| \right],$$

where $\lambda_j^{(\ell)} = (\partial/\partial t)\rho(t; \lambda)|_{t=|\hat{\beta}_j^{(\ell-1)}|}$ ($j = 1, \dots, p$). This procedure may be regarded as a multi-stage extension of the adaptive Lasso, which corresponds to the stage-2 solution $\hat{\boldsymbol{\beta}}^{(2)}$ with the bridge penalty. Unlike results for the adaptive Lasso, the results in [47, 49] for the multistage relaxation method allow $\min_{j \in \text{supp}(\boldsymbol{\beta})} |\beta_j|$ to achieve the optimal order of λ_{univ} , which match those of [43] and improve upon [19]. Moreover, only $\ell = O(\ln(\|\boldsymbol{\beta}\|_0))$ stages is necessary in order to achieve model selection consistency and oracle properties. It is worth pointing out that the multi-stage procedure can also be adapted to work with the Dantzig selector formulation [24].

For large p , the global solution of a nonconvex regularization method is hard to compute, so that local solutions are often used instead. Therefore theoretical analysis of nonconvex regularization has so far focused on specific numerical procedures that can find local solutions. For the ℓ_0 penalty, the penalized loss in (2) is typically evaluated for a subset of

the 2^p possible models $\text{supp}(\mathbf{b})$ such as those generated in stepwise regression. For smooth concave penalties, iterative algorithms can be used to find local minima of the penalized loss in (2) for a set of penalty levels [20, 52, 47, 5, 26, 49]. For the MCP and other quadratic spline concave penalties, a path following algorithm can be used to find local minima for an interval of penalty levels [43].

Advances have been carried out in the analysis of nonconvex regularization methods in multiple fronts [15, 16, 51, 19, 48, 43, 47, 6]. For concave penalized loss in (2), local minimizers exist with the oracle property (3) under mild conditions [15, 16]. However, it remains unclear whether there exist computationally efficient procedures that can find local minimizers investigated in [15, 16]. For the MCP, the local minima generated by the path following algorithm controls the estimation error and model size in the sense of (9) under an ℓ_2 regularity condition on \mathbf{X} [43]. Under an additional uniform signal strength condition

$$(12) \quad \min_{\beta_j \neq 0} |\widehat{\beta}_j^o| \geq \gamma \lambda_{univ} \geq \sup \{t : (\partial/\partial t)\rho(t; \lambda) \neq 0\}$$

with $\lambda_{univ} = \sigma \sqrt{(2/n) \ln p}$ and a certain constant $\gamma > 1$, the same path following solution has the oracle property (3) and thus the sign-consistency property [43]. Similar results hold for the SCAD and certain other quadratic spline penalties [43]. Under (12) and ℓ_2 regularity conditions on \mathbf{X} , the oracle property (3) and model selection consistency has also been established for a specific forward/backward stepwise regression scheme [48] that can be regarded as an approximate ℓ_0 penalty minimization algorithm. As we have mentioned earlier, the multi-stage relaxation scheme for minimizing (2) also leads to oracle inequality and model selection consistency under (12) and ℓ_2 regularity conditions on \mathbf{X} [47, 49].

While a number of specialized results were obtained for specialized numerical procedures under appropriate conditions, it is not clear what are the relationship among these solutions. For example, it is not clear whether the global solution of (2) is unique and whether it corresponds to solutions of various numerical procedures studied in the literature. This leads to a conceptual gap in the sense that it is not clear whether we should study specific local solutions as in the above mentioned previous work or we should try to solve (2) as accurately as possible (with the hope of finding the global solution). It is worth mentioning that related to this question, oracle inequalities involving global solutions with nonconvex penalties have been studied in the literature (for example, see related sections in [6]). However, such oracle inequalities do not lead to results comparable to those of [43, 47, 49]. Another relevant study is [21], which showed that in the lower dimensional scenario with $p \leq n$, the global solution of (2) agrees with the oracle estimator $\widehat{\beta}^o$ for the SCAD penalty when $\min_{\beta_j \neq 0} |\widehat{\beta}_j^o|$ is sufficiently large, and some other appropriate assumptions hold. However, their analysis does not directly generalize to the more complex high dimensional setting.

The purpose of the remaining of this paper is to present some general results showing that under appropriate ℓ_2 -regularity conditions, the global solution of an appropriate nonconvex regularization method leads to desirable recovery performance; moreover, under suitable

conditions, the global solution corresponds to the unique sparse local solution, which can be obtained via different numerical procedures. This leads to a unified view of concave high dimensional sparse estimation methods that can serve as a guideline for developing additional numerical algorithms.

3. A GENERAL THEORY

As we have discussed in our brief survey, concave regularized methods have been proven to control the estimation error and the dimension of the selected model (9) under ℓ_2 regularity conditions and possess the oracle property (3) or the sign-consistency property under the additional assumption (12). However, these results are established for specific local solutions of (2) with specific penalties. For $p > n$ it is still unclear if the global minimizer in (2) is identical to these local solutions or controls estimation and selection errors in a similar way.

In this paper, we unify the aforementioned results with the global solution of (2). We are mainly interested in two situations: ℓ_0 regularization where $\rho(t; \lambda)$ is discontinuous at $t = 0$, and smooth regularization which is continuous for all $t \geq 0$ and piece-wise differentiable. However, our basic results require only sub-additivity and monotonicity of $\rho(t; \lambda)$ in t in $[0, \infty)$. While the theorems are very technical, we present the general aims and a summary of main results at the beginning of each subsection. Before going into the main results, we will give some assumptions and definitions that are needed in the analysis.

3.1 General Assumptions and Definitions

In this subsection, we describe and discuss general conditions imposed in the rest of the paper. As we have pointed out, the key regularity conditions required in our analysis are expressed in terms of the sparse eigenvalues in (4) or invertibility factors RIF and CIF defined in (15) and (6). For the sake of clarity, we assume that these quantities are all constants, and this requirement is an ℓ_2 regularity condition. Another condition required by our analysis is called *null-consistency* (NC), which requires that if $\beta = 0$, then the global minimizer of (2) is achievable at $\hat{\beta} = 0$ (the actual condition, given in Assumption 2, is slightly stronger). Clearly this condition depends both on the matrix \mathbf{X} and on the noise vector ε . It is shown that under the standard sub-Gaussian noise assumption (see Assumption 1), the null-consistency condition is ℓ_2 regular. In summary, all assumptions on \mathbf{X} needed in our analysis are ℓ_2 regular; with this in mind, we may examine the technical details of the definitions and assumptions.

We first consider conditions on the regularizer $\rho(t; \lambda)$. We assume throughout the sequel the following conditions on the penalty function:

- (i) $\rho(0; \lambda) = 0$;
- (ii) $\rho(-t; \lambda) = \rho(t; \lambda)$;
- (iii) $\rho(t; \lambda)$ is non-decreasing in t in $[0, \infty)$;
- (iv) $\rho(t; \lambda)$ is subadditive with respect to t , $\rho(x + y; \lambda) \leq \rho(x; \lambda) + \rho(y; \lambda)$ for all $x, y \geq 0$.

This family of penalties is closed under the summation and maximization operations and includes all functions increasing and concave in $|t|$. Although we are mainly interested in the case where $\rho(t; \lambda)$ is concave in $|t|$, all of our results hold under the above specified weaker conditions, sometimes with side conditions such as the monotonicity of $\rho(t; \lambda)/t$ for $t > 0$ and the continuity of $\rho(t; \lambda)$ at $t = 0$. Therefore we will mention explicitly when such side conditions are needed.

We are particularly interested in the ℓ_0 regularization $\rho(t; \lambda) = (\lambda^2/2)I(t \neq 0)$ which is discontinuous at $t = 0$. In addition, we are interested in regularizer $\rho(t; \lambda)$ that is continuous in $t \geq 0$ and piece-wise differentiable. With such regularizers, local solutions of (2) can be defined as solutions with gradient zero. A local solution can be obtained using standard numerical procedures such as gradient descent.

Given a regularizer $\rho(t; \lambda)$ and any fixed $\lambda > 0$, we define the threshold level of the penalty as

$$(13) \quad \lambda^* := \inf_{t>0} \{t/2 + \rho(t; \lambda)/t\}.$$

The quantity λ^* is a function of λ that provides a natural normalization of λ . We call λ^* the threshold level since $\arg \min_t \{(z - t)^2/2 + \rho(t; \lambda)\} = 0$ iff $|z| \leq \lambda^*$. This can be easily seen from $(z - t)^2/2 + \rho(t; \lambda) - z^2/2 = t\{t/2 + \rho(t; \lambda)/t - z\}$. If $\rho(t; \lambda)$ is continuous at $t = 0$ and concave in $t \in (0, \infty)$, then $\lambda^* \leq \lim_{t \rightarrow 0+} (\partial/\partial t)\rho(t; \lambda)$. For simplicity, we may also require that $\rho(t; \lambda)$ be chosen such that $\lambda^* = \lambda$, which holds for the ℓ_0 , bridge, SCAD, MCP, and capped- ℓ_1 penalties. See Table 1 and Figure 1.

In the following, we will use the short-hand notation

$$\|\rho(\mathbf{b}; \lambda)\|_1 = \sum_{j=1}^p \rho(b_j; \lambda), \quad \forall \mathbf{b} = (b_1, \dots, b_p)^\top.$$

DEFINITION 2. *The following quantity bounds a general penalty via ℓ_1 penalty for sparse vectors:*

$$(14) \quad \Delta(a, k; \lambda) = \sup \left\{ \|\rho(\mathbf{b}; \lambda)\|_1 : \|\mathbf{b}\|_1 \leq ak, \|\mathbf{b}\|_0 = k \right\}.$$

PROPOSITION 1. *Let $\rho^*(t; \zeta) = \zeta|t| + (\zeta - |t|/2)_+^2/2$. Let λ^* be as in (13). Then,*

$$\min \{ \lambda^*|t|/2, (\lambda^*)^2/2 \} \leq \rho(t; \lambda) \leq \rho^*(t; \lambda^*).$$

Moreover, with $\gamma^* := \max_t \rho(t; \lambda)/(\lambda^*)^2$,

$$\Delta(a, k; \lambda) \leq k \min \{ \rho^*(a; \lambda^*), \gamma^*(\lambda^*)^2 \} \leq k\lambda^* \min \{ \max(a, 2\lambda^*), \gamma^*\lambda^* \}.$$

It follows from Proposition 1 that given a threshold level λ^* , all penalty functions satisfying general conditions (i)-(iv) are bounded by a capped- ℓ_1 penalty from below and the

maximum of the ℓ_0 and ℓ_1 penalties from above, up to a factor of 2. The function $\rho^*(t; \zeta)$ is a convex quadratic spline fit of $\max(\zeta^2/2, \zeta|t|)$, the maximum of the ℓ_0 and ℓ_1 penalties with threshold level ζ .

The trivial upper bound $\Delta(a, k; \lambda) \leq k\gamma^*(\lambda^*)^2$ is useful only for bounded penalties. For the ℓ_0 , capped- ℓ_1 , SCAD penalties and the MCP, the value of γ^* is given in Table 1. If $\rho(t; \lambda)$ is concave in $t \in [0, \infty)$, then $\Delta(a, k; \lambda) \leq k\rho(a; \lambda)$ by the Jensen inequality. For $a \geq 2\lambda^*$, $\Delta(a, k; \lambda) \leq a\lambda^*k$ matches the trivial bound for the ℓ_1 penalty, for which $\lambda = \lambda^*$.

Next, we consider conditions on the design matrix \mathbf{X} . Recall that \mathbf{X} is column normalized to $\|\mathbf{x}_j\|_2^2 = n$ for simplicity. Our analysis also depends on the sparse eigenvalues defined in (4) and the restricted invertibility factor defined as follows.

DEFINITION 3. For $q \geq 1$, $\xi > 0$ and $S \subset \{1, \dots, p\}$, we define the restricted invertibility factor as

$$(15) \quad \text{RIF}_q(\xi, S) = \inf \left\{ \frac{|S|^{1/q} \|\mathbf{X}^\top \mathbf{X} \mathbf{u}\|_\infty}{n \|\mathbf{u}\|_q} : \|\rho(\mathbf{u}_{S^c}; \lambda)\|_1 < \xi \|\rho(\mathbf{u}_S; \lambda)\|_1 \right\}.$$

The restricted invertibility factor is the quantity needed to separate conditions on \mathbf{X} and ε in our analysis. For $1 \leq q \leq 2$, sparse eigenvalues can be used to find lower bounds of $\text{RIF}_q(\xi, S)$.

PROPOSITION 2. Let CIF_q be as in (6). If $t/\rho(t; \lambda)$ is increasing in $t \in (0, \infty)$, then

$$(16) \quad \text{RIF}_q(\xi, S) \geq \inf_{|A|=|S|} \text{CIF}_q(\xi, A).$$

For the ℓ_1 penalty, $\text{RIF}_q = \text{CIF}_q$. If $\rho(t; \lambda)$ is concave in $t \in [0, \infty)$, then $t/\rho(t; \lambda)$ is increasing in t . Thus, Proposition 2 is applicable to all penalty functions discussed in Subsection 2.2, including the ℓ_0 , bridge, SCAD, MCP, and capped- ℓ_1 penalties.

REMARK 1. The CIF can be uniformly bounded from below in terms of sparse eigenvalues:

$$(17) \quad \text{CIF}_q(\xi, S) \geq \frac{I\{1 \leq q \leq 2\} \{\kappa_-(k + \ell) - (\xi/2)(k/\ell)^{1/2} \theta_{k+\ell, 4\ell}\}}{(1 + \xi)^{2/q-1} (1 + \xi^2 k / (4\ell))^{1-1/q} (1 + \ell/k)^{1/2}}$$

with $\theta_{k,\ell} \leq \kappa_+(k + \ell)$, for all $1 \leq \ell \leq (p - |S|)/5$ by Proposition 5 and (21) in [42], where $k = |S|$, $\theta_{k,\ell}$ is as in (5), and $\kappa_-(m)$ and $\kappa_+(m)$ are as in (4). For $(\xi, \ell, q) = (1, k/4, 2)$, $\text{CIF}_2(\xi, S) \geq \{\kappa_-(1.25k) - \theta_{1.25k, k}\} / \sqrt{2.5}$ bounds the UUP condition. For $(\xi, \ell, q) = (2, 2k, 2)$,

$$\text{CIF}_2(\xi, S) \geq \{\kappa_-(3k) - \theta_{3k, 8k} / \sqrt{2}\} / \sqrt{4.5}.$$

REMARK 2. *It follows from Proposition 2 and Remark 1 that conditions $\text{RIF}_q(\xi, S) > 0$ and $1/\text{RIF}_q(\xi, S) = O(1)$ are both ℓ_2 -regularity conditions on \mathbf{X} for $1 \leq q \leq 2$. Moreover, $\text{rank}(\mathbf{X}) = p$ implies $\text{RIF}(\xi, S) > 0$. To check the ℓ_2 regularity of these conditions, we suppose that the rows of \mathbf{X} are iid from $N(0, \Sigma)$ with all eigenvalues of Σ in $[c_1, c_2] \subset (0, \infty)$. Then, $c_1/2 \leq \kappa_-(m)$ and $\kappa_+(m) \leq 2c_2$ with at least probability $1 - \delta \in [0, 1)$ for $m \leq c_3 n / \ln(p/\delta)$ for a certain $c_3 > 0$. Let $c_4 = \{c_1/(\xi c_2)\}^2$. In this event, setting $k = s^*$ and $\ell = (m - s^*)/5$ in (17) yields*

$$\min_{|S| \leq s^*} \text{RIF}_2(\xi, S) \geq \min_{|S| \leq s^*} \text{CIF}_2(\xi, S) \geq (c_1/4) / \sqrt{(1 + \xi^2 c_4/4)(1 + 1/c_4)}$$

when $5s^*/(m - s^*) < c_4$ for some $m \leq c_3 n / \ln(p/\delta)$, which holds when $(s^*/n) \ln(p/\delta) \leq c_3/(1 + 5/c_4)$.

Finally, we consider conditions on the error vector.

ASSUMPTION 1. *An error vector $\boldsymbol{\varepsilon}$ is sub-Gaussian with noise level $\tilde{\sigma}$ if for all $t \geq 0$:*

$$P(|\mathbf{u}^\top \boldsymbol{\varepsilon}| > \tilde{\sigma} t) \leq \exp(-t^2/2)$$

for all vector \mathbf{u} with $\|\mathbf{u}\|_2 = 1$.

The above sub-Gaussian assumption implies that there exists a universal constant $c_0 \geq 1$ such that for $\sigma = c_0 \tilde{\sigma}$ we have:

$$P(\|\mathbf{P}_A \boldsymbol{\varepsilon}\|_2 / |A|^{1/2} > \sigma(1 + t)) \leq \exp(-|A|t^2/2)$$

for all subsets $A \subset \{1, \dots, p\}$, where \mathbf{P}_A is the orthogonal projection to the range of \mathbf{X}_A (that is, $\mathbf{P}_A = \mathbf{X}_A \mathbf{X}_A^\dagger$, where \mathbf{X}_A^\dagger is the Moore-Penrose generalized inverse of \mathbf{X}_A).

The above sub-Gaussian condition holds with $\boldsymbol{\varepsilon} \sim N(0, \tilde{\sigma}^2 \mathbf{I}_{n \times n})$ and $\sigma = \tilde{\sigma}$. Although the second part of the assumption follows from the first part, we list it in the assumption for convenience. Moreover, we will ignore the constant c_0 in the subsequent discussion and simply replace the noise level $\tilde{\sigma}$ by σ in the following. As we have mentioned in Section 3, what we really need is a null-consistency condition, which we give below. The sub-Gaussian condition will be used to verify the NC condition.

ASSUMPTION 2. *Let $\eta \in (0, 1]$. We say that the regularization method (2) satisfies the η null-consistency condition (η -NC) if the following equality holds:*

$$(18) \quad \min_{\mathbf{b} \in \mathbb{R}^p} \left(\|\boldsymbol{\varepsilon}/\eta - \mathbf{X}\mathbf{b}\|_2^2 / (2n) + \|\rho(\mathbf{b}; \lambda)\|_1 \right) = \|\boldsymbol{\varepsilon}/\eta\|_2^2 / (2n).$$

Given $\eta = 1$, the NC condition means that if $\boldsymbol{\beta} = 0$, then the global minimizer of (2) is achievable at $\hat{\boldsymbol{\beta}} = 0$. This requirement is clearly necessary for the global minimizer of (2) to satisfy the error bound (20) in Theorem 1 below for $|S| = 0$. Here, we also allow a slightly stronger condition with $\eta < 1$, which requires $\hat{\boldsymbol{\beta}} = 0$ for $\boldsymbol{\beta} = 0$ when the noise $\boldsymbol{\varepsilon}$ is proportionally inflated by $1/\eta$.

PROPOSITION 3. Suppose that ε is sub-Gaussian with noise level σ , $0 < \delta \leq 1$ and $\zeta_0 > 0$. Suppose $\rho(t; \lambda) \geq ((\lambda^*)^2/2) \wedge (\lambda^*|t|)$ with $\lambda^* \geq (1 + \zeta_0)(\sigma/\eta)n^{-1/2}(1 + \sqrt{2 \ln(2p/\delta)})$. Then, (2) satisfies the η -NC condition with at least probability $2 - e^{\delta/2} - \exp(-n(1 - 1/\sqrt{2})^2)$, provided that

$$(19) \quad \max \left\{ \lambda_{\max}^{1/2}(\mathbf{X}_B^\top \mathbf{P}_A \mathbf{X}_B/n) : \begin{array}{l} B \cap A = \emptyset, |A| = \text{rank}(\mathbf{P}_A) = |B| = k, \\ k(1 + \zeta_0)^2(1 + \sqrt{2 \ln(2p/\delta)})^2 \leq 2n \end{array} \right\} \leq \zeta_0.$$

Moreover, (19) holds with no smaller probability than $1 - \delta^4/(16p^2)$ if the rows of \mathbf{X} are iid from $N(0, \Sigma)$ and $\sqrt{8}\lambda_{\max}^{1/2}(\Sigma) \leq \zeta_0(1 + \zeta_0)$. This means that under the sub-Gaussian condition on ε , the η -NC condition is ℓ_2 -regular.

REMARK 3. The condition $\rho(t; \lambda) \geq \min(\lambda^2/2, \lambda|t|)$ holds for the ℓ_0 , ℓ_1 , SCAD, and capped ℓ_1 penalties in Table 1, so that Proposition 3 is directly applicable with $\lambda = \lambda^*$. In general, the condition of Proposition 3 holds for all penalties considered in this paper when the threshold level in (13) satisfies $\lambda^* \geq 2(1 + \zeta_0)(\sigma/\eta)n^{-1/2}(1 + \sqrt{2 \ln(2p/\delta)})$, in view of the lower bound of $\rho(t; \lambda)$ in Proposition 1. For ℓ_0 and ℓ_1 penalties, we may set $\zeta_0 = 0$ in Proposition 3 (the extra condition (19) is not necessary). The simplified condition for ℓ_0 penalty is explicitly given in Theorem 3. For the ℓ_1 penalty, the η -NC condition is equivalent to $\|\mathbf{X}^\top \varepsilon\|_\infty \leq \eta\lambda n$.

3.2 Basic Properties of the Global Solution

We now turn our attention to the global solution of (2) with a general subadditive nondecreasing regularizer $\rho(t; \lambda)$.

Theorem 1 below gives ℓ_q -norm error bounds for $\|\widehat{\beta} - \beta\|_q$ and a bound of the prediction error $\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta\|_2$ that are comparable with known results for ℓ_1 regularization. This means that under appropriate ℓ_2 regularity conditions, the global solution of concave regularization problems are no worse than the Lasso in terms of the order of estimation and prediction errors. Theorem 2 below shows that the global optimal solution of (2) is sparse, and under appropriate ℓ_2 regularity conditions, the sparsity is of the same order as $\|\beta\|_0$; that is, $\|\widehat{\beta}\|_0 = O(\|\beta\|_0)$. Thus, (9) holds for the global solution of (2). Moreover, if the second order derivative of $\rho(t; \lambda)$ with respect to t is sufficiently small, then the global solution is also the unique sparse local solution of (2). That is, if a vector $\widetilde{\beta}$ is a local solution of (2) which is sparse: $\|\widetilde{\beta}\|_0 = O(\|\beta\|_0)$, then $\widetilde{\beta}$ is the global solution of (2).

For uniformly bounded penalty functions, the prediction error bound requires only the η -NC condition and the sparsity of the global solution requires only an additional condition on the upper sparse eigenvalue of \mathbf{X} . Thus, by Proposition 3, no condition on the lower sparse eigenvalue is required for the prediction error bound and the sparsity of the global solution. These results are stated as Corollary 1 and Corollary 2 (i).

None of the results in this section require that $\min_{\beta_j \neq 0} |\widehat{\beta}_j^o|$ to be bounded away from zero. Furthermore, since these results require only ℓ_2 regularity conditions, they apply to the case of $p \gg n$ as long as $s^*(\ln p)/n$ is small.

We now present detailed technical statements. First we consider the estimation of $\mathbf{X}\beta$ and β .

THEOREM 1. *Let $S = \text{supp}(\beta)$, $\widehat{\beta}$ be as in (2), λ^* as in (13), and $\text{RIF}_q(\xi, S)$ as in (15). Consider $\eta \in (0, 1)$, and $\xi = (\eta + 1)/(1 - \eta)$, and assume that the η -NC condition (18) holds. Then for all $q \geq 1$:*

$$(20) \quad \|\widehat{\beta} - \beta\|_q \leq (1 + \eta)\lambda^*|S|^{1/q}/\text{RIF}_q(\xi, S),$$

and with $a_1 = (1 + \eta)/\text{RIF}_1(\xi, S)$, γ^* in Proposition 1, and $\Delta(a, k; \lambda)$ in (14),

$$(21) \quad \|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta\|_2^2/n \leq 2\xi\Delta(a_1\lambda^*, |S|; \lambda) \leq 2\xi \min\{(a_1 \vee 2), \gamma^*\}(\lambda^*)^2|S|.$$

By using the bound $\Delta(a_1\lambda^*, |S|; \lambda) \leq |S|\gamma^*(\lambda^*)^2$, we obtain the following corollary.

COROLLARY 1. *Consider penalties $\rho(t; \lambda)$ indexed by the threshold level; $\lambda^* = \lambda$ in (13). Suppose (18) holds. Let $S = \text{supp}(\beta)$ and $\gamma^* = \max_t \rho(t; \lambda)/\lambda^2$. Then,*

$$\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta\|_2^2/n \leq 2\{(1 + \eta)/(1 - \eta)\}\gamma^*\lambda^2|S|.$$

REMARK 4. *Corollary 1 can be readily applied to the ℓ_0 , capped- ℓ_1 , MCP, and SCAD penalties described in Table 1, where γ^* is also specified. It is worthwhile to note that the prediction error bound in Corollary 1 does not depend on \mathbf{X} , provided that penalty is large enough to guarantee NC. For the ℓ_0 penalty, the NC requires only $\|\mathbf{x}_j\|_2 = \sqrt{n}$ on \mathbf{X} , which we assume anyway. For other concave penalties in Corollary 1, we are only able to provide NC in Proposition 3 under a mild condition on the upper eigenvalue of $\mathbf{X}_B^\top \mathbf{P}_A \mathbf{X}_B/n$, but not on the sparse lower eigenvalue of the Gram matrix. In contrast, sparse lower eigenvalue condition of the Gram matrix (such as RE_1 or CIF_1) is needed for the Lasso. This presents a benefit of concave regularization in prediction, compared with ℓ_1 , at least from an analytical point of view.*

Next we provide an upper bound for the sparseness of $\widehat{\beta}$ based on Theorem 1 and the maximum sparse eigenvalue $\kappa_+(m)$. We denote by $\dot{\rho}(t; \lambda) = (\partial/\partial t)\rho(t; \lambda)$ any value between the left- and right- derivatives of $\rho(\cdot; \lambda)$ and assume the left- and right-differentiability of $\rho(\cdot; \lambda)$ whenever the notation $\dot{\rho}(t; \lambda)$ is invoked. For example, if $\rho(t; \lambda) = \lambda|t|$, then $\dot{\rho}(0\pm; \lambda) = \pm\lambda$ and $\dot{\rho}(0; \lambda)$ can be any value in $[-\lambda, \lambda]$ (which in all of our results, can be chosen as the most favorable value unless explicitly mentioned otherwise).

THEOREM 2. *Let $\{S, \widehat{\beta}, \lambda^*, \eta, \xi, a_1\}$ and $\Delta(a, k; \lambda)$ be as in Theorem 1, and $\widehat{S} = \text{supp}(\widehat{\beta})$. Suppose that the η -NC condition (18) holds. Consider $t_0 \geq 0$ and integer $m_0 \geq 0$ satisfying $m_0 = 0$ for $t_0 = 0$ and*

$$(22) \quad \sqrt{2\xi\kappa_+(m_0)\Delta(a_1\lambda^*, |S|; \lambda)/m_0} + \|\mathbf{X}^\top \varepsilon/n\|_\infty < \inf_{0 < s < t_0} \dot{\rho}(s; \lambda)$$

for $t_0 > 0$. Then,

$$(23) \quad |\widehat{S} \setminus S| < m := m_0 + \lfloor \xi \Delta(a_1 \lambda^*, |S|; \lambda) / \rho(t_0; \lambda) \rfloor.$$

The η -NC condition implies $\|\mathbf{X}^\top \boldsymbol{\varepsilon} / n\|_\infty \leq \eta \lambda^*$ by Lemma 1 in Section 5. If $\rho(t; \lambda)$ is concave in $t > 0$, then the right-hand side of (22) can be replaced by $\dot{\rho}(t_0; \lambda)$ and $\rho(t_0; \lambda) \geq t_0 \dot{\rho}(t_0; \lambda)$. These facts give the following corollary for ℓ_∞ bounded and ℓ_1 penalties.

COROLLARY 2. (i) Let $\rho(t; \lambda)$ and γ^* be as in Corollary 1. Suppose (2) is η -NC in the sense of (18) and $\dot{\rho}(a_0 \lambda; \lambda) \geq \lambda(1 - a_1/\gamma)$ for some $a_0 > 0$ and $a_1 \geq 0$. If $m_0 = \alpha |S|$ is an integer and $2\gamma^* \kappa_+(\alpha |S|) / \alpha < (1 - a_1/\gamma - \eta)^2(1 - \eta)/(1 + \eta)$, then

$$(24) \quad |\widehat{S} \setminus S| < m := \left(\alpha + \frac{\gamma^*/a_0}{1 - a_1/\gamma} \right) |S|.$$

(ii) Let $\widehat{S}^{(\ell_1)} = \text{supp}(\widehat{\boldsymbol{\beta}}^{(\ell_1)})$ with the Lasso (8) and CIF_q as in (6). In the event $\|\mathbf{X}^\top \boldsymbol{\varepsilon} / n\|_\infty \leq \eta \lambda$,

$$(25) \quad \frac{2\kappa_+(\alpha |S|) / \alpha}{\text{CIF}_1((1 + \eta)/(1 - \eta), S)} < \frac{(1 - \eta)^3}{(1 + \eta)^2} \Rightarrow |\widehat{S}^{(\ell_1)} \setminus S| < m := \alpha |S|.$$

REMARK 5. Theorem 2 and Corollary 2 imply that the global solution $\widehat{\boldsymbol{\beta}}$ in (2) is sparse under appropriate assumptions. Since the η -NC condition follows from a bound on the upper sparse eigenvalue in Proposition 3, Corollary 2 (i) asserts that for concave penalties with finite γ^* , the sparsity of the global solution does not require a condition on the lower sparse eigenvalue. For ℓ_0 regularization, we may take $m_0 = t_0 = 0$ with the convention $\kappa_+(0)/0 = 0$ in (22). The Lasso also satisfies the dimension bound $|\widehat{S} \setminus S| < m \vee 1$ under the SRC: $\{\kappa_+(m + |S|) / \kappa_-(m + |S|) - 1\} / (2 - 2a_0) \leq m / |S|$ with an $a_0 \in (0, 1)$, provided that $\lambda \geq (1 + o(1)) \{\kappa_+^{1/2}(m) / a_0\} \sigma \sqrt{(2/n) \ln p}$ [43]. An advantage of (25) is to allow an λ not dependent on the upper sparse eigenvalue of the design for sub-Gaussian $\boldsymbol{\varepsilon}$.

REMARK 6. Let $\kappa^* = \sup_{0 < s < t} \{\dot{\rho}(t; \lambda) - \dot{\rho}(s; \lambda)\} / (t - s)$ be the maximum concavity of the penalty. Suppose $\kappa_-(|S| + m + \tilde{m} - 2) > \kappa^*$. Then, the penalized loss $L_\lambda(\mathbf{b})$ in (2) is convex in all models $\text{supp}(\mathbf{b}) = A$ with $|A \setminus S| \leq m + \tilde{m} - 2$. This condition has been called sparse convexity [43]. If m is as in (23) or (24) and $\widehat{\boldsymbol{\beta}}$ is a local solution of (2) with $\#\{j \notin S : \tilde{\beta}_j \neq 0\} < \tilde{m}$, then the local solution must be identical to the global solution.

REMARK 7. Consider penalties with $\lambda^* = \lambda$ which holds for all penalties in Table 1. Let $\eta \in (0, 1)$ and $\lambda_* > 0$ be fixed. Suppose Theorem 2 or Corollary 2 is applicable with $m \leq \alpha^* |S|$ for a fixed constant α^* and all $\lambda \geq \lambda_*$. Suppose in addition $\dot{\rho}(t; \lambda)$ is continuous in $1/\lambda \in [0, 1/\lambda_*]$ uniformly in bounded sets of t . Under the sparse convexity condition $\kappa_-(|S| + m - 1) \geq \kappa_* > 0$, with the maximum concavity κ^* in Remark 6, the global solution

forms a continuous path in \mathbb{R}^p as a function of $1/\lambda \geq 1/\lambda_*$. This path is identical to the output of the path following algorithm in [43] if it starts with $\widehat{\boldsymbol{\beta}} = 0$ at $1/\lambda = 0$. We will show in Theorem 7 that gradient algorithms beginning from the Lasso may also yield the global solution under the sparse convexity condition.

As simple examples to illustrate Corollaries 1 and 2, we consider the capped- ℓ_1 penalty and the MCP with $\lambda = \lambda^*$ and γ^* in Table 1. For the capped- ℓ_1 penalty with $a_0 = \gamma/2$ in Corollary 2,

$$\begin{aligned} \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|_2^2/n &\leq \lambda^2|S|\gamma(1+\eta)/(1-\eta), \\ \gamma\kappa_+(\alpha|S|) &\leq \alpha(1-\eta)^3/(1+\eta) \Rightarrow |\widehat{S} \setminus S| < (\alpha+1)|S|. \end{aligned}$$

For the MCP with $\alpha_0 = \gamma/3$, the same prediction bound holds and

$$\gamma\kappa_+(\alpha|S|) \leq \alpha(2/3 - \eta)^2(1-\eta)/(1+\eta) \Rightarrow |\widehat{S} \setminus S| < (\alpha + 9/4)|S|.$$

Note that generally speaking, unless stronger conditions are imposed, Theorem 2 only implies that $|\widehat{S} \setminus S| = O(|S|)$ but not $|\widehat{S} \setminus S| = 0$ required for model selection consistency. The model selection consistency will be studied later in the paper.

3.3 The Global Solution of ℓ_0 Regularization

This subsection considers the global optimal solution $\widehat{\boldsymbol{\beta}}^{(\ell_0)}$ of ℓ_0 regularization in (11). We are interested in two results: sparsity of the global solution and its model selection quality. For clarity, the two results are separately stated in two theorems. First, it is shown that the global solution of ℓ_0 regularization is sparse. Moreover, with sub-Gaussian noise, the prediction error bound for ℓ_0 penalty in Theorem 3 below does not depend on properties of the design matrix \mathbf{X} . This significantly improves upon the corresponding existing results for the Lasso and Dantzig selector, which requires a non-trivial RE_1 or CIF_1 condition on the design matrix \mathbf{X} . It also improves upon Corollary 1 since the η -NC condition holds explicitly for $\lambda \geq (\sigma/\eta)(1 + \sqrt{2 \ln(p/\delta)})/\sqrt{n}$. If a certain lower sparse eigenvalue of $\mathbf{X}^\top \mathbf{X}/n$ is bounded from below and the uniform signal strength condition (12) holds, then we obtain in Theorem 4 below the selection consistency for ℓ_0 regularization, which implies the oracle property.

We can now describe our first result, which says that under appropriate conditions, the global solution of ℓ_0 regularization is sparse.

THEOREM 3. *If for all $\mathbf{b} \in \mathbb{R}^p$: $\boldsymbol{\varepsilon}^\top \mathbf{X}\mathbf{b} \leq \lambda\eta\sqrt{n}\|\mathbf{b}\|_0\|\mathbf{X}\mathbf{b}\|_2$ for some $\eta < 1$, then (11) satisfies the η -NC condition. It implies that the global optimal solution of (11) satisfies*

$$\|\widehat{\boldsymbol{\beta}}^{(\ell_0)}\|_0 \leq \frac{1+\eta^2}{1-\eta^2}\|\boldsymbol{\beta}\|_0, \quad \|\mathbf{X}\widehat{\boldsymbol{\beta}}^{(\ell_0)} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq \frac{(1+\eta)\lambda^2\|\boldsymbol{\beta}\|_0}{1-\eta}.$$

We also have the following result about model selection quality for ℓ_0 regularization.

THEOREM 4. *Assume that the assumption of Theorem 3 holds. Let $s = 2\|\boldsymbol{\beta}\|_0/(1 - \eta^2)$ and $\widehat{\boldsymbol{\beta}}^o$ be as in (3). Suppose $\|\mathbf{X}^\top(\mathbf{P}_S\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon})\|_\infty/n \leq \sqrt{2\kappa_-(s)}\lambda$, where \mathbf{P}_S is the orthogonal projection to the range of \mathbf{X}_S . Let $S = \text{supp}(\boldsymbol{\beta})$, $\delta^o = \#\{j \in S : |\widehat{\beta}_j^o| < \lambda\sqrt{2/\kappa_-(s)}\}$, and $\widehat{S} = \text{supp}(\widehat{\boldsymbol{\beta}}^{(\ell_0)})$. Then,*

$$|S - \widehat{S}| + 0.5|\widehat{S} - S| \leq 2\delta^o, \quad \|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{(\ell_0)} - \widehat{\boldsymbol{\beta}}^o)\|_2^2 \leq 2\lambda^2\delta^o.$$

If the error $\boldsymbol{\varepsilon}$ is sub-Gaussian in the sense of Assumption 1, then the condition of Theorems 3 and 4 holds with at least probability $2 - e^\delta$ for $\lambda \geq (\sigma/\eta)(1 + \sqrt{2\ln(p/\delta)})/\sqrt{n}$. Theorem 4 implies that model selection consistency can be achieved if the condition $\min_{j \in \text{supp}(\boldsymbol{\beta})} |\widehat{\beta}_j^o| \geq \lambda/\sqrt{\kappa_-(s)}$ holds, which implies that $\delta^o = 0$.

3.4 Approximate Local Solutions

This section considers penalties $\rho(t; \lambda)$ which are both left- and right-differentiable, for which one can define (approximate) local solutions that are what numerical optimization procedures compute. We provide sufficient conditions for the uniqueness of the sparse local solution of (2) and its equality to both the oracle least squares and global solutions.

Theorem 5 below considers the distance between two approximate local solutions. An immediate consequence of the result says that under appropriate assumptions, there is a unique sparse local solution of (2) that corresponds to the oracle least squares solution $\widehat{\boldsymbol{\beta}}^o$. Therefore the unique local solution has the oracle property. Moreover, this unique local solution has to be the global optimal solution according to Theorem 2. While Theorem 5 shows that it is possible for a penalty that is not second order differentiable to have a unique sparse local solution, it requires the uniform signal strength condition (12) for such penalties. In contrast, with a second order differentiable concave penalty, (12) is not needed in Theorem 5 for sparse local solutions to be unique. This suggests an advantage for using smooth concave penalties which may lead to fewer local solutions under certain conditions.

Theorem 6 below gives sufficient conditions under which the global optimal solution of (2) achieves model selection consistency. These sufficient conditions generalize the irrerepresentable condition (10) for the model selection consistency of the Lasso. However, unlike the irrerepresentable condition for the Lasso, which is not an ℓ_2 regularity condition, for a concave penalty where $(\partial/\partial t)\rho(t; \lambda) = 0$ for sufficiently large t , the generalized irrerepresentable condition required in Theorem 6 automatically holds when $\min_{\beta_j \neq 0} |\widehat{\beta}_j^o|$ is not too small and thus it is trivially an ℓ_2 regular condition. Moreover, for appropriate nonconvex penalties, it is possible to achieve a selection threshold of optimal order as in the uniform signal strength condition (12).

Suppose $\rho(t; \lambda)$ is both left- and right-differentiable. Given an excess $\nu \geq 0$, a vector $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$ is an approximate local solution (ALS) of (2) if

$$(26) \quad \|\mathbf{X}^\top(\mathbf{X}\widetilde{\boldsymbol{\beta}} - \mathbf{y})/n + \dot{\rho}(\widetilde{\boldsymbol{\beta}}; \lambda)\|_2^2 \leq \nu.$$

This $\tilde{\beta}$ is a local solution if $\nu = 0$. Note that by convention, $\dot{\rho}(t; \lambda)$ can be chosen to be any value between $\dot{\rho}(t_-; \lambda)$ and $\dot{\rho}(t_+; \lambda)$ to satisfy the equation. In this subsection, we provide estimates of distances among the ALS of (2) and use them to prove the equality of oracle approximate local and global solutions of (2). This gives the selection consistency of the global solution studied in Subsection 4.2. The oracle LSE is considered as an ALS. In addition, we define a sufficient condition for the existence of a sign consistent local solution which generalizes the irrepresentable condition for the Lasso selection and becomes an ℓ_2 regularity condition on \mathbf{X} for a broad class of concave penalties.

We first provide estimates of distances among the ALS of (2). We use the following function $\theta(t, \kappa)$ to measure the degree of nonconvexity of a regularizer $\rho(t; \lambda)$ at $t \in \mathbb{R}$. To our knowledge, this is the first time that it is introduced explicitly.

DEFINITION 4. For $\kappa \geq 0$ and $t \in \mathbb{R}$, define

$$\theta(t, \kappa) := \sup_s \{-\text{sgn}(s - t)(\dot{\rho}(s; \lambda) - \dot{\rho}(t; \lambda)) - \kappa|s - t|\}.$$

Moreover, given $\mathbf{u} = (u_1, \dots, u_p)^\top \in \mathbb{R}^p$, we let $\theta(\mathbf{u}, \kappa) = [\theta(u_1, \kappa), \dots, \theta(u_p, \kappa)]$.

We are mostly interested in values of $\theta(t, \kappa)$ that achieves zero. We note that $\theta(t, \kappa) = 0$ for convex $\rho(t, \lambda)$ with $\kappa \geq 0$. More generally, let κ^* be the maximum concavity as in Remark 6. Then, $\theta(t, \kappa) = 0$ for all t iff $\kappa \geq \kappa^*$. For $\dot{\rho}(t_+; \lambda) < \dot{\rho}(t_-; \lambda)$, $\theta(t, \kappa) > 0$ for all finite κ . However, we only need $\theta(t, \kappa) = 0$ for a proper set of t in our selection consistency theory. As an example, for $\kappa = 2/\gamma$, the capped- ℓ_1 penalty $\rho(t; \lambda) = \min(\gamma\lambda^2/2, \lambda|t|)$ gives $\theta(t, \kappa) = 0$ when either $t = 0\pm$ or $|t| \geq \gamma\lambda$.

The following theorem shows that under appropriate assumptions, two sparse approximate local solutions $\tilde{\beta}^{(1)}$ and $\tilde{\beta}^{(2)}$ are close.

THEOREM 5. Let $\tilde{\beta}^{(j)}$ be ALS of (2) with excess $\nu^{(j)}$ and $\mathbf{\Delta} = \tilde{\beta}^{(1)} - \tilde{\beta}^{(2)}$. Let $\kappa_{\pm}(\cdot)$ be the sparse eigenvalues in (4) and $\tilde{S}^{(j)} := \text{supp}(\tilde{\beta}^{(j)})$. Consider any $S \subset \{1, \dots, p\}$ with $k = |S|$, integer m such that $m + k \geq |\tilde{S}^{(1)} \cup \tilde{S}^{(2)}|$, and $0 < \kappa < \kappa_-(m + k)$. Then,

$$(27) \quad \|\mathbf{X}\mathbf{\Delta}\|_2^2/n \leq \frac{2\kappa_-(m + k)}{(\kappa_-(m + k) - \kappa)^2} \left\{ \|\theta(|\tilde{\beta}_{\tilde{S}^{(1)}}^{(1)}|, \kappa)\|_2^2 + |\tilde{S}^{(2)} \setminus \tilde{S}^{(1)}| \theta^2(0+, \kappa) + \nu \right\}$$

with $\nu = \{(\nu^{(1)})^{1/2} + (\nu^{(2)})^{1/2}\}^2$, and

$$(28) \quad |S \setminus \tilde{S}^{(2)}| \leq \inf_{\lambda_0 > 0} \left[\#\left\{ j \in S : |\tilde{\beta}_j^{(1)}| < \lambda_0 / \sqrt{\kappa_-(m + k)} \right\} + \|\mathbf{X}\mathbf{\Delta}\|_2^2 / (\lambda_0^2 n) \right].$$

If in addition $\theta(0+, \kappa) = 0$ and $\dot{\rho}(0+; \lambda) > \|\mathbf{X}_{S^c}^\top(\mathbf{X}\tilde{\beta}^{(1)} - \mathbf{y})/n\|_\infty$ with $S \supseteq S^{(1)}$ and $|S| \geq k$, then

$$(29) \quad |\tilde{S}^{(2)} \setminus S| \leq \frac{3[\{\kappa^2/\kappa_-(m + k) + \kappa_+(m)\} \|\mathbf{X}\mathbf{\Delta}\|_2^2/n + \tilde{\nu}^{(2)}]}{\{\dot{\rho}(0+; \lambda) - \|\mathbf{X}_{S^c}^\top(\mathbf{X}\tilde{\beta}^{(1)} - \mathbf{y})/n\|_\infty\}^2}.$$

Let $S = \text{supp}(\boldsymbol{\beta})$. For comparison between a sparse local or global solution $\tilde{\boldsymbol{\beta}}^{(2)}$ with $|\tilde{S}^{(2)} \setminus S| \leq m$ and an oracle solution $\tilde{\boldsymbol{\beta}}^{(1)}$ with $\tilde{S}^{(1)} = S$, the sparse convexity condition implies $\tilde{\boldsymbol{\beta}}^{(2)} = \tilde{\boldsymbol{\beta}}^{(1)}$ when $\kappa^* < \kappa_-(|S| + m)$ as in Remark 6. However, since $\kappa^* = \infty$ when $\dot{\rho}(t+; \lambda) < \dot{\rho}(t-; \lambda)$ at a point $t > 0$, the sparse convexity argument requires the continuity of $\dot{\rho}(t; \lambda)$ for $t > 0$. This does not apply to the capped- ℓ_1 penalty. In Theorem 5, if $\theta(0+, \kappa) = \theta(\tilde{\boldsymbol{\beta}}_S^{(1)}, \kappa) = 0$ with $\kappa < \kappa_-(|S| + m)$, then $\mathbf{X}\boldsymbol{\Delta} = 0$, and hence $\tilde{\boldsymbol{\beta}}^{(2)} = \tilde{\boldsymbol{\beta}}^{(1)}$ (since $\kappa_-(|\tilde{S}^{(1)} \cup \tilde{S}^{(2)}|) > 0$). Thus, the sparse convexity condition is much weakened to cover all left- and right-differentiable penalties such as the capped- ℓ_1 . On the other hand, Theorem 5 does not weaken the sparse convexity condition for the MCP, for which $\theta(0+; \kappa) = 0$ iff $\kappa \geq \kappa^* = 1/\gamma$ iff $\theta(t; \kappa) = 0$ for all $t > 0$. It is worth pointing out that for a piecewise differentiable penalty that is not second order differentiable, the condition $\theta(\tilde{\boldsymbol{\beta}}_S^{(1)}, \kappa) = 0$ (thus, the uniqueness of local solution) typically requires $|\tilde{\beta}_j^{(1)}|$ to be large to avoid the discontinuities of $\dot{\rho}(t; \lambda)$ when $j \in S$. As pointed out in Remark 6, this is not necessary when the penalty is second order differentiable. This means that there can be advantages of using smooth penalty terms that may have fewer local minimizers under certain conditions.

As a simple working example to illustrate Theorem 5, we consider the capped ℓ_1 penalty. Let $S = \text{supp}(\boldsymbol{\beta})$. Assume that $\kappa = \kappa_-(m + |S|)/2 \geq 2/\gamma$. Then $\theta(t, \kappa) = 0$ when either $t = 0 \pm$ or $|t| \geq \gamma\lambda$. Therefore, if we define $\tilde{\boldsymbol{\beta}}^{(1)}$ as $\tilde{\beta}_j^{(1)} = \hat{\beta}_j^o$ when $|\hat{\beta}_j^o| \geq \gamma\lambda$ and $\tilde{\beta}^{(j)} = 0$ otherwise, then

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/n \leq \frac{8\nu}{\kappa_-(m + |S|)},$$

and by taking $\lambda_0 = \gamma\lambda\sqrt{\kappa_-(m + |S|)}$, we have

$$|S \setminus \tilde{S}^{(2)}| \leq \frac{\|\mathbf{X}\boldsymbol{\Delta}\|_2^2}{\gamma^2\lambda^2\kappa_-(m + |S|)n}, \quad |\tilde{S}^{(2)} \setminus S| \leq \frac{3[1.25\kappa_+(m)\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/n + \tilde{\nu}^{(2)}]}{\{\lambda - \|\mathbf{X}_{S^c}^\top(\mathbf{X}\tilde{\boldsymbol{\beta}}^{(1)} - \mathbf{y})/n\|_\infty\}^2}.$$

We now consider selection consistency of the global solution (2) by comparing it with an oracle solution with Theorem 5. For this purpose, we treat the oracle LSE as an ALS by finding its excess ν in (26), and provide a sufficient condition for the existence of a sign consistent oracle local solution. This sufficient condition is characterized by the following extension of the quantities θ_1^* and θ_2^* in (10) from the ℓ_1 to general penalty:

$$\begin{aligned} \theta_1 &= \inf \{ \theta : \|(\mathbf{X}_S^\top \mathbf{X}_S/n)^{-1} \dot{\rho}(\mathbf{v}_S + \hat{\boldsymbol{\beta}}_S^o; \lambda)\|_\infty \leq \theta\lambda^*, \forall \|\mathbf{v}_S\|_\infty \leq \theta\lambda^* \}, \\ \theta_2 &= \sup \{ \|\mathbf{X}_{S^c}^\top \mathbf{X}_S (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} \dot{\rho}(\mathbf{v}_S + \hat{\boldsymbol{\beta}}_S^o; \lambda)\|_\infty / \lambda^* : \|\mathbf{v}_S\|_\infty \leq \theta_1\lambda^* \}, \end{aligned}$$

where $S = \text{supp}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}^o$ is the oracle LSE in Definition 1 (d). Note that when $\dot{\rho}(\hat{\boldsymbol{\beta}}_S^o; \lambda) = 0$, $\theta_1 = 0$ is attained with $\mathbf{v}_S = 0$ and consequently $\theta_2 = 0$.

THEOREM 6. (i) Let $S = \text{supp}(\boldsymbol{\beta})$ and \mathbf{P}_S be the projection to the column space of \mathbf{X}_S . Suppose $\rho(t; \lambda)$ is left- and right-differentiable in $t > 0$ and $\|\mathbf{X}_{S^c}^\top \mathbf{P}_S^\perp \boldsymbol{\varepsilon}\|_\infty \leq \dot{\rho}(0+; \lambda)$. Then,

the oracle LSE $\widehat{\boldsymbol{\beta}}^\circ$ satisfies the ALS condition (26) with $\nu = \|\dot{\rho}(\widehat{\boldsymbol{\beta}}_S^\circ; \lambda)\|^2$. If in addition the η -NC condition (18) holds and $\nu = 0 = \theta(\widehat{\boldsymbol{\beta}}^\circ, \kappa)$ with a certain $\kappa < \kappa_-(m + |S|)$ and m in (23) or (24), then $\widehat{\boldsymbol{\beta}}^\circ$ is the global solution of (2).

(ii) Suppose $\dot{\rho}(t; \lambda)$ is uniformly continuous in t in the region $\cup_{j \in S} [\widehat{\beta}_j^\circ - \theta_1, \widehat{\beta}_j^\circ + \theta_1]$. Suppose

$$(30) \quad \text{sgn}(\widehat{\boldsymbol{\beta}}^\circ) = \text{sgn}(\boldsymbol{\beta}), \quad \min_{j \in S} |\widehat{\beta}_j^\circ| > \theta_1 \lambda^*, \quad \lambda^* \geq \|\mathbf{X}^\top \mathbf{P}_S^\perp \boldsymbol{\varepsilon}/n\|_\infty / (1 - \theta_2)_+.$$

Then, there exists a local solution $\widetilde{\boldsymbol{\beta}}^\circ$ of (2) satisfying $\text{sgn}(\widetilde{\boldsymbol{\beta}}^\circ) = \text{sgn}(\boldsymbol{\beta})$ and $\|\widetilde{\boldsymbol{\beta}}^\circ - \widehat{\boldsymbol{\beta}}^\circ\|_\infty \leq \theta_1 \lambda^*$. If in addition (18) holds and $\theta(\boldsymbol{\beta}^\circ, \kappa) = 0$ with a certain $\kappa < \kappa_-(m + |S|)$ and m in (23) or (24). Then, $\widetilde{\boldsymbol{\beta}}^\circ$ is the global solution of (2).

REMARK 8. (i) Consider the capped- ℓ_1 , MCP, and SCAD (Table 1). For the capped- ℓ_1 penalty, $\theta(\widehat{\boldsymbol{\beta}}^\circ, \kappa) = 0$ for $\kappa \geq 2/\gamma$ and $\min_{j \in S} |\widehat{\beta}_j^\circ| > \gamma\lambda$, and $\nu = 0$ for $\min_{j \in S} |\widehat{\beta}_j^\circ| > \gamma\lambda/2$. For the MCP, $\theta(\cdot, \kappa) = 0$ for $\kappa \geq 1/\gamma$, and $\nu = 0$ for $\min_{j \in S} |\widehat{\beta}_j^\circ| > \gamma\lambda$. For the SCAD penalty, $\theta(\cdot, \kappa) = 0$ for $\kappa \geq 1/(\gamma - 1)$, and $\nu = 0$ for $\min_{j \in S} |\widehat{\beta}_j^\circ| > \gamma\lambda$. (ii) For the ℓ_1 penalty, $\dot{\rho}(\mathbf{b}) = \text{sgn}(\mathbf{b})$ so that (30) is identical to (10) for the Lasso selection consistency. For concave penalties, $|\dot{\rho}(t; \lambda)|$ is small for large $|t|$, so that $\{\theta_1, \theta_2\}$ are typically smaller than $\{\theta_1^*, \theta_2^*\}$ for strong signals. In such cases, (30) is much weaker than (10).

For a nonconvex penalties such that $\dot{\rho}(t; \lambda) = 0$ when $|t| > a_0\lambda$ for some constant $a_0 > 0$, we automatically have $\dot{\rho}(\widehat{\boldsymbol{\beta}}_S^\circ; \lambda) = 0$ when $\min_{j \in S} |\widehat{\beta}_j^\circ| > a_0\lambda$, which implies that $\theta_1 = \theta_2 = 0$. This special case gives the following easier to interpret corollary as a direct consequence of Theorems 5 and 6.

COROLLARY 3. Let $S = \text{supp}(\boldsymbol{\beta})$ and \mathbf{P}_S be the projection to the column space of \mathbf{X}_S . Suppose $\rho(t; \lambda)$ is left- and right-differentiable in $t > 0$ and $\|\mathbf{X}_{S^c}^\top \mathbf{P}_S^\perp \boldsymbol{\varepsilon}/n\|_\infty \leq \dot{\rho}(0+; \lambda)$. If (18) holds and $\dot{\rho}(\widehat{\boldsymbol{\beta}}_S^\circ; \lambda) = 0$, and $\theta(\widehat{\boldsymbol{\beta}}^\circ, \kappa) = 0$ with a certain $\kappa < \kappa_-(m + |S|)$ and m in (23) or (24), then $\widehat{\boldsymbol{\beta}}^\circ$ is the global solution of (2). Moreover, for any other exact local solution $\widetilde{\boldsymbol{\beta}}$ of (2) that is sparse with $|\text{supp}(\widetilde{\boldsymbol{\beta}}) \setminus S| \leq m$, we have $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^\circ$.

Consider the simple examples of the capped- ℓ_1 penalty and MCP. For the capped- ℓ_1 penalty $\rho(t; \lambda) = \min(\gamma\lambda^2/2, \lambda|t|)$, we pick a sufficiently large γ such that $\gamma > 2/\kappa_-(|S|+m)$ for the m in (23) or (24). This will be possible with $m \asymp |S|$ when $\kappa_-(m)$ is uniformly bounded away from zero for small $m(\ln p)/n$ and $|S|(\ln p)/n$ is even smaller. For the MCP, we pick $\gamma > 1/\kappa_-(|S|+m)$ for the m in (23) or (24). If $\min_{j \in S} |\widehat{\beta}_j^\circ| \geq \gamma\lambda$, then the conditions of Corollary 3 are automatically satisfied for both penalties when $\|\mathbf{X}_{S^c}^\top \mathbf{P}_S^\perp \boldsymbol{\varepsilon}/n\|_\infty < \lambda$ (which holds for sub-Gaussian errors and $\lambda = \sigma\sqrt{(n/2)\log p}$). It follows that in this case, $\widehat{\boldsymbol{\beta}}^\circ$ is the global solution of (2), and there is no other local solution with no more than m nonzero-elements out of S . The essential condition here is the η -NC condition (18), which is an ℓ_2 condition. Note that in view of Corollary 2, the RIF condition is not essential for the equality of the global and oracle solutions in these examples, both with finite $\gamma^* = \gamma/2$. A

similar result hold for the SCAD penalty, with somewhat different constant factors. The requirement of $\min_{j \in S} |\hat{\beta}_j^o| \geq \gamma\lambda$ is natural for variable selection, and it directly follows (with probability $1 - \delta$) from the condition of $\min_{j \in S} |\beta_j| > \gamma\lambda + \sigma(1 + \sqrt{2 \ln(|S|/\delta)}) \lambda_{\min}^{-1/2}(\mathbf{X}_S^\top \mathbf{X}_S)$ for sub-Gaussian errors under Assumption 1.

3.5 Approximate Global Solutions

We have mentioned in Remark 7 that gradient algorithm from the Lasso may yield the global solution of (2) for general $\rho(t; \lambda)$ under a sparse convexity condition or its generalization. Here we provide sufficient conditions for this to happen. This is done via a notion of approximate global solution.

The results in Subsection 3.4 show that if one can find a local solution of (2) and the solution is sparse, then under appropriate conditions, it is the global solution of (2) and it is close to the oracle least squares solution $\hat{\beta}^o$. It is possible to design numerical procedures that find a sparse local solution of (2). For such a procedure, results of Subsection 3.4 directly applies. This section further develops along this line of thinking. Theorem 7 shows that if a local solution is also an approximate global solution, then it is sparse. This fact can be combined with results in Subsection 3.4 to imply that under appropriate conditions, this particular local solution is the unique sparse local solution (which is also the global solution). Moreover, such a solution can be obtained via the Lasso followed by gradient descent, as it can be shown that the Lasso is a sufficiently accurate approximate global solution of (2) for the result to apply.

Given $\nu \geq 0$ and $\mathbf{b} \in \mathbb{R}^p$, we say that a vector $\tilde{\beta} \in \mathbb{R}^p$ is a $\{\nu, \mathbf{b}\}$ approximate global solution of (2) if

$$(31) \quad \left[\frac{1}{2n} \|\mathbf{X}\tilde{\beta} - \mathbf{y}\|_2^2 + \|\rho(\tilde{\beta}; \lambda)\|_1 \right] - \left[\frac{1}{2n} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \|\rho(\mathbf{b}; \lambda)\|_1 \right] \leq \nu.$$

To align different penalties at the same threshold level, we assume throughout this subsection that λ^* depends on $\rho(t; \lambda)$ only through λ in (13), e.g. $\lambda^* = \lambda$.

One method of finding a sparse local solution is to find a local solution that is also an approximate global solution. This can be achieved with the following simple procedure. First, we find the Lasso solution $\hat{\beta}^{(\ell_1)}$ of (8). The following theorem shows that it is a $\{\nu, \beta\}$ approximate global solution of (2) with a relatively small ν under proper conditions. Now we can start with this solution $\hat{\beta}^{(\ell_1)}$ and use gradient descent to find a local solution $\tilde{\beta}$ of (2) that is also an approximate global solution. The following theorem then shows that under appropriate conditions, this local solution is sparse. Therefore results from Subsections 4.2 and 4.4 can be applied to relate it to the true global solution of (2).

THEOREM 7. *Consider a penalty functions $\rho(t; \lambda)$ with $\lambda = \lambda^*$ in (13). Suppose the η -NC condition (18) for $\rho(t; \lambda)$ with $0 < \eta < 1$.*

(i) Suppose $m = O(|S|)$ in (25) or under the SRC in Remark 5 for the Lasso $\hat{\beta}^{(\ell_1)}$ in (8).

Then, the Lasso $\widehat{\beta}^{(\ell_1)}$ is a $\{\nu, \beta\}$ approximate global solution for the penalty $\rho(t; \lambda)$ with $\nu \lesssim \lambda^2 |S|$.

(ii) Assume that $\rho(t; \lambda)$ is continuous at $t = 0$. Let $\widetilde{\beta}$ be an local solution of (2) that is also a $\{\nu, \beta\}$ approximate global solution. Let $\xi^l = 2/(1 - \eta)$. Consider $t_0 > 0$ and integer $m_0 > 0$ such that $\{2\kappa_+(m_0)b/m_0\}^{1/2} + \|\mathbf{X}^\top \varepsilon/n\|_\infty < \inf_{0 < s < t_0} \dot{\rho}(s; \lambda)$, where $b = \xi^l \max\{\nu, \Delta(a'_1 \lambda_1^*, |S|; \lambda)\}$ with $a'_1 := (1 + \eta)/\text{RIF}_1(\xi^l, S)$ and $\lambda_1^* := \sup_{t \geq 0} |\dot{\rho}(t; \lambda)|$. Then,

$$\#\{j \notin S : \widetilde{\beta}_j \neq 0\} < \widetilde{m} := m_0 + \lfloor b/\rho(t_0; \lambda) \rfloor.$$

REMARK 9. If $\rho(t; \lambda)$ is concave in t , then $\lambda_1^* = \dot{\rho}(0+; \lambda)$, and $\inf_{0 < s < t_0} \dot{\rho}(s; \lambda)$ can be replaced by $\dot{\rho}(t_0; \lambda)$ for choosing (t_0, m_0) . Theorem 7 applies to the ℓ_1 , capped- ℓ_1 , MCP and SCAD penalties with $\lambda = \lambda^*$ and $b = \xi^l \max(\nu, |S|\gamma^* \lambda^2)$, but not to the bridge penalty for which $\lambda_1^* = \gamma^* = \infty$.

Theorem 7 shows that the ℓ_1 solution $\widehat{\beta}^{(\ell_1)}$ is $\{\nu, \beta\}$ approximately global optimal with $\nu = O(|S|(\lambda^*)^2)$ in (31), and that a local solution $\widetilde{\beta}$ which is also approximate global optimal is a sparse local solution. Thus, with $b = O((\lambda^*)^2 |S|)$ and $\rho(t_0; \lambda) \asymp (\lambda^*)^2 \asymp (\lambda_1^*)^2$, the local solution $\widetilde{\beta}$ obtained with gradient descent from $\widehat{\beta}^{(\ell_1)}$ is sparse with $\#\{j \notin S : \widetilde{\beta}_j \neq 0\} = O(|S|)$. Here we assume that a line-search is performed in the gradient descent procedure so that the objective function always decreases (and thus each step leads to an $\{\nu, \beta\}$ approximate global optimal solution). Now Remark 6 can be applied to this sparse local solution, providing suitable conditions for this solution to be identical to the global optimal solution. If $\min_{j \in S} |\beta_j| > C\lambda_{univ}$ for a sufficiently large C , Corollary 3 (or Theorems 5 plus Theorem 6) can be applied to identify this local solution as the oracle LSE (or penalized LSE) and the global solution.

It is worth pointing out that results of this paper concerning the global solution can be applied under the NC condition. For a general penalty function, this requires the condition (19) to hold for the upper sparse eigenvalue. Although this is an ℓ_2 condition, it is not needed for either ℓ_1 or ℓ_0 penalty as pointed out in Remark 3. In fact, this condition is also not needed if we consider local solution obtained with more specific numerical procedures such as [43, 49] that lead to specific sparse local solutions with oracle properties. Nevertheless, it is useful to observe that if the extra condition (19) holds, then such a local solution is also the unique global solution, and it can be obtained via other numerical procedures.

4. DISCUSSION

This paper gave a general survey of previous results for high dimensional sparsity analysis using convex penalty (Lasso) and various concave penalties. In particular, specific local solutions for certain concave penalties were studied previously. However, the relationship of these local solutions and their relationship to the global solution was unknown. In the context of these earlier results, we presented a general theory of concave penalty which tries to answer the following questions:

- What’s the relationship among local minima from different procedures?
- What’s the property of global optimal solution?
- Can we find global optimal solution efficiently of nonconvex sparse regularization under ℓ_2 conditions?

Our results answer the above questions. While the theory developed in the paper is general, we are specially interested in ℓ_0 regularization (which is natural for sparse recovery problems) and smooth regularization (for which local minimum can be naturally defined). For ℓ_0 regularization, we show that the global solution is sparse. Moreover, under appropriate ℓ_2 conditions, it recovers the oracle least squares solution, and thus is model selection consistent with oracle property.

Since ℓ_0 penalty is discontinuous at zero, it is difficult to solve with traditional numerical algorithms. For this reason, we consider smooth penalties that have well defined local minimum solutions. For such penalties, this paper provides affirmative answers to the above questions under appropriate ℓ_2 conditions. Specifically, we proved the following results:

- The global solution is sparse.
- Approximate global solution is sparse if it is also a local solution.
- There is a unique sparse local solution that has the oracle property.
- The approximate global solution for nonconvex penalty can be achieved by the Lasso.

This motivates the following numerical procedure. We first start with the Lasso solution, and then use gradient descent to decrease the nonconvex objective value with appropriate concave penalty until it converges to a local minimum solution. Our theory shows that under appropriate ℓ_2 conditions, the solution from this procedure converges to the unique global solution that is sparse, and thus it has the oral property.

In summary, our results imply the following: under appropriate ℓ_2 regularity conditions, plus appropriate assumptions on the penalty $\rho(t; \lambda)$, procedures considered earlier such as MCP [43] or multi-stage convex relaxation [52, 47] give the same local solution that is also the global minimizer of (2). Moreover, other procedures (such as the Lasso followed by gradient descent) can be designed to obtain the same solution. Therefore these results present a coherent view of concave regularization by unifying a number of earlier approaches and by extending a number of previous results. This unified theory presents a more satisfactory treatment of concave high dimensional sparse estimation procedures.

Supplementary Material

Supplementary material for “A general theory of concave regularization for high dimensional sparse estimation problems”

(<http://lib.stat.cmu.edu/aoas/???/???>) Due to space considerations, the proofs in this paper are all given in the supplementary document [45].

REFERENCES

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiadó, Budapest, 1973.
- [2] A. Antoniadis. Comments on: ℓ_1 -penalization for mixture regression models. *TEST*, 19:257–258, 2010.
- [3] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
- [4] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.
- [5] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5:232–253, 2011.
- [6] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York, 2011.
- [7] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [8] T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58:1300–1308, 2010.
- [9] E. Candes and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37:2145–2177, 2009.
- [10] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2404, 2007.
- [11] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.
- [12] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001.
- [13] K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook on the Geometry of Banach Spaces*, volume 1. 2001.
- [14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 32:407–499, 2004.
- [15] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.
- [16] J. Fan and H. Peng. On non-concave penalized likelihood with diverging number of parameters. *Annals of Statistics*, 32:928–961, 2004.
- [17] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.
- [18] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.
- [19] J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.
- [20] D. Hunter and R. Li. Variable selection using MM algorithms. *Annals of Statistics*, 33:1617–1642, 2005.
- [21] Y. Kim, H. Choi, and H.-S. Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of American Statistical Association*, 103:1665–1673, 2008.
- [22] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.
- [23] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009.
- [24] J. Liu, P. Wonka, and J. Ye. Multi-stage Dantzig selector. In *NIPS 10*. 2010.
- [25] C. Mallows. Some comments on Cp. *Technometrics*, 12:661–675, 1973.
- [26] R. Mazumder, J. Friedman, and T. Hastie. Sparsenet : Coordinate descent with non-convex penalties. *Journal of American Statistical Association*, page in press, 2011.
- [27] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

- [28] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.
- [29] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.
- [30] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.
- [31] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. Technical report, University of California, Berkeley, 2009.
- [32] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- [33] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1 -penalization for mixture regression models (with discussion). *Test*, 2:209–285, 2010.
- [34] T. Sun and C.-H. Zhang. Comments on: ℓ_1 -penalization for mixture regression models. *Test*, 2:270–275, 2010.
- [35] T. Sun and C.-H. Zhang. Scaled sparse linear regression. Technical Report arXiv:1104.4595, arXiv, 2011.
- [36] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1996.
- [37] J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006.
- [38] S. van de Geer. The deterministic Lasso. Technical Report 140, ETH Zurich, Switzerland, 2007.
- [39] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008.
- [40] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- [41] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.
- [42] F. Ye and C.-H. Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.
- [43] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010.
- [44] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.
- [45] C.-H. Zhang and T. Zhang. Supplementary material for “a general theory of concave regularization for high dimensional sparse estimation problems”. 2012.
- [46] T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.
- [47] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1087–1107, 2010.
- [48] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57:4689–4708, 2011.
- [49] T. Zhang. Multi-stage convex relaxation for feature selection. Technical Report arXiv:1106.0565, arXiv, 2011.
- [50] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.
- [51] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.
- [52] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics*, 36(4):1509–1533, 2008.