The Annals of Statistics
2019, Vol. 47, No. 6, 3069–3098
https://doi.org/10.1214/18-AOS1759

SORTED CONCAVE PENALIZED REGRESSION

BY LONG FENG AND CUN-HUI ZHANG[1]

City University of Hong Kong and Rutgers University

The Lasso is biased. Concave penalized least squares estimation (PLSE) takes advantage of signal strength to reduce this bias, leading to sharper error bounds in prediction, coefficient estimation and variable selection. For prediction and estimation, the bias of the Lasso can be also reduced by taking a smaller penalty level than what selection consistency requires, but such smaller penalty level depends on the sparsity of the true coefficient vector. The sorted ℓ_1 penalized estimation (Slope) was proposed for adaptation to such smaller penalty levels. However, the advantages of concave PLSE and Slope do not subsume each other. We propose sorted concave penalized estimation to combine the advantages of concave and sorted penalizations. We prove that sorted concave penalties adaptively choose the smaller penalty level and at the same time benefits from signal strength, especially when a significant proportion of signals are stronger than the corresponding adaptively selected penalty levels. A local convex approximation for sorted concave penalties, which extends the local linear and quadratic approximations for separable concave penalties, is developed to facilitate the computation of sorted concave PLSE and proven to possess desired prediction and estimation error bounds. Our analysis of prediction and estimation errors requires the restricted eigenvalue condition on the design, not beyond, and provides selection consistency under a required minimum signal strength condition in addition. Thus, our results also sharpens existing results on concave PLSE by removing the upper sparse eigenvalue component of the sparse Riesz condition.

1. Introduction. The purpose of this paper is twofold. First, we provide a unified treatment of prediction, coefficient estimation and variable selection properties of concave penalized least squares estimation (PLSE) in high-dimensional linear regression under the restrictive eigenvalue (RE) condition on the design matrix. Second, we propose sorted concave PLSE to combine the advantages of concave and sorted penalties, and to prove its superior theoretical properties and computational feasibility under the RE condition. Local convex approximation (LCA) is proposed and studied as a solution for the computation of sorted concave PLSE.

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{1.1}$$

Received November 2017; revised June 2018.

[1]Supported in part by NSF Grants IIS-1407939, DMS-1513378, DMS-1721495 and IIS-1741390.

MSC2010 subject classifications. Primary 62J05, 62J07; secondary 62H12.

Key words and phrases. Penalized least squares, sorted penalties, concave penalties, Slope, local convex approximation, restricted eigenvalue, minimax rate, signal strength.

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a noise vector and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is an unknown coefficient vector. For simplicity, we assume throughout the paper that the design matrix is column normalized with $\|\boldsymbol{x}_j\|_2^2 = n$.

Our study focuses on local and approximate solutions for the minimization of penalized loss functions of the form

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \mathrm{Pen}(\boldsymbol{\beta}) \tag{1.2}$$

with a penalty function $\mathrm{Pen}(\cdot)$ satisfying certain minimum penalty level and maximum concavity conditions. The PLSE can be viewed as a statistical choice among local minimizers of the penalized loss.

Among PLSE methods, the Lasso [27] with the ℓ_1 penalty $\mathrm{Pen}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ is the most widely used and extensively studied. The Lasso is relatively easy to compute as it is a convex minimization problem, but it is well known that the Lasso is biased. A consequence of this bias is the requirement of a neighborhood stability/strong irrepresentable condition on the design matrix $\boldsymbol{X}$ for the selection consistency of the Lasso [16, 28, 30, 37]. Fan and Li [11] proposed a concave penalty to remove the bias of the Lasso and proved an oracle property for one of the local minimizers of the resulting penalized loss. Zhang [33] proposed a path finding algorithm PLUS for concave PLSE and proved the selection consistency of the PLUS-computed local minimizer under a rate optimal signal strength condition on the coefficients and the sparse Riesz condition (SRC) [34] on the design. The SRC, which requires bounds on both the lower and upper sparse eigenvalues of the Gram matrix and is closely related to the restricted isometry property (RIP) [8], is substantially weaker than the strong irrepresentable condition. This advantage of concave PLSE over the Lasso has since become well understood.

For prediction and coefficient estimation, the existing literature somehow presents an opposite story. Consider hard sparse coefficient vectors satisfying $|\mathrm{supp}(\boldsymbol{\beta}^*)| \le s$ with $\log(p/s) \asymp \log p$ and small $(s/n)\log p$. Although rate minimax error bounds were proved under the RIP and SRC for the Dantzig selector and Lasso, respectively, in [7] and [34], Bickel, Ritov and Tsybakov [5] sharpened their results by weakening the RIP and SRC to the RE condition, and van de Geer and Bühlmann [29] proved comparable prediction and ℓ_1 estimation error bounds under an even weaker compatibility or ℓ_1 RE condition. Meanwhile, rate minimax error bounds for concave PLSE still require two-sided sparse eigenvalue conditions like the SRC [12, 31, 33, 36] or a proper known upper bound for the ℓ_1 norm of the true coefficient vector [15]. It turns out that the difference between the SRC and RE conditions are quite significant as Rudelson and Zhou [23] proved that the RE condition is a consequence of a lower sparse eigenvalue condition alone. This seems to suggest a theoretical advantage of the Lasso, in addition to its relative computational simplicity, compared with concave PLSE.

Emerging from the above discussion, an interesting question is whether the RE condition alone on the design matrix is also sufficient for the above discussed

results for concave penalized prediction, coefficient estimation and variable selection, provided proper conditions on the true coefficient vector and the noise. An affirmative answer to this question, which we provide in this paper, amounts to the removal of the upper sparse eigenvalue condition on the design matrix and actually also a relaxation of the lower sparse eigenvalue condition or the restricted strong convexity (RSC) condition [17] imposed in [15]; or equivalently, to the removal of the remaining analytical advantage of the Lasso as far as error bounds for the aforementioned aims are concerned. Specifically, we prove that when the true $\boldsymbol{\beta}$ is sparse, concave PLSE achieves rate minimaxity in prediction and coefficient estimation under the RE condition on the design. Furthermore, the selection consistency of concave PLSE is also guaranteed under the same RE condition and an additional uniform signal strength condition on the nonzero coefficients, and these results also cover nonseparable multivariate penalties imposed on the vector $\boldsymbol{\beta}$ as a whole, including sorted and mixed penalties such as the spike-and-slab Lasso [22].

In addition to the above conservative prediction and estimation error bounds for the concave PLSE that are comparable with those for the Lasso in both rates and regularity conditions on the design, we also prove faster rates for concave PLSE when the signal is partially strong. A short version of this result (cf. Corollary 1 in Section 2) can be stated as follows.

THEOREM 1. *Suppose* $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I})$ *in* (1.1). *Let* $\widehat{\boldsymbol{\beta}}^o$ *be the oracle least squares estimator of* $\boldsymbol{\beta}^*$ *based on the extra knowledge of the model* $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*)$, *and* $s_1 = \#\{j : 0 < |\beta_j^*| < \gamma\sigma\sqrt{(2/n)\log p}\}$ *with* $\gamma > 1$. *Then, under a restricted eigenvalue condition on the design matrix,*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o\|_2^2/n + \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\|_2^2 = O_{\mathbb{P}}(\sigma^2(s_1/n)\log p),$$

where $\widehat{\boldsymbol{\beta}}$ *is a statistical choice of a local minimizer of* (1.2) *with a proper concave penalty. Consequently, under the "beta-min" condition* $s_1 = 0$,

$$\mathbb{P}\{\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^o, \operatorname{sgn}(\widehat{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta}^*)\} \to 1.$$

Moreover, the solution $\widehat{\boldsymbol{\beta}}$ *can be computed in polynomial time.*

Because the prediction and ℓ_2 estimation error of the oracle $\widehat{\boldsymbol{\beta}}^o$ is of the order $\sigma^2 s/n$, the prediction rate for concave PLSE is $\sigma^2(s + s_1 \log p)/n$ by Theorem 1, which is of smaller order than the better known rate $\sigma^2(s/n)\log p$ for the Lasso when $s_1 \ll s$. Moreover, the ℓ_2 error bound automatically yields selection consistency for the concave PLSE under the beta-min condition. Thus, concave PLSE adaptively benefits from signal strength with no harm to the performance in the worst case scenario where all signals are just below the radar screen. This advantage of concave PLSE is known in the existing literature under the sparse Riesz and comparable conditions, but not under the RE condition as presented in this paper.

The bias of the Lasso can be also reduced by taking a smaller penalty level than those required for variable selection consistency, regardless of signal strength. In the literature, PLSE is typically studied in a standard setting at penalty level $\lambda \geq \lambda_* = (\sigma/\eta)\sqrt{(2/n)\log p}$, with $0 < \eta \leq 1$. This lower bound has been referred to as the universal penalty level. However, as the bias of the Lasso is proportional to its penalty level, rate minimaxity in prediction and coefficient estimation requires smaller $\lambda \asymp \sigma\sqrt{(2/n)\log(p/s)}$ [3, 26]. Unfortunately, this smaller penalty level depends on $s = \|\boldsymbol{\beta}^*\|_0$, which is typically unknown. For the ℓ_1 penalty, a remedy for this issue is to apply the Slope or a Lepski-type procedure [3, 24]. However, it is unclear from the literature whether the same can be done with concave penalties.

We propose a class of sorted concave penalties to combine the advantages of concave and sorted penalties. This extends the Slope beyond ℓ_1. Under an RE condition, we prove that the sorted concave PLSE inherits the benefits of both concave and sorted PLSE, namely bias reduction through signal strength and adaptation to the smaller penalty level. A short version of the resulting error bounds (cf. Corollary 3 in Section 3) can be stated as follows.

THEOREM 2. *Let $\{\varepsilon, \boldsymbol{\beta}^*, s_1\}$ be as in Theorem 1 and $s = \|\boldsymbol{\beta}^*\|_0$. Then, under a restricted eigenvalue condition on the design matrix,*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_{\mathbb{P}}(\sigma^2\{s + s_1\log(p/s)\}/n),$$

where $\widehat{\boldsymbol{\beta}}$ is a certain statistical choice of an approximate local minimizer of (1.2) with a proper sorted concave penalty. Moreover, the solution $\widehat{\boldsymbol{\beta}}$ can be computed in polynomial time.

We recall that under the same condition, the error bound for the Slope or the Lasso with Lepski adaptation is of the order $\sigma(s/n)\log(p/s)$ [3, 24]. We would like to mention here that based on empirical evidence, the least squares estimator (LSE) after model selection by the Lasso is commonly believed to reduce the bias of the Lasso. However, to the best of our knowledge, such bias reduction has not yet been theoretically quantified without imposing strong conditions to guarantee model selection consistency. In fact, existing error bounds for LSE after Lasso [4, 25] is typically no superior to those for the Lasso.

To prove the computational feasibility of our theoretical results in polynomial time, we develop an LCA algorithm for a large class of multivariate concave PLSE to produce approximate local solutions to which our theoretical results apply. The LCA is a majorization-minimization (MM) algorithm and is closely related to the local quadratic approximation (LQA) [11] and the local linear approximation (LLA) [38] algorithms. The development of the LCA is needed as the LLA does not majorize sorted concave penalties in general. Our analysis of the LCA can be viewed as an extension of the results in [1, 12, 14, 15, 17, 31, 36] where separable penalties are considered, typically at larger penalty levels.

The rest of this paper is organized as follows. In Section 2, we develop a unified treatment of prediction, coefficient estimation and variable selection properties of concave PLSE under the RE condition at penalty levels required for variable selection consistency. In Section 3, we provide error bounds for approximate solutions with sorted or smaller penalties. In Section 4, we introduce the LCA for sorted penalties. Section 5 contains discussion and a generalization of the results in Section 2 to multivariate penalty functions. We provide the detailed technical proofs in the Supplementary Material [13].

Notation: We denote by $\boldsymbol{\beta}^*$ the true regression coefficient vector, $\overline{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$ the sample Gram matrix, $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*)$ the support set of the coefficient vector, $s = |\mathcal{S}|$ the size of the support and $\Phi(\cdot)$ the standard Gaussian cumulative distribution function. For vectors $\boldsymbol{v} = (v_1, \ldots, v_p)$, we denote by $\|\boldsymbol{v}\|_q = \sum_j (|v_j|^q)^{1/q}$ the ℓ_q norm, with $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ and $\|\boldsymbol{v}\|_0 = \#\{j : v_j \neq 0\}$, and by $\boldsymbol{v}^\#$ the vector with components $v_j^\#$ as the jth largest among $\{|v_1|, \ldots, |v_p|\}$. For symmetric matrices $\boldsymbol{M}$, $\phi_{\min}(\boldsymbol{M})$ and $\phi_{\max}(\boldsymbol{M})$, respectively, denote the minimum and maximum eigenvalues. Moreover, $x_+ = \max(x, 0)$.

2. PLSE with separable concave penalties. In this section, we present our results for concave PLSE at a sufficiently high penalty level to allow selection consistency. Smaller and sorted penalties are considered in Section 3. We divide the section into three subsections to describe the collection of PLSE under consideration, conditions on the design matrix and error bounds for prediction, coefficient estimation and variable selection.

2.1. *Concave PLSE.* In general, separable penalty functions can be written as a sum of penalties on individual variables,

$$\operatorname{Pen}(\boldsymbol{b}) = \sum_{j=1}^{p} \rho_j(b_j; \lambda_j). \tag{2.1}$$

When $\rho_j(\cdot)$ and λ_j are fixed across $j = 1, \ldots, p$, $\operatorname{Pen}(\boldsymbol{b})$ reduces to usual form

$$\operatorname{Pen}(\boldsymbol{b}) = \rho(\boldsymbol{b}; \lambda) = \sum_{j=1}^{p} \rho(b_j; \lambda). \tag{2.2}$$

We assume that $\rho(t; \lambda)$ is a parametric family of concave penalties that is symmetric about 0, $\rho(t; \lambda) = \rho(-t; \lambda)$, with $\rho(0; \lambda) = 0$, and monotone, $\rho(t; \lambda) \uparrow |t|$. Moreover, we assume that $\rho(t; \lambda)$ is indexed by its penalty level in the sense of $\lambda = \dot{\rho}(0+; \lambda)$. Here, $\dot{\rho}(t; \lambda)$ is defined as any value between the left and right derivatives of $\rho(t; \lambda)$ with respect to t. We denote by $\partial\rho(t; \lambda)$ the set of all $\dot{\rho}(t; \lambda)$ and by $\partial \operatorname{Pen}(\boldsymbol{b})$ the set of all possible choices of

$$\dot{\operatorname{Pen}}\big((b_1, \ldots, b_p)^T\big) = \big(\dot{\rho}_1(b_1; \lambda_1), \ldots, \dot{\rho}_p(b_p; \lambda_p)\big)^T.$$

Unless otherwise stated, given a mathematical expression where $\dot{\mathrm{Pen}}(\boldsymbol{b})$ appears, $\dot{\mathrm{Pen}}(\boldsymbol{b})$ denotes the most favorable member of $\partial\,\mathrm{Pen}(\boldsymbol{b})$ for the expression to hold. We define the concavity of $\rho(t;\lambda)$ as

$$\overline{\kappa}(t;\rho,\lambda)=\sup_{t'>t}\{\dot{\rho}(t';\lambda)-\dot{\rho}(t;\lambda)\}/(t-t'), \tag{2.3}$$

where the supreme is taken over all possible choices of $\dot{\rho}(t;\lambda)$ and $\dot{\rho}(t';\lambda)$. Further, we define the overall maximum concavity of $\rho(t;\lambda)$ as

$$\overline{\kappa}(\rho)=\max_{t\geq 0,\lambda>0}\overline{\kappa}(t;\rho,\lambda). \tag{2.4}$$

Our analysis is applicable to many popular penalty functions, such as the ℓ_1 penalty $\rho(t;\lambda)=\lambda|t|$ for the Lasso with $\overline{\kappa}(\rho)=0$, the SCAD (smoothly clipped absolute deviation) penalty [11] with $\rho(t;\lambda)=\int_0^{|t|}\{\lambda-\overline{\kappa}(x-\lambda)_+\}_+\,dx$ and $\overline{\kappa}(\rho)=\overline{\kappa}$, and the MCP (minimax concave penalty) [33] with $\rho(t;\lambda)=\int_0^{|t|}(\lambda-\overline{\kappa}x)_+\,dx$ and $\overline{\kappa}(\rho)=\overline{\kappa}$.

Given a penalty (2.2), local minimizers of the penalized loss (1.2) must satisfy the following (local) Karush–Kuhn–Tucker (KKT) condition

$$\boldsymbol{X}_j^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n=\dot{\rho}(\widehat{\beta}_j;\lambda) \tag{2.5}$$

for a certain $\dot{\rho}(\widehat{\beta}_j;\lambda)\in\partial\rho(\widehat{\beta}_j;\lambda)$. Our analysis also works with approximate local minimizers of (1.2) with general penalties of the form (2.1) and a common penalty level $\lambda=\lambda_j$. Such approximate solutions can be written as

$$\boldsymbol{X}_j^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n=\dot{\rho}_j(\widehat{\beta}_j;\lambda)+\nu_j, \tag{2.6}$$

where $\boldsymbol{v}=(\nu_1,\ldots,\nu_p)^T$ is an approximation error satisfying $\widehat{\beta}_j\nu_j\geq 0$. This extension from (2.5), which explicitly quantifies the approximation error, brings true practical benefits as many algorithms only provide approximate solutions for computational efficiency. We note that the condition $\widehat{\beta}_j\nu_j\geq 0$, which the approximate solution (2.6) is required to satisfy, is verifiable when the solution is computed. It requires the approximation error not to reduce the penalty. We also assume in (2.11) below that $\|\boldsymbol{v}\|_\infty/\lambda$ is not too large. In Section 3, we consider solutions with less restrictive approximation errors. If we confine our attention to penalty functions $\rho_j(\cdot)$ with upper bounded concavity $\overline{\kappa}(\rho_j)\leq\kappa_*$, the collection of local solutions of form (2.6) is characterized by

$$\begin{cases}(\lambda-\kappa_*|\widehat{\beta}_j|)_+\leq\operatorname{sgn}(\widehat{\beta}_j)\boldsymbol{X}_j^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\leq\lambda+\operatorname{sgn}(\widehat{\beta}_j)\nu_j, & \widehat{\beta}_j\neq 0,\\ |\boldsymbol{X}_j^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n|\leq\lambda+|\nu_j|, & \widehat{\beta}_j=0,\end{cases} \tag{2.7}$$

as solutions of (2.7) can be constructed with separable penalties $\mathrm{Pen}(\boldsymbol{b})=\sum_{j=1}^p\rho_j(b_j;\lambda)$ with a common penalty level λ and potentially different concavity satisfying $\overline{\kappa}(\rho_j)\leq\kappa_*$. Compared with (2.6), (2.7) is more explicit. However, unfortunately, they do not cover solutions with sorted penalties.

We study solutions of (2.7) by comparing them with an oracle coefficient vector $\boldsymbol{\beta}^o$. We assume that for a certain sparse subset $\mathcal{S}$ of $\{1, \ldots, p\}$, $\lambda_* > 0$ and $\eta \in [0, 1)$, the oracle $\boldsymbol{\beta}^o$ satisfies the following:

$$\operatorname{supp}(\boldsymbol{\beta}^o) \subseteq \mathcal{S}, \qquad \|\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}^o)/n\|_\infty < \eta\lambda_*. \tag{2.8}$$

We may take $\boldsymbol{\beta}^o \in \mathbb{R}^p$ as the true coefficient vector $\boldsymbol{\beta}^*$ with $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*)$, or the oracle LSE $\widehat{\boldsymbol{\beta}}^o$ given by

$$\widehat{\boldsymbol{\beta}}^o_{\mathcal{S}} = (\boldsymbol{X}^T_{\mathcal{S}}\boldsymbol{X}_{\mathcal{S}})^{-1}\boldsymbol{X}^T_{\mathcal{S}}\boldsymbol{y}, \qquad \widehat{\boldsymbol{\beta}}^o_{\mathcal{S}^c} = \boldsymbol{0}. \tag{2.9}$$

When $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* \sim N(\boldsymbol{0}, \boldsymbol{V})$ with $\phi_{\max}(\boldsymbol{V}) \vee \max_{j\le p} \boldsymbol{x}_j^T\boldsymbol{V}\boldsymbol{x}_j/n \le \sigma^2$,

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^o \sim N(\boldsymbol{0}, \boldsymbol{V}^o) \quad \text{with } \phi_{\max}(\boldsymbol{V}^o) \vee \max_{j\le p} \boldsymbol{x}_j^T\boldsymbol{V}^o\boldsymbol{x}_j/n \le \sigma^2 \tag{2.10}$$

for such $\boldsymbol{\beta}^o$, so that (2.8) holds with at least probability $1 - \sqrt{2/(\pi \log p)}$ for

$$\lambda_* = (\sigma/\eta)\sqrt{(2/n)\log p}.$$

Given an oracle $\boldsymbol{\beta}^o$ satisfying (2.8), we consider solutions of (2.7) with penalties satisfying the following conditions:

$$\lambda \ge \lambda_*, \qquad \|\boldsymbol{v}_{\mathcal{S}}\|_\infty/\lambda \le (1-\eta)\xi - (1+\eta), \tag{2.11}$$

for some λ_* and η satisfying (2.8) and $\xi > 0$. Let $\mathscr{B}(\lambda_*, \kappa_*)$ be the set of all local solutions satisfying (2.7) with penalty level and approximation errors satisfying (2.11),

$$\mathscr{B}(\lambda_*, \kappa_*) = \{\widehat{\boldsymbol{\beta}} : (2.7) \text{ holds with } \lambda \text{ and } \boldsymbol{v} \text{ satisfying } (2.11)\}. \tag{2.12}$$

Given a penalty $\operatorname{Pen}(\boldsymbol{b}) = \sum_{j=1}^p \rho_j(b_j; \lambda)$ satisfying $\overline{\kappa}(\rho_j) \le \kappa_*$ and (2.11), $\mathscr{B}(\lambda_*, \kappa_*)$ contains all local minimizers of the penalized loss (1.2). Our theory is applicable to the subclass

$$\mathscr{B}_0(\lambda_*, \kappa_*) = \{\widehat{\boldsymbol{\beta}} : \widehat{\boldsymbol{\beta}} \text{ and } \boldsymbol{0} \text{ are connected in } \mathscr{B}(\lambda_*, \kappa_*)\}. \tag{2.13}$$

Here, $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{0}$ are connected if there exist $\widehat{\boldsymbol{\beta}}^{(k)} \in \mathscr{B}(\lambda_*, \kappa_*)$, $k = 1, \ldots, k^*$ with penalty levels $\lambda^{(k)}$ such that for the $a_0 > 0$ in Proposition 2 below,

$$\widehat{\boldsymbol{\beta}}^{(0)} = 0, \qquad \widehat{\boldsymbol{\beta}}^{(k^*)} = \widehat{\boldsymbol{\beta}}, \qquad \|\widehat{\boldsymbol{\beta}}^{(k)} - \widehat{\boldsymbol{\beta}}^{(k-1)}\|_1 \le a_0\lambda^{(k)}. \tag{2.14}$$

This condition will be further relaxed in Section 3.

By definition, $\mathscr{B}_0(\lambda_*, \kappa_*)$ contains the set of all local solutions (2.7) computable by path following algorithms starting from the origin, with constraints on the penalty levels, concavities and approximation errors. This is a large class of statistical solutions as it includes all local solutions connected to the origin regardless of the specific algorithms used to compute the solution, and different types of penalties can be used in a single solution path. For example, the Lasso estimator belongs

to the class as it is connected to the origin through the LARS algorithm [10, 19, 20]. The SCAD and MCP solutions with $\lambda \geq \lambda_*$ and $\overline{\kappa} \leq \kappa_*$ belong to the class if they are computed by the PLUS algorithm [33] or by a continuous path following algorithm from the Lasso solution. More important, many iterative algorithms generating approximate solutions of Lasso, SCAD or MCP also belong to the class, for example, [1, 12, 31]. As $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ is the sparsest solution, $\mathscr{B}_0(\lambda_*, \kappa_*)$ can be viewed as the sparse branch of the solution space $\mathscr{B}(\lambda_*, \kappa_*)$.

We prove that all solutions in (2.13) belong to the following cone:

$$\mathscr{C}(\mathcal{S}; \xi) = \{\boldsymbol{u} : \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq \xi \|\boldsymbol{u}_{\mathcal{S}}\|_1\}, \tag{2.15}$$

when the RE below is no smaller than κ_* in (2.7).

2.2. *Restricted eigenvalue condition.* The RE condition, proposed in [5], is arguably the weakest available on the design to guarantee rate minimax performance in prediction and coefficient estimation for the Lasso. The RE for the ℓ_2 estimation loss can be defined as follows. For $\mathcal{S} \subset \{1, \ldots, p\}$ and $\xi > 0$,

$$\mathrm{RE}_2(\mathcal{S}; \xi) = \inf\left\{ \frac{(\boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u})^{1/2}}{\|\boldsymbol{u}\|_2} : \boldsymbol{u} \in \mathscr{C}(\mathcal{S}; \xi) \right\} \tag{2.16}$$

with $\inf \varnothing = \phi_{\min}(\overline{\boldsymbol{\Sigma}})$ for $\mathcal{S} = \varnothing$ and $\mathscr{C}(\mathcal{S}; \xi)$ as in (2.15). The RE condition refers to the property that $\mathrm{RE}_2(\mathcal{S}; \xi)$ is no smaller than a certain positive constant for all design matrices under consideration. For prediction and ℓ_1 estimation, it suffices to impose a somewhat weaker compatibility condition [29]. The compatibility coefficient, also called ℓ_1-RE [29], is defined as

$$\mathrm{RE}_1(\mathcal{S}; \xi) = \inf\left\{ \frac{(\boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u})^{1/2}}{\|\boldsymbol{u}_{\mathcal{S}}\|_1 / |\mathcal{S}|^{1/2}} : \boldsymbol{u} \in \mathscr{C}(\mathcal{S}; \xi) \right\}. \tag{2.17}$$

In addition to the RE above, we define a relaxed cone invertibility factor (RCIF) for prediction as

$$\mathrm{RCIF}_{\mathrm{pred}}(\mathcal{S}; \eta, \boldsymbol{w}) = \inf\left\{ \frac{\|\overline{\boldsymbol{\Sigma}} \boldsymbol{u}\|_\infty^2 |\mathcal{S}|}{\boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u}} : \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < -\frac{\boldsymbol{w}_{\mathcal{S}}^T \boldsymbol{u}_{\mathcal{S}}}{1-\eta} \right\}, \tag{2.18}$$

with $\eta \in [0, 1)$ and a vector $\boldsymbol{w} \in \mathbb{R}^p$, and a RCIF for the ℓ_q estimation as

$$\mathrm{RCIF}_{\mathrm{est},q}(\mathcal{S}; \eta, \boldsymbol{w}) = \inf\left\{ \frac{\|\overline{\boldsymbol{\Sigma}} \boldsymbol{u}\|_\infty |\mathcal{S}|^{1/q}}{\|\boldsymbol{u}\|_q} : \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < -\frac{\boldsymbol{w}_{\mathcal{S}}^T \boldsymbol{u}_{\mathcal{S}}}{1-\eta} \right\}. \tag{2.19}$$

The RCIF is a relaxation of the cone invertibility coefficient [32] for which the constraint $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < \xi \|\boldsymbol{u}_{\mathcal{S}}\|_1$ is imposed.

The choices of ξ, η and $\boldsymbol{w}$ depend on the problem under consideration, but in our analysis of (2.6), we may let η and ξ satisfy (2.8) and (2.11) and

$$\begin{aligned} \boldsymbol{w} &= \{\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o) + \boldsymbol{v} - \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^o)/n\}/\lambda \\ &= \big(\dot{\rho}_j(\beta_j^o; \lambda) + v_j - \boldsymbol{X}_j^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^o)/n, \, j = 1, \ldots, p\big)^T, \end{aligned} \tag{2.20}$$

where $\mathrm{Pen}(\cdot)$ can be any penalty function of the form (2.1) satisfying $\lambda_j = \lambda$ and $\kappa(\rho_j) \le \kappa_*$ and (2.11). It follows that $\|\boldsymbol{w}_S\|_\infty \le (1-\eta)\xi$, so that the minimization in (2.18) and (2.19) is taken over a smaller cone than (2.15). In the presence of partial signal strength, $\|\boldsymbol{w}_S\|_2$ can be much smaller than $|\mathcal{S}|^{1/2}(1-\eta)\xi$. Moreover, for selection consistency, it is feasible to have $\boldsymbol{w}_S = \boldsymbol{0}$ under a beta-min condition when $\boldsymbol{v} = 0$. In the following subsection, we use an RE condition to prove cone membership of the estimation error of the concave PLSE and the RCIF to bound the prediction and coefficient estimation errors. The following proposition, which follows from the analysis in Section 3.2 of [32], shows that the RCIF may provide sharper bounds than the RE does.

PROPOSITION 1. *If* $\|\boldsymbol{w}_S\|_\infty \le (1-\eta)\xi$, *then*

$$\begin{aligned}
&\mathrm{RCIF}_{\mathrm{pred}}(\mathcal{S};\eta,\boldsymbol{w}) \ge \mathrm{RE}_1^2(\mathcal{S};\xi)/(1+\xi)^2,\\
&\mathrm{RCIF}_{\mathrm{est},1}(\mathcal{S};\eta,\boldsymbol{w}) \ge \mathrm{RE}_1^2(\mathcal{S};\xi)/(1+\xi)^2,\\
&\mathrm{RCIF}_{\mathrm{est},2}(\mathcal{S};\eta,\boldsymbol{w}) \ge \mathrm{RE}_1(\mathcal{S};\xi)\,\mathrm{RE}_2(\mathcal{S};\xi)/(1+\xi).
\end{aligned} \tag{2.21}$$

2.3. *Error bounds.* Let $\mathscr{B}(\lambda_*,\kappa_*)$ be as in (2.12), $\mathscr{B}_0(\lambda_*,\kappa_*)$ as in (2.13), $\mathscr{C}(\mathcal{S};\xi)$ be as in (2.15). We define a set of "good solutions" for PLSE with separable penalties as

$$\mathscr{B}_0^*(\lambda_*,\kappa_*) = \mathscr{B}_0(\lambda_*,\kappa_*) \cup \{\widehat{\boldsymbol{\beta}} \in \mathscr{B}(\lambda_*,\kappa_*) : \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}(\mathcal{S};\xi)\}. \tag{2.22}$$

Here, under an RE condition on the design matrix, we provide prediction and coefficient estimation error bounds for all good solutions and prove the sign consistency for exact solutions under a "beta-min" condition.

THEOREM 3. *Suppose* (2.8) *holds for certain* $\boldsymbol{\beta}^o \in \mathbb{R}^p$ *and* $\mathrm{RE}_2^2(\mathcal{S};\xi) \ge \kappa_*$. *Let* $\widehat{\boldsymbol{\beta}} \in \mathscr{B}_0^*(\lambda_*,\kappa_*)$ *be a solution of* (2.7) *with penalty level* λ. *Then*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^o\|_2^2/n \le \min\left\{\frac{\overline{\lambda}^2|\mathcal{S}|}{\mathrm{RCIF}_{\mathrm{pred}}(\mathcal{S};\eta,\boldsymbol{w})}, \frac{(\overline{\lambda}+\eta\xi\lambda_*)^2|\mathcal{S}|}{\mathrm{RE}_1^2(\mathcal{S};\xi)}\right\}, \tag{2.23}$$

with $\overline{\lambda} = \lambda + \eta\lambda_* + \|\boldsymbol{v}\|_\infty \le \lambda - \eta\xi\lambda_*$ *and the* $\boldsymbol{w}$ *in* (2.20), *and*

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \le \min\left\{\frac{\overline{\lambda}|\mathcal{S}|}{\mathrm{RCIF}_{\mathrm{est},1}(\mathcal{S};\eta,\boldsymbol{w})}, \frac{(\overline{\lambda}+\eta\xi\lambda_*)(1+\xi)|\mathcal{S}|}{\mathrm{RE}_1^2(\mathcal{S};\xi)}\right\},\\
&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_2 \le \min\left\{\frac{\overline{\lambda}|\mathcal{S}|^{1/2}}{\mathrm{RCIF}_{\mathrm{est},2}(\mathcal{S};\eta,\boldsymbol{w})}, \frac{(\overline{\lambda}+\eta\xi\lambda_*)|\mathcal{S}|^{1/2}}{\mathrm{RE}_1(\mathcal{S};\xi)\,\mathrm{RE}_2(\mathcal{S};\xi)}\right\},\\
&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_q \le \frac{\overline{\lambda}|\mathcal{S}|^{1/q}}{\mathrm{RCIF}_{\mathrm{est},q}(\mathcal{S};\eta,\boldsymbol{w})}, \quad \forall q \ge 1.
\end{aligned} \tag{2.24}$$

Compared with Theorem 3, existing results on statistical solutions of concave PLSE [12, 31, 33] cover only separable penalties of form (2.2), require stronger conditions on the design (such as upper sparse eigenvalue) and offer less explicit error bounds. While the error bounds for concave penalty in Theorem 3 match existing ones for the Lasso [5, 29] and other methods [7, 9] up to a constant factor, they also hold when $\mathrm{RE}_1(\mathcal{S};\xi)$ and $\mathrm{RE}_2(\mathcal{S};\xi)$ in (2.23) and (2.24) are replaced by their larger version with the constraint $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \le \xi\|\boldsymbol{u}_{\mathcal{S}}\|_1$ replaced by the more stringent $(1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \le -\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}}$ in their definition, as $-\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}}$ could be much smaller than $(1-\eta)\xi\|\boldsymbol{u}_{\mathcal{S}}\|_1$ when a significant proportion of the signals are strong. We describe the benefit of concave PLSE over the Lasso in such scenarios in the following two theorems.

THEOREM 4. *Suppose* (2.8) *holds for an oracle solution* $\boldsymbol{\beta}^o$ *of* (2.6) *with* $\boldsymbol{v} = 0$ *and some* $\rho_j(\cdot)$. *Assume that* $\mathrm{RE}_2^2(\mathcal{S};\xi) \ge \kappa_* \ge \max_j \overline{\kappa}(\rho_j)$. *Let* $\widehat{\boldsymbol{\beta}} \in \mathscr{B}_0^*(\lambda_*,\kappa_*)$ *be a solution of* (2.6) *with* $\boldsymbol{v} = 0$ *and the same* $\rho_j(\cdot)$. *Then*

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} = \boldsymbol{0} \qquad \textit{and} \qquad \mathrm{sgn}(\widehat{\beta}_j)\,\mathrm{sgn}(\beta_j^o) \ge 0 \quad \forall j \in \mathcal{S}. \tag{2.25}$$

If in addition $\max_{j\in\mathcal{S}} \overline{\kappa}(0;\rho_j,\lambda) < \phi_{\min}(\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}})$, *for example,* $\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}^c} \ne \boldsymbol{0}$ *a priori, then*

$$\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}^o). \tag{2.26}$$

Theorem 4 provides selection consistency of concave PLSE under the restricted eigenvalue condition, compared with the required irrepresentable condition for the Lasso [16, 28, 37]. This is also new as existing results require the stronger sparse Riesz condition [33] or other combination of lower and upper sparse eigenvalue conditions on the design [12, 31, 35] for selection consistency.

The proof of Theorem 4 unifies the analysis for prediction, estimation and variable selection, as the proof of selection consistency is simply done by inspecting the case of $\boldsymbol{w}_{\mathcal{S}} = 0$ in the prediction and estimation error bounds. For $\boldsymbol{v} = 0$, the $\boldsymbol{\beta}^o$ in (2.8) is a solution of (2.6) iff $\boldsymbol{w}_{\mathcal{S}} = 0$, and this is the case for the oracle LSE $\boldsymbol{\beta}^o = \widehat{\boldsymbol{\beta}}^o$ in (2.9) under the beta-min condition $\min_{j\in\mathcal{S}}|\widehat{\beta}_j^o| \ge \gamma\lambda$ when $\dot{\rho}_j(t;\lambda) = 0$ for all $|t| \ge \gamma\lambda$. Regarding the additional condition for the sign consistency, we note that $\mathrm{RE}_2^2(\mathcal{S};\xi) < \mathrm{RE}_2^2(\mathcal{S};0) = \phi_{\min}(\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}})$ when $\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}^c} \ne \boldsymbol{0}$ and $\overline{\kappa}(\rho_j) \le \kappa_* \le \mathrm{RE}_2^2(\mathcal{S};\xi)$ by the RE condition.

THEOREM 5. *Suppose* (2.8) *holds for* $\boldsymbol{\beta}^o \in \mathbb{R}^p$ *and* $\kappa_* \le \mathrm{RE}_2^2(\mathcal{S};\xi)$. *Let* $\widehat{\boldsymbol{\beta}} \in \mathscr{B}_0^*(\lambda_*,\kappa_*)$ *be a solution of* (2.6) *with penalty level* λ *and* $\rho_j(\cdot)$ *satisfying* $\overline{\kappa}(\rho_j) \le \kappa_*$ *and* $\max_j \overline{\kappa}(\widehat{\beta}_j;\rho_j,\lambda) \le (1-1/C_0)\,\mathrm{RE}_2^2(\mathcal{S};\xi)$. *Then*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^o\|_2^2/n \le (C_0\lambda)^2 \sup_{\boldsymbol{u}\ne 0} \frac{[-\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1]_+^2}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}} \tag{2.27}$$

with the $\boldsymbol{w}$ *in* (2.20), *and for any seminorm* $\|\cdot\|$ *as a loss function*

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\| \le C_0\lambda \sup_{\boldsymbol{u}\neq 0} \frac{\|\boldsymbol{u}\|[-\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}}-(1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1]}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}}. \tag{2.28}$$

COROLLARY 1. *Let* $\boldsymbol{\beta}^o=\widehat{\boldsymbol{\beta}}^o$ *be the oracle LSE in* (2.9). *Suppose* (2.8) *holds with high probability and other conditions of Theorem* 5 *hold with* $C_0^2/\mathrm{RE}_2^2(\mathcal{S};\xi)=O(1)$ *and* $\lambda=(\sigma/\eta)\sqrt{(2/n)\log p}$. *Let* $s=|\mathcal{S}|$ *and* $s_1=\|(\dot{\mathrm{Pen}}(\boldsymbol{\beta}^o)+\boldsymbol{v})_{\mathcal{S}}/\lambda\|_2$ *with* $\mathrm{Pen}(\cdot)=\sum_j\rho_j(\cdot;\lambda)$. *Then*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o\|_2^2/n+\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^o\|_2^2+\|\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^o\|_1^2/s=O_{\mathbb{P}}(\sigma^2/n)s_1\log p,$$

implying $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}^o$ *when* $s_1=0$, *and for the true* $\boldsymbol{\beta}^*$

$$\begin{aligned}&\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n+\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2+\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1^2/s\\&\quad=O_{\mathbb{P}}(\sigma^2/n)(s_1\log p+s).\end{aligned} \tag{2.29}$$

For $\boldsymbol{v}=\boldsymbol{0}$ *and* $\rho_j(\cdot;\lambda)$ *satisfying* $\mathrm{supp}(\dot{\rho}_j(t;\lambda))\subseteq[-\gamma\lambda,\gamma\lambda]$,

$$s_1\le \#\{j\in\mathcal{S}:|\widehat{\beta}_j^o|<\lambda\gamma\}, \tag{2.30}$$

Theorem 5 and Corollary 1 demonstrate the benefits of concave PLSE, as $s_1=s=|\mathcal{S}|$ in (2.29) for the Lasso but s_1 could be much smaller than s for concave penalties. For usual penalties with $\rho_1(\cdot)=\cdots=\rho_p(\cdot)$, (2.30) holds with $\gamma=1+1/\overline{\kappa}$ for the SCAD (2.4) and $\gamma=1/\overline{\kappa}$ for the MCP (2.5).

For the exact Lasso solution with $\overline{\kappa}(\rho)=0$, $C_0=1$ and $\|\boldsymbol{w}_{\mathcal{S}}\|_\infty\le 1+\eta$, Theorem 5 yields the sharpest possible prediction and estimation error bounds based on the basic inequality $\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}+(1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1\le(1+\eta)\|\boldsymbol{u}_{\mathcal{S}}\|_1$ with $\boldsymbol{u}=(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)/\lambda$, as stated in the following corollary.

COROLLARY 2. *Let* $\widehat{\boldsymbol{\beta}}$ *be the Lasso estimator with penalty level* $\lambda\ge\lambda_*$. *If* (2.8) *holds for a coefficient vector* $\boldsymbol{\beta}^o\in\mathbb{R}^p$, *then*

$$\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}^o\|_2^2}{n(1+\eta)^2\lambda^2}\le\sup_{\boldsymbol{u}\neq 0}\frac{\psi^2(\boldsymbol{u})}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}}=\max_{0<t<1}\frac{|\mathcal{S}|(1-t)^2}{\mathrm{RE}_1^2(\mathcal{S};t\xi)}$$

with $\psi(\boldsymbol{u})=[\|\boldsymbol{u}_{\mathcal{S}}\|_1-\|\boldsymbol{u}_{\mathcal{S}^c}\|_1/\xi]_+$ *and* $\xi=(1+\eta)/(1-\eta)$,

$$\frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_2}{(1+\eta)\lambda}\le\sup_{\boldsymbol{u}\neq 0}\frac{\|\boldsymbol{u}\|_2\psi(\boldsymbol{u})}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}}=\max_{0<t<1}\frac{|\mathcal{S}|^{1/2}(1-t)}{\mathrm{RE}_{1,2}^2(\mathcal{S};t\xi)}$$

with $\mathrm{RE}_{1,2}(\mathcal{S};\xi)=\inf_{\|\boldsymbol{u}_{\mathcal{S}^c}\|<\xi\|\boldsymbol{u}_{\mathcal{S}}\|_1}\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}/(\|\boldsymbol{u}\|_2\|\boldsymbol{u}_{\mathcal{S}}\|_1/|\mathcal{S}|^{1/2})$, *and*

$$\frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1}{(1+\eta)\lambda}\le\sup_{\boldsymbol{u}\neq 0}\frac{\|\boldsymbol{u}\|_1\psi(\boldsymbol{u})}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}}=\max_{0<t<1}\frac{|\mathcal{S}|^{1/2}(1+t\xi)(1-t)}{\mathrm{RE}_1^2(\mathcal{S};t\xi)}.$$

As Theorems 3–5 deal with the same estimator under the same RE conditions on the design, they give a unified treatment of the prediction, coefficient estimation and variable selection performance of the PLSE, including the ℓ_1 and concave penalties. For prediction and coefficient estimation, (2.23) and (2.24) match those of state-of-art for the Lasso in both the convergence rate and the regularity condition on the design, while (2.27), (2.28) and Corollary 1 demonstrate the advantages of concave penalization when s_1 is much smaller than s. Meanwhile, for selection consistency, Theorem 4 weakens existing conditions on the design to the same RE condition as required for ℓ_2 estimation with the Lasso. These RE-based results are significant as the existing theory for concave penalization, which requires substantially stronger conditions on the design, leaves a false impression that the Lasso has a technical advantage in prediction and parameter estimation by requiring much weaker conditions on the design than the concave PLSE.

The following lemma, which can be viewed as a basic inequality for analyzing concave PLSE, is the beginning point of our analysis.

LEMMA 1. *Let $\widehat{\boldsymbol{\beta}}$ be as in (2.7), $\boldsymbol{\beta}^o$ as in (2.8), $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$, $\boldsymbol{w}$ as in (2.20), and $\overline{\lambda} = \lambda + \eta\lambda_* + \|\boldsymbol{v}\|_\infty$ as in Theorem 3. Then*

$$\boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h} \le \min\bigl\{-\lambda\boldsymbol{h}^T\boldsymbol{w} + \kappa_*\|\boldsymbol{h}\|_2^2, \overline{\lambda}\|\boldsymbol{h}_{\mathcal{S}}\|_1 + \eta\lambda_*\|\boldsymbol{h}_{\mathcal{S}^c}\|_1\bigr\}. \tag{2.31}$$

Moreover, for a proper choice of $\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o) \in \partial\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o)$ in (2.20),

$$\boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h} + \bigl\{\lambda - \bigl\|\boldsymbol{X}_{\mathcal{S}^c}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^o)/n\bigr\|_\infty\bigr\}\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le -\lambda\boldsymbol{h}_{\mathcal{S}}^T\boldsymbol{w}_{\mathcal{S}} + \kappa_*\|\boldsymbol{h}\|_2^2. \tag{2.32}$$

Next, we prove that the solutions in $\mathscr{B}_0^*(\lambda_*, \kappa_*)$ in (2.22) and other approximate local solutions (2.7) in $\mathscr{B}(\lambda_*, \kappa_*)$ are separated by a gap of size $a_0\lambda_*$ in the ℓ_1 distance for some $a_0 > 0$. Consequently, $\widehat{\boldsymbol{\beta}} \in \mathscr{B}_0(\lambda_*, \kappa_*)$ implies $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}(\mathcal{S}; \xi)$ for the $\mathscr{B}_0(\lambda_*, \kappa_*)$ in (2.13) and $\mathscr{C}(\mathcal{S}; \xi)$ in (2.15).

PROPOSITION 2. *Let $\mathscr{B}(\lambda_*, \kappa_*)$ as in (2.12) and $\mathscr{C}(\mathcal{S}; \xi)$ as in (2.15). Suppose $\mathrm{RE}_2^2(\mathcal{S}; \xi) \ge \kappa_*$. Let $a_1 = \eta - \|\boldsymbol{z}_{\mathcal{S}^c}\|_\infty/\lambda_*$, $a_2 = a_1\xi/[2\max_j \overline{\kappa}(\widetilde{\beta}_j; \rho_j, \widetilde{\lambda}) \times (\xi+1)]$ and $a_3 = a_1(1-\eta)\xi/\{(1-\eta)(\xi+1)+a_1\}$. Let $\widehat{\boldsymbol{\beta}} \in \mathscr{B}(\lambda_*, \kappa_*)$ with penalty level λ, and $\widetilde{\boldsymbol{\beta}} \in \mathscr{B}(\lambda_*, \kappa_*)$ with $\widetilde{\lambda}$. Then*

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}(\mathcal{S}; \xi) \quad \text{and} \quad \|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_1 \le \widetilde{\lambda}a_0 \quad \text{imply } \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}(\mathcal{S}; \xi),$$

with $a_0 = \min[a_2, a_2a_3/\{(1-\eta)(\xi \vee 1)\}]$. Consequently,

$$\mathscr{B}_0^*(\lambda_*, \kappa_*) = \bigl\{\widehat{\boldsymbol{\beta}} \in \mathscr{B}(\lambda_*, \kappa_*) : \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}(\mathcal{S}; \xi)\bigr\}.$$

It follows from Proposition 2 that our theoretical results are applicable to statistical choices of approximate local solutions (2.7) computable through a discrete path of solutions $\widehat{\beta}^{(t)}$ satisfying $\|\widehat{\boldsymbol{\beta}}^{(t)} - \widehat{\boldsymbol{\beta}}^{(t-1)}\|_1 \le a_0\lambda^{(t)}$ and beginning from $\mathbf{0}$ or the Lasso solution, as $\{\mathbf{0}\}$ and the Lasso solution both satisfy the condition $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}(\mathcal{S}; \xi)$.

3. Sorted and smaller penalties. We have studied in Section 2 exact solutions (2.5) and approximate solutions (2.6) and (2.7) for no smaller penalty level than λ_* in the event where λ_* is a strict upper bound of the supreme norm of $\boldsymbol{z} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^o)/n$ as in (2.8). Such penalty or threshold levels are commonly used in the literature to study regularized methods in high-dimensional regression, but this is conservative and may yield poor numerical results. Under the normality assumption (2.10), (2.8) requires $\lambda \geq \lambda_* = (\sigma/\eta)\sqrt{(2/n)\log p}$, but the rate optimal penalty level is the smaller $\lambda \geq \sigma\sqrt{(2/n)\log(p/s)}$ for prediction and coefficient estimation with $s = |\mathcal{S}|$. It is known that for $\log(p/s) \ll \log p$, rate optimal performance in prediction and coefficient estimation can be guaranteed by the Lasso with the nonadaptive smaller penalty level depending on s [3, 26] or the Slope to achieve adaptation to the smaller penalty level [3, 24]. However, it is unclear from the literature whether the same can be done with concave penalties to also take advantage of signal strength, and under what conditions on the design. Moreover, for computational considerations, it is desirable to relax the condition imposed in Section 2 on the approximation error. In this section, we consider approximate solutions for general sorted concave penalties and unsorted nonadaptive smaller penalties. This section is divided into four subsections to discuss respectively sorted penalties, collection of PLSE with sorted and smaller penalties under consideration, a relaxed RE condition and error bounds for prediction and coefficient estimation.

3.1. *Sorted concave penalties.* Given a sequence of sorted penalty levels $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, the sorted ℓ_1 penalty [24] is defined as

$$\text{Pen}(\boldsymbol{b}) = \sum_{j=1}^{p} \lambda_j b_j^{\#}, \tag{3.1}$$

where $b_j^{\#}$ is the jth largest value among $|b_1|, \ldots, |b_p|$.

Here, we extend the sorted penalty beyond ℓ_1. Given a family of univariate penalty functions $\rho(t;\lambda)$ and a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$ with nonincreasing nonnegative elements, we define the associated sorted penalty as

$$\text{Pen}(\boldsymbol{b}) = \rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda}) = \sum_{j=1}^{p} \rho\big(b_j^{\#};\lambda_j\big). \tag{3.2}$$

Although (3.2) seems to be a superficial extension of (3.1), it brings upon potentially significant benefits and its properties are nontrivial. We say that the sorted penalty is concave if $\rho(t;\lambda)$ is concave in t in $[0,\infty)$. We will prove that under an RE condition, the sorted concave penalty inherits the benefits of both the concave and sorted penalties, namely bias reduction for strong signal components and adaptation of the penalty level to the unknown sparsity of $\boldsymbol{\beta}$.

Penalty level and concavity of univariate penalty functions are well understood as we briefly described below (2.2). However, for nonseparable penalties such as

sorted concave penalties, their penalty level and concavity need to be studied carefully. This is done in Section 5. Here, we provide the results for sorted concave penalties in the following proposition. For the sorted penalty (3.1), we define its subdifferential, denoted by $\partial \operatorname{Pen}(\boldsymbol{b})$ or $\partial\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$, as the set of all vectors $\boldsymbol{g}$ satisfying

$$\begin{cases} g_{k_j} = \dot\rho(b_{k_j};\lambda_j), & |b_{k_1}| \ge \cdots \ge |b_{k_s}| > 0, \quad j \le s_{\boldsymbol{b}}, \\ |g_{k_j}| \le \lambda_j, & \{k_{s+1},\dots,k_p\} = \mathcal{S}_{\boldsymbol{b}}^c, \quad j > s_{\boldsymbol{b}}, \end{cases} \tag{3.3}$$

where $\mathcal{S}_{\boldsymbol{b}} = \operatorname{supp}(\boldsymbol{b})$ and $s_{\boldsymbol{b}} = |\mathcal{S}_{\boldsymbol{b}}|$. We denote such vectors $\boldsymbol{g}$ as $\dot{\operatorname{Pen}}(\boldsymbol{b})$ or $\dot\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$. As in the case separable penalties, $\dot{\operatorname{Pen}}(\boldsymbol{b})$ is taken as the favorable choice unless otherwise stated.

PROPOSITION 3. *Let* $\operatorname{Pen}(\boldsymbol{b}) = \rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$ *be as in* (3.2) *with* $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\#}$. *Suppose* $\rho(t;\lambda) = \int_0^{|t|} \dot\rho(x;\lambda)\,dx$. *If* $\dot\rho(x;\lambda)$ *is continuous in* $x > 0$, *then*

$$\liminf_{t\to 0+} t^{-1}\{\operatorname{Pen}(\boldsymbol{b}+t\boldsymbol{u}) - \operatorname{Pen}(\boldsymbol{b})\} \ge \boldsymbol{g}^T\boldsymbol{u} \quad \forall \boldsymbol{u}\in\mathbb{R}^p,\ \boldsymbol{g}\in\partial\operatorname{Pen}(\boldsymbol{b}). \tag{3.4}$$

If $\dot\rho(x;\lambda)$ *is nondecreasing in* λ *almost everywhere in positive* x, *then*

$$\boldsymbol{h}^T\{\dot\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda}) - \dot\rho_{\#}(\boldsymbol{b}+\boldsymbol{h};\boldsymbol{\lambda})\} \le \overline{\kappa}(\rho)\|\boldsymbol{h}\|_2^2 \quad \forall \boldsymbol{b},\boldsymbol{h}. \tag{3.5}$$

The monotonicity condition on $\dot\rho(x;\lambda)$ holds for the ℓ_1, SCAD and MCP. In general, we may define the concavity of $\dot\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$ at $\boldsymbol{b}$ as

$$\overline{\kappa}(\boldsymbol{b}) = \overline{\kappa}(\boldsymbol{b};\boldsymbol{\lambda}) = \sup_{\boldsymbol{h}\ne 0}\{\boldsymbol{h}^T(\dot\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda}) - \dot\rho_{\#}(\boldsymbol{b}+\boldsymbol{h};\boldsymbol{\lambda}))/\|\boldsymbol{h}\|_2^2\}. \tag{3.6}$$

See Section 5. Thus, (3.5) asserts that sorting the penalty does not increase the maximum concavity of a penalty family. For the penalty level, (3.3) asserts that the maximum penalty level at each index $j \in \mathcal{S}_{\boldsymbol{b}}^c$ is λ_{s+1}. Although the penalty level does not reach λ_{s+1} simultaneously for all $j \in \mathcal{S}_{\boldsymbol{b}}^c$, we still take λ_{s+1} as the penalty level for the sorted penalty $\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$. This is especially reasonable when λ_j decreases slowly in j. In Section 3.4, we show that this weaker version of the penalty level is adequate for Gaussian errors, provided a certain minimum penalty level condition in (3.13) below. More important, (3.3) shows sorted penalties automatically pick penalty level λ_{s+1} from the sequence $\{\lambda_j\}$ without requiring the knowledge of s.

A key element of the proof of Proposition 3 is to write (3.2) as

$$\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda}) = \max\left\{\sum_{j=1}^p \rho(b_j;\lambda_{k_j}) : (k_1,\dots,k_p)^T \in \operatorname{perm}(p)\right\}, \tag{3.7}$$

where $\operatorname{perm}(p)$ is the set of all vectors generated by permuting $(1,\dots,p)^T$. For the Slope with $\rho(t;\lambda) = \lambda|b|$, (3.7) follows directly from the rearrangement inequality. The monotonicity of $\dot\rho(t;\lambda)$ in λ, imposed in Proposition 3, is actually also necessary for the equivalence of (3.7) and (3.2) for all $(\boldsymbol{b},\boldsymbol{\lambda})$.

3.2. *PLSE with sorted and smaller penalties*. For general penalties $\mathrm{Pen}(\cdot)$, we consider local approximate solutions satisfying

$$\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n=\dot{\mathrm{P}}\mathrm{en}(\widehat{\boldsymbol{\beta}})+\boldsymbol{v}, \tag{3.8}$$

including nonseparable sorted penalties and separable penalties (2.6). However, instead of imposing ℓ_∞ bounds on the approximation error as in (2.11), we impose in this section a more practical ℓ_∞-ℓ_2 split bound as in (3.12) below. Computational algorithms for approximate solutions and statistical properties of the resulting estimators have been considered in [1, 12, 15, 17, 31] among others. However, these studies of approximate solutions all focus on separable penalties with penalty level (2.8) or higher.

Given a sorted penalty with $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_p\ge 0$, define a sorted norm

$$\|\boldsymbol{b}\|_{\#,s}=\sum_{j=1}^{p-s}(\lambda_{s+j}/\lambda_{s+1})b_j^{\#}\quad\forall\boldsymbol{b}\in\mathbb{R}^{p-s},\ s=|\mathcal{S}|, \tag{3.9}$$

a standardized dual norm induced by the set of $\boldsymbol{g}_{\mathcal{S}^c}$ given in (3.3). Here, $b_j^{\#}$ is the jth largest value among $|b_1|,\ldots,|b_{p-s}|$. Note that $\|\boldsymbol{b}\|_{\#,s}$ reduces to $\|\boldsymbol{b}_{\mathcal{S}^c}\|_1$ when $\lambda_{s+1}=\cdots=\lambda_p$.

Define $\lambda=\lambda_{s+1}$ as the effective penalty level for sorted penalties. Let $\boldsymbol{z}=\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^o)/n$. For positive r and γ and $\{\boldsymbol{w},\boldsymbol{v}\}\subset\mathbb{R}^p$, define

$$\Delta(r,\boldsymbol{w},\boldsymbol{v})=\sup_{\boldsymbol{u}\ne 0}\frac{\boldsymbol{u}_{\mathcal{S}^c}^T\boldsymbol{z}_{\mathcal{S}^c}/\lambda-\eta\|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}-\boldsymbol{u}_{\mathcal{S}}^T\boldsymbol{w}_{\mathcal{S}}-\boldsymbol{u}^T\boldsymbol{v}/\lambda}{r\max\{\|\boldsymbol{u}\|_2,\|\boldsymbol{X}\boldsymbol{u}\|_2\sqrt{\gamma/n}\}}. \tag{3.10}$$

We assume that for a certain constant $\xi>0$,

$$\Delta(r_1,\boldsymbol{w},\boldsymbol{v})<1\quad\text{for some } r_1\le(1-\eta)\xi|\mathcal{S}|^{1/2} \tag{3.11}$$

for each approximate solutions (3.8) under consideration, where $\boldsymbol{v}$ is as in (3.8) and $\boldsymbol{w}=\{\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o)-\boldsymbol{z}\}/\lambda$ with a favorable $\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o)$ as in (3.3). While $(1-\eta)\xi|\mathcal{S}|^{1/2}$ is imposed as an upper bound for r_1 for all approximate solutions under consideration, (3.11) also allows individual approximate solutions to have a r_1 of smaller order, for example, to have smaller $\boldsymbol{w}_{\mathcal{S}}$ and/or $\boldsymbol{v}$. The seemingly complicated (3.10), which summarize conditions in our analysis on the noise vector $\boldsymbol{\varepsilon}$, the penalty level λ and approximation error $\boldsymbol{v}$, is actually not hard to decipher. Consider $\boldsymbol{v}$ satisfying

$$\boldsymbol{u}^T\boldsymbol{v}\le\eta_1\lambda_{s+1}\|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}+r_2\lambda_{s+1}\|\boldsymbol{u}\|_2\quad\forall\boldsymbol{u}\in\mathbb{R}^p \tag{3.12}$$

with $\eta_1\in(0,\eta)$ and $r_2>0$. Proposition 4 below shows that under the normality assumption (2.10) and the minimum penalty level condition

$$\lambda_j\ge\lambda_{*,j}=A_0\sigma\sqrt{(2/n)\log\bigl(p/(\alpha j)\bigr)},\quad j=1,\ldots,p, \tag{3.13}$$

with $A_0>1>\alpha$, (3.11) holds with high probability.

In the simpler case with unsorted smaller penalties, we impose the same condition (3.11). By Proposition 10 in [26] and some algebra, (3.11) also holds with high probability under the same normality assumption (2.10) and the minimum penalty level condition

$$\lambda \geq \lambda_* = (\eta - \eta_1)^{-1}\sigma L/n^{1/2} \tag{3.14}$$

with $\Phi(-L) \leq s/p$, for example, $L = \sqrt{2\log(p/s)}$, when $\|\boldsymbol{w}_{\mathcal{S}}\|_\infty \leq (1-\eta)\xi'$ with $\xi' < \xi$, $s = |\mathcal{S}|$ and $r_2/s^{1/2}$ is sufficiently small.

Building upon (3.11) and the approximate KKT condition (3.8), we define a solution set as in (2.12) as follows:

$$\mathscr{B}(\lambda_*, \kappa_*, \gamma) = \big\{\widehat{\boldsymbol{\beta}} : (3.8) \text{ and } (3.11) \text{ hold}, \overline{\kappa}(\widehat{\boldsymbol{\beta}}) \leq \kappa_*\big\}, \tag{3.15}$$

where $\overline{\kappa}(\boldsymbol{b})$ is the concavity of the sorted penalty as defined in (3.6). Here, λ_* is a minimum penalty level requirement implicit in (3.11), for example, (3.13) for sorted penalties and (3.14) for unsorted smaller penalties. As in Section 2, we define below a subclass of (3.15) to which our analysis applies.

We first relax (2.14), the condition that requires the solution to be connected to $\mathbf{0}$. Let

$$\|\boldsymbol{u}\|_{\#,*} = \|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s} + |\mathcal{S}|^{1/2}\max\big\{\|\boldsymbol{u}\|_2, \|\boldsymbol{X}\boldsymbol{u}\|_2\sqrt{\gamma/n}\big\}. \tag{3.16}$$

We say that two solutions $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ in $\mathscr{B}(\lambda_*, \kappa_*, \gamma)$ are connected by an a_0-chain if there exist $\widehat{\boldsymbol{\beta}}^{(k)} \in \mathscr{B}(\lambda_*, \kappa_*, \gamma)$ with sorted penalty levels $\{\lambda_1^{(k)}, \ldots, \lambda_p^{(k)}\}$ such that for the $a_0 > 0$ in Proposition 5 below:

$$\widehat{\boldsymbol{\beta}}^{(0)} = \widetilde{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}}^{(k^*)} = \widehat{\boldsymbol{\beta}}, \qquad \big\|\widehat{\boldsymbol{\beta}}^{(k)} - \widehat{\boldsymbol{\beta}}^{(k-1)}\big\|_{\#,*} \leq a_0|\mathcal{S}|\lambda_{s+1}^{(k)}, \tag{3.17}$$

$k = 1, \ldots, k^*$. This condition holds if $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are connected through a continuous path in $\mathscr{B}(\lambda_*, \kappa_*, \gamma)$. Similar to (2.13), we define the sparse branch of the solution set $\mathscr{B}(\lambda_*, \kappa_*, \gamma)$ as

$$\mathscr{B}_0(\lambda_*, \kappa_*, \gamma) = \big\{\widehat{\boldsymbol{\beta}} \in \mathscr{B}(\lambda_*, \kappa_*, \gamma) : (3.17) \text{ holds with } \widetilde{\boldsymbol{\beta}} = \mathbf{0}\big\}. \tag{3.18}$$

3.3. *RE condition for sorted and smaller penalties.* As in Section 2, we shall prove that all solutions in the sparse branch (3.18) belong to the following cone:

$$\mathscr{C}_\#(\mathcal{S}; \xi, \gamma) = \big\{\boldsymbol{u} : \|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s} \leq \xi|\mathcal{S}|^{1/2}\max\big(\|\boldsymbol{u}\|_2, (\gamma\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u})^{1/2}\big)\big\}, \tag{3.19}$$

when the RE below is no smaller than the κ_* in (3.15).

The RE for sorted penalties is defined as

$$\mathrm{RE}_\#(S; \xi, \gamma) = \inf\left\{\frac{(\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u})^{1/2}}{\|\boldsymbol{u}\|_2} : \mathbf{0} \neq \boldsymbol{u} \in \mathscr{C}_\#(S; \xi, \gamma)\right\}. \tag{3.20}$$

As $\mathscr{C}_{\#}(S;\xi,\gamma)$ in (3.19) depends on the sorted $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_p)^T$, the infimum in (3.20) is taken over all $\boldsymbol{\lambda}$ under consideration. We note that

$$\mathrm{RE}_{\#}(S;\xi,\gamma)\geq\left[\inf\left\{\frac{(\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u})^{1/2}}{\|\boldsymbol{u}\|_2}:\|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}<\xi|\mathcal{S}|^{1/2}\|\boldsymbol{u}\|_2\right\}\right]\wedge\frac{1}{\gamma}.$$

When the cone is confined to $\lambda_1=\lambda_p$, that is, $\|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}=\|\boldsymbol{u}_{\mathcal{S}^c}\|_1$, the RE condition on $\mathrm{RE}^2_{\#}(S;\xi,\gamma)$ is equivalent to the restricted strong convexity condition [17] as $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1<\xi|\mathcal{S}|^{1/2}\|\boldsymbol{u}\|_2$ implies $\|\boldsymbol{u}\|_1<(\xi+1)|\mathcal{S}|^{1/2}\|\boldsymbol{u}\|_2$. Compared with (2.16), the RE in (3.20) is smaller due to the use of a larger cone. However, this is hard to avoid because the sorted penalty may not control the ℓ_∞ measure of the noise as in (2.8) and we do not wish to impose uniform bound on the approximation error $\boldsymbol{v}$.

3.4. *Error bounds.* Let $\mathscr{B}(\lambda_*,\kappa_*,\gamma)$ be as in (3.15), $\mathscr{B}_0(\lambda_*,\kappa_*,\gamma)$ as in (3.18), $\mathscr{C}_{\#}(\mathcal{S};\xi,\gamma)$ be as in (3.20). We define the good solution set as

$$\begin{aligned}&\mathscr{B}_0^*(\lambda_*,\kappa_*,\gamma)\\&\quad=\mathscr{B}_0(\lambda_*,\kappa_*,\gamma)\cup\big\{\widehat{\boldsymbol{\beta}}\in\mathscr{B}(\lambda_*,\kappa_*,\gamma):\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\in\mathscr{C}_{\#}(\mathcal{S};\xi,\gamma)\big\}.\end{aligned}\tag{3.21}$$

This is the set of approximate solutions (3.8) with the estimation error inside the cone or connected to the origin through a chain. Now we provide predication and estimation error bounds for the good solution under the RE condition.

THEOREM 6. *Suppose* $\mathrm{RE}^2_{\#}(\mathcal{S};\xi,\gamma)\geq\kappa_*>0$. *Let* $\widehat{\boldsymbol{\beta}}\in\mathscr{B}_0^*(\lambda_*,\kappa_*,\gamma)$ *satisfying* $\overline{\kappa}(\widehat{\boldsymbol{\beta}})\leq(1-1/C_0)\,\mathrm{RE}^2_{\#}(\mathcal{S};\xi,\gamma)$ *and condition* (3.11) *with a certain* $r_1\leq(1-\eta)\xi|\mathcal{S}|^{1/2}$ *and* λ *for either a sorted* ($\lambda=\lambda_{s+1}$) *or separable penalty. Then, for any seminorm* $\|\cdot\|$,

$$\|\boldsymbol{h}\|\leq C_0\lambda\sup_{\boldsymbol{u}\neq0}\frac{\|\boldsymbol{u}\|\{r_1F(\boldsymbol{u})-(1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}\}}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}},\tag{3.22}$$

where $\boldsymbol{h}=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o$ *and* $F(\boldsymbol{u})=\|\boldsymbol{u}\|_2\vee(\gamma\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u})^{1/2}$. *In particular,*

$$\frac{\|\boldsymbol{h}_{\mathcal{S}}\|_1+\|\boldsymbol{h}_{\mathcal{S}^c}\|_{\#,s}}{(1+\xi)|\mathcal{S}|^{1/2}}\leq F(\boldsymbol{h})\leq\frac{(\boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h})^{1/2}}{\mathrm{RE}_{\#}(\mathcal{S};\xi,\gamma)}\leq\frac{C_0r_1\lambda}{\mathrm{RE}^2_{\#}(\mathcal{S};\xi,\gamma)}.$$

As an extension of Theorem 5, Theorem 6 provides prediction and ℓ_2 estimation error bounds in the same form along with a comparable sorted ℓ_1 error bound. It demonstrates the benefit of sorted concave penalization as $r_1^2\asymp s_1+s/\log^2(p/s)+r_2^2$ in standard settings as described in Theorem 7 and Corollary 3 below, where $s=|\mathcal{S}|$, s_1 can be viewed as the number of small nonzero coefficients, and r_2 is the ℓ_2 component of the approximation error as in (3.12).

Theorem 6 applies to approximate solutions of PLSE with sorted penalties that adaptively choose penalty level λ_{s+1} from assigned sorted sequence $\lambda_1\geq\cdots\geq\lambda_p$,

or unsorted smaller penalties ($\lambda_j = \lambda\ \forall j$). However, it does not guarantee the selection consistency of (3.8) as a false positive cannot be ruled out at the smaller penalty level or with general approximation errors. On the other hand, as properties of the Lasso at smaller penalty levels and the Slope have been studied in [3, 24, 26] among others, Theorem 6 can be viewed as an extension of their results to concave and/or sorted penalties.

It is also possible to derive error bounds in the case of $C_0 = \infty$ as in Theorem 3 if the ℓ_∞ norm in the definition of the RCIF is replaced by the norm $\|\cdot\|_{\#,*}$ in (3.16). We omit details for the sake of space.

We still need to prove that (3.11) holds with high probability under the normality assumption (2.10) and the minimum penalty level condition (3.13) and (3.14), respectively for sorted and fixed penalty levels. To this end, we need to define a few more quantities. For nonnegative integer s and $0 < \alpha < 1 < A$, define $p_{\alpha,A} = 2\alpha \sum_{k=0}^{\infty} \alpha^{(A-1)A^k}$, $q_{\alpha,A} = (1 - \sqrt{2p_{\alpha,A}})_+$, $x_1 = s/q_{\alpha,A}$, $L_x = \sqrt{2\log(p/(\alpha x))}$, and

$$\mu_{\#,s} = \left\{ \frac{2s(x_1/p)^{A^2-1}(2/q_{\alpha,A})^2}{A^2 L_{s+1}^2 (A^2 L_{x_1}^2 + 2)} \right\}^{1/2} I_{\{s=0\}}. \tag{3.23}$$

We assume here that $x_1 \le p$. This is reasonable when p/s is large.

THEOREM 7. *Suppose the normality condition* (2.10) *on the noise and condition* (3.12) *on the approximation error with* $\eta_1 \in (0, \eta)$ *and* $r_2 > 0$. *Let* $A > 1$. *Suppose* (3.13) *holds for sorted penalties with* $\alpha \in (0, 1/4)$ *and* $A_0 = A/(\eta - \eta_1)$, *and that* (3.14) *holds for unsorted smaller penalties. Let* $\boldsymbol{\beta}^o = \widehat{\boldsymbol{\beta}}^o$ *be the oracle LSE as in* (2.9) *and* $\mathscr{B}_0^*(\lambda_*, \kappa_*, \gamma)$ *be the solution set in* (3.21). *Let* $F(\boldsymbol{u}) = \max\{\|\boldsymbol{u}\|_2, (\gamma \boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u})^{1/2}\}$, $\eta_* = \{\sum_{j=1}^{s} \lambda_j^2/(\lambda_{s+1}^2(s \vee 1))\}^{1/2}$,

$$\xi = \{(1 - 1/a)(1 - \eta)\}^{-1}\big[\{(\eta - \eta_1)^2 \mu_{\#,s}^2 + \eta_*^2\}^{1/2} + r_2/s^{1/2}\big]$$

with the $\mu_{\#,s}$ *in* (3.23) *and* $a \in (0, 1)$. *Suppose* $\kappa_* \le (1 - 1/C^*)\,\mathrm{RE}_{\#}^2(\mathcal{S}; \xi, \gamma)$. *Then there exists an event* Ω *such that* $\mathbb{P}\{\Omega\} \ge 1 - e^{-\xi_* s \log(p/(\alpha(s+1)))}$ *with* $\xi_* = (1 - \eta)^2 \xi^2 \gamma / \{a(\eta - \eta_1)\}^2$, *and that in the event* Ω

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\| \le (1 - \eta) C^* \lambda \sup_{\boldsymbol{u} \ne 0} \frac{\|\boldsymbol{u}\| \{\xi s^{1/2} F(\boldsymbol{u}) - \|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}\}}{\boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u}} \tag{3.24}$$

for all approximate solutions $\widehat{\boldsymbol{\beta}} \in \mathscr{B}_0^*(\lambda_*, \kappa_*, \gamma)$ *and seminorms* $\|\cdot\|$, *where* $\lambda = \lambda_{s+1}$ *for sorted penalties. In addition, for all sorted or unsorted penalties satisfying* $\|\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o)/\lambda\|_2 \le s_1$,

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\| \le C^* \lambda \sup_{\boldsymbol{u} \ne 0} \frac{\|\boldsymbol{u}\| \{r_1 F(\boldsymbol{u}) - (1 - \eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_{\#,s}\}}{\boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u}} \tag{3.25}$$

with $r_1 = (1 - 1/a)^{-1}[\{(\eta - \eta_1)^2 \mu_{\#,s}^2 + s_1^2\}^{1/2} + r_2]$ *and at least probability* $\mathbb{P}\{\Omega\} - e^{-\xi_*' r_1 \log(p/(\alpha(s+1)))}$ *with* $\xi_*' = \gamma/\{a(\eta - \eta_1)\}^2$.

COROLLARY 3. *Suppose* $\lambda = \lambda_{s+1} \asymp \sigma\sqrt{(2/n)\log(p/s)}$ *in* (3.25) *and* $\overline{\kappa}(\widehat{\boldsymbol{\beta}}) \le (1 - 1/C_0)\,\mathrm{RE}^2_{\#}(\mathcal{S};\xi,\gamma)$ *with* $C_0^2/\,\mathrm{RE}^2_{\#}(\mathcal{S};\xi,\gamma) = O(1)$. *Then*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o\|_2^2/n + \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\|_2^2 + \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^o\|_{\#,s}^2/s = O_{\mathbb{P}}(\sigma^2/n) r_1^2 \log(p/s)$$

with $r_1^2 = s_1 + s/\log^2(p/s) + r_2^2$, *where* $s_1 = \|\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{\beta}^o)/\lambda\|_2^2$, *and*

$$\begin{aligned} &\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 + \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^*\|_{\#,s}^2/s \\ &\quad = O_{\mathbb{P}}(\sigma^2/n)\{(s_1 + r_2^2)\log(p/s) + s\}. \end{aligned} \tag{3.26}$$

Moreover, for sorted concave penalty (3.2) *with* $\mathrm{supp}(\dot{\rho}(\cdot;\lambda)) \subseteq [-\gamma\lambda, \gamma\lambda]$,

$$s_1 \le \#\{j \le |\mathcal{S}| : (\boldsymbol{\beta}^o)^{\#}_j \le \gamma\lambda_j\}.$$

Corollary 3 extends Corollary 1 to smaller penalty levels $\lambda = \lambda_{s+1} \ge A_0\sigma\sqrt{(2/n)\log(p/\alpha(s+1))}$ and sorted penalties. In the worst case scenario where $s_1 \asymp s$, the error bounds in Corollary 3 attain the minimax rate [3, 32].

Theorem 7 and Corollary 3 provide sufficient conditions to guarantee simultaneous adaptation of sorted concave PLSE: (a) picking level λ_{s+1} automatically from $\{\lambda_1, \ldots, \lambda_p\}$, and (b) partially removing the bias of the Slope [3, 24] when $s_1 \ll s$, without requiring the knowledge of s or s_1.

Theorem 7 is a direct consequence of Theorem 6 and the following proposition.

PROPOSITION 4. *Suppose the normality assumption* (2.10) *holds. Let* $\eta > \eta_1$ *and* $\{\lambda_j\}$ *be as in* (3.13) *for all sorted penalties. Let* $\{\sum_{j=1}^s \lambda_j^2/(\lambda_{s+1}^2(s \vee 1))\}^{1/2}$, $\boldsymbol{z} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n$, $\boldsymbol{w} = \{\dot{\mathrm{P}}\mathrm{en}(\widehat{\boldsymbol{\beta}}^o) - \boldsymbol{z}\}/\lambda_{s+1}$, $\mu_{\#,s}$ *be as in* (3.23) *with* $q_{\alpha,A} > 0$ *and* $L = \sqrt{2\log(p/(\alpha(s+1)))}$. *Suppose* $\mathrm{supp}(\boldsymbol{z}) \subseteq \mathcal{S}^c$. *Then*

$$\mathbb{P}\{(3.12) \text{ and } \|\boldsymbol{w}_{\mathcal{S}}\|_2 \le w \text{ imply } \Delta(r_1, \boldsymbol{w}, \boldsymbol{v}) < 1\} \ge \Phi\Big(\frac{r_1 L\sqrt{\gamma}}{a(\eta - \eta_1)}\Big) \tag{3.27}$$

when $(1 - 1/a)r_1 \ge \{(\eta - \eta_1)^2\mu_{\#,s}^2 + w^2\}^{1/2} + r_2$ *for* $a > 0$. *In particular, when* $w = \eta_* s^{1/2}$, $\|\boldsymbol{w}_{\mathcal{S}}\|_2 \le w$ *holds automatically and*

$$\mathbb{P}\{(3.12) \text{ implies } (3.11)\} \ge \Phi\Big(\frac{(1-\eta)\xi s^{1/2} L}{a(\eta - \eta_1)\gamma^{-1/2}}\Big), \tag{3.28}$$

with $\xi \ge \{(1 - 1/a)(1 - \eta)\}^{-1}[\{(\eta - \eta_1)^2\mu_{\#,s}^2/s + \eta_*^2\}^{1/2} + r_2/s^{1/2}]$. *Moreover, when* (3.8) *is confined to penalties with the minimum fixed penalty level* $\lambda \ge \lambda_*$ *as in* (3.14), (3.28) *holds with the* L *in* (3.14) *and* $\mu_{\#,s} = \sqrt{4s/(L^4 + 2L^2)}$.

The following lemma provides the basic inequality for analyzing PLSE with sorted and smaller penalties.

LEMMA 2. *Let* $\widehat{\boldsymbol{\beta}}$ *be a solution of* (3.8), $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$, $\boldsymbol{z} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^o)/n$, $\boldsymbol{w} = \{\dot{\mathrm{Pen}}(\boldsymbol{\beta}^o) - \boldsymbol{z}\}/\lambda$ *with* λ *being the fixed penalty level for unsorted penalties or* $\lambda = \lambda_{s+1}$ *for sorted penalties. Then*

$$\begin{aligned} &\boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h} + (1-\eta)\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_{\#,s} - \overline{\kappa}(\widehat{\boldsymbol{\beta}})\|\boldsymbol{h}\|_2^2 \\ &\quad\le \big(\boldsymbol{h}_{\mathcal{S}^c}^T\boldsymbol{z}_{\mathcal{S}^c} - \eta\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_{\#,s}\big) - \lambda\boldsymbol{h}_{\mathcal{S}}^T\boldsymbol{w}_{\mathcal{S}} - \boldsymbol{h}^T\boldsymbol{v}. \end{aligned} \tag{3.29}$$

Next, we provide conditions under which the error $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$ belongs to the cone $\mathscr{C}_{\#}(\mathcal{S};\xi,\gamma)$ for the solutions in $\mathscr{B}_0^*(\lambda_*,\kappa_*,\gamma)$ in (3.18).

PROPOSITION 5. *Let* $\overline{\kappa}(\boldsymbol{b})$ *be as in* (3.6), $\mathscr{B}(\lambda_*,\kappa_*,\gamma)$ *as in* (3.15) *and* $\mathscr{C}_{\#}(\mathcal{S};\xi,\gamma)$ *as in* (3.19). *Suppose* $\mathrm{RE}_{\#}^2(\mathcal{S};\xi,\gamma) \ge \kappa_*$ *and*

$$\Delta\big((1-\eta-a_1)\xi|\mathcal{S}|^{1/2}, \boldsymbol{w}, \boldsymbol{v}\big) \le 1, \quad 0 < a_1 < 1-\eta, \tag{3.30}$$

as in (3.11) *for all* $\{\lambda, \boldsymbol{w}, \boldsymbol{v}\}$ *associated with solutions in* $\mathscr{B}(\lambda_*,\kappa_*,\gamma)$. *Let* $a_2 = a_1\xi^2/\{2\overline{\kappa}(\widetilde{\boldsymbol{\beta}})\}$, $a_3 = a_1(1-\eta)\xi/\{(1-\eta-a_1)(\xi+1)+a_1\}$ *and* $a_0 = \min[a_2, a_2a_3/\{(1-\eta)(\xi\vee 1)\}]$. *Then*

$$\mathscr{B}_0^*(\lambda_*,\kappa_*,\gamma) = \big\{\widehat{\boldsymbol{\beta}} \in \mathscr{B}(\lambda_*,\kappa_*,\gamma) : \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \in \mathscr{C}_{\#}(\mathcal{S};\xi,\gamma)\big\}.$$

4. Local convex approximation. In this section, we develop a local convex approximation (LCA) algorithm for general penalized optimization problem (1.2), especially for sorted penalties. The LCA is a majorization-minimization (MM) algorithm, and is closely related to and in fact very much inspired by the LQA [11] and LLA [14, 36, 38]. We describe the LCA algorithm and how it can be applied to sorted penalizations in the first subsection, and prove the error bounds for estimation and prediction in the second subsection.

4.1. *The LCA algorithm.* Consider a general minimization problem

$$L(\boldsymbol{b}) + \mathrm{Pen}(\boldsymbol{b}), \quad \boldsymbol{b} \in \mathbb{R}^p, \tag{4.1}$$

with $L(\boldsymbol{b})$ a differentiable loss function and $\mathrm{Pen}(\boldsymbol{b})$ a multivariate penalty function. Suppose for a certain continuously differentiable convex $\mathrm{Pen}_-(\boldsymbol{b})$,

$$\mathrm{Pen}_+(\boldsymbol{b}) = \mathrm{Pen}(\boldsymbol{b}) + \mathrm{Pen}_-(\boldsymbol{b}) \tag{4.2}$$

is convex. The LCA algorithm can be written as

$$\boldsymbol{b}^{(\mathrm{new})} = \arg\min_{\boldsymbol{b}}\big\{L(\boldsymbol{b}) + \mathrm{Pen}_+(\boldsymbol{b}) - \boldsymbol{b}^T\dot{\mathrm{Pen}}_-\big(\boldsymbol{b}^{(\mathrm{old})}\big)\big\}. \tag{4.3}$$

This is clearly a MM-algorithm: Because

$$\mathrm{Pen}^{(\mathrm{new})}(\boldsymbol{b}) = \mathrm{Pen}_+(\boldsymbol{b}) - \mathrm{Pen}_-\big(\boldsymbol{b}^{(\mathrm{old})}\big) - \big(\boldsymbol{b} - \boldsymbol{b}^{(\mathrm{old})}\big)^T\dot{\mathrm{Pen}}_-\big(\boldsymbol{b}^{(\mathrm{old})}\big)$$

is a convex majorization of $\mathrm{Pen}(\boldsymbol{b})$ with $\mathrm{Pen}^{(\mathrm{new})}(\boldsymbol{b}^{(\mathrm{old})}) = \mathrm{Pen}(\boldsymbol{b}^{(\mathrm{old})})$,

$$\begin{aligned} L\big(\boldsymbol{b}^{(\mathrm{new})}\big) + \mathrm{Pen}\big(\boldsymbol{b}^{(\mathrm{new})}\big) &\le L\big(\boldsymbol{b}^{(\mathrm{new})}\big) + \mathrm{Pen}^{(\mathrm{new})}\big(\boldsymbol{b}^{(\mathrm{new})}\big) \\ &\le L\big(\boldsymbol{b}^{(\mathrm{old})}\big) + \mathrm{Pen}^{(\mathrm{new})}\big(\boldsymbol{b}^{(\mathrm{old})}\big) \\ &= L\big(\boldsymbol{b}^{(\mathrm{old})}\big) + \mathrm{Pen}\big(\boldsymbol{b}^{(\mathrm{old})}\big). \end{aligned} \tag{4.4}$$

Let $\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$ be the sorted concave penalty in (3.2) with a penalty family $\rho(t;\lambda)$ and a vector of sorted penalty levels $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_p)^T$. Suppose $\dot\rho(x;\lambda) = (\partial/\partial x)\rho(x;\lambda)$ is nondecreasing in λ almost everywhere in positive x, so that Proposition 3 applies. Suppose for a certain continuously differentiable convex function $\rho_-(t)$,

$$\rho_+(t;\lambda_j) = \rho(t;\lambda_j) + \rho_-(t) \quad \text{is convex in } t \text{ for } j = 1,\ldots,p. \tag{4.5}$$

By (3.7), $\rho_{+,\#}(\boldsymbol{b};\boldsymbol{\lambda}) = \rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda}) + \rho_-(\boldsymbol{b})$, the sorted penalty with $\rho_+(t;\lambda)$, is convex in $\boldsymbol{b}$, so that the LCA algorithm for $\rho_{\#}(\boldsymbol{b};\boldsymbol{\lambda})$ can be written as

$$\boldsymbol{b}^{(\mathrm{new})} = \arg\min_{\boldsymbol{b}}\big\{L(\boldsymbol{b}) + \rho_{+,\#}(\boldsymbol{b};\boldsymbol{\lambda}) - \boldsymbol{b}^T\dot\rho_-\big(\boldsymbol{b}^{(\mathrm{old})}\big)\big\}, \tag{4.6}$$

where $\dot\rho_-(\boldsymbol{b})$ is the gradient of $\rho_-(\boldsymbol{b}) = \sum_{j=1}^p \rho_-(b_j)$. The simplest version of LCA takes $\rho_-(t) = t^2\overline{\kappa}(\rho)/2$ with the maximum concavity defined in (2.4), but this is not necessary as (4.5) is only required to hold for the given $\boldsymbol{\lambda}$.

Figure 1 demonstrates that for $p = 1$, the LCA with $\rho_-(t) = t^2\overline{\kappa}(\rho)/2$ also majorizes the LLA with $\mathrm{Pen}^{(\mathrm{new})}(b) = \rho(|b^{(\mathrm{old})}|;\lambda) + \dot\rho(|b^{(\mathrm{old})}|;\lambda)(|b| - |b^{(\mathrm{old})}|)$. With $\rho_-(t) = \lambda|t| - \rho(t;\lambda)$ in (4.5), the LCA is identical to an unfolded LLA with $\mathrm{Pen}^{(\mathrm{new})}(b) = \lambda|b| + \{\dot\rho(b^{(\mathrm{old})};\lambda) - \lambda\,\mathrm{sgn}(b^{(\mathrm{old})})\}(b - b^{(\mathrm{old})})$. The situation is the

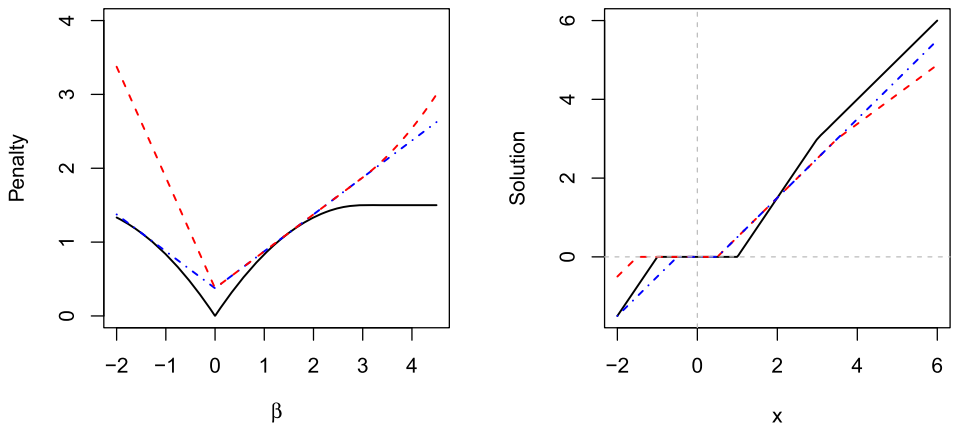

FIG. 1. *Local convex approximation* (*red dashed*), *local linear approximation* (*blue mixed*) *and original penalty* (*black solid*) *for MCP with* $\lambda = 1$ *and* $\overline{\kappa} = 1/3$ *at* $b^{(\mathrm{old})} = 1.5$. *Left*: *penalty function and its approximations*; *Right*: $\arg\min_b\{(x-b)^2/2 + \mathrm{Pen}(b)\}$.

Algorithm 1 FISTA for LCA

Initialization:	$\boldsymbol{x}^1 = \boldsymbol{b}^0 = \boldsymbol{b}^{(\mathrm{old})}$, $t_1 = 1$
Iteration:	$\boldsymbol{b}^k = \mathrm{prox}(\boldsymbol{x}^k - t_* \nabla L(\boldsymbol{x}^k) + t_* \dot{\rho}_-(\boldsymbol{b}^{(\mathrm{old})}); t_* \rho_{+,\#}(\cdot; \boldsymbol{\lambda}))$
	$t_{k+1} = \{1 + (1 + 4t_k^2)^{1/2}\}/2$
	$\boldsymbol{x}^{k+1} = \boldsymbol{b}^k + \{(t_k - 1)/t_{k+1}\}(\boldsymbol{b}^k - \boldsymbol{b}^{k-1})$

same for separable penalties, that is, $\lambda_1 = \lambda_p$. However, the LLA is not feasible for truly sorted concave penalties with $\lambda_1 > \lambda_p$. As the LCA also majorized the LLA, it imposes larger penalty on solutions with larger step size compared with the LLA, but this has little effect in our theoretical analysis in Sections 4.2.

As in (3.8), we may write approximate solutions of (4.3) and (4.6) as

$$\mathbf{0} = \dot{L}\big(\boldsymbol{b}^{(\mathrm{new})}\big) + \dot{\mathrm{Pen}}_+\big(\boldsymbol{b}^{(\mathrm{new})}\big) - \dot{\mathrm{Pen}}_-\big(\boldsymbol{b}^{(\mathrm{old})}\big) + \boldsymbol{v}. \tag{4.7}$$

Such approximate solutions can be viewed as output of iterative algorithms such as the proximal gradient algorithms [2, 18, 21], which approximate $L(\boldsymbol{b})$ by $L(\boldsymbol{x}) + (\boldsymbol{b} - \boldsymbol{x})^T \nabla L(\boldsymbol{x}) + \|\boldsymbol{b} - \boldsymbol{x}\|_2^2/(2t)$ around $\boldsymbol{x}$. For example, for (4.6) with sorted concave penalty, the FISTA [2] can be written as Algorithm 1 where $\rho_{+,\#}(\boldsymbol{b}; \boldsymbol{\lambda})) = \sum_{j=1}^p \rho_+(b_j^{\#}; \lambda_j)$, t_* is the reciprocal of a Lipschitz constant for ∇L or determined in the iteration by backtracking, and

$$\mathrm{prox}(\boldsymbol{x}; \mathrm{Pen}) = \arg\min_{\boldsymbol{b}} \big\{ \|\boldsymbol{b} - \boldsymbol{x}\|_2^2/2 + \mathrm{Pen}(\boldsymbol{b}) \big\} \tag{4.8}$$

is the proximal mapping for convex Pen, for example, $\mathrm{Pen}(\boldsymbol{b}) = t\rho_{+,\#}(\boldsymbol{b}; \boldsymbol{\lambda})$.

For sorted penalties $\rho_{\#}(\boldsymbol{b}; \boldsymbol{\lambda})$, the proximal mapping is not separable but still preserves the sign and ordering in absolute value of the input. Thus, after removing the sign and sorting the input and output simultaneously, it can be solved with the isotonic proximal mapping,

$$\mathrm{iso.prox}(\boldsymbol{x}; \mathrm{Pen}) = \arg\min_{\boldsymbol{b}} \big\{ \|\boldsymbol{b} - \boldsymbol{x}\|_2^2/2 + \mathrm{Pen}(\boldsymbol{b}) : b_j \downarrow \text{ in } j \big\}, \tag{4.9}$$

with $\mathrm{Pen}(\boldsymbol{b}) = t\rho_+(\boldsymbol{b}; \boldsymbol{\lambda})$. Moreover, similar to the Slope [6], (4.9) can be computed by Algorithm 2 as a pool adjacent violators algorithm.

For the MCP, $\rho_+(t; \lambda) = \rho(t; \lambda) + \overline{\kappa} t^2/2 = \int_0^{|t|} \{\lambda \vee (\overline{\kappa} x)\}\, dx$, the univariate

$$\mathrm{prox}\big(x; t\rho_+(\cdot; \lambda)\big) = \mathrm{sgn}(x) \min\big\{ \big(|x| - t\lambda\big)_+, |x|/(1 + t\overline{\kappa}) \big\},$$

is a combination of the soft threshold and shrinkage estimators. Figure 2 plots this univariate proximal mapping for a specific (λ, t). The computational cost of Algorithm 2 for the MCP is no greater than $O(p \log p)$, the same as sorting its input, because the while loop is to run no more than p times and for each merge the cost of the search for the cutoff is $O(\log p)$. We formally state the above discussion in the following proposition.

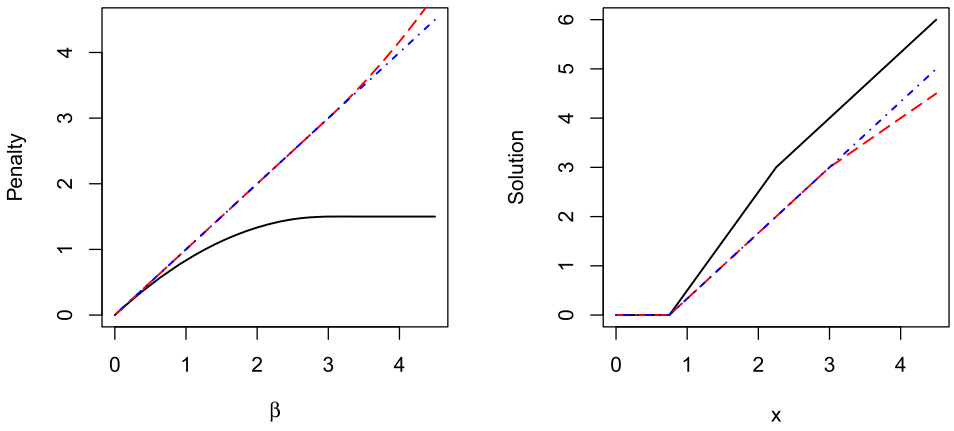

FIG. 2. $\rho_+(\cdot;\lambda) = \rho(t;\lambda) + \overline{\kappa}t^2/2$ *for LCA* (*red dashed*), *Lasso* (*blue mixed*) *and MCP* (*black solid*) *with* $\lambda = 1$ *and* $\overline{\kappa} = 1/3$. *Left*: *penalties*; *Right*: *proximal mappings*.

PROPOSITION 6. *Let* $\boldsymbol{\lambda}$ *be a vector with components* $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

(i) *Let* $\rho_\#(\boldsymbol{b};\boldsymbol{\lambda}) = \sum_{j=1}^p \rho(b_j^\#;\lambda_j)$ *as in* (3.2), *and* $\boldsymbol{b} = \text{prox}(\boldsymbol{x};\rho_\#(\cdot;\boldsymbol{\lambda}))$. *Then* $\text{sgn}(b_j) = \text{sgn}(x_j)$, $|x_j| \geq |x_k|$ *implies* $|b_j| \geq |b_k|$, *and*

$$(b_1^\#, \ldots, b_p^\#)^T = \text{iso.prox}(\boldsymbol{x}^\#;\rho(\cdot;\boldsymbol{\lambda})). \tag{4.10}$$

(ii) $\text{iso.prox}(\boldsymbol{x};t\rho_+(\cdot;\boldsymbol{\lambda}))$ *is solved by Algorithm 2 when* $x_1 \geq \cdots \geq x_p \geq 0$.

(iii) *For the MCP, the computational cost of Algorithm 2 is* $O(p\log p)$ *in the number of operations, and the cost of each iteration of Algorithm 1 is a sum of* $O(p\log p)$ *and the cost of computing the gradient* $\nabla L(\boldsymbol{x}^k)$.

In linear regression, the cost of computing $\nabla L(\boldsymbol{x}^k)$ exactly is of the order np, and the cost of uniformly approximating $\nabla L(\boldsymbol{x}^k)$ in stochastic gradient descent is of no smaller order than $p\log p$. As $\log p \lesssim n$ is required for the RE condition to

Algorithm 2 $\text{iso.prox}(\boldsymbol{x};t\rho_+(\cdot;\boldsymbol{\lambda}))$ with monotone input

Input: $\boldsymbol{\lambda}\downarrow, \boldsymbol{x}\downarrow$

Compute $b_j = \arg\min_b\{(x_j - b)^2/2 + t\rho_+(b;\lambda_j)\}$

While $\boldsymbol{b}$ is not nonincreasing **do**

Identify blocks of violators of the monotonicity constraint,

$b_{j'-1} > b_{j'} \leq b_{j'+1} \leq \cdots \leq b_{j''} > b_{j''+1}, b_{j'} < b_{j''}$

Replace b_j, $j' \leq j \leq j''$, with common value b as follows:

For the MCP, $\rho_+(b;\lambda) = \int_0^{|t|}\{\lambda \vee (\overline{\kappa}x)\}\,dx$, b solves

$\sum_{j=j'}^{j''}\{b - x_j + t\max(\lambda_j, \overline{\kappa}b)\} = 0$

Else $b = \arg\min_{t\geq 0}\sum_{j=j'}^{j''}\{(x_j - t)^2/2 + t\rho_+(t;\lambda_j)\}$

hold uniformly with $|\mathcal{S}| = 2$, the cost of each iteration in Algorithm 1 for sorted MCP is dominated by that of computing or approximating the gradient, required for all penalties.

The computation of the isotonic proximal mapping for SCAD can be carried out in a similar but more complicated fashion, as the region for shrinkage is broken by an interval for soft thresholding.

4.2. *Error bounds of the LCA.* Let $\rho_{\#}(\boldsymbol{b}; \boldsymbol{\lambda})$ be the sorted concave penalty in (3.7), $\rho_{+,\#}(\boldsymbol{b}; \boldsymbol{\lambda})$ also given by (3.7) with $\rho(t; \lambda)$ replaced by the $\rho_+(t; \lambda)$ in (4.5), and $\rho_-(\boldsymbol{b}) = \sum_{j=1}^p \rho_-(b_j)$. For sorted concave PLSE, we write the approximate solution (4.7) of an LCA step as

$$\boldsymbol{X}^T\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(\mathrm{new})}\big)/n = \dot{\rho}_{+,\#}\big(\boldsymbol{b}^{(\mathrm{new})}; \boldsymbol{\lambda}^{(\mathrm{new})}\big) - \dot{\rho}_-\big(\boldsymbol{b}^{(\mathrm{old})}\big) + \boldsymbol{v}^{(\mathrm{new})}, \tag{4.11}$$

as $L(\boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2/(2n)$, where $\dot{\rho}_{+,\#}(\boldsymbol{b}; \boldsymbol{\lambda})$ and $\dot{\rho}_-(\boldsymbol{b})$ can be any members of the respective subdifferentials as characterized in (3.3) and Proposition 3. We study in this subsection statistical properties of estimates generated by iterative applications of (4.11).

We first provide a basic inequality for (4.11) as in Lemmas 1 and 2.

LEMMA 3. *Let* $\boldsymbol{b}^{(\mathrm{new})}$ *be as in* (4.11), $\boldsymbol{w} = (\dot{\rho}_{\#}(\boldsymbol{\beta}^o; \boldsymbol{\lambda}) - \boldsymbol{z})/\lambda_{s+1}^{(\mathrm{new})}$, $\lambda = \lambda_{s+1}^{(\mathrm{new})}$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(\mathrm{new})}$, $\boldsymbol{h} = \boldsymbol{b}^{(\mathrm{new})} - \boldsymbol{\beta}^o$ *and* $\boldsymbol{v}_{\mathrm{carry}} = \dot{\rho}_-(\boldsymbol{b}^{(\mathrm{old})}) - \dot{\rho}_-(\boldsymbol{\beta}^o)$ *be the carryover error in the gradient. Then*

$$\begin{aligned} &\boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h} + (1-\eta)\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_{\#,s} \\ &\quad\le \boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h} + (1-\eta)\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_{\#,s} + D_+\big(\boldsymbol{b}^{(\mathrm{new})}, \boldsymbol{\beta}^o\big) \\ &\quad= \big(\boldsymbol{h}_{\mathcal{S}^c}^T\boldsymbol{z}_{\mathcal{S}^c} - \eta\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_{\#,s}\big) - \lambda\boldsymbol{h}_{\mathcal{S}}^T\boldsymbol{w}_{\mathcal{S}} - \boldsymbol{h}^T\big(\boldsymbol{v}^{(\mathrm{new})} - \boldsymbol{v}_{\mathrm{carry}}\big) \end{aligned} \tag{4.12}$$

when the favorable $\dot{\rho}_{\#}(\boldsymbol{\beta}^o; \boldsymbol{\lambda})$ *is taken from the subdifferential* (3.3), *where* $D_+(\boldsymbol{b}, \boldsymbol{\beta}) = (\boldsymbol{b} - \boldsymbol{\beta})^T\{\dot{\rho}_{+,\#}(\boldsymbol{b}; \boldsymbol{\lambda}) - \dot{\rho}_{+,\#}(\boldsymbol{\beta})\}$ *is the symmetric Bregman divergence for* $\rho_{+,\#}(\boldsymbol{b}; \boldsymbol{\lambda})$.

Lemma 3 asserts that the basic inequality (4.6) holds for the LCA with $\overline{\kappa}(\widehat{\boldsymbol{\beta}}) = 0$ and an extra carryover term $\boldsymbol{h}^T\boldsymbol{v}_{\mathrm{carry}}$. Therefore, when

$$\Delta\big((1-\eta)\xi|\mathcal{S}|^{1/2}, \boldsymbol{w}, \boldsymbol{v}^{(\mathrm{new})} - \boldsymbol{v}_{\mathrm{carry}}\big) < 1 \tag{4.13}$$

as in (3.11), Theorem 6 is applicable to the LCA with $\boldsymbol{h} = \boldsymbol{b}^{(\mathrm{new})} - \boldsymbol{\beta}^o$. We can always take $\rho_-(t)$ with $|(\partial/\partial t)^2\rho_-(t)| \le \overline{\kappa}(\rho)$, so that

$$\|\boldsymbol{v}_{\mathrm{carry}}\|_2 \le \big\|\dot{\rho}_-\big(\boldsymbol{b}^{(\mathrm{old})}\big) - \dot{\rho}_-\big(\boldsymbol{\beta}^o\big)\big\|_2 \le \overline{\kappa}(\rho)\big\|\boldsymbol{b}^{(\mathrm{old})} - \boldsymbol{\beta}^o\big\|_2. \tag{4.14}$$

Next, we summarize the above discussion to show that Theorem 6 can be iteratively applied to the LCA. Let

$$\boldsymbol{b}^{(t)} \leftarrow \mathrm{LCA}(\boldsymbol{b}^{(t-1)}, \rho^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{v}^{(t)}), \tag{4.15}$$

where $\boldsymbol{b}^{(\mathrm{new})} \leftarrow \mathrm{LCA}(\boldsymbol{b}^{(\mathrm{old})}, \rho, \boldsymbol{\lambda}^{(\mathrm{new})}, \boldsymbol{v}^{(\mathrm{new})})$ is the one-step approximate LCA (4.11) with a sorted concave penalty (3.7) generated from $\rho = \rho(t; \lambda)$.

THEOREM 8. *Let $\boldsymbol{b}^{(t)}$ be as in (4.15) with penalty level $\boldsymbol{\lambda}^{(t)}$ and concavity $\kappa(\rho^{(t)}) \le \kappa_0$. Let $\lambda^{(t)} = \lambda_{s+1}^{(t)}$ and $\boldsymbol{w}^{(t)} = (\dot\rho_{\#}^{(t)}(\boldsymbol{\beta}^o; \boldsymbol{\lambda}^{(t)}) - \boldsymbol{z})/\lambda^{(t)}$. Suppose*

$$\Delta\big(r_1^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{v}^{(t)}\big) \le 1, \quad t = 1, \ldots, t_{\mathrm{fin}}, \tag{4.16}$$

with the $\Delta(r, \boldsymbol{w}, \boldsymbol{v})$ in (3.10) and certain $r_1^{(t)} > 0$. Let $\boldsymbol{h}^{(t)} = \boldsymbol{b}^{(t)} - \boldsymbol{\beta}^o$ and $v_0 = \kappa_0 \|\boldsymbol{h}^{(0)}\|_2$ be the initial carryover error. Suppose the RE condition

$$\mathrm{RE}_{\#}(\mathcal{S}; \xi, \gamma) \ge \kappa_0\big\{\lambda^{(t)}/\lambda^{(t+1)}\big\}\big\{r_1^{(t)}\lambda^{(1)}/v_0 + 1\big\}$$

with $(1-\eta)\xi s^{1/2} > v_0/\lambda^{(1)} + r_1^{(t)}$, $t = 1, \ldots, t_{\mathrm{fin}}$. Then, for $t \le t_{\mathrm{fin}}$

$$F\big(\boldsymbol{h}^{(t)}\big) \le \frac{r_1^{(t)}\lambda^{(t)}}{\mathrm{RE}_{\#}^2(\mathcal{S}; \xi, \gamma)} + \theta_0 \big\|\boldsymbol{h}^{(t-1)}\big\|_2 \le \sum_{k=1}^{t} \frac{\theta_0^{t-k} r_1^{(k)} \lambda^{(k)}}{\mathrm{RE}_{\#}^2(\mathcal{S}; \xi, \gamma)} + \theta_0^t \big\|\boldsymbol{h}^{(0)}\big\|_2 \tag{4.17}$$

with $F(\boldsymbol{u}) = \max\{\|\boldsymbol{u}\|_2, (\gamma \boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u})^{1/2}\}$ and $\theta_0 = \kappa_0/\mathrm{RE}_{\#}^2(\mathcal{S}; \xi, \gamma)$.

We note that Theorem 8 is also applicable to the separable penalty (2.2) as $\lambda_1^{(t)} = \cdots = \lambda_p^{(t)} = \lambda^{(t)}$ is allowed here.

To find an approximate solution of PLSE with sorted penalty $\rho_{\#}(\boldsymbol{b}; \boldsymbol{\lambda}_*)$, we may implement (4.15) with a fixed penalty family $\rho(x; \lambda)$ and $\boldsymbol{\lambda}^{(t)} \to \boldsymbol{\lambda}_*$,

$$\boldsymbol{X}^T\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}^{(t)}\big)/n = \dot\rho_{+,\#}\big(\boldsymbol{b}^{(t)}; \boldsymbol{\lambda}^{(t)}\big) - \dot\rho_{-}\big(\boldsymbol{b}^{(t-1)}\big) + \boldsymbol{v}^{(t)},$$

$t = 1, \ldots, t_{\mathrm{fin}}$, where $\boldsymbol{\lambda}_* = (\lambda_{*,1}, \ldots, \lambda_{*,p})^T$ is a vector of target penalty levels, sorted or fixed. For example, we may take

$$\begin{aligned} &\lambda_j^{(t)} = \lambda_{*,j} \vee \big(\theta^{t-1}\lambda^{(1)}\big), \qquad \lambda^{(1)} = \big\|\boldsymbol{X}^T\boldsymbol{y}/n\big\|_\infty, \\ &\boldsymbol{b}^{(0)} = \boldsymbol{0}, \qquad \theta \in (0, 1). \end{aligned} \tag{4.18}$$

Alternatively, we may move gradually from the Lasso to $\rho_{\#}(\boldsymbol{b}; \boldsymbol{\lambda}_*)$, with

$$\begin{cases} \lambda_j^{(t)} = \max\big(\lambda_{*,j}, \theta^t \lambda_{*,1}\big), \quad j \le p, \\ \rho^{(t)}\big(\boldsymbol{b}; \boldsymbol{\lambda}^{(t)}\big) = \theta^t \lambda_{*,1}\|\boldsymbol{b}\|_1 + \big(1 - \theta^t\big)\rho_{\#}\big(\boldsymbol{b}; \boldsymbol{\lambda}^{(t)}\big), \end{cases} \tag{4.19}$$

for a suitable $\theta < 1$. We recall that the prediction and squared ℓ_2 estimation error bounds are of the order $(\sigma^2/n)s \log p$ for the Lasso and $(\sigma^2/n)\{s + s_1 \log(p/s)\}$ for sorted concave penalties, where s_1 can be understood as the number of small nonzero coefficients.

COROLLARY 4. *Let $\boldsymbol{\lambda}_* = (\lambda_{*,1}, \ldots, \lambda_{*,p})^T$ be given by (3.13). Suppose the conditions of Corollary 3 hold with $r_2^2 \lesssim s_1$. Suppose the conditions of Theorem 8*

hold for the penalty sequences (4.18) *and* (4.19). *If the LCA* (4.15) *is implemented with the penalty sequence* (4.18), *then the right-hand side of* (4.17) *is of no greater order than* (3.26) *when*

$$t_{\text{fin}} \geq 3\left\lceil \frac{\log(\lambda^{(1)}/\lambda_{*,p})}{\log(1/\theta)} \vee \frac{\log(n\|\boldsymbol{\beta}^*\|_2^2/\{\sigma^2(s+s_1\log(p/s))\})}{2\log(1/\theta_0)} \right\rceil. \tag{4.20}$$

If the LCA (4.15) *is implemented with the penalty sequence* (4.19), *then the right-hand side of* (4.17) *is of no greater order than* (3.26) *when*

$$t_{\text{fin}} \geq 3\left\lceil \frac{\log((\log p)/\log(p/s))}{2\log(1/\theta)} \vee \frac{\log((s\log p)/(s+s_1\log(p/s)))}{2\log(1/\theta_0)} \right\rceil.$$

The cost of computing a sorted concave PLSE by the LCA is determined by the number of required LCA steps, for example, t_{fin} in Corollary 4, the number of proximal-gradient iterations in each LCA step, for example, as in Algorithm 1, and the cost per proximal-gradient iteration. By Proposition 3, the cost per proximal-gradient iteration for sorted MCP is dominated by the cost of computing the gradient for the squared loss, which does depend on the choice of penalty. By a variation of the analysis in [31], the number of required proximal-gradient iterations in each LCA step is bounded by $C_1\log(C_2\sqrt{s})$ for some constants C_1 and C_2. Thus, in view of (4.20), the cost of computing the sorted PLSE with the MCP is of the same order as that of computing the Lasso and other PLSE as in [31].

5. Discussion. In Sections 2 and 3, we have studied separable and sorted penalties in (2.1) and (3.2), respectively. However, our analysis does not require the penalty to have these specific forms as long as the penalty level and concavity of the penalty are controlled within proper levels. Such a more general version of our theory can be found in the first version of this paper on arXiv.

5.1. *Subdifferential, penalty level and concavity.* As in the cases of separable and sorted penalties, we may define the subdifferential $\partial\,\text{Pen}(\boldsymbol{b})$ of a penalty $\text{Pen}(\boldsymbol{b})$ as the set of all vectors $\boldsymbol{g}$ satisfying (3.4). As $\boldsymbol{g}^T\boldsymbol{u}$ is continuous in $\boldsymbol{g}$, $\partial\,\text{Pen}(\boldsymbol{b})$ is always a closed convex set.

Suppose the loss function $L(\boldsymbol{b})$ in (4.1) is differentiable with derivative $\dot{L}(\boldsymbol{b})$. It follows immediately from the definition of the subdifferential that

$$\liminf_{t\to 0+}\frac{1}{t}\big[\{L(\widehat{\boldsymbol{\beta}}+t\boldsymbol{u})+\text{Pen}(\widehat{\boldsymbol{\beta}}+t\boldsymbol{u})\}-\{L(\widehat{\boldsymbol{\beta}})+\text{Pen}(\widehat{\boldsymbol{\beta}})\}\big]\geq 0$$

for all $\boldsymbol{u}\in\mathbb{R}^p$ iff $-\dot{L}(\widehat{\boldsymbol{\beta}})\in\partial\,\text{Pen}(\widehat{\boldsymbol{\beta}})$. This includes all local minimizers. Let $\dot{\text{P}}\text{en}(\boldsymbol{b})$ denote a member of $\partial\,\text{Pen}(\boldsymbol{b})$. We say that $\widehat{\boldsymbol{\beta}}$ is a local solution for minimizing (4.1) iff the following estimating equation is feasible:

$$-\dot{L}(\widehat{\boldsymbol{\beta}})=\dot{\text{P}}\text{en}(\widehat{\boldsymbol{\beta}}). \tag{5.1}$$

For simplicity, we define the penalty level of $\mathrm{Pen}(\cdot)$ at a point $\boldsymbol{b} \in \mathbb{R}^p$ as

$$\lambda(\boldsymbol{b}) = \sup\big[\lambda : \big\{\boldsymbol{g}_{\mathcal{S}_{\boldsymbol{b}}^c} : \boldsymbol{g} \in \partial\,\mathrm{Pen}(\boldsymbol{b})\big\} \supseteq [-\lambda, \lambda]^{|\mathcal{S}_{\boldsymbol{b}}^c|}\big]. \tag{5.2}$$

This definition is designed to achieve sparsity for solutions of (5.1). Although $\lambda(\boldsymbol{b})$ is a function of $\boldsymbol{b}$ in general, it depends solely on $\mathrm{Pen}(\cdot)$ for many commonly used penalty functions. For sorted penalties, $\lambda(\boldsymbol{b}) = \lambda_{s+1}$ is a somewhat weaker penalty level as we discussed in Section 3.1.

We define the concavity of $\mathrm{Pen}(\cdot)$ at $\boldsymbol{b}$, relative to a target $\widetilde{\boldsymbol{b}}$, as

$$\overline{\kappa}(\boldsymbol{b}) = \overline{\kappa}(\boldsymbol{b}, \widetilde{\boldsymbol{b}}) = \sup\big\{(\widetilde{\boldsymbol{b}} - \boldsymbol{b})^T\big(\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{b}) - \dot{\mathrm{P}}\mathrm{en}(\widetilde{\boldsymbol{b}})\big)/\|\boldsymbol{b} - \widetilde{\boldsymbol{b}}\|_2^2\big\} \tag{5.3}$$

with the convention $0/0 = 0$, where the supremum is taken over all choices $\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{b}) \in \partial\,\mathrm{Pen}(\boldsymbol{b})$ and $\dot{\mathrm{P}}\mathrm{en}(\widetilde{\boldsymbol{b}}) \in \partial\,\mathrm{Pen}(\widetilde{\boldsymbol{b}})$. We use $\overline{\kappa} = \overline{\kappa}(\mathrm{Pen}) = \sup_{\boldsymbol{b},\widetilde{\boldsymbol{b}}} \overline{\kappa}(\boldsymbol{b}, \widetilde{\boldsymbol{b}})$ to denote the maximum concavity of $\mathrm{Pen}(\cdot)$. This definition of concavity includes (2.3) for separable penalties and (3.5) for sorted penalties. However, we may also consider a relaxed concavity, given $s \geq \|\widetilde{\boldsymbol{b}}\|_0$ and $\xi > 0$, as

$$\overline{\kappa}_{1,2}(\boldsymbol{b}; \xi) = \inf\big\{\overline{\kappa}_2(\boldsymbol{b}) + (1+\xi)^2 s\overline{\kappa}_1(\boldsymbol{b})\big\}, \tag{5.4}$$

where infimum is taken over all nonnegative $\overline{\kappa}_1(\boldsymbol{b})$ and $\overline{\kappa}_2(\boldsymbol{b})$ satisfying

$$(\boldsymbol{b} - \widetilde{\boldsymbol{b}})^T\big(\dot{\mathrm{P}}\mathrm{en}(\widetilde{\boldsymbol{b}}) - \dot{\mathrm{P}}\mathrm{en}(\boldsymbol{b})\big) \leq \overline{\kappa}_1(\boldsymbol{b})\|\boldsymbol{h}\|_1^2 + \overline{\kappa}_2(\boldsymbol{b})\|\boldsymbol{h}\|_2^2 \tag{5.5}$$

for all choices of $\dot{\mathrm{P}}\mathrm{en}(\boldsymbol{b})$ and $\dot{\mathrm{P}}\mathrm{en}(\widetilde{\boldsymbol{b}})$. This notion of concavity is more relaxed than the ℓ_2 one in (5.3) because $\overline{\kappa}_{1,2}(\boldsymbol{b}; \xi) \leq \overline{\kappa}(\boldsymbol{b}, \widetilde{\boldsymbol{b}})$ always holds due to the option of picking $\overline{\kappa}_1(\boldsymbol{b}) = 0$.

5.2. *Multivariate mixed penalties.* For $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T \in [0, \infty)^p$, let $\rho(\boldsymbol{b}; \boldsymbol{\lambda}) = \sum_{j=1}^p \rho(b_j; \lambda_j)$ be a separable penalty function with different penalty levels for different coefficients b_j. A multivariate mixed penalty can be constructed by mixing penalties $\rho(\boldsymbol{b}; \boldsymbol{\lambda})$ with a probability measure $\nu(d\boldsymbol{\lambda})$ and a real r_n as follows:

$$\rho_\nu(\boldsymbol{b}) = -r_n^{-1} \log \int \exp\big\{-r_n \rho(\boldsymbol{b}; \boldsymbol{\lambda})\big\}\nu(d\boldsymbol{\lambda}), \tag{5.6}$$

with the convention $\rho_\nu(\boldsymbol{b}) = \int \rho(\boldsymbol{b}; \boldsymbol{\lambda})\nu(d\boldsymbol{\lambda})$ for $r_n = 0$. In particular, for $\rho(t; \lambda) = \lambda|t|$, $r_n = n$ and mixing distributions ν giving i.i.d. two-point components λ_j, (5.6) gives the spike-and-slab Lasso penalty as in [22].

By definition, the subdifferential of (5.6) can be written as

$$\partial\rho_\nu(\boldsymbol{b}) = \left\{\frac{\int \dot{\rho}(\boldsymbol{b}; \boldsymbol{\lambda}) \exp\{-r_n \rho(\boldsymbol{b}; \boldsymbol{\lambda})\}\nu(d\boldsymbol{\lambda})}{\int \exp\{-r_n \rho(\boldsymbol{b}; \boldsymbol{\lambda})\}\nu(d\boldsymbol{\lambda})} : \dot{\rho}(\boldsymbol{b}; \boldsymbol{\lambda}) \in \partial\rho(\boldsymbol{b}; \boldsymbol{\lambda})\right\} \tag{5.7}$$

with $\partial\rho(\boldsymbol{b}; \boldsymbol{\lambda})$ being the set of all vectors $\dot{\rho}(\boldsymbol{b}; \boldsymbol{\lambda}) = (\dot{\rho}(b_1; \lambda_1), \ldots, \dot{\rho}(b_p; \lambda_p))^T$. If we treat $\exp\{-r_n\rho(t; \boldsymbol{\lambda})\}\nu(d\boldsymbol{\lambda})/\int \exp\{-r_n\rho(t; x)\}\nu(dx)$ as conditional density of $\boldsymbol{\lambda}$ under a joint probability $\mathbb{P}_\nu$, we may write (5.7) as

$$\partial\rho_\nu(\boldsymbol{b}) = \big\{\mathbb{E}_\nu\big[\dot{\rho}(\boldsymbol{b}; \boldsymbol{\lambda})|\boldsymbol{b}\big] : \dot{\rho}_j(\boldsymbol{b}; \boldsymbol{\lambda}) = \big(\dot{\rho}(b_1; \lambda_1), \ldots, \dot{\rho}(b_p; \lambda_p)\big)^T\big\}. \tag{5.8}$$

PROPOSITION 7. *Let $\rho_v(\boldsymbol{b})$ be a mixed penalty in (5.6). Let $\mathcal{S}_{\boldsymbol{b}} = \operatorname{supp}(\boldsymbol{b})$ with $s_{\boldsymbol{b}} = |\mathcal{S}_{\boldsymbol{b}}| < p$. Then the concavity of $\rho_v(\boldsymbol{b})$ satisfies*

$$\overline{\kappa}(\boldsymbol{b}) \le \overline{\kappa}(\rho) + \sup_{\boldsymbol{u}\in[\widetilde{\boldsymbol{b}},\boldsymbol{b}]} \phi_{\max}\big(r_n \operatorname{Cov}_v\big(\dot\rho(\boldsymbol{u};\boldsymbol{\lambda}), \dot\rho(\boldsymbol{u};\boldsymbol{\lambda})|\boldsymbol{u}\big)\big), \tag{5.9}$$

where $[\widetilde{\boldsymbol{b}}, \boldsymbol{b}] = \{\boldsymbol{u} = t\boldsymbol{b} + (1-t)\widetilde{\boldsymbol{b}} : 0 \le t \le 1\}$, and (5.5) holds with

$$\overline{\kappa}_2(\boldsymbol{b}) \le \overline{\kappa}(\rho), \qquad \overline{\kappa}_1(\boldsymbol{b}) \le (r_n \vee 0) \sup_{\boldsymbol{u}} \max_{1\le j\le p} \operatorname{Var}_v\big(\dot\rho(u_j;\lambda_j)|\boldsymbol{u}\big).$$

If the components of $\boldsymbol{\lambda}$ are independent given θ, then (5.5) holds with

$$\overline{\kappa}_2(\boldsymbol{b}) \le \overline{\kappa}(\rho) + (r_n \vee 0) \sup_{\boldsymbol{u}} \max_{1\le j\le p} \mathbb{E}_v\big[\operatorname{Var}\big(\dot\rho(u_j;\lambda_j)|\boldsymbol{u},\theta\big)|\boldsymbol{u}\big],$$

$$\overline{\kappa}_1(\boldsymbol{b}) \le (r_n \vee 0) \sup_{\boldsymbol{u}} \max_{1\le j\le p} \operatorname{Var}_v\big(\mathbb{E}_v\big[\dot\rho(u_j;\lambda_j)|\boldsymbol{u},\theta\big]\big|\boldsymbol{u}\big).$$

If in addition $\boldsymbol{\lambda}$ is exchangeable under $v(d\boldsymbol{\lambda})$, the penalty level of (5.6) is

$$\lambda(\boldsymbol{b}) = \mathbb{E}[\lambda_j|\boldsymbol{b}] \quad \forall j \notin \mathcal{S}_{\boldsymbol{b}}. \tag{5.10}$$

Interestingly, (5.9) indicates that mixing $\rho(\boldsymbol{b};\boldsymbol{\lambda})$ with $r_n < 0$ makes the penalty more convex.

For the nonseparable spike-and-slab Lasso [22], the prior is hierarchical where $\beta_j|\boldsymbol{\lambda} \sim (r_n\lambda_j/2)e^{-|t|r_n\lambda_j}$ are independent, $\lambda_j|\theta$ are i.i.d. with $\pi(\lambda_j = \lambda'|\theta) = \theta = 1-\pi(\lambda_j = \lambda''|\theta)$ for some given constants λ' and λ'', and $\theta \sim \pi(d\theta)$. As $\pi(\boldsymbol{0}|\theta) = \{(r_n/2)(\theta\lambda' + (1-\theta)\lambda'')\}^p$ and $\pi(\boldsymbol{0}) = \int \pi(\boldsymbol{0}|\theta)\pi(d\theta)$, the penalty can be written as

$$\rho_v(\boldsymbol{b}) = \frac{-1}{r_n}\log\frac{\pi(\boldsymbol{b})}{\pi(\boldsymbol{0})} = \frac{-1}{r_n}\log\int \exp\Big\{-r_n\sum_{j=1}^p \lambda_j|b_j|\Big\}v(d\boldsymbol{\lambda}),$$

where $\lambda_j \in \{\lambda', \lambda''\}$ are i.i.d. given θ with $v(\lambda_j = \lambda'|\theta) = \theta\lambda'/\{\theta\lambda' + (1-\theta)\lambda''\}$ and $v(d\theta) = \pi(\boldsymbol{0}|\theta)\pi(d\theta)/\pi(\boldsymbol{0})$. The penalty level is given by

$$\lambda(\boldsymbol{b}) = \frac{\int \lambda_\theta \exp\{-r_n\sum_{j\in\mathcal{S}_{\boldsymbol{b}}}\rho_\theta(b_j)\}v(d\theta)}{\int \exp\{-r_n\sum_{j\in\mathcal{S}_{\boldsymbol{b}}}\rho_\theta(b_j)\}v(d\theta)},$$

with $\rho_\theta(t) = \{\theta\lambda' e^{-r_n\lambda'|t|} + (1-\theta)\lambda'' e^{-r_n\lambda''|t|}\}/\{\theta\lambda' + (1-\theta)\lambda''\}$ and $\lambda_\theta = \{\theta(\lambda')^2 + (1-\theta)(\lambda'')^2\}/\{\theta\lambda' + (1-\theta)\lambda''\}$, and the relaxed concavity in (5.4)–(5.5) are bounded by $\overline{\kappa}_2(\boldsymbol{b}) = 0$, $\overline{\kappa}_1(\boldsymbol{b}) \le r_n(\lambda' - \lambda'')^2/4$.

SUPPLEMENTARY MATERIAL

Supplement to "Sorted concave penalized regression" (DOI: 10.1214/18-AOS1759SUPP; .pdf). The Supplementary Material contains detailed proofs for Lemmas 1–3, Propositions 2–7, Theorems 3–6, 8 and Corollary 4. We omit proofs of Theorems 1, 2 and 7 and Proposition 1 as explained above or below their statements.

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609

[2] BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527

[3] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663

[4] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

[5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[6] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[7] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[8] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215. MR2243152

[9] DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp pac-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.

[10] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166

[11] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[12] FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841. MR3782385

[13] FENG, L. and ZHANG, C.-H. (2019). Supplement to "Sorted concave penalized regression." DOI:10.1214/18-AOS1759SUPP.

[14] HUANG, J. and ZHANG, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.* **13** 1839–1864. MR2956344

[15] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[16] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[17] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

[18] NESTEROV, Y. (2007). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. MR3071865

[19] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–403. MR1773265

[20] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

[21] PARIKH, N. and BOYD, S. (2013). Proximal algorithms. In *Foundations and Trends in Optimization*.

[22] ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476
[23] RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* **59** 3434–3447. MR3061256
[24] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953
[25] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166
[26] SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466
[27] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[28] TROPP, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** 1030–1051. MR2238069
[29] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316
[30] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873
[31] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. MR3269977
[32] YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192
[33] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701
[34] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448
[35] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135
[36] ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825
[37] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449
[38] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

SCHOOL OF DATA SCIENCE
CITY UNIVERSITY OF HONG KONG
83 TAT CHEE AVENUE, KOWLOON TONG
HONG KONG
E-MAIL: lfengstat@gmail.com

DEPARTMENT OF STATISTICS AND BIOSTATISTICS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: czhang@stat.rutgers.edu