Statistica Sinica Preprint No: SS-2024-0412	
Title	Nonconvex Penalised Regression and Post-Selection Least Squares Estimation under High Dimensions: a Local Asymptotic Perspective
Manuscript ID	SS-2024-0412
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0412
Complete List of Authors	Xiaoya Xu and Stephen M. S. Lee
Corresponding Authors	Xiaoya Xu
E-mails	xuxiaoya@connect.hku.hk
Notice: Accepted author version.	

Statistica Sinica

Nonconvex Penalised Regression and Post-Selection Least Squares Estimation under High Dimensions: a Local Asymptotic Perspective

Xiaoya XU and Stephen M. S. LEE

Institute of Applied Mathematics, Shenzhen Polytechnic University

Department of Statistics and Actuarial Science, The University of Hong Kong

Abstract:

In the realm of high-dimensional linear regression, nonconvex penalised estimators have enjoyed increasing popularity due to their much acclaimed oracle property, which holds under assumptions weaker than those typically required for convex penalised estimators to enjoy the same property. However, validity of such oracle property of nonconvex penalisation and the accompanying inference tools is questionable in the presence of many weak signals and/or a few moderate signals, which may incur substantial biases. To address this issue, we first provide a more holistic assessment of the selection and convergence properties of nonconvex penalised estimators from a local asymptotic perspective, under a framework which accommodates existence of many weak signals and heavy tail conditions on covariates and random errors. We then show that post-selection least squares estimation has the beneficial effect of removing the bias incurred by nonconvex penalisation of mod-

The work by Xiaoya Xu was supported by Shenzhen Polytechnic University [Project No. 6025310026K]. The work by Stephen M.S. Lee was supported by the General Research Fund grant number 17307321.

erate signals. Post-selection least squares estimators acquire convergence properties more desirable than nonconvex penalised estimators and, in the case of multiple solutions to the nonconvex optimisation program, are ratewise more robust against the choice of selected sets. Empirical results obtained from large-scale simulation studies corroborate our theoretical findings. In particular, the post-selection least squares method is found to improve on nonconvex penalised estimation, especially under heavy-tailed settings.

Key words and phrases: High-dimension; Local asymptotics; nonconvex penalised regression; SCAD; Post-selection least squares estimation; Weak sparsity.

1. Introduction

Consider a linear regression model $Y = \boldsymbol{X}^\top \boldsymbol{\beta}_0 + \epsilon$ for the covariate-response pair $(\boldsymbol{X}, Y) = \left([X^{(1)}, \ldots, X^{(p)}]^\top, Y\right) \in \mathbb{R}^p \times \mathbb{R}$, where $\boldsymbol{\beta}_0 = [\beta_0^{(1)}, \ldots, \beta_0^{(p)}]^\top$ and ϵ denote a p-dimensional vector of unknown parameters and a zero-mean random error independent of $\boldsymbol{X}$, respectively. Let $\mathcal{D}_n = \{(\boldsymbol{X}_i, Y_i) : 1 \leq i \leq n\}$ be a random sample of n replicates of $(\boldsymbol{X}, Y)$. We adopt a local asymptotic framework under which the signal strength of each covariate $X^{(j)}$ is measured by the asymptotic order, scaled by n, of its associated coefficient $\beta_0^{(j)}$. In particular, the biggest order $1 = n^0$ identifies any signal of strength bounded away from 0, while the smallest order $0 = n^{-\infty}$ identifies a non-existent signal.

We envisage a high-dimensional scenario where p may grow, possibly at a much faster rate, with n, and are interested in making inference about rela-

tionships between Y and the "strong" signals, based on the data $\mathcal{D}_n$. Sparsity assumptions on $\boldsymbol{\beta}_0$ help relieve a high-dimensional problem by prompting an easily interpretable model. Penalised regression methods have been regarded as effective devices to select variables and estimate $\boldsymbol{\beta}_0$ simultaneously by shrinking some coefficients towards zero. One prominent example is the LASSO, which has been well received owing to its computational feasibility and desirable theoretical properties. Yet its non-standard, quite possibly intractable, sampling distribution is not amenable to direct statistical inference unless we are willing to impose stringent assumptions on signal strengths and covariate correlations (Lockhart et al., 2014). To address the above problems, Belloni and Chernozhukov (2013) and Javanmard and Montanari (2013) apply ordinary least squares to a strong signal set selected by LASSO estimators. The resulting post-LASSO estimator performs as well as, or even better than, the original LASSO estimator, converging at a possibly faster rate (Belloni and Chernozhukov, 2013). Javanmard and Montanari (2013) prove the two-stage method correctly recovers the active set.

Meanwhile, much progress has been made on the theoretical front to provide important insights into the workings of nonconvex penalised regression as an alternative to LASSO under weaker conditions for support recovery guarantees. Apart from its well-known oracle properties (Fan and Li, 2001; Zhang, 2010;

Fan and Lv, 2011), many useful results have been established on the properties of local minima and their relationships with sparsity, the global minimum and the oracle estimator. Kim and Kwon (2012) state a uniqueness condition under which a nonconvex penalised squared loss possesses a unique sparse local minimum, which reduces to the oracle estimator under a more stringent beta-min condition. Loh and Wainwright (2015) establish conditions for consistency of all local minima, and derive bounds on their convergence rates. Loh and Wainwright (2017) show under mild design conditions that nonconvex regularisation allows for a unique local minimum with a guaranteed l_∞-convergence rate. They show further that for sub-Gaussian errors and covariates, this minimum is global under an incoherence condition, and is oracle under a beta-min condition.

Perhaps the most advocated advantage of nonconvex penalised methods lies in their oracle properties, which are often perceived as a theoretical guarantee for standard least squares inference under the selected model. This also explains why bias correction has never been deemed necessary for nonconvex penalised regression, unlike the case of LASSO for which much effort has been spent on finding ways of "debiasing". However, the limitation of nonconvex penalised methods have been revealed by Leeb and Pötscher (2008) and Pötscher and Leeb (2009). We shall show under our general local asymptotic framework that such oracle properties may provide a false sense of security, especially when the

beta-min condition is violated. Indeed, substandard empirical performances of nonconvex penalised estimators have urged development of post-selection remedies. Xiao and Sun (2019) show for finite samples that the post-SCAD estimator, tuned using BIC, outperforms other penalised estimators and their corresponding post-selection variants. Ratnasingam and Ning (2021) develop quantile regression with a post-SCAD method and give a real-data demonstration. Fousekis and Grigoriadis (2022) apply the post-SCAD method directly to conduct statistical inference. Other applications of the post-selection remedy for nonconvex penalised estimators can be found in economics and finance (Uematsu and Tanaka, 2019; Xiao and Sun, 2020; Bonaccolto, 2021). Despite its growing prominence in applications, theoretical properties of the post-selection estimator built upon nonconvex penalisation, especially under high dimensions, have been rarely investigated. This motivates us to fill the gap.

In reality, strict sparsity of $\boldsymbol{\beta}_0$, a common assumption made for high-dimensional penalised methods, may easily be violated by the existence of many weak signals among the covariate set $\boldsymbol{X}$, which correspond to coefficients $\beta_0^{(j)}$ close to, but not exactly equal to, zero. This calls for a practically more relevant view of signal strength that spans a continuous spectrum, over which may scatter many weak signals at different positions. Such a view has, however, received only sporadic attention in the literature (e.g. Belloni et al., 2012; Javanmard and Lee, 2020).

For example, Horowitz and Huang (2013) set up a generalised sparsity condition to accommodate weak signals and address the issue of strong signal selection. Zhang and Zhang (2014) propose a thresholded low dimensional projection estimator under a weaker assumption which they term the capped l_1 sparsity condition. Qu and Shi (2016) require weak signals to share a specific asymptotic order determined by regularisation parameters, and strive to make inferences for both weak and strong signals. Zhao et al. (2018) also allow for the presence of a small number of weak signals, but have overstated the capacity of MCP for bias reduction, as the presence of even a few weak signals may significantly impair estimation accuracy of strong signals, as will be made clear by our local asymptotic theory. Other examples include Shao and Deng (2012), who consider a non-sparse parameter vector which may contain many small components, and Liu et al. (2017), who introduce cliff-weak-sparsity for a bootstrap procedure for constructing confidence intervals.

Under our local asymptotic framework, a more holistic insight can be gained into the asymptotic properties of a nonconvex penalised estimator $\hat{\boldsymbol{\beta}}$ and the corresponding post-selection least squares estimator derived from the selected model, without assuming stringent sparsity or eigenvalue conditions typically imposed in previous studies. Thus, our focus is not so much on establishing sufficient conditions for oracle performance of high-dimensional nonconvex pe-

nalised regression as on a holistic assessment of its performance under general, practically relevant, circumstances. Our contributions can be summarised briefly as follows.

1. We catalogue a series of changing configurations of signal strengths over the asymptotic spectrum, under which the asymptotic properties of $\hat{\boldsymbol{\beta}}$ undergo gradual phase changes, ranging from inconsistency to exhibition of a kind of oracle property, generalised under a local asymptotic framework.

2. We detail the componentwise asymptotic properties of $\hat{\boldsymbol{\beta}}$ under all general configurations of signal strengths within a continuous spectrum delimited by the conditions $\{|\beta_0^{(j)}| = O(1) : 1 \leq j \leq p\}$. Our theory therefore provides a microscopic view of (local) asymptotic properties which cannot be revealed by non-asymptotic bounds on estimation errors measured in vector norms.

3. We show that post-selection least squares estimation is a worthwhile strategy for improving the accuracy of $\hat{\boldsymbol{\beta}}$. Specifically, it engenders a bias correction for nonconvex penalised methods during a transition phase of signal strength configurations and, unlike the debiased LASSO, retains the benefits brought by sparsity in model interpretation and subsequent inference tasks.

2. Consistent nonconvex penalised estimator

2.1 Preliminaries and notation

Before detailing our framework, we introduce the notation used throughout the paper. For any $\boldsymbol{v} = [v_1, \ldots, v_q]^\top \in \mathbb{R}^q$ and $\mathbb{A} \subset \{1, \ldots, q\}$, write $\boldsymbol{v}^{\mathbb{A}}$ for the subvector of $\boldsymbol{v}$ indexed by $\mathbb{A}$, $\mathrm{diag}(\boldsymbol{v})$ for the diagonal matrix with diagonal elements $\boldsymbol{v}$ and $\mathrm{supp}(\boldsymbol{v}) = \{j : v_j \neq 0\}$. For any matrix $\mathcal{M} \in \mathbb{R}^{d_1 \times d_2}$ and any subsets $\mathbb{A} \subset \{1, \ldots, d_1\}$ and $\mathbb{B} \subset \{1, \ldots, d_2\}$, write $\mathcal{M}_{\mathbb{A}\mathbb{B}}$ for the submatrix of $\mathcal{M}$ whose rows and columns are indexed by $\mathbb{A}$ and $\mathbb{B}$ respectively. For any, possibly random, real sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \prec b_n$, $a_n \succ b_n$, $a_n \preceq b_n$ and $a_n \succeq b_n$ for $a_n = o_p(b_n)$, $b_n = o_p(a_n)$, $a_n = O_p(b_n)$ and $b_n = O_p(a_n)$, respectively. Write $a_n \asymp b_n$ if $a_n \succeq b_n$ and $a_n \preceq b_n$. More generally, for a vector or matrix sequence $\{a_n\}$, the above relations are interpreted componentwise. For any $r, s \in \mathbb{R}$, define $(r)_+ = \max\{0, r\}$, $r \vee s = \max\{r, s\}$, $r \wedge s = \min\{r, s\}$ and, for $r \neq 0$, $\mathrm{sgn}(r) = \mathbf{1}\{r > 0\} - \mathbf{1}\{r < 0\}$, where the indicator function is denoted by $\mathbf{1}\{\cdot\}$. For any index subset $\mathbb{A} \subset \{1, \ldots, p\}$, write $\mathbb{A}^c$ for $\{1, \ldots, p\} \backslash \mathbb{A}$ and $|\mathbb{A}|$ for the cardinality of $\mathbb{A}$. The l_r-norm on $\mathbb{R}^q$ is denoted by $\|\cdot\|_r$, for $0 \leq r \leq \infty$.

Our analysis revolves around a local asymptotic framework where the strength of signals can vary and is measured by its asymptotic order with respect to the

sample size n. We define below several key quantities, central to characterising the phase transitions in the asymptotic properties presented in our main theorems.

1. The *oracle active set* $\mathcal{A}_0 = \{j : |\beta_0^{(j)}| \succeq \lambda/n\}$ contains the "strong" signals, whose magnitudes reach or exceed the effective penalty level asymptotically.

2. The total strength of "weak" signals (those not in $\mathcal{A}_0$) is measured by $B_0 = \left\|\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\right\|_1$.

3. The magnitude of the weakest "strong" signal is $B_U = \min\left\{|\beta_0^{(j)}| : j \in \mathcal{A}_0\right\}$ or ∞ if $\mathcal{A}_0$ is empty.

4. The above measures are drawn upon to define a critical quantity $\psi = (\lambda/n)\{1 - B_U/(\alpha\kappa)\}_+ \vee B_0$, which can be viewed as a measure of proximity between the strong and weak groups of signals and plays a key role in determining the asymptotic behavior of the penalised estimators.

2.2 Problem setting

For tuning parameters $\kappa, \lambda > 0$, consider a nonconvex penalised estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$, which is sparse and locally minimises

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 + \lambda\kappa \sum_{j=1}^{p} q(|\beta_j|/\kappa) \tag{2.1}$$

over $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top \in \mathbb{R}^p$, in the sense that $\hat{\boldsymbol{\beta}}$ satisfies the subgradient conditions

$$\sum_{i=1}^{n} \boldsymbol{X}_i^{\{j\}} \big(Y_i - \boldsymbol{X}_i^{\hat{\mathcal{A}}\top} \hat{\boldsymbol{\beta}}^{\hat{\mathcal{A}}}\big) \begin{cases} \in [-\lambda/2,\, \lambda/2], & j \in \hat{\mathcal{A}}^c, \\ = (\lambda/2)\, \mathrm{sgn}(\hat{\beta}_j) q'\big(|\hat{\beta}_j|/\kappa\big), & j \in \hat{\mathcal{A}}, \end{cases} \tag{2.2}$$

where $\hat{\mathcal{A}} = \mathrm{supp}(\hat{\boldsymbol{\beta}})$ satisfies $|\hat{\mathcal{A}}| \preceq 1$, and define, for some $\alpha > 2$, a nonconvex penalty function

$$q(y) = \begin{cases} y, & 0 \le y \le 1, \\ y - 2^{-1}(y-1)^2/(\alpha - 1), & 1 < y < \alpha, \\ (\alpha+1)/2, & y \ge \alpha. \end{cases}$$

The dimension p is allowed to grow with n and the parameters α, κ, λ may depend on n. It can be shown that under the condition $\lambda \asymp n\alpha\kappa$, the nonconvex penalty q satisfies the amenability condition introduced by Loh and Wainwright (2017). In particular, the general formulation (2.1) reduces to MCP (Zhang, 2010) if we push $(\alpha, \kappa) \to (\infty, 0)$ with $\alpha\kappa = \gamma\lambda/(2n)$ for a regularisation parameter $\gamma > 0$. Furthermore, setting $\gamma = 1/2$ gives rise to the hard thresholding rule with threshold $\lambda/(4n)$ (Antoniadis and Fan, 2001). On the other hand, if we set $\kappa = \lambda/(2n)$ with a fixed $\alpha \in (2, \infty)$, then (2.1) reduces to SCAD (Fan and Li, 2001). Though not of our primary interest, setting $(\alpha, \kappa) = (\infty, 1)$ reduces (2.1) to the LASSO objective with a convex penalty.

We adopt a local asymptotic framework in our study. Within this framework, all $\beta_0^{(j)}$'s are assumed to be $O(1)$ and each $\beta_0^{(j)}$ takes on a specific asymptotic order $\mathscr{O}_j$, where $\mathscr{O}_j$ either converges to 0 at a fixed rate or equals 1 identically as $n \to \infty$. If $\beta_0^{(j)} \neq 0$, its order $\mathscr{O}_j$ is designated by a convergent sequence, commonly of the form $n^{-\omega}$ for some $\omega \geq 0$, such that $\lim_{n\to\infty} \mathscr{O}_j^{-1}\beta_0^{(j)} \in \mathbb{R} \setminus \{0\}$. The faster $\mathscr{O}_j$ converges to 0, the weaker is the j-th signal. We set $\mathscr{O}_j \equiv 0$ for a "non-existent signal" with $\beta_0^{(j)} = 0$. This framework allows us to study a continuum of signal strengths, from strong signals ($\mathscr{O}_j \equiv 1$) through weak signals ($0 < \mathscr{O}_j \prec 1$) to non-existent signals ($\mathscr{O}_j \equiv 0$). Define $\mathcal{A}_0 = \{j : |\beta_0^{(j)}| \succeq \lambda/n\}$, $B_0 = \left\|\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\right\|_1$,

$$B_U = \begin{cases} \min\left\{|\beta_0^{(j)}| : j \in \mathcal{A}_0\right\}, & \mathcal{A}_0 \neq \emptyset, \\ \infty, & \mathcal{A}_0 = \emptyset. \end{cases} \quad \text{and} \quad \psi = \frac{\lambda}{n}\Big(1 - \frac{B_U}{\alpha\kappa}\Big)_+ \vee B_0.$$

Existence of a consistent sparse local minimum of (2.1) is guaranteed by the weak conditions ($A2$), ($A3$) and ($A4$), to be introduced in Section 2.3. For any local minimum $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \ldots, \hat{\beta}_p]^\top$ consistent for $\boldsymbol{\beta}_0$, let $r_j \to \infty$ denote the sharpest convergence rate such that the scaled componentwise error $U_j = r_j(\hat{\beta}_j - \beta_0^{(j)})$ satisfies $U_j \asymp 1$, for $j = 1, \ldots, p$. We assume without loss of generality that $r_j = r_k$ if and only if $r_j \asymp r_k$, for any $j, k \in \{1, \ldots, p\}$. Denote by $\hat{\mathscr{K}}$ the collection of selected sets $\hat{\mathcal{A}} = \mathrm{supp}(\hat{\boldsymbol{\beta}})$, each formed by a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ satisfying (2.2).

Under the conventional fixed-parameter asymptotic framework, it is typically assumed that $B_0 = 0$ and $B_U \asymp 1$. Recent literature has seen a variety of weak oracle properties, established under less restrictive conditions. For example, Fan and Lv (2011) consider the case where $B_U \succ \lambda/n$ and $B_0 = 0$, and obtain for $\hat{\boldsymbol{\beta}}$ an l_∞-loss of order between B_U and λ/n, which they term a weak oracle property. Kim et al. (2016) give sufficient conditions for the existence of a local minimum which is oracle. For SCAD and MCP, the conditions require $B_0 = 0$ and $B_U \succ (\lambda/n) \vee \sqrt{\|\boldsymbol{\beta}_0\|_0/n}$. Both papers, among many others, do not consider cases where there may exist coefficients with $0 \neq \beta_0^{(j)} \preceq \lambda/n$. Our local asymptotic spectrum allows for all general configurations of the $\beta_0^{(j)}$'s, with each $\beta_0^{(j)}$ characterised by a precise order $\mathscr{O}_j$ within the spectrum.

Instead of providing non-asymptotic bounds on norms of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, as is common in the literature on high-dimensional regression, our focus is on the precise componentwise convergence rates r_j and the form of the weak limits of $(U_1, \ldots, U_p)$.

2.3 Theory

Define $\hat{C} = [\hat{C}_{st} : s, t = 1, \ldots, p] = n^{-1} \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$ and $\mathcal{C} = [\mathcal{C}_{st} : s, t = 1, \ldots, p] = \mathbb{E}\big[\boldsymbol{X}\boldsymbol{X}^\top\big]$. For the sake of generality, we refrain from imposing on $\mathcal{C}$ restrictive conditions such as the irrepresentable condition, the sparse Riesz con-

dition or various restricted eigenvalue conditions, which are typically required to ensure variable selection consistency of lasso-type methods. Instead we make only a mild boundedness assumption on each $\mathcal{C}_{ss}$:

($C1$) there exist constants $\underline{\mathcal{C}}, \bar{\mathcal{C}} > 0$ such that $\underline{\mathcal{C}} \leq \mathcal{C}_{ss} \leq \bar{\mathcal{C}}$ for any $s \in \{1, \ldots, p\}$.

Assume that $(\boldsymbol{X}, Y)$ satisfies either one of the following tail conditions for all $j, j' \in \{1, \ldots, p\}$.

($\mathcal{T}_1$) $\mathbb{P}(|\epsilon| > t) + \mathbb{P}(|X^{(j)} - \mathbb{E}X^{(j)}| > t) \preceq e^{-ct^2}$ as $t \to \infty$, for some $c > 0$.

($\mathcal{T}_2$) $\mathbb{P}(|\epsilon| > t) \asymp t^{-\gamma}$ and $\mathbb{P}(|X^{(j)} - \mathbb{E}X^{(j)}| > t) \preceq t^{-2\gamma}$ as $t \to \infty$, for some $\gamma > 2$.

Note that ($\mathcal{T}_1$) and ($\mathcal{T}_2$) amount to a sub-Gaussian and a moderately heavy tail conditions, respectively, both supporting a finite variance. A third condition, which imposes a heavier tail with infinite variance, is included in a generalised version of our theory discussed in Appendix 1. Writing $\epsilon_i = Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0$, define $\boldsymbol{W} = [W_1, \ldots, W_p]^\top = n^{-1/2} \sum_{i=1}^n \epsilon_i \boldsymbol{X}_i$. Note that for $\lambda \succ \sqrt{n}$, the signals captured by the index set $\mathcal{A}_0$ are not confounded with sampling noise.

Local asymptotic properties of $\hat{\mathcal{A}}$ and $\hat{\boldsymbol{\beta}}$ may take on numerous possible forms depending on the precise componentwise orders $\mathscr{O}_j$ of $\beta_0^{(j)}$ and their relationships with $\mathcal{C}$. For a general, seamless, exposition of the phase changes

undergone by our local asymptotic theory, we shall impose conditions to exclude pathological cases of $(\boldsymbol{\beta}_0, \mathcal{C})$. Define, for $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_p]^\top \in \{1,2,3\}^p$, $k_0 = \pm 1$ and any $\mathcal{B} \subset \{1, \ldots, p\}$ with $|\mathcal{B}| = O(1)$,

$$C_0(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}) = \Big[\mathcal{C}_{\mathcal{B}\mathcal{B}} - \frac{\lambda}{2n(\alpha-1)\kappa}\mathrm{diag}\big(\mathbf{1}\{\nu_j = 3\} : j \in \mathcal{B}\big)\Big],$$

$$f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0) = C_0(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu})^{-1}\Big\{ - \mathcal{C}_{\mathcal{B}\mathcal{B}^c}\boldsymbol{\beta}_0^{\mathcal{B}^c}$$
$$+ \frac{\lambda}{2n}\Big[k_0\big(\mathbf{1}\{\nu_j = 2\} + \frac{\alpha}{\alpha-1}\mathbf{1}\{\nu_j = 3\}\big) - \frac{\beta_0^{(j)}}{(\alpha-1)\kappa}\mathbf{1}\{\nu_j = 3\} : j \in \mathcal{B}\Big]\Big\}.$$

We assume the following condition.

$(A1)$ For any $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_p]^\top \in \{1,2,3\}^p$, $k_0 = \pm 1$, $\mathcal{B} \subset \{1, \ldots, p\}$ with $|\mathcal{B}| = O(1)$, $j \in \mathcal{B}$ and $j' \in \mathcal{B}^c$,

$$\lim_{n\to\infty} \det\big(C_0(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu})\big) \neq 0,$$
$$\lim_{n\to\infty} f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}/\beta_0^{(j)} \neq 1 \text{ if } \mathscr{O}_j \asymp \lambda/n,$$
$$\lim_{n\to\infty} \big|\beta_0^{(j)} - f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}\big|/(\alpha\kappa) \neq 1,$$
$$\lim_{n\to\infty} \big|\beta_0^{(j)} - f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}\big|/\kappa \neq 1 \text{ if } \kappa \succeq \lambda/n,$$
$$\lim_{n\to\infty} (2n/\lambda)\big|\mathcal{C}_{\{j'\}\mathcal{B}} f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0) + \mathcal{C}_{\{j'\}\mathcal{B}^c}\boldsymbol{\beta}_0^{\mathcal{B}^c}\big| \neq 1.$$

The condition $(A1)$ rules out cases which lie on the boundaries between contiguous phases of asymptotic behaviour displayed by $(\hat{\mathcal{A}}, \hat{\boldsymbol{\beta}})$, thereby preventing them from obtruding on the general trend portrayed in our local asymptotic the-

ory. From a measure-theoretic perspective, if we embed the sequence

$$\big(\{\mathscr{O}_1^{-1}\beta_0^{(1)}\}_{n=1}^{\infty}, \{\mathscr{O}_2^{-1}\beta_0^{(2)}\}_{n=1}^{\infty}, \ldots, \mathcal{C}_{11}, \mathcal{C}_{21}, \mathcal{C}_{22}, \mathcal{C}_{31}, \ldots\big)$$

in a product space formed by an infinite collection of convergent real sequences endowed with the sup norm, then the boundary cases excluded by $(A1)$ constitute only a null set, omission of which does not undermine generality of our theory to all intents and purposes. In a similar vein, Zhang and Zhang (2012) introduce for their general theory of nonconvex penalised regression an abstract ℓ_2 regularity condition on the covariates to match the sparsity level of $\boldsymbol{\beta}_0$, which is more restrictive than our assumptions $(C1)$ and $(A1)$.

To ensure existence of a consistent sparse local minimum of (2.1) we impose further the conditions

$(A2)$ $\lim_{n\to\infty} \dfrac{\lambda}{n(\alpha-1)\kappa} < 2\underline{\mathcal{C}}$ and $\sqrt{n} \prec \lambda \prec n$;

$(A3)$ $|\mathcal{A}_0| \vee B_0 \preceq 1$.

Note that $(A3)$ amounts to a kind of weak sparsity condition, which requires that the number of strong signals in $\mathcal{A}_0$ be bounded but does not restrict the number of weak signals outside $\mathcal{A}_0$, provided they have a bounded total strength.

In order to study the asymptotic behaviour of $\hat{\mathcal{A}}$, we introduce a penalised parameter $\boldsymbol{\beta}^* = [\beta_1^*, \ldots, \beta_p^*]^\top$, a population analogue of $\hat{\boldsymbol{\beta}}$ defined to locally

minimise $\mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\beta})^2 + n^{-1}\lambda\kappa\sum_{j=1}^{p} q(|\beta_j|/\kappa)$ in the sense that it satisfies

$$\begin{cases} (2n/\lambda)\left|\mathcal{C}_{\{j\}\mathcal{A}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathcal{A}} - \mathcal{C}_{\{j\}\mathcal{A}^c}\boldsymbol{\beta}_0^{\mathcal{A}^c}\right| \le 1, & j \in \mathcal{A}^c, \\ \mathcal{C}_{\{j\}\mathcal{A}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathcal{A}} = -(2n)^{-1}\lambda\,\mathrm{sgn}(\beta_j^*)q'(|\beta_j^*|/\kappa) + \mathcal{C}_{\{j\}\mathcal{A}^c}\,\boldsymbol{\beta}_0^{\mathcal{A}^c}, & j \in \mathcal{A}. \end{cases} \tag{2.3}$$

Denote by $\mathscr{K}_n$ the collection of selected sets $\mathcal{A} = \mathrm{supp}(\boldsymbol{\beta}^*)$, each corresponding to a sparse and consistent local minimum $\boldsymbol{\beta}^*$ with $|\mathcal{A}| \preceq 1$. We assume

$(A4)$ $\limsup_{n\to\infty} \mathscr{K}_n \neq \emptyset$,

which implies that at least one index subset is captured by the $\mathscr{K}_n$'s infinitely often. Different notions of penalised parameters have been considered in the literature. For example, Greenshtein and Ritov (2004) study persistence of predictor selection procedures with respect to an "oracle" penalised parameter which minimises $\mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\beta})^2$ subject to $\|\cdot\|_0$- or $\|\cdot\|_1$-sparsity constraints. Yu et al. (2019) consider inferences about both $\boldsymbol{\beta}_0$ and the penalised parameter which minimises $\mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\beta})^2$ subject to a general convex or nonconvex penalty.

We examine in what follows the componentwise convergence rates and weak limits of all sparse and consistent local minima $\hat{\boldsymbol{\beta}}$'s. In the following lemma, we write $(\hat{\mathscr{K}}_n, \hat{\mathcal{A}}_n)$ for $(\hat{\mathscr{K}}, \hat{\mathcal{A}})$ to highlight their dependence on n.

Lemma 1. *Assume the conditions $(C1)$, $(A1)$, $(A2)$, $(A3)$, $(A4)$. Assume either $(\mathcal{T}_1)$ holds with $p \prec e^{\lambda^2/n}$ or $(\mathcal{T}_2)$ holds with $p \prec \lambda^2/n$. Then,*

(i) for every sequence $\mathcal{A}_n \in \mathscr{K}_n$*, there exists a sequence of selected sets* $\hat{\mathcal{A}}_n \in \hat{\mathscr{K}}_n$ *such that* $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}_n) \to 1$ *as* $n \to \infty$*;*

(ii) for every sequence of selected sets $\hat{\mathcal{A}}_n \in \hat{\mathscr{K}}_n$*, there exists a sequence* $\mathcal{A}_n \in \mathscr{K}_n$ *such that* $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}_n) \to 1$ *as* $n \to \infty$.

Remark 1. The conditions $p \prec e^{\lambda^2/n}$ and $p \prec \lambda^2/n$ place constraints on the rate at which the dimension p can grow with the sample size n, which explicitly relates the permissible dimensionality to the tail condition. Under a sub-Gaussian tail ($\mathcal{T}_1$), the condition $p \prec e^{\lambda^2/n}$ allows p to grow exponentially fast. On the other hand, the moderate tail condition ($\mathcal{T}_2$) restricts p to grow at a polynomial rate.

Remark 2. Under the conditions of Lemma 1, each sparse local minimum $\hat{\boldsymbol{\beta}}$ is supported, with probability converging to one, on a non-random set $\mathcal{A}_n \in \mathscr{K}_n$. Parts (i) and (ii) of Lemma 1 together imply that $\mathbb{P}(\hat{\mathscr{K}}_n = \mathscr{K}_n) \to 1$ as $n \to \infty$.

Remark 3. Under the fixed-parameter asymptotic framework, $\mathscr{K}_n = \{\mathcal{A}_0\}$ is a fixed singleton and the above result reduces to the conventional definition of selection consistency.

It follows from Lemma 1(i) that for any $\mathcal{A} \in \limsup_{n\to\infty} \mathscr{K}_n$, there exists a subsequence of selected sets $\hat{\mathcal{A}}_{n_k} \in \hat{\mathscr{K}}_{n_k}$ satisfying $\mathbb{P}(\hat{\mathcal{A}}_{n_k} = \mathcal{A}) \to 1$. We shall characterise in Proposition 1 the componentwise convergence rates r_j of

$\hat{\boldsymbol{\beta}}$ supported on $\hat{\mathcal{A}}_{n_k}$, and show that $\hat{\mathcal{A}}_{n_k}$ successfully identifies all sufficiently strong signals. For brevity we write $\hat{\mathcal{A}}$ for $\hat{\mathcal{A}}_{n_k}$ in what follows.

Proposition 1. *Under the conditions of Lemma 1, the following holds for any consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ with support $\hat{\mathcal{A}} = \mathcal{A}$ almost surely, for some $\mathcal{A} \in \limsup_{n\to\infty} \mathscr{K}_n$.*

(i) $r_j \asymp |\beta_0^{(j)}|^{-1} \succeq n/\lambda$ *for* $j \in \mathcal{A}^c$ *and, for some* $r_0 \succeq n/\lambda$, $r_j = r_0$ *for* $j \in \mathcal{A}$.

(ii) $B_0 \preceq \lambda/n$.

(iii) $\mathbb{P}\big(\hat{\mathcal{A}} \supseteq \{j : |\beta_0^{(j)}| \succ \lambda/n\}\big) \to 1$.

Remark 4. It follows from Proposition 1(ii) that if $B_0 \succ \lambda/n$ under $(A2)$, then $\hat{\boldsymbol{\beta}}$ is inconsistent for $\boldsymbol{\beta}_0$ in general.

Remark 5. Proposition 1(iii) suggests that the selected set $\hat{\mathcal{A}}$ captures, with large probability, all strong signals bigger than the order λ/n, but may be conservative with the inclusion of weak signals.

Proposition 1 provides only a conservative lower bound n/λ on the convergence rate r_0 of $\hat{\boldsymbol{\beta}}^{\mathcal{A}}$ under mild conditions on (B_0, B_U). We state in the next theorem a necessary and sufficient condition on (B_0, B_U) for the existence of a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ whose nonzero components converge at a

rate faster than n/λ and succeed in selecting, with large probability, only strong signals of order at least λ/n. The theorem also provides an explicit expression for the weak limit of this $\hat{\boldsymbol{\beta}}$, and compares it with other consistent sparse local minima.

Theorem 1. *Suppose that $|\mathcal{A}_0| \asymp 1$ and the conditions of Proposition 1 hold. Then, a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ exists with a componentwise convergence rate $r_0 \succ n/\lambda$ if and only if $\psi \prec \lambda/n$. In this case, we have $r_0 \asymp \sqrt{n} \wedge \psi^{-1}$ and, with probability converging to one, supp$(\hat{\boldsymbol{\beta}}) = \mathcal{A}_0$ and*

$$\begin{cases} \hat{\boldsymbol{\beta}}^{\mathcal{A}_0^c} = \mathbf{0}, \\ \hat{\boldsymbol{\beta}}^{\mathcal{A}_0} = \boldsymbol{\beta}_0^{\mathcal{A}_0} + \Big\{ \hat{C}_{\mathcal{A}_0\mathcal{A}_0} - \dfrac{\lambda}{2n(\alpha-1)\kappa} \Lambda_{\mathcal{A}_0\mathcal{A}_0} \Big\}^{-1} \\ \qquad \times \Big(n^{-1/2} \boldsymbol{W}^{\mathcal{A}_0} + \hat{C}_{\mathcal{A}_0\mathcal{A}_0^c} \boldsymbol{\beta}_0^{\mathcal{A}_0^c} - \phi^{\mathcal{A}_0}/r_0 \Big), \end{cases} \tag{2.4}$$

where

$$\Lambda = \mathrm{diag}\big(\mathbf{1}\{|\beta_0^{(j)}| < \alpha\kappa\} : j = 1, \ldots, p\big),$$
$$\phi = \frac{r_0 \lambda\, \alpha}{2n(\alpha-1)} \Big[\Big(1 - \frac{|\beta_0^{(j)}|}{\alpha\kappa} \Big)_+ \mathrm{sgn}(\beta_0^{(j)}) : j = 1, \ldots, p \Big].$$

If, in addition, $B_U \succ \lambda/n$, then the above local minima are the only consistent sparse local minima selecting $\mathcal{A}_0$ with probability converging to one, while any other consistent sparse local minima converge at a slower rate n/λ and are less sparse with supports $\supsetneq \mathcal{A}_0$ asymptotically.

If, in addition, $\lim_{n\to\infty} B_U/(\alpha\kappa) > 1$*, then the above minima yield an objective function* (2.1) *strictly smaller than that yielded by any other consistent sparse local minima.*

Remark 6. Theorem 1 suggests that if the group of strong signals is sufficiently distinct from the group of weak signals such that $\psi \prec \lambda/n$, then the set of consistent sparse local minima can be classified into two subsets, having a fast convergence rate $\sqrt{n} \wedge \psi^{-1}$ and a slow convergence rate n/λ, respectively, with the former being non-empty. With large probability, the fast converging local minima select the strong signal set $\mathcal{A}_0$, while the slow converging local minima, if any, select signal sets containing $\mathcal{A}_0 \setminus \{j : |\beta_0^{(j)}| \asymp \lambda/n\}$.

Remark 7. If the strong signals are further away from the weak signals such that $B_U \succ \lambda/n$, then the local minima which select $\mathcal{A}_0$ are sparsest and have the fastest convergence rate among all consistent sparse local minima. If, in addition, $\lim_{n\to\infty} B_U/(\alpha\kappa) > 1$, then they are also the unique local minima which yield the smallest value of the objective function (2.1). Note that the latter condition follows immediately from $B_U \succ \lambda/n$ if $\alpha\kappa \asymp \lambda/n$, which is satisfied by, for example, SCAD and MCP.

Remark 8. Similar to our Proposition 1, Loh and Wainwright (2015) establish a conservative slow rate n/λ for the convergence of every local minimum as a

statistical guarantee for the latter. Our results in Theorem 1 clarify the conditions for the existence of local minima converging at a faster rate $\sqrt{n} \wedge \psi^{-1}$.

Remark 9. If $\hat{\boldsymbol{\beta}}$ selects $\mathcal{A}_0$ with probability converging to one, then the magnitudes of its nonzero components provide a correct ordering of all sufficiently strong signals along the local asymptotic spectrum, with probability converging to one.

Remark 10. If $\alpha\kappa \succ \lambda/n$, which excludes the common SCAD and MCP methods, then $\psi \prec \lambda/n$ implies $B_U \succ \lambda/n$, and $\hat{\boldsymbol{\beta}}^{\mathcal{A}_0}$ in (2.4) reduces to $\boldsymbol{\beta}_0^{\mathcal{A}_0} + \hat{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\{(1/\sqrt{n})\boldsymbol{W}^{\mathcal{A}_0} + \hat{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} - r_0^{-1}\phi^{\mathcal{A}_0}\}$.

Remark 11. Theorem 1 provides a more holistic picture of the selection and convergence properties of consistent sparse local minima $\hat{\boldsymbol{\beta}}$ from a local asymptotic perspective, which covers as a special case the weak oracle property introduced by Fan and Lv (2011), under weaker conditions on B_0, B_U and the covariate design than those assumed by the said paper. In particular, we see from (2.4) that even when $\hat{\boldsymbol{\beta}}$ selects $\mathcal{A}_0$ correctly and converges at a fast rate, it is not necessarily asymptotically equivalent to an ordinary least squares (OLS) estimator derived from $\mathcal{A}_0$, casting doubt on our conventional interpretation of oracle properties of nonconvex penalised estimators.

Remark 12. In the special case $\alpha\kappa = \infty$, which holds for LASSO, the condition

$\psi \prec \lambda/n$ fails. A contraposition of Theorem 1 shows that the active $\hat{\beta}_j$'s have a slow convergence rate n/λ. A faster convergence rate may result under a fixed p if we set $\lambda \asymp \sqrt{n}$, as is typically adopted by LASSO. However, the latter condition fails to guarantee selection consistency in general, unless we impose further constraints on $\mathcal{C}$.

We may deduce from (2.4) a series of phase changes of the asymptotic behaviour of $U^{\mathcal{A}_0} = r_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{A}_0}$, when signal patterns undergo the following transition phases over the local asymptotic spectrum.

(a) If $\psi = (\lambda/n)\big(1 - B_U/(\alpha\kappa)\big)_+ \succ B_0$, setting $r_0 = \sqrt{n}\big\{1 \wedge (\sqrt{n}/\lambda)\big(1 - B_U/(\alpha\kappa)\big)_+^{-1}\big\}$ gives

$$U^{\mathcal{A}_0} = \Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1} \times \Big[\big\{1 \wedge (\sqrt{n}/\lambda)\big(1 - B_U/(\alpha\kappa)\big)_+^{-1}\big\}\boldsymbol{W}^{\mathcal{A}_0} - \phi^{\mathcal{A}_0}\Big] + o_p(1),$$

which has a non-random leading term

$$- \Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1}\phi^{\mathcal{A}_0} \tag{2.5}$$

if and only if $\big(1 - B_U/(\alpha\kappa)\big)_+ \succ \sqrt{n}/\lambda$.

(b) If $\psi = B_0 \succ (\lambda/n)\,(1 - B_U/(\alpha\kappa))_+$, setting $r_0 = \sqrt{n} \wedge B_0^{-1}$ gives

$$U^{\mathcal{A}_0} = \Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1} \times \Big(1 \wedge \frac{1}{\sqrt{n}B_0}\Big)\big\{\boldsymbol{W}^{\mathcal{A}_0} + \sqrt{n}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\big\} + o_p(1),$$

which has a non-random leading term

$$\Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1} B_0^{-1}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} \tag{2.6}$$

if and only if $B_0 \succ 1/\sqrt{n}$, or reduces to

$$\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\Big(1 \wedge \frac{1}{\sqrt{n}B_0}\Big)\big\{\boldsymbol{W}^{\mathcal{A}_0} + \sqrt{n}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\big\} + o_p(1)$$

if $B_U \geq \alpha\kappa$.

(c) If $\psi \prec 1/\sqrt{n}$, setting $r_0 = \sqrt{n}$ gives

$$U^{\mathcal{A}_0} = \Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1),$$

which has a random leading term.

Given its prominence in the literature, the conventional oracle property, generalised under our local asymptotic framework, is given below as an immediate corollary to Theorem 1, which is obtained by intersecting the above phases (b) and (c) into a final oracle phase.

Corollary 1. *(Generalised oracle property) Assume the conditions of Proposition 1,* $|\mathcal{A}_0| \asymp 1$, $\psi \prec 1/\sqrt{n}$, *and that either* $\lambda/n \prec \alpha\kappa$ *or* $B_U \geq \alpha\kappa$. *Then a generalised oracle estimator* $\hat{\boldsymbol{\beta}}_{go}$ *exists with*

$$\mathbb{P}\big(\hat{\boldsymbol{\beta}}_{go}^{\mathcal{A}_0^c} = 0\big) \to 1 \quad \textit{and} \quad \sqrt{n}(\hat{\boldsymbol{\beta}}_{go} - \boldsymbol{\beta}_0)^{\mathcal{A}_0} = \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1).$$

If, in addition, $B_U \succ \lambda/n$, *then* $\hat{\boldsymbol{\beta}}_{go}$ *is the only consistent sparse local minimum selecting* $\mathcal{A}_0$ *with probability converging to one. Any other consistent sparse local minima necessarily converge at a slower rate* n/λ *and are less sparse with supports* $\hat{\mathcal{A}} \supsetneq \mathcal{A}_0$ *asymptotically.*

For generalised versions of Theorem 1 and Corollary 1 which cover a heavy tail condition, see Theorem A.1 and Corollary A.1, respectively, in Appendix 1.

Remark 13. With probability converging to one, the generalised oracle estimator $\hat{\boldsymbol{\beta}}_{go}$ estimates the coefficients of weak signals (in $\mathcal{A}_0^c$) to be zero and those of strong signals (in $\mathcal{A}_0$) by ordinary least squares. If $(\boldsymbol{X}, Y)$ satisfies tail conditions $(\mathcal{T}_1)$ or $(\mathcal{T}_2)$, then $n^{1/2}(\hat{\boldsymbol{\beta}}_{go} - \boldsymbol{\beta}_0)^{\mathcal{A}_0}$ is asymptotically zero-mean Gaussian.

Remark 14. Loh and Wainwright (2017) show, under the tail condition $(\mathcal{T}_1)$, a sparse Riesz condition on $\mathcal{C}$ and a betamin condition $B_0 = 0$, that $\hat{\boldsymbol{\beta}}_{go}$ is the unique local, hence global, minimum. Assuming a weaker bound $|\mathcal{A}_0| \preceq n/\lambda$ than ours, they establish a conservative convergence rate of order n/λ for $\hat{\boldsymbol{\beta}}_{go}$, which is slower than the rate $\sqrt{n}$ shown in Corollary 1.

For completeness we conclude this section with a theorem about the properties of $\hat{\boldsymbol{\beta}}$ in the absence of strong signals, that is when $\mathcal{A}_0 = \emptyset$.

Theorem 2. *Suppose that $\mathcal{A}_0 = \emptyset$ and the conditions of Proposition 1 hold. For any consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ with $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ for some $\mathcal{A} \in \limsup_{n\to\infty} \mathscr{K}_n$, its componentwise estimation error satisfies*

$$\hat{\beta}_j - \beta_0^{(j)} \asymp \begin{cases} \lambda/n \succ |\beta_0^{(j)}|, & j \in \mathcal{A}, \\ |\beta_0^{(j)}| \prec \lambda/n, & j \in \mathcal{A}^c. \end{cases}$$

If, in addition, $B_0 \prec \lambda/n$, then a zero local minimum $\hat{\boldsymbol{\beta}} = \mathbf{0}$ exists and uniquely minimises the objective function (2.1) over all consistent sparse local minima.

2.4 Schematic illustration

To further elucidate our theory established in Section 2.3, Figure 1 provides a graphical illustration of the asymptotic properties of a generic consistent local minimum $\hat{\boldsymbol{\beta}}$ under five different signal patterns, which exemplify the main phase changes in the asymptotics of $\hat{\boldsymbol{\beta}}$. For the sake of illustration it suffices to consider the case $\lambda/n \prec \alpha\kappa \prec 1$ under the conditions of Lemma 1.

With the omission of some trivial variants, signal patterns 1–5 shown in Figure 1 sketch out all general scenarios where a consistent sparse $\hat{\boldsymbol{\beta}}$ exists. Patterns 1 and 2 epitomise a confused phase covered by Proposition 1, which

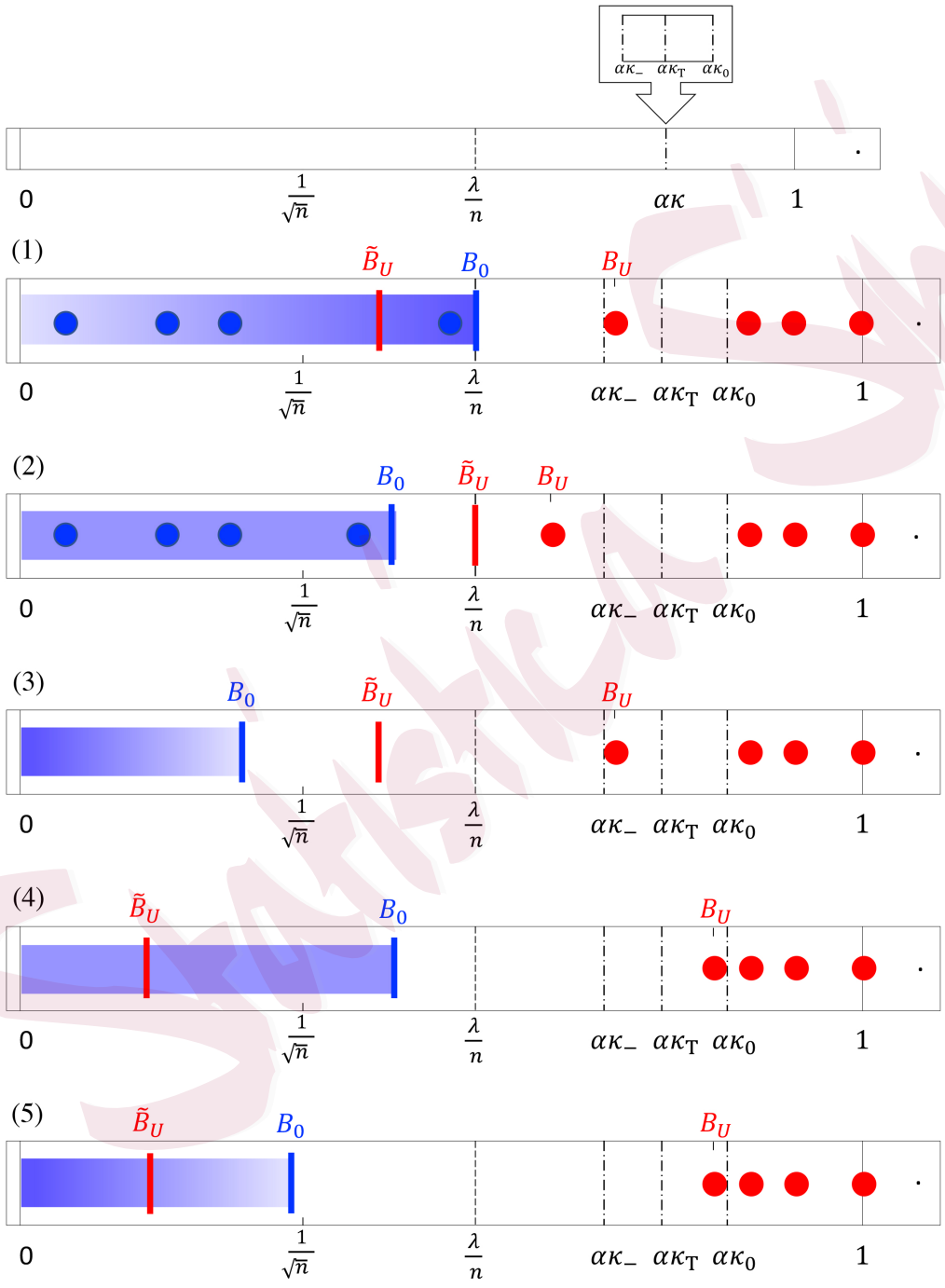

Figure 1: Phase changes in selection properties of consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ across five signal patterns, under $\alpha\kappa \succ \lambda/n$ and conditions of Lemma 1.

assumes the weakest conditions on the configurations of the $\beta_0^{(j)}$'s. Theorem 1 is typified by patterns 3 and 4, while pattern 5 reveals the oracle phase discussed in Corollary 1.

For each pattern, the local asymptotic spectrum is represented by a ruler marked with a scale of asymptotic orders: $0 \prec 1/\sqrt{n} \prec \lambda/n \prec \alpha\kappa \prec 1$, with ∞ indicated by a tiny dot. We write for convenience $\tilde{B}_U = (\lambda/n)\{1-B_U/(\alpha\kappa)\}_+$, so that $\psi = B_0 \vee \tilde{B}_U$. Note that B_U and $\tilde{B}_U$ are related in such a way that they move in opposite directions on the ruler. As B_U decreases but remains at the order $\alpha\kappa$, $\tilde{B}_U$ has an order increasing from 0 to λ/n. In order to differentiate transition phases arising from a change of order of $\tilde{B}_U$ between 0 and λ/n, the accompanying locations of B_U are represented by an interval ranging from $\alpha\kappa_-$ through $\alpha\kappa_T$ to $\alpha\kappa_0$. The interval is integrated into the spectrum ruler to provide a magnified view of the order $\alpha\kappa$. For a generic coefficient β of magnitude $\alpha\kappa(1-\Delta) \asymp \alpha\kappa$, the upper end $\alpha\kappa_0$ and the lower end $\alpha\kappa_-$ signify the cases $1 \asymp \Delta \leq 0$ and $1 \asymp \Delta > 0$, respectively. The sub-intervals $(\alpha\kappa_T, \alpha\kappa_0)$ and $(\alpha\kappa_-, \alpha\kappa_T)$ refer to the intermediate cases $0 < \Delta \prec \sqrt{n}/\lambda$ and $\sqrt{n}/\lambda \prec \Delta \prec 1$, respectively. Thus, when positioned at $\alpha\kappa_T$, β satisfies $0 < \Delta \asymp \sqrt{n}/\lambda$, that is $(\lambda/n)\{1-|\beta|/(\alpha\kappa)\}_+ \asymp 1/\sqrt{n}$.

On each spectrum shown in Figure 1, the orders of the coefficients in $\mathcal{A}_0$ and $\mathcal{A}_0^c$ are marked in red and blue on either side of λ/n, respectively, together

with the corresponding orders of B_0, B_U and $\tilde{B}_U$. Note that $B_0 \preceq \lambda/n$ by Proposition 1(ii) and $B_U \succeq \lambda/n$ by definition. We allow for $|\mathcal{A}_0^c| \succ 1$, that is a growing number of weak signals, which are merged into a blue horizontal bar with darker shades indicating higher concentrations. The sparsity condition $|\mathcal{A}_0| \preceq 1$ is exemplified by a number of isolated red dots to the right of λ/n. The orders of signals selected by $\hat{\boldsymbol{\beta}}$ are shown by dots in either colour. Since $|\hat{\mathcal{A}}| \preceq 1$, all coloured dots are isolated.

When the order of B_U lies between λ/n and $\alpha\kappa_-$, the order of $\tilde{B}_U$ remains unchanged at λ/n (pattern 2). When B_U moves from $\alpha\kappa_-$ to $\alpha\kappa_T$, the order of $\tilde{B}_U$ moves from λ/n to $1/\sqrt{n}$ in the opposite direction (patterns 1 and 3). When B_U moves from $\alpha\kappa_T$ to $\alpha\kappa_0$, the order of $\tilde{B}_U$ moves from $1/\sqrt{n}$ to 0 (patterns 4 and 5). When B_U increases beyond $\alpha\kappa_0$, $\tilde{B}_U$ stays unchanged at 0. Combined with the position of B_0, the order of the critical measure $\psi = B_0 \vee \tilde{B}_U$ then emerges.

Patterns 1 and 2 correspond to the case $\psi \asymp \lambda/n$, where B_U and B_0 are too close for the strong signals to be easily differentiated from the weak ones. In this case, all strong signals, indicated by the red dots, as well as a few weak signals, indicated by the blue dots, are detected by $\hat{\boldsymbol{\beta}}$, which exemplifies Proposition 1(iii).

Patterns 3 and 4 correspond to the case $1/\sqrt{n} \prec \psi \prec \lambda/n$, where B_U

and B_0 become more widely apart than those under patterns 1 or 2. It follows by Theorem 1 that a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ exists with a unique fast convergence rate and selects only the strong signals (red dots). In particular, we may deduce from (2.5) that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{A}_0}$ has a non-random leading term $-\tilde{B}_U \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\phi^{\mathcal{A}_0} \asymp \tilde{B}_U$ under pattern 3, and from (2.6) that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{A}_0}$ has a non-random leading term $\mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} \asymp B_0$ under pattern 4.

Finally, pattern 5 spotlights a desirable configuration with sufficiently small B_0 and $\tilde{B}_U$ such that $\psi \prec 1/\sqrt{n}$. In this case, Corollary 1 implies the existence of a generalised oracle estimator $\hat{\boldsymbol{\beta}}_{go}$ supported on $\mathcal{A}_0$, with $\sqrt{n}(\hat{\boldsymbol{\beta}}_{go} - \boldsymbol{\beta}_0)^{\mathcal{A}_0} = \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1)$.

3. Post-selection OLS estimator

Rewriting the subgradient conditions (2.2) as

$$\begin{cases} (2n/\lambda)\big|n^{-1/2}W_j - \hat{C}_{\{j\}\hat{\mathcal{A}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\hat{\mathcal{A}}} + \hat{C}_{\{j\}\hat{\mathcal{A}}^c}\boldsymbol{\beta}_0^{\hat{\mathcal{A}}^c}\big| \le 1, \ j \in \hat{\mathcal{A}}^c & (3.7) \\ \hat{C}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big)^{\hat{\mathcal{A}}} = n^{-1/2}\boldsymbol{W}^{\hat{\mathcal{A}}} + \mathscr{B}_1 + \mathscr{B}_2, & (3.8) \end{cases}$$

where $\mathscr{B}_1 = -(2n)^{-1}\lambda\big[\mathrm{sgn}(\hat{\beta}_j)q'(|\hat{\beta}_j|/\kappa) : j \in \hat{\mathcal{A}}\big]$ and $\mathscr{B}_2 = \hat{C}_{\hat{\mathcal{A}}\hat{\mathcal{A}}^c}\boldsymbol{\beta}_0^{\hat{\mathcal{A}}^c}$, we see that the generalised oracle property (Corollary 1) of $\hat{\boldsymbol{\beta}}_{go}$, which follows essentially by applying the central limit theorems to $\boldsymbol{W}^{\hat{\mathcal{A}}}$, is impaired in the transition phases by additional biases stemming from $\mathscr{B}_1$ and $\mathscr{B}_2$, which are

made non-negligible by a small B_U and a large B_0, respectively. We propose in this section a simple strategy to remove the bias due to $\mathscr{B}_1$.

Define, for any $\mathcal{B} \subset \{1, \ldots, p\}$ with $|\mathcal{B}| = O(1)$, $\hat{\boldsymbol{b}}(\mathcal{B}) = n^{-1}\hat{C}_{\mathcal{B}\mathcal{B}}^{-1}\sum_{i=1}^{n} Y_i \boldsymbol{X}_i^{\mathcal{B}}$, that is the sparse OLS estimator restricted to the submodel containing only variables $\boldsymbol{X}^{\mathcal{B}}$. The bias term $\mathscr{B}_1$ can be offset by the post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$, for any selected set $\hat{\mathcal{A}}$ derived from a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ satisfying (3.7) and (3.8). This follows immediately by noting that $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ satisfies

$$\hat{C}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\beta}_0\}^{\hat{\mathcal{A}}} = n^{-1/2}\boldsymbol{W}^{\hat{\mathcal{A}}} + \mathscr{B}_2, \tag{3.9}$$

which has $\mathscr{B}_1$ removed from (3.8). We now detail the convergence properties of $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ as an estimator of $\boldsymbol{\beta}_0$, under mild conditions on design and signal strength.

Theorem 3. *Suppose that $|\mathcal{A}_0| \preceq 1$ and the conditions of Proposition 1 hold. If $\psi \prec \lambda/n$, then there exists a consistent sparse post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ which is supported on $\mathcal{A}_0$ with probability converging to one and converges at a rate $B_0^{-1} \wedge \sqrt{n}$.*

If, in addition, $B_U \succ \lambda/n$, then any consistent sparse post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ is supported on $\hat{\mathcal{A}} \supseteq \mathcal{A}_0$ with probability converging to one and converges at a rate within the range $\left[B_0^{-1} \wedge \sqrt{n},\ \sqrt{n}\right]$.

Remark 15. As has been discussed in Remark 6, under the condition $\psi \prec \lambda/n \prec B_U$, a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ converges either at a fast rate

$\sqrt{n} \wedge \psi^{-1}$ with selected set $\mathcal{A}_0$ or at a slow rate n/λ with selected set $\supsetneq \mathcal{A}_0$. Under the same signal pattern, any post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ has a convergence rate $\succeq B_0^{-1} \wedge \sqrt{n}$, which is at least as fast as that of any fast-converging $\hat{\boldsymbol{\beta}}$ and strictly faster than the rate of any slowly-converging $\hat{\boldsymbol{\beta}}$. If, in addition, $B_0 \vee 1/\sqrt{n} \prec (\lambda/n)\{1 - B_U/(\alpha\kappa)\}_+$, which is satisfied by, for example, pattern 3 in Figure 1, then any post-selection OLS estimator converges strictly faster than any local minimum $\hat{\boldsymbol{\beta}}$, provided they are consistent and sparse.

Remark 16. In general, assuming the conditions of Proposition 1, $|\mathcal{A}_0| \preceq 1$, $B_U \succ \lambda/n$ and $B_0 \prec 1/\sqrt{n}$, we have, for any $\hat{\mathcal{A}} \in \hat{\mathscr{K}}$ and $\mathcal{A} \in \limsup_{n\to\infty} \mathscr{K}_n$ with $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$, that $\mathcal{A} \supseteq \mathcal{A}_0$ and that the corresponding post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ converges at the fastest possible rate $\sqrt{n}$ and satisfies $\sqrt{n}\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\beta}_0\}^{\mathcal{A}} = \mathcal{C}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{W}^{\mathcal{A}} + o_p(1)$.

Corollary 2. *(Generalised oracle property) Assume the conditions of Proposition 1,* $|\mathcal{A}_0| \preceq 1$, $\{1 - B_U/(\alpha\kappa)\}_+ \prec 1$ *and* $B_0 \prec 1/\sqrt{n}$. *Then, a sequence of selected sets* $\hat{\mathcal{A}} \in \hat{\mathscr{K}}$ *exists such that*

$$\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}_0) \to 1 \quad \textit{and} \quad \sqrt{n}\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\beta}_0\}^{\mathcal{A}_0} = \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1). \tag{3.10}$$

Remark 17. As has been shown in Corollary 1, existence of a generalised oracle $\hat{\boldsymbol{\beta}}_{go}$ requires more restrictive conditions on B_U, namely $\{1 - B_U/(\alpha\kappa)\}_+ \prec \sqrt{n}/\lambda$ if $\alpha\kappa \succ \lambda/n$ or $B_U \geq \alpha\kappa$ if $\alpha\kappa \asymp \lambda/n$, compared to those required

by Corollary 2. If, in addition, $B_U \succ \lambda/n$, then all the post-selection OLS estimators converge at the fastest rate $\sqrt{n}$, while the corresponding local minima $\hat{\boldsymbol{\beta}}$ except $\hat{\boldsymbol{\beta}}_{go}$ all converge at the slowest rate n/λ.

In summary, by removing the bias term $\mathscr{B}_1$, the post-selection OLS estimators $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ acquire convergence properties more desirable than the local minima $\hat{\boldsymbol{\beta}}$ and, in the case of multiple solutions to the nonconvex optimisation program (2.2), ratewise more robust against the choice of strong signal sets $\hat{\mathcal{A}}$.

Theorem 3 and Corollary 2 are extended in Appendix 1 to Theorem A.2 and Corollary A.2, respectively, to accommodate a heavy tail condition.

4. Extension and simulation study

By including an additional heavy-tailed setting, we generalise in Appendix 1 the theoretical results contained in Sections 2.3 and 3, with technical proofs given in Appendix 2. From a predictive perspective, it may be of interest to draw inference about the effects of strong signals after adjusting for the omission of weak signals under a weakly sparse model. Define an "oracle" target to be $\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\theta})^2 : \boldsymbol{\theta}^{\mathcal{A}_0^c} = \boldsymbol{0}\}$, which can be interpreted as the effects of strong signals in $\mathcal{A}_0$ adjusted for the omission of weak signals in $\mathcal{A}_0^c$. In Section A.2 of Appendix 1, we revisit the asymptotic properties of both the nonconvex penalised estimator and its post-selection counterpart when applied

to estimate adjusted effects of strong signals for the purpose of prediction.

For a numerical illustration of our theoretical findings, a simulation study has been conducted to compare the empirical performance of SCAD with post-SCAD OLS under both Gaussian and heavy-tailed settings. A detailed description of the simulation study, including methodology, performance metrics, implementation and results, can be found in Section A.3.3 of Appendix 1.

5. Conclusion

Under the conventional, fixed-parameter, asymptotic framework where each coefficient $\beta_0^{(j)}$ is either zero or a nonzero constant, it follows immediately by Corollary 1 that a nonconvex penalised estimator $\hat{\boldsymbol{\beta}}_{go}$ exists and enjoys the generalised oracle property, which is in this case equivalent to the conventional oracle property discussed in the literature. Within this oracle phase, post-selection least squares estimation does not make further improvement by reducing the asymptotic error of $\hat{\boldsymbol{\beta}}_{go}$. In this paper, a local asymptotic framework is adopted to allow for the existence of many mild signals with $0 \neq \beta_0^{(j)} \prec \lambda/n$. This broadens the scope of our asymptotic investigation and complements the oracle phase with several transition phases in the asymptotics of $\hat{\boldsymbol{\beta}}$ to cover a more complete, yet practically relevant, range of signal strength configurations. Assuming a bounded number of strong signals, $|\mathcal{A}_0| \preceq 1$, and mild conditions on

covariate correlations, we study all scenarios where a consistent sparse $\hat{\boldsymbol{\beta}}$ exists and derive its selection and convergence properties accordingly. We show that phase changes in the asymptotics of $\hat{\boldsymbol{\beta}}$ are determined critically by (B_0, B_U), and provide a necessary and sufficient condition, namely $\psi \equiv B_0 \vee \{(\lambda/n)(1 - B_U/(\alpha\kappa))\}_+ \prec \lambda/n$, for the existence of a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ which selects $\mathcal{A}_0$ consistently and has a fast componentwise convergence rate $\sqrt{n} \wedge \psi^{-1}$. It follows that when B_U is not large enough or B_0 is not small enough, exemplified by patterns 3 and 4 in Figure 1, the generalised oracle property fails to hold for any consistent sparse local minimum $\hat{\boldsymbol{\beta}}$.

We have also clarified the effects of post-selection least squares estimation on nonconvex penalised regression. In general, the post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ is freed from a penalty-driven bias and is ratewise more robust than $\hat{\boldsymbol{\beta}}$ against selection of the set $\hat{\mathcal{A}}$ of strong signals. Specifically, any $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ has a convergence rate at least as fast as that of $\hat{\boldsymbol{\beta}}$, under the same signal pattern. Indeed, under signal pattern 3 in Figure 1, $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ enjoys the generalised oracle property, which does not hold for $\hat{\boldsymbol{\beta}}$. If $B_U \succ \lambda/n$, our Corollary 2 shows that the fastest rate $\sqrt{n}$ is achieved by all choices of $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$, while the corresponding local minima $\hat{\boldsymbol{\beta}}$ except $\hat{\boldsymbol{\beta}}_{go}$ converge only at the slowest rate n/λ. This provides $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ with a desirable theoretical guarantee in the presence of multiple local minima.

We argue in Appendix 1 that from a predictive perspective, adjusting $\boldsymbol{\beta}_0$ for the omission of weak signals makes for a practically more relevant target $\boldsymbol{\theta}_0$ under a weakly sparse model. With this change of target, under tail conditions $(\mathcal{T}_1)$ or $(\mathcal{T}_2)$, we may weaken the condition on B_0 from $B_0 \prec 1/\sqrt{n}$ to $B_0 \prec \lambda/n$ for $\hat{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ to satisfy the generalised oracle property.

We have conducted elaborate simulation studies to compare SCAD with post-SCAD OLS by a variety of numerical and graphical measures, and reported the results in Appendix 1. Admittedly, a foolproof method for deriving every consistent sparse local minimum remains beyond our reach, not least because of the ambiguity inherent in any practical interpretation of what we mean by a sparse solution. Nevertheless, a simple, practically viable, approach is to run a standard computational algorithm (e.g. the R package `ncvreg`) multiple times based on distinct choices of initial guesses to acquire multiple solutions. We may then apply OLS to the active set selected by the solution which incurs the smallest empirical loss. The numerical findings corroborate our theory in general, suggesting that post-SCAD OLS successfully reduces the bias of SCAD and displays a more robust performance. The improvement made by post-SCAD OLS is especially significant under a heavy-tailed setting, which calls for a heavier SCAD penalty weight for consistent selection.

Going forward, the local asymptotic results established in this paper set an

important stage for the development of theoretically tractable bootstrap post-selection inference procedures for high-dimensional nonconvex penalised regression. We shall pursue this in a future work.

References

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. Journal of the American Statistical Association 96(455), 939–967.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. Bernoulli 19(2), 521 – 547.

Bonaccolto, G. (2021). Quantile–based portfolios: post–model–selection estimation with alternative specifications. Computational Management Science 18(3), 355–383.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96(456), 1348–1360.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. IEEE Transactions on Information Theory 57(8), 5467–5484.

Fousekis, P. and V. Grigoriadis (2022). Conditional tail price risk spillovers in coffee markets across quality, physical space, and time: Empirical analysis with penalized quantile regressions. Economic Modelling 106, 105691.

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli 10(6), 971–988.

Horowitz, J. L. and J. Huang (2013). Penalized estimation of high-dimensional models under a generalized sparsity condition. Statistica Sinica, 725–748.

Javanmard, A. and J. D. Lee (2020, 05). A flexible framework for hypothesis testing in high dimensions. Journal of the Royal Statistical Society Series B: Statistical Methodology 82(3), 685–718.

Javanmard, A. and A. Montanari (2013). Model selection for high-dimensional regression under the generalized irrepresentability condition. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), Advances in Neural Information Processing Systems, Volume 26. Curran Associates, Inc.

Kim, Y., J.-J. Jeon, and S. Han (2016). A necessary condition for the strong oracle property. Scandinavian Journal of Statistics 43(2), 610–624.

Kim, Y. and S. Kwon (2012). Global optimality of nonconvex penalized estimators. Biometrika 99(2), 315–325.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. Journal of Econometrics 142(1), 201–211.

Liu, H., X. Xu, and J. J. Li (2017, 06). A bootstrap lasso + partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. Statistica Sinica.

Lockhart, R., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2014). A significance test for the lasso. Annals of statistics 42(2), 413.

Loh, P.-L. and M. J. Wainwright (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. The Journal of Machine Learning Research 16(1), 559–616.

Loh, P.-L. and M. J. Wainwright (2017). Support recovery without incoherence: A case for nonconvex regularization. The Annals of Statistics 45(6), 2455–2482.

Pötscher, B. M. and H. Leeb (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. Journal of Multivariate Analysis 100(9), 2065–2082.

Qu, A. and P. Shi (2016). Weak signal identification and inference in penalized model selection. Annals of Statistics.

Ratnasingam, S. and W. Ning (2021). Sequential change point detection for high-dimensional data using nonconvex penalized quantile regression. Biometrical Journal 63(3), 575–598.

Shao, J. and X. Deng (2012). Estimation in high-dimensional linear models with deterministic design matrices. The Annals of Statistics 40(2), 812–831.

Uematsu, Y. and S. Tanaka (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. The Econometrics Journal 22(1), 34–56.

Xiao, H. and Y. Sun (2019). On tuning parameter selection in model selection and model averaging: A monte carlo study. Journal of Risk and Financial Management 12(3).

Xiao, H. and Y. Sun (2020). Forecasting the returns of cryptocurrency: A model averaging approach. Journal of Risk and Financial Management 13(11), 278.

Yu, G., L. Yin, S. Lu, and Y. Liu (2019). Confidence intervals for sparse penalized regression with random designs. Journal of the American Statistical Association.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics 38(2), 894–942.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(1), 217–242.

Zhang, C.-H. and T. Zhang (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. Statistical Science 27(4), 576–593.

Zhao, T., H. Liu, and T. Zhang (2018). Pathwise coordinate optimization for sparse learning: Algorithm and theory. The Annals of Statistics 46(1), 180–218.

Institute of Applied Mathematics, Shenzhen Polytechnic University

E-mail: (xuxiaoya@connect.hku.hk)

Department of Statistics and Actuarial Science, The University of Hong Kong

E-mail: (smslee@hku.hk)