The Annals of Statistics
2014, Vol. 42, No. 6, 2164–2201
DOI: 10.1214/14-AOS1238

OPTIMAL COMPUTATIONAL AND STATISTICAL RATES OF CONVERGENCE FOR SPARSE NONCONVEX LEARNING PROBLEMS

BY ZHAORAN WANG, HAN LIU[1] AND TONG ZHANG[2]

Princeton University, Princeton University and Rutgers University

We provide theoretical analysis of the statistical and computational properties of penalized M-estimators that can be formulated as the solution to a possibly nonconvex optimization problem. Many important estimators fall in this category, including least squares regression with nonconvex regularization, generalized linear models with nonconvex regularization and sparse elliptical random design regression. For these problems, it is intractable to calculate the global solution due to the nonconvex formulation. In this paper, we propose an approximate regularization path-following method for solving a variety of learning problems with nonconvex objective functions. Under a unified analytic framework, we simultaneously provide explicit statistical and computational rates of convergence for any local solution attained by the algorithm. Computationally, our algorithm attains a global geometric rate of convergence for calculating the full regularization path, which is optimal among all first-order algorithms. Unlike most existing methods that only attain geometric rates of convergence for one single regularization parameter, our algorithm calculates the full regularization path with the same iteration complexity. In particular, we provide a refined iteration complexity bound to sharply characterize the performance of each stage along the regularization path. Statistically, we provide sharp sample complexity analysis for all the approximate local solutions along the regularization path. In particular, our analysis improves upon existing results by providing a more refined sample complexity bound as well as an exact support recovery result for the final estimator. These results show that the final estimator attains an oracle statistical property due to the usage of nonconvex penalty.

1. Introduction. This paper considers the statistical and computational properties of a family of penalized M-estimators that can be formulated as

$$\widehat{\boldsymbol{\beta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \{\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta})\}, \tag{1.1}$$

Received February 2014; revised May 2014.

[1]Supported by NSF Grants III-1116730 and NSF III-1332109, NIH R01MH102339, NIH R01GM083084 and NIH R01HG06841 and FDA HHSF223201000072C.

[2]Supported by Grants NSF IIS-1016061, NSF DMS-10-07527 and NSF IIS-1250985.

MSC2010 subject classifications. Primary 62F30, 90C26; secondary 62J12, 90C52.

Key words and phrases. Nonconvex regularized M-estimation, path-following method, geometric computational rate, optimal statistical rate.

where $\mathcal{L}(\boldsymbol{\beta})$ is a loss function, while $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is a penalty function with regularization parameter λ. A familiar example is the Lasso estimator [Tibshirani (1996)], in which $\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2/(2n)$ and $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$. Here $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ is the design matrix, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the response vector, $\|\cdot\|_2$ is the Euclidean norm and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^d |\beta_j|$ is the ℓ_1 norm of $\boldsymbol{\beta}$. In general, we prefer the settings where both the loss function $\mathcal{L}(\boldsymbol{\beta})$ and the penalty term $\mathcal{P}_\lambda(\boldsymbol{\beta})$ in (1.1) are convex, since convexity makes both statistical and computational analysis convenient.

Significant progress has been made on understanding convex penalized M-estimators [Bickel, Ritov and Tsybakov (2009), Koltchinskii (2009), Negahban et al. (2012), Raskutti, Wainwright and Yu (2011), Rothman et al. (2008), van de Geer (2000, 2008), Wainwright (2009), Zhang (2009)]. Meanwhile, penalized M-estimators with nonconvex loss or penalty functions have recently attracted much interest because of their more attractive statistical properties. For example, unlike the ℓ_1 penalty, which induces significant estimation bias for parameters with large absolute values [Zhang and Huang (2008)], nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li (2001)] and minimax concave penalty (MCP) [Zhang (2010a)] can eliminate this estimation bias and attain more refined statistical rates of convergence. As another example of penalized M-estimators with nonconvex loss functions, we consider a semiparametric variant of the penalized least squares regression. Recall that a penalized least squares regression estimator can be formulated as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_\lambda &\in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\biggl\{\frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \mathcal{P}_\lambda(\boldsymbol{\beta})\biggr\}\\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\biggl\{\frac{1}{2}(1, -\boldsymbol{\beta}^T)\widehat{\mathbf{S}}(1, -\boldsymbol{\beta}^T)^T + \mathcal{P}_\lambda(\boldsymbol{\beta})\biggr\},\end{aligned}$$

where $\widehat{\mathbf{S}} = (\mathbf{y}, \mathbf{X})^T(\mathbf{y}, \mathbf{X})/n$ is the sample covariance matrix of a random vector $(Y, \mathbf{X}^T)^T \in \mathbb{R}^{d+1}$. When the design matrix $\mathbf{X}$ contains heavy-tail data, we may resort to elliptical random design regression, which is a semiparametric extension of Gaussian random design regression. In detail, we replace the sample covariance matrix $\widehat{\mathbf{S}}$ with a possibly indefinite covariance matrix estimator $\widehat{\mathbf{K}}$ (to be defined in Section 2.2), which is more robust within the elliptical family. Since $\widehat{\mathbf{K}}$ does not guarantee to be positive semidefinite, the loss function $\mathcal{L}(\boldsymbol{\beta}) = (1, -\boldsymbol{\beta}^T)\widehat{\mathbf{K}}(1, -\boldsymbol{\beta}^T)^T/2$ could be nonconvex.

Though the global solutions of these nonconvex M-estimators enjoy nice statistical properties, it is in general computationally intractable to obtain the global solutions. Instead, a more realistic approach is to directly leverage standard optimization procedures to obtain a local solution $\widehat{\boldsymbol{\beta}}_\lambda$ that satisfies the first-order Karush–Kuhn–Tucker (KKT) condition

$$\mathbf{0} \in \partial\bigl\{\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda) + \mathcal{P}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\bigr\}, \tag{1.2}$$

where $\partial(\cdot)$ denotes the subgradient operator.

In the context of least squares regression with nonconvex penalties, several numerical procedures have been proposed to find the local solutions, including local quadratic approximation (LQA) [Fan and Li (2001)], minorize–maximize (MM) algorithm [Hunter and Li (2005)], local linear approximation (LLA) [Zou and Li (2008)], concave convex procedure (CCCP) [Kim, Choi and Oh (2008)] and coordinate descent [Breheny and Huang (2011), Mazumder, Friedman and Hastie (2011)]. The theoretical properties of the local solutions obtained by these numerical procedures are in general unestablished. Only recently Zhang and Zhang (2012) showed that the gradient descent method initialized at a Lasso solution attains a unique local solution that has the same statistical properties as the global solution; Fan, Xue and Zou (2014) proved that the LLA algorithm initialized with a Lasso solution attains a local solution with oracle statistical properties. The same conclusion was also obtained by Zhang (2010b, 2013), where the LLA algorithm was referred to as multi-stage convex relaxation. In recent work, Wang, Kim and Li (2013) proposed a calibrated concave-convex procedure (CCCP) along with a high-dimensional BIC criterion that can achieve the oracle estimator. However, these works mainly focused on statistical recovery results, while the corresponding computational complexity results remain unclear. Also, they did not consider nonconvex loss functions. In addition, their analysis relies on the assumption that all the computation (e.g., solving an optimization problem) can be carried out exactly, which is unrealistic in practice, since practical computational procedures can only attain finite numerical precision in finite time. Moreover, our method only requires the weakest possible minimum signal strength to attain the oracle estimator [Zhang and Zhang (2012)], while the procedures in Fan, Xue and Zou (2014), Wang, Kim and Li (2013) rely on a stronger signal strength which is suboptimal. See Section 6 for a more detailed discussion.

In this paper, we propose an approximate regularization path-following method for solving a general family of penalized M-estimators with possibly nonconvex loss or penalty functions. Our algorithm leverages the fast local convergence in the proximity of sparse solutions, which is also observed by Agarwal, Negahban and Wainwright (2012), Nesterov (2013), Wright, Nowak and Figueiredo (2009), Xiao and Zhang (2013). More specifically, we consider a decreasing sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$, where λ_0 corresponds to an all-zero solution, and $\lambda_N = \lambda_{\mathrm{tgt}}$ is the target regularization parameter that ensures the obtained estimator to achieve the optimal statistical rate of convergence. For each λ_t, we construct a sequence of local quadratic approximations of the loss function $\mathcal{L}(\boldsymbol{\beta})$, and utilize a variant of Nesterov's proximal-gradient method [Nesterov (2013)], which iterates over the updating step

$$\begin{aligned}\boldsymbol{\beta}_t^{k+1} \leftarrow \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\Big\{&\mathcal{L}(\boldsymbol{\beta}_t^k) + \nabla\mathcal{L}(\boldsymbol{\beta}_t^k)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_t^k)\\ &+ \frac{L_t^k}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_t^k\|_2^2 + \mathcal{P}_{\lambda_t}(\boldsymbol{\beta})\Big\},\end{aligned} \tag{1.3}$$

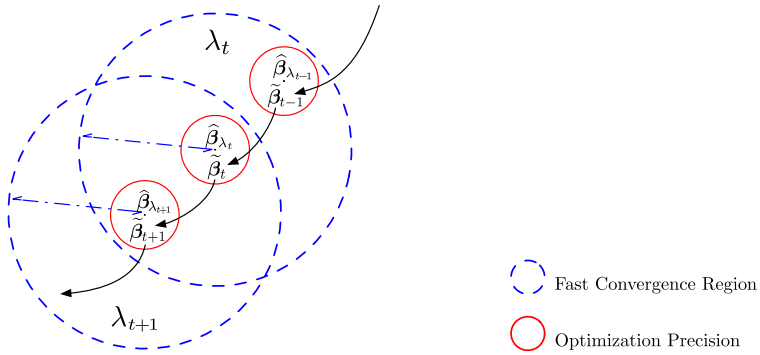

FIG. 1. *For regularization parameter λ_t, $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is an exact local solution satisfying (1.2) with $\lambda = \lambda_t$. Within the tth path-following stage, our algorithm achieves an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$, which approximates the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ up to certain optimization precision. Our approximate path-following algorithm ensures that $\widetilde{\boldsymbol{\beta}}_t$ is sparse, and therefore falls into the fast convergence region corresponding to regularization parameter λ_{t+1}.*

where $k = 1, 2, \ldots$. Here $\boldsymbol{\beta}_t^k$ and L_t^k correspond to the kth iteration of the proximal-gradient method for regularization parameter λ_t. Here L_t^k is chosen by an adaptive line-search method, which will be specified in Section 3.2. Let $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ be an exact local solution satisfying (1.2) with $\lambda = \lambda_t$. As illustrated in Figure 1, for each λ_t, our algorithm calculates an approximation $\widetilde{\boldsymbol{\beta}}_t$ of the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ up to certain optimization precision. Such approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ guarantees to be sparse, and therefore falls into the fast convergence region corresponding to λ_{t+1}. Consequently, the resulting procedure achieves a geometric rate of convergence within each path-following stage, and therefore attains a global geometric rate of convergence for calculating the entire regularization path. Moreover, we establish the nonasymptotic statistical rates of convergence and oracle properties for all the approximate and exact local solutions along the full regularization path.

The idea of path following has been well studied for sparse recovery problems [Breheny and Huang (2011), Efron et al. (2004), Friedman, Hastie and Tibshirani (2010), Hastie et al. (2004), Mairal and Yu (2012), Mazumder, Friedman and Hastie (2011), Park and Hastie (2007), Rosset and Zhu (2007), Xiao and Zhang (2013), Zhao and Yu (2007)]. Compared with these previous works, we consider a broader family of nonconvex M-estimators, including nonconvex penalty functions, such as SCAD and MCP, as well as nonconvex loss functions, such as semiparametric elliptical design loss. Moreover, we provide sharp computational and statistical analysis for all the approximate and exact local solutions attained by the proposed approximate path-following method along the regularization path.

The contributions of this paper are twofold:

- Computationally, we propose an optimization algorithm that ensures a global geometric rate of convergence for nonconvex sparse learning problems. In de-

tail, recall that N is the total number of path-following stages. Within the Nth path-following stage, we denote by ε_{opt} the desired optimization precision of the approximate local solution $\widetilde{\boldsymbol{\beta}}_N$. We need no more than a logarithmic number of the proximal-gradient update iterations defined in (1.3) to calculate the entire path,

$$\text{Total \# of proximal-gradient iterations} \leq C \log\biggl(\frac{1}{\varepsilon_{\text{opt}}}\biggr),$$

where $C > 0$ is a constant. This global geometric rate of convergence is optimal among all first-order methods because it attains the lower bound for first-order methods on strongly convex and smooth objective function [Nesterov (2004), Theorem 2.1.12], which is a subclass of the possibly nonconvex objective functions considered in this paper.

- Statistically, we prove that along the full regularization path, all the approximate local solutions obtained by our algorithm enjoy desirable statistical rates of convergence for estimating the true parameter vector $\boldsymbol{\beta}^*$. In detail, let s^* be the number of nonzero entries of $\boldsymbol{\beta}^*$, and the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$'s satisfy

$$\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\|_2 \leq C\lambda_t\sqrt{s^*} \qquad \text{for } t = 1, \ldots, N \tag{1.4}$$

with high probability. In particular, within the Nth path-following stage, we have $\lambda_N = \lambda_{\text{tgt}} = C'\sqrt{\log d/n}$. Here C and C' are positive constants that do not dependent on d and n. In the $d \gg n$ regime, the final approximate local solution $\widetilde{\boldsymbol{\beta}}_N$ achieves the optimal statistical rate of convergence. Furthermore, we prove that, within the tth path-following stage, the iterative solution sequence $\{\boldsymbol{\beta}_t^k\}_{k=0}^{\infty}$ produced by (1.3) converges toward a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$, which enjoys a more refined oracle statistical property. More specifically, let s_1^* be the number of "large" nonzero coefficients of $\boldsymbol{\beta}^*$ and $s_2^* = s^* - s_1^*$ be the number of "small" nonzero coefficients (detailed definitions of s_1^* and s_2^* are provided in Theorem 4.8), we have

$$\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{\frac{s_1^*}{n}} + C'\sqrt{s_2^*}\lambda_t \qquad \text{for } t = 1, \ldots, N \tag{1.5}$$

with high probability. In particular, for the final stage we have $\lambda_N = \lambda_{\text{tgt}} = C''\sqrt{\log d/n}$. Here C, C' and C'' are positive constants. Note that the oracle statistical property in (1.5) is significantly sharper than the rate of convergence in (1.4); for example, when $s^* = s_1^*$ and $t = N$, the right-hand side of (1.4) is of the order of $\sqrt{s^* \log d/n}$, while the right-hand side of (1.5) is of the order of $\sqrt{s^*/n}$. Moreover, we prove that when the absolute values of the nonzero coefficients of $\boldsymbol{\beta}^*$ are larger than $C'''\sqrt{\log d/n}$, $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ exactly recovers the support of $\boldsymbol{\beta}^*$, that is,

$$\operatorname{supp}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) = \operatorname{supp}(\boldsymbol{\beta}^*).$$

In summary, our joint analysis of the statistical and computational properties provides a theoretical characterization of the entire regularization path.

In independent work, Loh and Wainwright (2013) discussed similar problems. In detail, they provided sufficient conditions under which local optima have desired theoretical properties, and verified that the approximate local solution attained by the composite gradient descent method satisfies these conditions. Our work differs from theirs in three aspects:

(i) Our statistical recovery result in (1.4) covers all the approximate local solutions along the entire regularization path. They provided a similar statistical result, but only for the target regularization parameter, that is, $\lambda_N = \lambda_{\mathrm{tgt}}$ in (1.4).

(ii) As results of independent interest, we prove the oracle statistical properties of the exact local solutions along the regularization path, including the refined statistical rates of convergence in (1.5) and the guarantee of exact support recovery, while they did not provide such results. Since the statistical result in (1.4) is also achievable using convex regularization, for example, the ℓ_1 penalty, these oracle properties are essential for justifying the benefits of using nonconvex penalty functions.

(iii) Our analysis technique is different from theirs. In detail, our statistical analysis is embedded in the analysis of the optimization procedure. In particular, we provide fine-grained analysis of the sparsity pattern of all the intermediate solutions obtained from the proximal-gradient iterations. In contrast, they provided characterizations of local solutions under a global restricted strongly convex/smoothness condition.

The rest of this paper is organized as follows. First we briefly introduce some useful notation. In Section 2 we introduce M-estimators with possibly nonconvex loss and penalty functions. In Section 3 we present the proposed approximate regularization path-following method. In Section 4 we present the main theoretical results on the computational efficiency and statistical accuracy of the proposed procedure. In Section 5 we prove the theoretical results in Section 4. In Section 6 we provide a detailed comparison between our method and the existing nonconvex procedures. Numerical results are presented in Section 7.

Notation: For $q \in [1, +\infty)$, the ℓ_q norm of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T \in \mathbb{R}^d$ is denoted by $\|\boldsymbol{\beta}\|_q = (\sum_{j=1}^d |\beta_j|^q)^{1/q}$. Specifically, we define $\|\boldsymbol{\beta}\|_\infty = \max_{1 \le j \le d}\{|\beta_j|\}$ and $\|\boldsymbol{\beta}\|_0 = \mathrm{card}\{\mathrm{supp}(\boldsymbol{\beta})\}$, where $\mathrm{supp}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$ and $\mathrm{card}\{\cdot\}$ is the cardinality of a set. Correspondingly, we denote the ℓ_q ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_q \le R\}$ by $B_q(R)$. For a set S, we denote its cardinality by $|S|$ and its complement by $\bar{S}$. For $S, \bar{S} \subseteq \{1, \ldots, d\}$, we define $\boldsymbol{\beta}_S \in \mathbb{R}^d$ and $\boldsymbol{\beta}_{\bar{S}} \in \mathbb{R}^d$ as $(\boldsymbol{\beta}_S)_j = \mathbb{1}(j \in S) \cdot \beta_j$ and $(\boldsymbol{\beta}_{\bar{S}})_j = \mathbb{1}(j \notin S) \cdot \beta_j$ for $j = 1, \ldots, d$, where $\mathbb{1}(\cdot)$ is the indicator function. We denote all-zero matrices by $\mathbf{0}$. For notational simplicity, we use generic absolute constants $C, C', \ldots,$ whose values may change from line to line.

Throughout, we denote the exact and approximate local solutions by $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, respectively. We index $\widehat{\boldsymbol{\beta}}$ with the corresponding regulation parameter λ, for

example, $\widehat{\boldsymbol{\beta}}_\lambda$. For the proposed path-following method, we use subscript t to index the path-following stages, for example, the approximate local solution obtained within the tth stage is denoted by $\widetilde{\boldsymbol{\beta}}_t$. Within the tth stage, we index the proximal-gradient iterations with superscript k, for example, $\boldsymbol{\beta}_t^k$.

2. Some nonconvex sparse learning problems. Many theoretical results on penalized M-estimators rely on the condition that the loss and penalty functions are convex, since convexity makes both computational and statistical analysis convenient. However, the statistical performance of the estimator obtained from these convex formulations could be suboptimal in some settings. In the following, we introduce several nonconvex sparse learning problems as motivating examples.

2.1. *Nonconvex penalty.* Throughout this paper, we consider decomposable penalty functions

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_\lambda(\beta_j),$$

for example, the ℓ_1 penalty $\lambda\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^d \lambda|\beta_j|$. When the minimum of $|\beta_j^*| > 0$ is not close to zero, the ℓ_1 penalty introduces large bias in parameter estimation. To remedy this effect, Fan and Li (2001) proposed the SCAD penalty

$$p_\lambda(\beta_j) = \lambda \int_0^{|\beta_j|} \left\{ \mathbb{1}(z \le \lambda) + \frac{(a\lambda - z)_+}{(a-1)\lambda} \mathbb{1}(z > \lambda) \right\} \mathrm{d}z, \qquad a > 2, \tag{2.1}$$

and Zhang (2010a) proposed the MCP penalty

$$p_\lambda(\beta_j) = \lambda \int_0^{|\beta_j|} \left(1 - \frac{z}{\lambda b}\right)_+ \mathrm{d}z, \qquad b > 0. \tag{2.2}$$

See Zhang and Zhang (2012) for a detailed survey. These nonconvex penalty functions are illustrated in Figure 2(a). In fact, these nonconvex penalties can be formulated as the sum of the ℓ_1 penalty and a concave part

$$p_\lambda(\beta_j) = \lambda|\beta_j| + q_\lambda(\beta_j). \tag{2.3}$$

The concave components $q_\lambda(\beta_j)$ of SCAD and MCP are illustrated in Figure 2(b), while the corresponding derivatives $q'_\lambda(\beta_j)$ are illustrated in Figure 2(c). See Section A.1 of the supplementary material [Wang, Liu and Zhang (2014)] for the detailed analytical forms of $p_\lambda(\beta_j)$ and $q_\lambda(\beta_j)$ for SCAD and MCP.

In fact, our method and theory are not limited to SCAD and MCP. More generally, we only rely on the following regularity conditions on the concave component $q_\lambda(\beta_j)$:

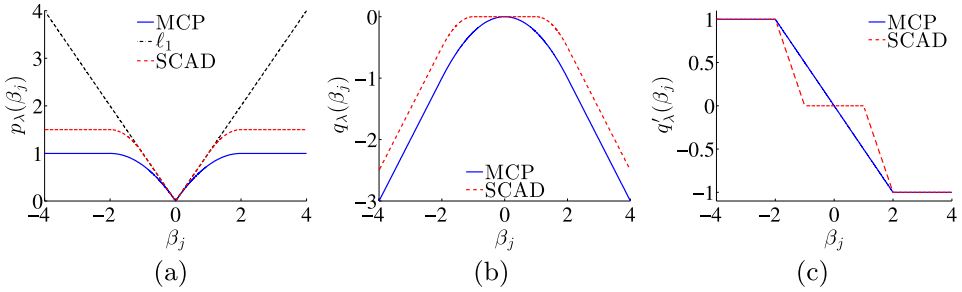

FIG. 2. *An illustration of nonconvex penalties*: (a) *plots of* $p_\lambda(\beta_j)$ *for MCP,* ℓ_1 *and SCAD*; (b) *plots of* $q_\lambda(\beta_j)$ *for MCP and SCAD*; (c) *plots of* $q'_\lambda(\beta_j)$ *for MCP and SCAD. Here* $p_\lambda(\beta_j)$ *is the penalty function evaluated at the* j*th dimension of* $\boldsymbol{\beta}$, $q_\lambda(\beta_j)$ *is the concave component of* $p_\lambda(\beta_j)$ *and* $q'_\lambda(\beta_j)$ *is the derivative of* $q_\lambda(\beta_j)$. *Here we set* $a = 2.1$ *for SCAD,* $b = 2$ *for MCP and* $\lambda = 1$.

Regularity conditions on nonconvex penalty.

(a) $q'_\lambda(\beta_j)$ is monotone and Lipschitz continuous, that is, for $\beta'_j > \beta_j$, there exist two constants $\zeta_- \geq 0$ and $\zeta_+ \geq 0$ such that:

$$-\zeta_- \leq \frac{q'_\lambda(\beta'_j) - q'_\lambda(\beta_j)}{\beta'_j - \beta_j} \leq -\zeta_+ \leq 0;$$

(b) $q_\lambda(\beta_j)$ is symmetric, that is, $q_\lambda(-\beta_j) = q_\lambda(\beta_j)$ for any β_j;
(c) $q_\lambda(\beta_j)$ and $q'_\lambda(\beta_j)$ pass through the origin, that is, $q_\lambda(0) = q'_\lambda(0) = 0$;
(d) $q'_\lambda(\beta_j)$ is bounded, that is, $|q'_\lambda(\beta_j)| \leq \lambda$ for any β_j;
(e) $q'_\lambda(\beta_j)$ has bounded difference with respect to λ: $|q'_{\lambda_1}(\beta_j) - q'_{\lambda_2}(\beta_j)| \leq |\lambda_1 - \lambda_2|$ for any β_j.

In regularity condition (a), ζ_- and ζ_+ are two parameters that control the concavity of $q_\lambda(\beta_j)$. Note that the second order derivative of a function characterizes its convexity/concavity. Taking $\beta'_j \to \beta_j$ in regularity condition (a), we have $q''_\lambda(\beta_j) \in [-\zeta_-, -\zeta_+]$ [ignoring those β_j's where $q''_\lambda(\beta_j)$ does not exist], which suggests larger ζ_- and ζ_+ allow $q_\lambda(\beta_j)$ to be more concave. For SCAD we have $\zeta_- = 1/(a-1)$ and $\zeta_+ = 0$, while for MCP we have $\zeta_- = 1/b$ and $\zeta_+ = 0$. In Figure 2(b) and (c), we can verify that regularity conditions (a)–(d) hold for MCP and SCAD. In addition, we illustrate regularity condition (e) for MCP and SCAD in Section A.2 of the supplementary material [Wang, Liu and Zhang (2014)].

From (2.3) we have $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d p_\lambda(\beta_j) = \lambda\|\boldsymbol{\beta}\|_1 + \sum_{j=1}^d q_\lambda(\beta_j)$. For notational simplicity, we define

$$\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j) = \mathcal{P}_\lambda(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1. \tag{2.4}$$

Hence $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ denotes the decomposable concave component of the nonconvex penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$.

2.2. *Nonconvex loss function*. In this paper, we focus on an example of nonconvex loss function named semiparametric elliptical design regression. More specifically, we have n pairs of observations $\mathbf{z}_1 = (y_1, \mathbf{x}_1^T)^T, \ldots, \mathbf{z}_n = (y_n, \mathbf{x}_n^T)^T$ of a random vector $\mathbf{Z} = (Y, \mathbf{X}^T)^T \in \mathbb{R}^{d+1}$ that follows a $(d+1)$-dimensional elliptical distribution. (See Section A.3 of the supplementary material [Wang, Liu and Zhang (2014)] for a detailed introduction to elliptical distribution.) Then we can verify that $(Y|\mathbf{X} = \mathbf{x})$ follows a univariate elliptical distribution. If we assume that $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}^*$, the population version of the semiparametric elliptical design regression estimator can be defined as

$$\begin{aligned}\breve{\boldsymbol{\beta}} &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\frac{1}{2}\mathbb{E}_{\mathbf{X},Y}\big((Y - \mathbf{X}^T\boldsymbol{\beta})^2\big) + \mathcal{P}_\lambda(\boldsymbol{\beta})\right\} \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\frac{1}{2}(1, -\boldsymbol{\beta}^T)\boldsymbol{\Sigma}_{\mathbf{Z}}(1, -\boldsymbol{\beta}^T)^T + \mathcal{P}_\lambda(\boldsymbol{\beta})\right\}.\end{aligned} \tag{2.5}$$

The above procedure is not practically implementable since the population covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Z}}$ in (2.5) is unknown. In practice, we need to estimate the population covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Z}}$. For this purpose, we propose a rank-based covariance matrix estimator $\widehat{\mathbf{K}}_{\mathbf{Z}}$, which is calculated by a two-step procedure described in Section A.4 of the supplementary material [Wang, Liu and Zhang (2014)]. Since $\widehat{\mathbf{K}}_{\mathbf{Z}}$ is not necessarily positive semidefinite, the loss function in semiparametric elliptical design regression, that is,

$$\mathcal{L}(\boldsymbol{\beta}) = \tfrac{1}{2}(1, -\boldsymbol{\beta}^T)\widehat{\mathbf{K}}_{\mathbf{Z}}(1, -\boldsymbol{\beta}^T)^T \tag{2.6}$$

is possibly nonconvex.

3. Approximate regularization path-following method. Before we go into details, we first present the high-level idea of approximate regularization path following. We then introduce the basic building block of our path-following method, a proximal-gradient method tailored to nonconvex problems.

3.1. *Approximate regularization path following*. Fast local geometric convergence in the proximity of sparse solutions has been observed by many authors [Agarwal, Negahban and Wainwright (2012), Blumensath and Davies (2009), Wright, Nowak and Figueiredo (2009), Xiao and Zhang (2013)]. We exploit such fast local convergence under an approximate path framework to achieve fast global convergence.

Initialization: In (1.1), when the regularization parameter λ is sufficiently large, the solution to sparse learning problems is an all-zero vector. Recall that any exact local solution $\widehat{\boldsymbol{\beta}}_\lambda$ satisfies the first-order optimality condition, $\mathbf{0} \in \partial\{\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda) + \mathcal{P}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\}$. Since the nonconvex penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$ can be formulated as

$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \mathcal{Q}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$, where $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is defined in (2.4), the first-order optimality condition implies there should exist some subgradient $\boldsymbol{\xi} \in \partial\|\widehat{\boldsymbol{\beta}}_\lambda\|_1$ such that

$$\mathbf{0} = \nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda) + \nabla\mathcal{Q}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) + \lambda\boldsymbol{\xi}. \tag{3.1}$$

Let λ be chosen such that $\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{0}$. Then regularity condition (c) implies $\nabla\mathcal{Q}_\lambda(\mathbf{0}) = \mathbf{0}$. Meanwhile, since $\boldsymbol{\xi} \in \partial\|\mathbf{0}\|_1$, we have $\|\boldsymbol{\xi}\|_\infty \le 1$, which implies $\|\nabla\mathcal{L}(\mathbf{0})\|_\infty \le \lambda$ in (3.1). Hence $\lambda_0 = \|\nabla\mathcal{L}(\mathbf{0})\|_\infty$ is the smallest regularization parameter such that any exact local solution $\widehat{\boldsymbol{\beta}}_\lambda$ to the minimization problem (1.1) is all-zero. We choose this λ_0 to be the initial parameter of our regularization path.

Approximate path following: Let $\lambda_{\mathrm{tgt}} \in (0, \lambda_0)$ be the target regularization parameter in (1.1). In practice, we may choose λ_{tgt} by cross-validation or the high-dimensional BIC criterion proposed by Wang, Kim and Li (2013). We consider a decreasing sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$, where

$$\lambda_t = \eta^t\lambda_0 \qquad (t = 0, \ldots, N), \qquad \lambda_N = \lambda_{\mathrm{tgt}} \quad \text{and} \quad \eta \in [0.9, 1). \tag{3.2}$$

Here η is an absolute constant that does not scale with sample size n and dimension d. In Sections 4 and 5 we will prove that $\eta \in [0.9, 1)$ ensures the global geometric rate of convergence. Consequently, since we have $\lambda_{\mathrm{tgt}} = \lambda_0\eta^N$ by (3.2), the number of path-following stages is

$$N = \frac{\log(\lambda_0/\lambda_{\mathrm{tgt}})}{\log(\eta^{-1})}. \tag{3.3}$$

Without loss of generality, we assume that η is properly chosen such that N is an integer. We will show in Section 4 that λ_{tgt} scales with sample size n and dimension d. Since η is a constant, the number of stages N also scales with n and d. Within the tth ($t = 1, \ldots, N$) path-following stage, we aim to obtain a local solution to the minimization problem $\min_{\boldsymbol{\beta}}\{\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_{\lambda_t}(\boldsymbol{\beta})\}$.

As shown in lines 5–9 of Algorithm 1, within the tth ($t = 1, \ldots, N-1$) path-following stage, we employ a variant of proximal-gradient method (Algorithm 3) to obtain an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ for regularization parameter $\lambda_t = \eta^t\lambda_0$. To ensure that each path-following stage enjoys a fast geometric rate of convergence, we propose an approximation path-following strategy. More specifically, we use the approximate local solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ obtained within the $(t-1)$th path-following stage to initialize the tth stage (lines 8 and 12 of Algorithm 1). Recall that we need to adaptively search for the best L_t^k ($k = 0, 1, \ldots$) in (1.3). To achieve computational efficiency, within the $(t-1)$th path-following stage, we store the chosen L_{t-1}^k at the last proximal-gradient iteration of the $(t-1)$th stage as L_{t-1}. Within the tth stage we initialize the search for L_t^0 with L_{t-1} (lines 8 and 12 of Algorithm 1), which will be explained in Section 3.2.

Configuration of optimization precision: We set the optimization precision ε_t for the tth ($t = 1, \ldots, N-1$) stage to be $\lambda_t/4$ (line 7 of Algorithm 1). Within the Nth path-following stage where $\lambda_N = \lambda_{\mathrm{tgt}}$ (line 10), we solve up to high optimization precision $\varepsilon_{\mathrm{opt}} \ll \lambda_{\mathrm{tgt}}/4$ (line 11). The intuition behind this configuration of optimization precision is explained as follows:

Algorithm 1 The approximate path-following method, which solves for a decreasing sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$. Within the tth path-following stage, we employ the proximal-gradient method illustrated in Algorithm 3 to achieve an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ for λ_t. This approximate local solution is then used to initialize the $(t+1)$th stage.

$\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N \leftarrow$ Approximate-Path-Following$(\lambda_{\text{tgt}}, \varepsilon_{\text{opt}})$
input: $\lambda_{\text{tgt}} > 0$, $\varepsilon_{\text{opt}} > 0$ {Here we set $\varepsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$.}
parameters: $\eta \in [0.9, 1)$, $R > 0$, $L_{\min} > 0$, $\lambda_0 = \|\nabla \mathcal{L}(\mathbf{0})\|_\infty$
{For logistic loss, we set $R \in (0, +\infty)$; For other loss functions, we set $R = +\infty$.}
{In practice, we set $L_{\min}$ to be a sufficiently small value, for example, 10^{-6}.}
initialize: $\widetilde{\boldsymbol{\beta}}_0 \leftarrow \mathbf{0}$, $L_0 \leftarrow L_{\min}$, $N \leftarrow \log(\lambda_0/\lambda_{\text{tgt}})/\log(\eta^{-1})$
for $t = 1, \ldots, N-1$ **do**
  $\lambda_t \leftarrow \eta^t \lambda_0$
  $\varepsilon_t \leftarrow \lambda_t/4$
  $\{\widetilde{\boldsymbol{\beta}}_t, L_t\} \leftarrow$ Proximal-Gradient$(\lambda_t, \varepsilon_t, \widetilde{\boldsymbol{\beta}}_{t-1}, L_{t-1}, R)$ as in Algorithm 3
end for
$\lambda_N \leftarrow \lambda_{\text{tgt}}$
$\varepsilon_N \leftarrow \varepsilon_{\text{opt}}$
$\{\widetilde{\boldsymbol{\beta}}_N, L_N\} \leftarrow$ Proximal-Gradient$(\lambda_N, \varepsilon_N, \widetilde{\boldsymbol{\beta}}_{N-1}, L_{N-1}, R)$
return $\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N$

- For $t = 1, \ldots, N-1$, recall the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is an estimator of the true parameter vector $\boldsymbol{\beta}^*$ corresponding to the regularization parameter λ_t. According to high-dimensional statistical theory, the statistical error of $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ should be upper bounded by $C\lambda_t\sqrt{s^*}$ with high probability, where $s^* = \|\boldsymbol{\beta}^*\|_0$. In Lemma 5.1 we will prove that if the optimization error of the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ is at most $\lambda_t/4$, then $\widetilde{\boldsymbol{\beta}}_t$ lies within a ball of radius $C'\lambda_t\sqrt{s^*}$ centered at $\boldsymbol{\beta}^*$ with high probability. That is to say, the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ has the same order of statistical error as the exact solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$, and therefore enjoys desired statistical recovery properties. In particular, in Theorem 5.5 we will prove that $\widetilde{\boldsymbol{\beta}}_t$ is guaranteed to be sparse, and thus falls into the fast convergence region of the next path-following stage.
- However, for $t = N$, we need to solve up to high optimization precision $\varepsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$. This is because even though $\widetilde{\boldsymbol{\beta}}_t$ and $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ both have statistical error of the order $\lambda_t\sqrt{s^*}$, in certain regimes (to be specified in Theorem 4.8), the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ can achieve an improved recovery performance [as shown in (1.5)] due to the usage of nonconvex penalties. Therefore, within the final stage we need to obtain an approximate solution $\widetilde{\boldsymbol{\beta}}_N$ as close to the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}$ as possible, so that $\widetilde{\boldsymbol{\beta}}_N$ has a sharper statistical rate of convergence.

In Algorithm 1, $R > 0$ (line 3) is a parameter that determines the radius of the constraint used in the proximal-gradient method (lines 8 and 12). For least squares loss and semiparametric elliptical design loss, we do not need any constraint. Therefore, we set $R = +\infty$. However, for logistic loss we need to impose an ℓ_2 constraint of radius $R \in (0, +\infty)$. Here $L_{\min}$ is a parameter used in the proximal-gradient method (line 3 of Algorithm 3), which is often set to be a sufficiently small value in practice, for example, $L_{\min} = 10^{-6}$; we will provide the details in Section 3.2.

3.2. *Proximal-gradient method for nonconvex problems.* Before we introduce our proximal-gradient method which is tailored to nonconvex problems, we first give a brief introduction to Nesterov's proximal-gradient method [Nesterov (2013)], which solves the following convex optimization problem:

$$\text{minimize}\, \phi_\lambda(\boldsymbol{\beta}) \qquad \text{where } \phi_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}), \boldsymbol{\beta} \in \Omega. \tag{3.4}$$

Here $\mathcal{L}(\boldsymbol{\beta})$ is convex and differentiable, $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is convex but possibly nonsmooth and Ω is a closed convex set.

Recall that $\boldsymbol{\beta}_t^k$ corresponds to the kth iteration of the proximal-gradient method within the tth path-following stage. Nesterov's proximal-gradient method updates $\boldsymbol{\beta}_t^k$ to be the minimizer of the following local quadratic approximation of $\phi_{\lambda_t}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_t^{k-1}$:

$$\begin{aligned}\psi_{L_t^k,\lambda_t}(\boldsymbol{\beta}; \boldsymbol{\beta}_t^{k-1}) = \mathcal{L}(\boldsymbol{\beta}_t^{k-1}) + \nabla\mathcal{L}(\boldsymbol{\beta}_t^{k-1})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{k-1}) \\ + \frac{L_t^k}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{k-1}\|_2^2 + \mathcal{P}_{\lambda_t}(\boldsymbol{\beta}),\end{aligned} \tag{3.5}$$

where $L_t^k > 0$ is chosen by line-search.

Nesterov's proximal-gradient method requires that both $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ in (3.4) are convex. However, in the optimization problem (1.1) considered in this paper, $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ may be no longer convex. In this case, directly plugging $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ into Nesterov's proximal-gradient might lead to the phenomenon of bad local optima under a path-following scheme, as observed by She (2009, 2012). To extend the proximal-gradient method to nonconvex settings, we adopt an alternative formulation of the objective function.

Recall that the nonconvex penalty can be decomposed as $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 + \mathcal{Q}_\lambda(\boldsymbol{\beta})$, where $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is defined in (2.4). For notational simplicity, we denote $\mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ by $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$. Therefore, the objective function $\phi_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$ can be reformulated as

$$\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{3.6}$$

where we can view $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ as a surrogate loss function and $\lambda\|\boldsymbol{\beta}\|_1$ as a new penalty function. This reformulation ensures the convexity of the new penalty function.

Moreover, in Lemma 5.1 we will prove that, the surrogate loss function $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is actually strongly convex on a sparse set. Correspondingly, we modify Nesterov's proximal-gradient method to minimize the local quadratic approximation defined as

$$\begin{aligned}\psi_{L_t^k,\lambda_t}(\boldsymbol{\beta};\boldsymbol{\beta}_t^{k-1}) &= \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{k-1}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{k-1})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_t^{k-1}) \\ &\quad + \frac{L_t^k}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_t^{k-1}\|_2^2 + \lambda_t\|\boldsymbol{\beta}\|_1.\end{aligned} \tag{3.7}$$

Note that, unlike (3.5), we use a quadratic approximation to the surrogate loss function $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ in (3.7), instead of the original loss function $\mathcal{L}(\boldsymbol{\beta})$. At the kth iteration of the proximal-gradient method, we update $\boldsymbol{\beta}_t^k$ to be the minimizer of the quadratic approximation defined in (3.7), that is,

$$\boldsymbol{\beta}_t^k \leftarrow \operatorname*{argmin}_{\boldsymbol{\beta}\in\Omega}\{\psi_{L_t^k,\lambda_t}(\boldsymbol{\beta};\boldsymbol{\beta}_t^{k-1})\}. \tag{3.8}$$

Now we specify the constraint set Ω in (3.8). For $\mathcal{L}(\boldsymbol{\beta})$ being least squares or semiparametric elliptical design loss, we set $\Omega=\mathbb{R}^d$. For logistic loss, we set $\Omega=B_2(R)$ with $R\in(0,+\infty)$, where $B_2(R)$ is a centered ℓ_2 ball of radius R. In Lemma 5.1 we will show that, in the setting of logistic loss, the boundedness of $\|\boldsymbol{\beta}_t^k\|_2$'s is essential for establishing the strong convexity of the surrogate loss function $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ along the full regularization path. To unify the notation, we consider $\Omega=B_2(R)$ throughout—when the constraint set $\Omega=\mathbb{R}^d$, we set $R=+\infty$. Correspondingly, we denote (3.8) by

$$\boldsymbol{\beta}_t^k \leftarrow \mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1};R). \tag{3.9}$$

In the sequel, we provide the closed-form expression of update scheme (3.9):

Update scheme of proximal-gradient method for nonconvex problems.

- For $\Omega=\mathbb{R}^d$, that is, $R=+\infty$, $\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1};+\infty)$ is a soft-thresholding operator taking the form of

$$\begin{aligned}&\big(\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1};+\infty)\big)_j \\ &\quad=\begin{cases}0, & \text{if } |\bar\beta_j|\le\lambda_t/L_t^k,\\ \operatorname{sign}(\bar\beta_j)(|\bar\beta_j|-\lambda_t/L_t^k), & \text{if } |\bar\beta_j|>\lambda_t/L_t^k,\end{cases}\end{aligned} \tag{3.10}$$

  for $j=1,\dots,d$, where

$$\begin{aligned}\bar{\boldsymbol{\beta}} &= \boldsymbol{\beta}_t^{k-1} - \frac{1}{L_t^k}\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{k-1}) \\ &= \boldsymbol{\beta}_t^{k-1} - \frac{1}{L_t^k}(\nabla\mathcal{L}(\boldsymbol{\beta}_t^{k-1}) + \nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}_t^{k-1})),\end{aligned} \tag{3.11}$$

  and $\bar\beta_j$ is the jth dimension of $\bar{\boldsymbol{\beta}}$.

- For $\Omega = B_2(R)$ with $R \in (0, +\infty)$, $\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; R)$ can be obtained by projecting $\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; +\infty)$ defined in (3.10) onto $B_2(R)$, that is,

$$\begin{aligned}&\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; R)\\ &= \begin{cases} \mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; +\infty), & \text{if } \|\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; +\infty)\|_2 < R,\\ \dfrac{R \cdot \mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; +\infty)}{\|\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; +\infty)\|_2}, & \text{if } \|\mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; +\infty)\|_2 \geq R. \end{cases}\end{aligned} \tag{3.12}$$

See Section B.2 of the supplementary material [Wang, Liu and Zhang (2014)] for a detailed derivation. In Section B.1 of the supplementary material [Wang, Liu and Zhang (2014)], we provide the specific forms of $\nabla\mathcal{L}(\boldsymbol{\beta})$ and $\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta})$ in (3.11) for the nonconvex problems discussed in Section 2.

Line-search method: Before we present the proposed proximal-gradient method in detail, we briefly introduce a line-search algorithm, which adaptively searches for the best quadratic coefficient L_t^k of the local quadratic approximation (3.7). As shown in lines 4–7 of Algorithm 2, the main idea of line-search is to iteratively increase L_t^k by a factor of two and compute the corresponding $\boldsymbol{\beta}_t^k$, until the local approximation $\psi_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^k; \boldsymbol{\beta}_t^{k-1})$ becomes a tight upper bound of the objective function $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k)$. We will theoretically characterize the computational complexity of this line-search algorithm in Remark 4.6, and specify the range of L_t^k in Theorem 5.5.

Stopping criterion: In the following, we introduce the stopping criterion of our proximal-gradient method. In other words, we specify the optimality conditions that should be satisfied by the approximate solution $\widetilde{\boldsymbol{\beta}}_t$ attained by our proximal-gradient method.

Algorithm 2 The line-search method used to search for the best L_t^k and compute the corresponding $\boldsymbol{\beta}_t^k$. Here $\phi_{\lambda_t}(\boldsymbol{\beta})$ is the objective function defined in (3.4), and $\psi_{L_t^k,\lambda_t}(\boldsymbol{\beta}; \boldsymbol{\beta}_t^{k-1})$ is the local quadratic approximation of $\phi_{\lambda_t}(\boldsymbol{\beta})$ defined in (3.7).

1: $\{\boldsymbol{\beta}_t^k, L_t^k\} \leftarrow$ line-Search$(\lambda_t, \boldsymbol{\beta}_t^{k-1}, L_{\text{init}}, R)$
2: **input:** $\lambda_t > 0$, $\boldsymbol{\beta}_t^{k-1} \in \mathbb{R}^d$, $L_{\text{init}} > 0$, $R > 0$
3: **initialize:** $L_t^k \leftarrow L_{\text{init}}$
4: **repeat**
5: $\quad\boldsymbol{\beta}_t^k \leftarrow \mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; R)$ as defined in (3.9)
6: $\quad$**if** $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k) > \psi_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^k; \boldsymbol{\beta}_t^{k-1})$ **then** $L_t^k \leftarrow 2L_t^k$
7: **until** $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k) \leq \psi_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^k; \boldsymbol{\beta}_t^{k-1})$
8: **return** $\{\boldsymbol{\beta}_t^k, L_t^k\}$

It is known that any exact local solution $\widehat{\boldsymbol{\beta}}_\lambda$ to the optimization problem

$$\text{minimize}\, \phi_\lambda(\boldsymbol{\beta}) \qquad \text{where } \phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \boldsymbol{\beta} \in \Omega$$

satisfies the optimality condition, that is, there exists some $\boldsymbol{\xi} \in \partial\|\widehat{\boldsymbol{\beta}}_\lambda\|_1$ such that

$$(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^T\big(\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) + \lambda\boldsymbol{\xi}\big) \le 0 \qquad \text{for any } \boldsymbol{\beta} \in \Omega. \tag{3.13}$$

We can understand this optimality condition as follows: Locally at $\widehat{\boldsymbol{\beta}}_\lambda$, any feasible direction pointed at $\widehat{\boldsymbol{\beta}}_\lambda$, that is, $(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})$ where $\boldsymbol{\beta} \in \Omega$, leads to a decrease in the objective function value $\phi_\lambda(\boldsymbol{\beta})$, because as shown in (3.13), such direction forms an obtuse angle with the (sub)gradient vector of $\phi_\lambda(\boldsymbol{\beta})$ evaluated at $\widehat{\boldsymbol{\beta}}_\lambda$. If $\widehat{\boldsymbol{\beta}}_\lambda$ lies in the interior of Ω, for example, $\Omega = \mathbb{R}^d$, then (3.13) reduces to the well-known first-order KKT condition,[3]

$$\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) + \lambda\boldsymbol{\xi} = \mathbf{0} \qquad \text{where } \boldsymbol{\xi} \in \partial\|\widehat{\boldsymbol{\beta}}_\lambda\|_1. \tag{3.14}$$

Based on the optimality condition in (3.13), we measure the suboptimality of a $\boldsymbol{\beta} \in \Omega$ with

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1}\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big) \right\}. \tag{3.15}$$

To understand this measure of suboptimality, first note that if $\boldsymbol{\beta}$ is an exact local solution, then we have $\omega_\lambda(\boldsymbol{\beta}) \le 0$ by (3.13). Otherwise, if $\boldsymbol{\beta}$ is close to some exact local solution, then $\omega_\lambda(\boldsymbol{\beta})$ is some small positive value. When $\boldsymbol{\beta}$ lies in the interior of Ω, then (3.15) reduces to a more straightforward

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \big\{\big\|\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big\|_\infty\big\}. \tag{3.16}$$

Because for any fixed $\mathbf{v} \in \mathbb{R}^d$, we have $(\boldsymbol{\beta} + C\mathbf{v}) \in \Omega$ for $C > 0$ sufficiently small. Setting $\boldsymbol{\beta}$ to be this value in (3.15), we have

$$\begin{aligned}\omega_\lambda(\boldsymbol{\beta}) &= \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\mathbf{v} \in \mathbb{R}^d} \left\{ \frac{\mathbf{v}^T}{\|\mathbf{v}\|_1}\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big) \right\} \\ &= \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \big\{\big\|\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big\|_\infty\big\},\end{aligned}$$

where the second equality follows from the duality between ℓ_1 and ℓ_∞ norm.

Equipped with the suboptimality measure $\omega_\lambda(\boldsymbol{\beta})$ defined in (3.15), now we can define the stopping criterion of our proximal-gradient method to be $\omega_{\lambda_t}(\boldsymbol{\beta}_t^k) \le \varepsilon_t$, where $\varepsilon_t > 0$ is the desired optimization precision within the tth path-following stage (line 9 of Algorithm 3). Therefore, the proximal-gradient method achieves

[3]Because given that $\widehat{\boldsymbol{\beta}}_\lambda$ lies in the interior of Ω, we have $(\widehat{\boldsymbol{\beta}}_\lambda + C\mathbf{v}) \in \Omega$ and $(\widehat{\boldsymbol{\beta}}_\lambda - C\mathbf{v}) \in \Omega$ for any fixed $\mathbf{v} \in \mathbb{R}^d$ and $C > 0$ sufficiently small. Setting $\boldsymbol{\beta}$ in (3.13) to be these two values, we obtain $\mathbf{v}^T(\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\xi}) = 0$, which implies (3.14) since $\mathbf{v}$ is arbitrarily chosen.

Algorithm 3 The proximal-gradient method for nonconvex problems, which iteratively leverages the line-search method illustrated in Algorithm 2 at each iteration.

1: $\{\widetilde{\boldsymbol{\beta}}_t, L_t\} \leftarrow$ Proximal-Gradient$(\lambda_t, \varepsilon_t, \boldsymbol{\beta}_t^0, L_t^0, R)$
2: **input:** $\lambda_t > 0, \varepsilon_t > 0, \boldsymbol{\beta}_t^0 \in \mathbb{R}^d, L_t^0 > 0, R > 0$
3: **parameter:** $L_{\min} > 0$
4: **initialize:** $k \leftarrow 0$
5: **repeat**
6: $\quad k \leftarrow k + 1$
7: $\quad L_{\text{init}} \leftarrow \max\{L_{\min}, L_t^{k-1}/2\}$
8: $\quad \boldsymbol{\beta}_t^k, L_t^k \leftarrow$ line-Search$(\lambda_t, \boldsymbol{\beta}_t^{k-1}, L_{\text{init}}, R)$ as in Algorithm 2
9: **until** $\omega_{\lambda_t}(\boldsymbol{\beta}_t^k) \leq \varepsilon_t$ as defined in (3.15)
10: $\widetilde{\boldsymbol{\beta}}_t \leftarrow \boldsymbol{\beta}_t^k$
11: $L_t \leftarrow L_t^k$
12: **return** $\{\widetilde{\boldsymbol{\beta}}_t, L_t\}$

an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ with suboptimality ε_t. Recall that within the tth path-following stage ($t = 1, \ldots, N-1$), we set ε_t to be $\lambda_t/4$ (line 7 of Algorithm 1), while within the Nth path-following stage, we set $\varepsilon_t = \varepsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$ (line 11 of Algorithm 1).

Proposed proximal-gradient method: We are now ready to present the proposed proximal-gradient method in detail. Recall that within the tth stage of our path-following algorithm, we employ the proximal-gradient method to obtain the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ (lines 8 and 12 of Algorithm 1). As shown in line 8 of Algorithm 3, at the kth iteration of our proximal-gradient method, we employ the line-search method (Algorithm 2) to search for the best L_t^k and calculate the corresponding $\boldsymbol{\beta}_t^k$.

At the kth iteration of the proximal-gradient method, we set the initial value L_{init} of line-search to be $\max\{L_{\min}, L_t^{k-1}/2\}$ (line 7 of Algorithm 3), where $L_{\min} > 0$ is used to prevent L_{init} from being too small. In practice, $L_{\min}$ is often set to be a sufficiently small value, for example, $L_{\min} = 10^{-6}$. The intuition behind such initialization can be understood as follows: As shown in (3.7), L_t^{k-1} and L_t^k are the quadratic coefficients of the local quadratic approximations of the objective function at $\boldsymbol{\beta}_t^{k-2}$ and $\boldsymbol{\beta}_t^{k-1}$, respectively. Intuitively speaking, $\boldsymbol{\beta}_t^{k-2}$ and $\boldsymbol{\beta}_t^{k-1}$ are close to each other, which implies that L_t^{k-1} is a good guess for L_t^k. Hence we can initialize the line-search method for L_t^k with a value slightly smaller than L_t^{k-1}, for example, $L_t^{k-1}/2$.

When the stopping criterion $\omega_{\lambda_t}(\boldsymbol{\beta}_t^k) \leq \varepsilon_t$ is satisfied (line 9 of Algorithm 3), the proximal-gradient method stops and outputs the approximate local solution $\widetilde{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t^k$ (line 10 of Algorithm 3). We also keep track of $L_t = L_t^k$ to accelerate the line-search procedure within the next path-following stage.

4. Theoretical results. We establish theoretical results on the iteration complexity and statistical performance of our approximate regularization path-following method for nonconvex learning problems.

4.1. *Assumptions*. We first list the required assumptions. The first assumption is about the relationship between λ_{tgt} and $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty$.

ASSUMPTION 4.1. For least squares loss and logistic loss, we set $\lambda_{\mathrm{tgt}} = C\sqrt{\log d/n}$. Meanwhile, for semiparametric elliptical design loss, we set $\lambda_{\mathrm{tgt}} = C'\|\boldsymbol{\beta}^*\|_1\sqrt{\log d/n}$. We assume

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le \lambda_{\mathrm{tgt}}/8. \tag{4.1}$$

Assumption 4.1 is a common condition that λ_{tgt} should be large enough to dominate the noise. For instance, for least squares loss we have

$$\nabla\mathcal{L}(\boldsymbol{\beta}^*) = \frac{1}{n}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}),$$

where $\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}$ is in fact the noise vector. In Lemma C.1 in Section C.1 of the supplementary material [Wang, Liu and Zhang (2014)] we will show that for least squares loss and logistic loss, we have that $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le C\sqrt{\log d/n}$ holds with high probability. Similarly, in Lemma C.2 in Section C.1 of the supplementary material [Wang, Liu and Zhang (2014)] we will prove that, for semiparametric elliptical design loss, $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le C'\|\boldsymbol{\beta}^*\|_1\sqrt{\log d/n}$ holds with high probability. Thus our assumption on λ_{tgt} and $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty$ holds with high probability.

In the sequel, we lay out another assumption on the sparse eigenvalues of $\nabla^2\mathcal{L}(\boldsymbol{\beta})$, which are defined as follows.

DEFINITION 4.2 (Sparse eigenvalues). Let s be a positive integer. The largest and smallest s-sparse eigenvalues of the Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ are

$$\rho_+(\nabla^2\mathcal{L}, s) = \sup\{\mathbf{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\mathbf{v} : \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1, \boldsymbol{\beta}\in\mathbb{R}^d\},$$
$$\rho_-(\nabla^2\mathcal{L}, s) = \inf\{\mathbf{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\mathbf{v} : \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1, \boldsymbol{\beta}\in\mathbb{R}^d\}.$$

For least squares loss and semiparametric elliptical design loss, $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$. However, for logistic loss we have

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \cdot \frac{1}{1+\exp(-\mathbf{x}_i^T\boldsymbol{\beta})}\cdot\frac{1}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}, \tag{4.2}$$

which depends on $\boldsymbol{\beta}$. Note in Definition 4.2, the smallest s-sparse eigenvalue $\rho_-(\nabla^2\mathcal{L}, s)$ is obtained by taking infimum over all $\boldsymbol{\beta}\in\mathbb{R}^d$. Consequently, for logistic loss, $\rho_-(\nabla^2\mathcal{L}, s)$ is always zero because in (4.2) we can take $\boldsymbol{\beta}$ such that $|\mathbf{x}_i^T\boldsymbol{\beta}| \to +\infty$ for all nonzero $\mathbf{x}_i$'s, which implies that $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ goes to an all-zero

matrix. To avoid this degenerate case, for logistic loss we define the sparse eigenvalues by taking infimum/supremum over all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_2$ bounded instead of over all $\boldsymbol{\beta} \in \mathbb{R}^d$.

DEFINITION 4.3 (Sparse eigenvalues for logistic loss). Let s be a positive integer. For logistic loss, we define the largest and smallest s-sparse eigenvalues of $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ to be

$$\rho_+(\nabla^2\mathcal{L}, s, R) = \sup\{\mathbf{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\mathbf{v} : \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1, \|\boldsymbol{\beta}\|_2 \le R\},$$
$$\rho_-(\nabla^2\mathcal{L}, s, R) = \inf\{\mathbf{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\mathbf{v} : \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1, \|\boldsymbol{\beta}\|_2 \le R\},$$

where $R \in (0, +\infty)$ is an absolute constant such that $\|\boldsymbol{\beta}^*\|_2 \le R$.

In Definition 4.3, we implicitly assume that $\|\boldsymbol{\beta}^*\|_2$ is upper bounded by some known absolute constant. Although it seems rather restrictive, this assumption is essential for logistic loss. Otherwise, $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$ may go to an all-zero matrix when $\|\boldsymbol{\beta}^*\|_2 \to +\infty$. In this case, the curvature of the objective function at $\boldsymbol{\beta}^*$ is zero, a consistent estimation of $\boldsymbol{\beta}^*$ is impossible. Although such assumption is necessary for theoretical purposes, we require no prior knowledge about the exact value of $\|\boldsymbol{\beta}^*\|_2$ in practice, since we can always set R to be a sufficiently large constant in our algorithm (line 3 of Algorithm 1). To unify the later analysis for different loss functions, we omit the extra term R in Definition 4.3 unless its necessary.

Recall that we impose an ℓ_2 constraint of radius R for all the proximal-gradient iterations within each path-following stage (lines 8 and 12 of Algorithm 1). Therefore, we have $\|\boldsymbol{\beta}_t^k\|_2 \le R$ during the whole iterative procedure [for least squares loss and semiparametric elliptical design loss, $R = +\infty$; for logistic loss, $R \in (0, +\infty)$]. Now we are ready to present the assumption on the sparse eigenvalues of the Hessian matrix.

ASSUMPTION 4.4. Let $s^* = \|\boldsymbol{\beta}^*\|_0$. We assume:

- There exists an integer $\widetilde{s} > Cs^*$ such that
$$\rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) < +\infty, \qquad \rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) > 0$$
are two absolute constants. The constant $C > 0$ is specified in (4.4).
- The concavity parameter ζ_- defined in regularity condition (a) satisfies
$$\zeta_- \le C'\rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) \tag{4.3}$$
with constant $C' < 1$.

In Assumption 4.4, the constant

$$C = 144\kappa^2 + 250\kappa, \tag{4.4}$$

where κ is a condition number defined as

$$\kappa = \frac{\rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) - \zeta_+}{\rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) - \zeta_-}. \tag{4.5}$$

The constant in (4.4) is rather large for practical purposes. We could expect it to be much smaller if we manage to get smaller constants in the technical proof. However, we mainly focus on providing novel theoretical insights in this paper, without paying too much attention to optimizing constants.

Recall that regularity condition (a) implies $\zeta_+ \le \zeta_-$. Meanwhile, we have $\rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) \le \rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s})$ by definition. Thus (4.3) implies

$$\zeta_+ \le C'\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big), \tag{4.6}$$

where $C' < 1$ is the same constant as in (4.3). Therefore, we have $\kappa \in [1, +\infty)$. Restrictions (4.3) and (4.6) on the concavity parameters suggest that the concavity of the concave component $\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j)$ of the nonconvex penalty should not outweigh the convexity of the loss function on a sparse set. It is also worth noting the concavity parameters are independent from the regularization parameter; for example, for MCP in (2.2), $b = 1/\zeta_-$ and λ are two independent parameters. Thus Assumption 4.4 does not depend on λ at all.

Assumption 4.4 is closely related to the restricted isometry property (RIP) condition proposed by Candès and Tao (2005). Similar conditions have been studied by Bickel, Ritov and Tsybakov (2009), Raskutti, Wainwright and Yu (2010), Negahban et al. (2012), Zhang (2010b, 2013) and Xiao and Zhang (2013). In detail, for least squares loss, the RIP condition assumes there exists an integer s and some constant $\delta \in (0, 1)$ such that

$$1 - \delta \le \rho_-\big(\nabla^2\mathcal{L}, s\big) \le \rho_+\big(\nabla^2\mathcal{L}, s\big) \le 1 + \delta. \tag{4.7}$$

Now we justify Assumption 4.4 for least squares loss with an example.

To show that Assumption 4.4 is well defined, we assume the RIP condition in (4.7) holds with $s = 877s^*$ and $\delta = 0.01$. We set the concavity parameters of the nonconvex penalty in (a) to be $\zeta_+ = 0$ and $\zeta_- = \rho_-(\nabla^2\mathcal{L}, s)/20$; for example, for MCP defined in (2.2), we take $b = 1/\zeta_- = 20/\rho_-(\nabla^2\mathcal{L}, s)$. In the following, we verify there exists an integer $\widetilde{s} = 438s^*$ that satisfies Assumption 4.4.

First, according to the RIP condition, we have

$$\begin{aligned}\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) &= \rho_+\big(\nabla^2\mathcal{L}, 877s^*\big) = \rho_+\big(\nabla^2\mathcal{L}, s\big)\\ &\le (1 + \delta) = 1.01 < +\infty,\end{aligned} \tag{4.8}$$

$$\begin{aligned}\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) &= \rho_-\big(\nabla^2\mathcal{L}, 877s^*\big) = \rho_-\big(\nabla^2\mathcal{L}, s\big)\\ &\ge (1 - \delta) = 0.99 > 0.\end{aligned} \tag{4.9}$$

Second, we calculate the value of $\widetilde{s}$ in detail. Since the condition number κ defined in (4.5) satisfies

$$1 \le \kappa = \frac{\rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) - \zeta_+}{\rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) - \zeta_-} = \frac{\rho_+(\nabla^2\mathcal{L}, s) - \zeta_+}{\rho_-(\nabla^2\mathcal{L}, s) - \zeta_-}$$
$$= \frac{20}{19} \cdot \frac{\rho_+(\nabla^2\mathcal{L}, s)}{\rho_-(\nabla^2\mathcal{L}, s)} \le \frac{20}{19} \cdot \frac{1+\delta}{1-\delta} < 1.08.$$

We now verify that $\widetilde{s}$ satisfies $\widetilde{s} > Cs^*$ in Assumption 4.4, where C is defined in (4.4). Plugging the range $1 \le \kappa < 1.08$ into the definition of C, we obtain $C = 144\kappa^2 + 250\kappa < 438$. Therefore, as long as the RIP condition holds with $s = 877s^*$ and $\delta = 0.01$, we can find an integer $\widetilde{s} = 438s^*$ that satisfies Assumption 4.4, which also implies Assumption 4.4 is a weaker assumption than the RIP condition. For least squares loss, the RIP condition is known to hold for a variety of design matrices with high probability, which implies that Assumption 4.4 also holds with high probability for these designs.

Furthermore, we will justify Assumption 4.4 for $\mathcal{L}(\boldsymbol{\beta})$ being semiparametric elliptical design loss and logistic loss in Section C.2 of the supplementary material [Wang, Liu and Zhang (2014)]. Also, in the discussion for logistic loss in Section C.2, we prove that the assumption of restricted strong convexity/smoothness in Loh and Wainwright (2013) is stronger than our Assumption 4.4.

Hereafter, we use the shorthand

$$\rho_+ = \rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}), \qquad \rho_- = \rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) \tag{4.10}$$

for notational simplicity.

4.2. *Main theorems.* We first provide the main results about the computational rate of convergence. We then establish the statistical properties of the local solutions obtained by our approximate path-following method.

4.2.1. *Computational theory.* The next theorem shows that the proposed approximate regularization path-following method achieves a global geometric rate of convergence for calculating the entire regularization path, which is the optimal rate among all first-order optimization methods.

Recall that $\varepsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$ is the desired optimization precision of the final path-following stage (line 12 of Algorithm 1), and $N = \log(\lambda_0/\lambda_{\text{tgt}})/\log(\eta^{-1})$ is the total number of approximate path-following stages, where $\eta \in [0.9, 1)$ is an absolute constant. Meanwhile, we remind the reader that $\rho_- = \rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) > 0$ is the smallest sparse eigenvalue specified in Assumption 4.4; As defined in regularity condition (a), $\zeta_- > 0$ is the concavity parameter of the nonconvex penalty, which satisfies (4.3) in Assumption 4.4.

THEOREM 4.5 (Geometric rate of convergence). *Under Assumptions* 4.1 *and* 4.4, *we have the following results*:

(1) Geometric rate of convergence within the tth stage: *Within the tth ($t = 1, \ldots, N$) path-following stage (lines 8 and 12 of Algorithm 1), the iterative sequence $\{\boldsymbol{\beta}_t^k\}_{k=0}^{\infty}$ produced by the proximal-gradient method (Algorithm 3) converges to a unique local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$.*

- *Within the tth path-following stage ($t = 1, \ldots, N-1$), the total number of proximal-gradient iterations (lines 5–9 of Algorithm 3) is no more than $C' \log(4C\sqrt{s^*})$.*
- *Within the Nth stage ($\lambda_N = \lambda_{\mathrm{tgt}}$), the total number of proximal-gradient iterations is no more than $\max\{0, C' \log(C\lambda_{\mathrm{tgt}}\sqrt{s^*}/\varepsilon_{\mathrm{opt}})\}$.*

Here $s^* = \|\boldsymbol{\beta}^*\|_0$ and

$$C = 2\sqrt{21} \cdot \sqrt{\kappa}(1+\kappa), \qquad C' = 2/\log\biggl(\frac{1}{1-1/(8\kappa)}\biggr), \tag{4.11}$$

where $\kappa \in [1, +\infty)$ is the condition number defined in (4.5).

(2) Geometric rate of convergence over the full path: *To compute the entire path, we need no more than*

$$\underbrace{(N-1)C' \log\bigl(4C\sqrt{s^*}\bigr)}_{1,\ldots,(N-1)\text{th stages}} + \underbrace{C' \log\biggl(\frac{C\lambda_{\mathrm{tgt}}\sqrt{s^*}}{\varepsilon_{\mathrm{opt}}}\biggr)}_{N\text{th stage}} \tag{4.12}$$

proximal-gradient iterations, where C, C' are specified in (4.11).

(3) Geometric rate of convergence of objective function value: *Let $\widetilde{\boldsymbol{\beta}}_t$ be the approximate local solution obtained within the tth stage.*

- *For $t = 0, \ldots, N-1$, the value of the objective function decays exponentially toward the value at the final exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}$, that is,*

$$\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_t) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}) \le C\lambda_0^2 s^* \cdot \eta^{2(t+1)}, \tag{4.13}$$

where $C = 105/(\rho_- - \zeta_-)$.

- *For $t = N$, we have*

$$\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_N) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}) \le \bigl(C'\lambda_{\mathrm{tgt}} s^*\bigr) \cdot \varepsilon_{\mathrm{opt}}, \tag{4.14}$$

where $C' = 21/(\rho_- - \zeta_-)$.

PROOF. See the next section for a detailed proof. □

Result (1) suggests that, within each path-following stage, the proximal-gradient algorithm attains a geometric rate of convergence. More specifically, within the tth ($t = 1, \ldots, N$) stage (lines 8 and 12 of Algorithm 1), we only need a logarithmic number of proximal-gradient update iterations (lines 5–9 of Algorithm 3) to compute an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$. Furthermore, within the tth path-following

stage, the iterative sequence $\{\boldsymbol{\beta}_t^k\}_{k=0}^{\infty}$ produced by Algorithm 3 converges toward a unique local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. In Theorem 4.8, we will show that $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ enjoys a more refined statistical rate of convergence due to the usage of nonconvex penalty.

Result (2) suggests that our approximate path-following method attains a global geometric rate of convergence. From the perspective of high-dimensional statistics, the total number of stages N scales with dimension d and sample size n, because $N = \log(\lambda_0/\lambda_{\mathrm{tgt}})/\log(\eta^{-1})$, where η is an absolute constant. From the perspective of optimization, given dimension d and sample size n, when the optimization precision $\varepsilon_{\mathrm{opt}}$ is sufficiently small such that in (4.12) the second term dominates its first term, then the total iteration complexity is $C\log(1/\varepsilon_{\mathrm{opt}})$. In other words, we only need to conduct a logarithmic number of proximal-gradient iterations to compute the full regularization path.

Recall that we measure the suboptimality of an approximate solution with $\omega_\lambda(\boldsymbol{\beta})$ defined in (3.15), which does not directly reflect the suboptimality of the objective function value. Hence we provide result (3) to characterize the decay of the objective gap $\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_t) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}})$. In detail, (4.13) illustrates the exponential decay of the objective gap along the regularization path, that is, $t = 1, \ldots, N-1$, while (4.14) suggests that, the final objective function value evaluated at $\widetilde{\boldsymbol{\beta}}_N$ is sufficiently close to the value at the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}$, as long as the optimization precision $\varepsilon_{\mathrm{opt}}$ is sufficiently small.

Recall the largest sparse eigenvalue $\rho_+ = \rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}) > 0$ is specified in Assumption 4.4; as defined in regularity condition (a), $\zeta_+ > 0$ is the concavity parameter of the nonconvex penalty, which satisfies (4.6) in Assumption 4.4; $L_{\min}$ is a parameter of Algorithm 3 (line 3).

REMARK 4.6. Nesterov (2013) proved that the total number of line-search steps (lines 4–7 of Algorithm 2) within the kth proximal-gradient iteration (line 8 of Algorithm 3) is no more than

$$2(k+1) + \max\left\{0, \frac{\log(\rho_+ - \zeta_+) - \log L_{\min}}{\log 2}\right\}.$$

Piecing the above results together, we conclude that the total number of line-search iterations (lines 4–7 of Algorithm 2) required to compute the full regularization path is of the same order as (4.12).

4.2.2. *Statistical theory.* We present two types of statistical results. Recall that $\widetilde{\boldsymbol{\beta}}_t$ is the approximate local solution obtained within the tth path-following stage, while $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is the corresponding exact local solution that satisfies the exact optimality condition in (3.13). In Theorem 4.7, we will provide a statistical characterization of all the approximate local solutions $\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N$ attained along the full regularization path. Recall that in Theorem 4.5 we prove that within the tth stage, the iterative sequence $\{\boldsymbol{\beta}_t^k\}_{k=0}^{\infty}$ produced by the proximal-gradient method converges

toward a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. In Theorem 4.8, we will provide more refined statistical properties of these exact local solutions $\{\widehat{\boldsymbol{\beta}}_{\lambda_t}\}_{t=1}^N$ along the full regularization path. Since $\widehat{\boldsymbol{\beta}}_{\lambda_N} = \widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}$, this result justifies the statistical property of the final estimator.

THEOREM 4.7 (Statistical rates of convergence of approximate local solutions). *Recall that $\widetilde{\boldsymbol{\beta}}_t$ is the approximate local solution obtained within the tth path-following stage (lines 8 and 12 of Algorithm 1). Under Assumptions 4.1 and 4.4, we have*

$$\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\|_2 \le C\lambda_t\sqrt{s^*} \qquad \text{for } t = 1, \ldots, N, \tag{4.15}$$

where $s^* = \|\boldsymbol{\beta}^*\|_0$ and $C = (21/8)/(\rho_- - \zeta_-)$.

PROOF. See the next section for a detailed proof. □

Theorem 4.7 provides statistical rates of convergence of all the approximate local solutions attained by our algorithm along the regularization path. Recall that in Assumption 4.1, we set $\lambda_{\mathrm{tgt}} = C\sqrt{\log d/n}$ for least squares and logistic loss, and $\lambda_{\mathrm{tgt}} = C'\|\boldsymbol{\beta}^*\|_1\sqrt{\log d/n}$ for semiparametric elliptical design loss. For least squares and logistic loss, taking $t = N$ in Theorem 4.7, we have

$$\|\widetilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*\|_2 \le \frac{21/8}{\rho_- - \zeta_-}\lambda_{\mathrm{tgt}}\sqrt{s^*} = \frac{21/8 \cdot C}{\rho_- - \zeta_-}\sqrt{\frac{s^*\log d}{n}}.$$

Hence the final approximate local solution $\widetilde{\boldsymbol{\beta}}_N$ attains the minimax rate of convergence for parameter estimation. Similarly, for semiparametric elliptical design loss, we have

$$\|\widetilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*\|_2 \le \frac{21/8 \cdot C'}{\rho_- - \zeta_-}\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{s^*\log d}{n}},$$

which suggests that the rate of convergence of the final approximate local solution is also optimal in the regime where $\|\boldsymbol{\beta}^*\|_1$ is upper bounded by a constant. Moreover, since η is an absolute constant, for $\widetilde{\boldsymbol{\beta}}_{N-K}$ with K being a positive integer constant, Theorem 4.7 gives

$$\|\widetilde{\boldsymbol{\beta}}_{N-K} - \boldsymbol{\beta}^*\|_2 \le \frac{21/8}{\rho_- - \zeta_-}\lambda_{N-K}\sqrt{s^*} \le \frac{21/8 \cdot \eta^{-K}}{\rho_- - \zeta_-}\lambda_{\mathrm{tgt}}\sqrt{s^*},$$

which suggests that the approximate local solution $\widetilde{\boldsymbol{\beta}}_{N-K}$ enjoys the same rate of convergence as the final approximate local solution $\widetilde{\boldsymbol{\beta}}_N$, but with a larger constant $C = (21/8)\cdot\eta^{-K}/(\rho_- - \zeta_-) > (21/8)/(\rho_- - \zeta_-)$.

In independent work, Theorem 1 and Corollaries 1–3 of Loh and Wainwright (2013) show that the approximate local solution $\widetilde{\boldsymbol{\beta}}$ attained by their optimization

procedure satisfies $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C\lambda_{\mathrm{tgt}}\sqrt{s^*}$. A comparison between Theorem 4.7 and their result suggests that our approximate local solution $\widetilde{\boldsymbol{\beta}}_N$ obtained within the final path-following stage has the same statistical rate of convergence as the approximate local solution attained by their procedure. Meanwhile, Theorem 4.7 provides additional statistical characterizations for the other regularization parameters along the regularization path, that is, $\lambda_1, \ldots, \lambda_{N-1}$.

In the next theorem, we provide a refined statistical rate of convergence. Recall within the tth path-following stage, the iterative sequence $\{\boldsymbol{\beta}_t^k\}_{k=0}^{\infty}$ produced by the proximal-gradient method converges toward a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. The next theorem states that $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ benefits from nonconvex regularization and possesses an improved statistical rate of convergence.

THEOREM 4.8 (Refined statistical rates of convergence of exact local solutions). *For the regularization parameter λ_t, we assume that the nonconvex penalty $\mathcal{P}_{\lambda_t}(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_{\lambda_t}(\beta_j)$ satisfies*

$$p'_{\lambda_t}(\beta_j) = 0 \qquad \text{for } |\beta_j| \geq \nu_t, \tag{4.16}$$

for some $\nu_t > 0$. Let $S_1^* \cup S_2^* = S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$ with $|S_1^*| = s_1^*$, $|S_2^*| = s_2^*$ and $|S^*| = s^* = s_1^* + s_2^*$. For $j \in S_1^* \subseteq S^*$, we assume $|\beta_j^*| \geq \nu_t$, while for $j \in S_2^* \subseteq S^*$, we assume $|\beta_j^*| < \nu_t$. Under Assumptions 4.1 and 4.4, we have

$$\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*\|_2 \leq \underbrace{C\|(\nabla\mathcal{L}(\boldsymbol{\beta}^*))_{S_1^*}\|_2}_{S_1^*:\ \text{Large } |\beta_j|\text{'s}} + \underbrace{C'\lambda_t\sqrt{s_2^*}}_{S_2^*:\ \text{Small } |\beta_j|\text{'s}} \qquad \text{for } t = 1, \ldots, N, \tag{4.17}$$

where $C = 1/(\rho_- - \zeta_-)$ and $C' = 3/(\rho_- - \zeta_-)$.

PROOF. See the next section for a detailed proof. □

In Theorem 4.8, the assumption in (4.16) applies to a variety of nonconvex penalty functions. For SCAD in (2.1), we have $\nu_t = a\lambda_t$; While for MCP in (2.2), we have $\nu_t = b\lambda_t$. Theorem 4.8 suggests that, for "small" coefficients such that $|\beta_j| < \nu_t$, the second part on the right-hand side of (4.17) has the same recovery performance as in Theorem 4.7, while for "large" coefficients such that $|\beta_j| \geq \nu_t$, the first part in (4.17) possesses a more refined rate of convergence. To understand this, we consider an example with $\mathcal{L}(\boldsymbol{\beta})$ being least squares loss. We assume that $(Y|\mathbf{X} = \mathbf{x}_i)$ follows a sub-Gaussian distribution with mean $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and variance proxy σ^2. Moreover, we assume that the columns of $\mathbf{X}$ are normalized in such a way that $\max_{j\in\{1,\ldots,d\}}\{\|\mathbf{X}_j\|_2\} \leq \sqrt{n}$. Then we have

$$\|(\nabla\mathcal{L}(\boldsymbol{\beta}^*))_{S_1^*}\|_2 \leq C\sigma\sqrt{\frac{s_1^*}{n}} \tag{4.18}$$

with high probability. Clearly, this $\sqrt{s_1^*/n}$ rate of convergence on the right-hand side of (4.18) is significantly faster than the usual $\sqrt{s^* \log d/n}$ rate, since it gets rid of the $\log d$ term, and $s_1^* \le s^*$. In fact, ν_t is the minimum signal strength above which we are able to obtain this refined rate of convergence. In the examples of SCAD and MCP, we have $\nu_t = C\lambda_t$. Recall that $\{\lambda_t\}_{t=0}^N$ is a decreasing sequence. Hence, we are able to achieve this more refined rate of convergence for smaller and smaller signal strength along the regularization path. Moreover, for $t = N$, the minimum signal strength $\nu_N = \lambda_N = \lambda_{\mathrm{tgt}} = C\sqrt{\log d/n}$. Hence the required minimum signal strength goes to zero as the sample size increases. Following a proof similar to Lemmas C.1 and C.2 in the supplementary material [Wang, Liu and Zhang (2014)], we can also obtain similar results for logistic loss and semiparametric elliptical design loss. This refined rate of convergence is sharper than the result in Theorem 4.7, which is also achievable via convex regularization, for example, the ℓ_1 penalty. Therefore, Theorem 4.8 clearly justifies the benefits of using nonconvex regularization. Moreover, in Section 6 we will show that our requirement on the minimum signal strength to achieve this refined rate of convergence is optimal and is a weaker requirement than the suboptimal requirements in Fan, Xue and Zou (2014), Wang, Kim and Li (2013).

In addition to the refined rate of convergence for parameter estimation in Theorem 4.8, in the next theorem we prove that the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ also recovers the support of $\boldsymbol{\beta}^*$. Before we present the next theorem, we introduce the definition of an oracle estimator, denoted by $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$. Recall that $S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$. The oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is defined as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{O}} = \mathop{\mathrm{argmin}}_{\substack{\mathrm{supp}(\boldsymbol{\beta}) \subseteq S^* \\ \boldsymbol{\beta} \in \Omega}} \mathcal{L}(\boldsymbol{\beta}), \tag{4.19}$$

where $\Omega = \mathbb{R}^d$ for least squares loss and semiparametric elliptical design loss, while $\Omega = B_2(R)$ for logistic loss with $R \ge \|\boldsymbol{\beta}^*\|_2$. In the next lemma, we show that $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is the unique global solution to the minimization problem in (4.19) even for nonconvex loss functions and has nice statistical properties.

LEMMA 4.9. *Under Assumption 4.4, the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is the unique global minimizer of (4.19). For $\mathcal{L}(\boldsymbol{\beta})$ being least squares loss, we assume that $(Y|\mathbf{X} = \mathbf{x}_i)$ follows a sub-Gaussian distribution with mean $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and variance proxy σ^2. Then the oracle estimator satisfies*

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}^*\|_\infty \le C\sigma\sqrt{2/\rho_-} \cdot \sqrt{\frac{\log s^*}{n}} \tag{4.20}$$

with high probability for some constant C.

PROOF. See the supplementary material [Wang, Liu and Zhang (2014)] for a detailed proof. □

Statistical recovery results similar to (4.20) also hold for logistic loss and semiparametric elliptical design loss under different conditions. These results are omitted here for simplicity. Lemma 4.9 suggests that, for a sufficiently large n and sufficient minimum signal strength, the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ exactly recovers the support of $\boldsymbol{\beta}^*$. More specifically, if the minimum signal strength satisfies $\min_{j\in S^*}|\beta_j^*| \geq 2\nu$ for $\nu > 0$, then with high probability

$$\min_{j\in S^*}|(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})_j| \geq \min_{j\in S^*}|\beta_j^*| - \|\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}^*\|_\infty \geq 2\nu - \sigma\sqrt{2/\rho_-}\cdot\sqrt{\frac{\log s^*}{n}},$$

which implies $\min_{j\in S^*}|(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})_j| \geq \nu > 0$ for n sufficiently large. Meanwhile, recall that $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) \subseteq S^*$ by definition. Hence we have $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) = S^*$.

The next theorem states that under the condition of sufficient minimum signal strength, $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is the oracle estimator, and exactly recovers the support of $\boldsymbol{\beta}^*$.

THEOREM 4.10 (Support recovery). *For the regularization parameter λ_t, suppose that the nonconvex penalty $\mathcal{P}_{\lambda_t}(\boldsymbol{\beta}) = \sum_{j=1}^d p_{\lambda_t}(\beta_j)$ satisfies (4.16) for some $\nu_t > 0$. For least squares loss, we assume that $(Y|\mathbf{X} = \mathbf{x}_i)$ follows a sub-Gaussian distribution with mean $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and variance proxy σ^2. Under Assumptions 4.1 and 4.4, if the minimum signal strength satisfies $\min_{j\in S^*}|\beta_j^*| \geq 2\nu_t$, then for n sufficiently large, $\widehat{\boldsymbol{\beta}}_{\lambda_t} = \widehat{\boldsymbol{\beta}}_{\mathrm{O}}$, and $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) = \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) = \mathrm{supp}(\boldsymbol{\beta}^*)$ with high probability.*

PROOF. See the next section for a detailed proof. □

Recall that the assumption in (4.16) applies to a variety of nonconvex penalties, including SCAD and MCP, for which we have $\nu_t = C\lambda_t$ with $C > 0$. Hence the minimum signal strength that is required to achieve the oracle estimator and exact support recovery actually shrinks with the decreasing sequence $\{\lambda_t\}_{t=0}^N$ along the regularization path. For least squares loss, we have $\nu_N = C\lambda_{\mathrm{tgt}} = C'\sqrt{\log d/n}$ for $t = N$. Hence within the final path-following stage, the required minimum signal strength goes to zero as sample size $n \to \infty$. Furthermore, such requirement on the minimum signal strength for achieving the oracle estimator is optimal; that is, no weaker requirement exists [Zhang and Zhang (2012)]. In Section 6 we will show that, for least squares loss, some recent works [Fan, Xue and Zou (2014), Wang, Kim and Li (2013)] require a stronger minimum signal strength to achieve the oracle estimator in the same setting of least squares regression. Similar results to Theorem 4.10 also hold for other loss functions, but under different conditions. They are omitted here for simplicity.

5. Proof of main results. In this section we present the proof sketch of the main results. The desired computational and statistical results rely on the strong convexity of the surrogate loss function $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$. For example, we need $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ to

be strongly convex to establish the geometric rate of convergence of the proximal-gradient method within each path-following stage. However, $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is nonconvex in general, since $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$, where $\mathcal{L}(\boldsymbol{\beta})$ is possibly nonconvex and $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is concave. In the following lemma, we prove that $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ is strongly convex for $\boldsymbol{\beta}$ on a sparse set. In a similar way, we establish the strong smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ on a sparse set.

Recall $\rho_- = \rho_-(\nabla^2\mathcal{L}, s^* + 2\widetilde{s})$ and $\rho_+ = \rho_+(\nabla^2\mathcal{L}, s^* + 2\widetilde{s})$ are the sparse eigenvalues specified in Assumption 4.4. As defined in regularity condition (a), $\zeta_-, \zeta_+ > 0$ are the concavity parameters of the nonconvex penalty, which satisfy (4.3) and (4.6).

LEMMA 5.1. *Let $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d$ be two sparse vectors, which satisfy $\|(\boldsymbol{\beta} - \boldsymbol{\beta}')_{\overline{S^*}}\|_0 \le 2\widetilde{s}$, where $\widetilde{s}$ is specified in Assumption 4.4 and $S^* = \operatorname{supp}(\boldsymbol{\beta}^*)$. For $\mathcal{L}(\boldsymbol{\beta})$ being logistic loss, we further assume $\|\boldsymbol{\beta}\|_2 \le R$ and $\|\boldsymbol{\beta}'\|_2 \le R$, where R is a constant specified in Definition 4.3. Then the surrogate loss function $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ satisfies the restricted strong convexity*

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') \ge \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2$$

and the restricted strong smoothness

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') \le \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_+ - \zeta_+}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2.$$

PROOF. See Section D.2 in the supplementary material [Wang, Liu and Zhang (2014)] for a detailed proof. □

A similar result has been discussed by Negahban et al. (2012). The main difference is that our constraint set, where $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is strongly convex/smooth, is a sparse subspace, while that of Negahban et al. (2012) is a cone.

Note that in Lemma 5.1, the strong convexity and smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ rely on the sparsity of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. Hence we need to establish results regarding the sparsity of $\boldsymbol{\beta}_t^k$ throughout the whole iterative procedure. In the sequel, we provide several important lemmas: Lemmas 5.2 and 5.3 characterize the statistical properties of any sparse $\boldsymbol{\beta}$; based on such statistical properties, Lemma 5.4 proves that, any proximal-gradient update iteration with a sparse input produces a sparse output. Equipped with these lemmas, we can establish the sparsity of the solution path by mathematical induction in Theorem 5.5.

The next lemma provides a characterization of any sparse $\boldsymbol{\beta}$ with certain suboptimality.

LEMMA 5.2. *We assume that $\boldsymbol{\beta}$ satisfies*

$$\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \le \widetilde{s}, \qquad \omega_\lambda(\boldsymbol{\beta}) \le \lambda/2 \tag{5.1}$$

with $\lambda \geq \lambda_{\mathrm{tgt}}$, *where* $\omega_\lambda(\boldsymbol{\beta})$ *is the measure of suboptimality defined in* (3.15). *For logistic loss, we assume* $\|\boldsymbol{\beta}\|_2 \leq R$, *where* $R > 0$ *is a constant specified in Definition* 4.3. *Under Assumptions* 4.1 *and* 4.4, $\boldsymbol{\beta}$ *satisfies*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C\lambda\sqrt{s^*} \qquad \text{where } C = \frac{21/8}{\rho_- - \zeta_-}.$$

Meanwhile, the objective function value evaluated at $\boldsymbol{\beta}$ *satisfies*

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq C'\lambda^2 s^* \qquad \text{where } C' = \frac{21/2}{\rho_- - \zeta_-}.$$

PROOF. See Section D.3 of the supplementary material [Wang, Liu and Zhang (2014)] for a detailed proof. □

Recall that we use the approximate local solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ obtained within the $(t-1)$th path-following stage to be the initialization of the tth stage (line 8 of Algorithm 1), that is, $\boldsymbol{\beta}_t^0 = \widetilde{\boldsymbol{\beta}}_{t-1}$. By setting $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_{t-1} = \boldsymbol{\beta}_t^0$ and $\lambda = \lambda_t$ in Lemma 5.2, we can see that if $\widetilde{\boldsymbol{\beta}}_{t-1}$ is sparse and $(\lambda_t/2)$-suboptimal, then the initial point $\boldsymbol{\beta}_t^0$ of the tth stage has nice statistical recovery performance. However, it is unclear whether the rest of $\boldsymbol{\beta}_t^k$'s $(k = 1, 2, \ldots)$ within the tth stage also have similar recovery performance. To prove this, we first present Lemma 5.3, which shows that under the condition that $\boldsymbol{\beta}$ is sparse and $\phi_\lambda(\boldsymbol{\beta})$ is close to $\phi_\lambda(\boldsymbol{\beta}^*)$, $\boldsymbol{\beta}$ has desired statistical properties. After Lemma 5.3, we will explain that if $\boldsymbol{\beta}_t^0$ satisfies this condition, then all the $\boldsymbol{\beta}_t^k$'s $(k = 1, 2, \ldots)$ within the same path-following stage also satisfy this condition and therefore enjoys nice statistical properties.

LEMMA 5.3. *Suppose that for* $\lambda \geq \lambda_{\mathrm{tgt}}$, $\boldsymbol{\beta}$ *satisfies*

$$\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}, \qquad \phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq C\lambda^2 s^* \qquad \text{where } C = \frac{21/2}{\rho_- - \zeta_-}.$$

For logistic loss, we further assume $\|\boldsymbol{\beta}\|_2 \leq R$, *where* R *is a constant specified in Definition* 4.3. *Under Assumptions* 4.1 *and* 4.4, *we have*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C'\lambda\sqrt{s^*} \qquad \text{where } C' = \frac{15/2}{\rho_- - \zeta_-}.$$

PROOF. See Section D.4 of the supplementary material [Wang, Liu and Zhang (2014)] for a detailed proof. □

Let $\lambda = \lambda_t$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_t^k$ in Lemma 5.3. It suggests that within the tth path-following stage, all $\boldsymbol{\beta}_t^k$'s $(k = 1, 2, \ldots)$ have nice statistical recovery performance under three sufficient conditions: (i) each $\boldsymbol{\beta}_t^k$ is sparse; (ii) the objective function value $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k)$ is close to $\phi_{\lambda_t}(\boldsymbol{\beta}^*)$; (iii) for logistic loss, we further need $\|\boldsymbol{\beta}_t^k\|_2 \leq R$.

For condition (ii), recall that if we set $\boldsymbol{\beta} = \boldsymbol{\beta}_t^0$ and $\lambda = \lambda_t$ in Lemma 5.2, then $\boldsymbol{\beta}_t^0$ being sparse and $(\lambda_t/2)$-suboptimal implies that $\phi_{\lambda_t}(\boldsymbol{\beta}_t^0)$ is close to $\phi_{\lambda_t}(\boldsymbol{\beta}^*)$. Since the proximal-gradient method ensures the monotone decrease of $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^k)\}_{k=0}^{\infty}$ within the tth stage (see Lemma D.1 of the supplementary material [Wang, Liu and Zhang (2014)]), condition (ii) also holds. Meanwhile, condition (iii) obviously holds because of the ℓ_2 constraint. To establish the statistical recovery performance of all the $\boldsymbol{\beta}_t^k$'s within the tth stage, we still need to establish the sparsity of $\boldsymbol{\beta}_t^k$'s to guarantee that condition (i) holds. To prove this, we present Lemma 5.4, which states that if $\boldsymbol{\beta}$ is sparse, then a proximal-gradient update operation (3.9) on $\boldsymbol{\beta}$ produces a sparse solution.

LEMMA 5.4. *Suppose that, for* $\lambda \geq \lambda_{\mathrm{tgt}}$, $\boldsymbol{\beta}$ *satisfies*

$$\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}, \qquad \phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq C\lambda^2 s^* \quad \text{and} \quad L < 2(\rho_+ - \zeta_+),$$

where $C = (21/2)/(\rho_- - \zeta_-)$. *For logistic loss, we assume* $\|\boldsymbol{\beta}\|_2 \leq R$, *where* R *is specified in Definition 4.3. Under Assumptions 4.1 and 4.4, the proximal-gradient update operation defined in* (3.9) *produces a sparse solution, that is,*

$$\|(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R))_{\overline{S^*}}\|_0 \leq \widetilde{s}.$$

Here we set $R = +\infty$ *if the domain* Ω *in* (3.8) *is* $\mathbb{R}^d$.

PROOF. See Section D.4 of the supplementary material [Wang, Liu and Zhang (2014)] for a detailed proof. □

For $\boldsymbol{\beta} = \boldsymbol{\beta}_t^{k-1}$, $\lambda = \lambda_t$ and $L = L_t^k$, Lemma 5.4 states that if $\boldsymbol{\beta}_t^{k-1}$ is sparse and the objective function value $\phi_{\lambda_t}(\boldsymbol{\beta}_t^{k-1})$ is close to $\phi_{\lambda_t}(\boldsymbol{\beta}^*)$, then $\boldsymbol{\beta}_t^k = \mathcal{T}_{L_t^k,\lambda_t}(\boldsymbol{\beta}_t^{k-1}; R)$ produced by the proximal-gradient update step (3.8) is also sparse. Within the tth path-following stage, if $\boldsymbol{\beta}_t^0$ is sparse, $\omega_{\lambda_t}(\boldsymbol{\beta}_t^0) \leq \lambda_t/2$, and for logistic loss $\|\boldsymbol{\beta}_t^0\|_2 \leq R$, then by Lemma 5.2 we have

$$\phi_{\lambda_t}(\boldsymbol{\beta}_t^0) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*.$$

Since $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^k)\}_{k=0}^{\infty}$ decreases monotonically, we have

$$\phi_{\lambda_t}(\boldsymbol{\beta}_t^k) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \phi_{\lambda_t}(\boldsymbol{\beta}_t^0) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^* \qquad \text{for } k = 1, 2, \ldots.$$

Assume that we have $L_t^k \leq 2(\rho_+ - \zeta_+)$ (which will be proved in Theorem 5.5). Applying Lemma 5.4 recursively, we obtain $\|(\boldsymbol{\beta}_t^k)_{\overline{S^*}}\|_0 \leq \widetilde{s}$ $(k = 1, 2, \ldots)$. Meanwhile, we have $\|\boldsymbol{\beta}_t^k\|_2 \leq R$ due to the ℓ_2 constraint. Then according to Lemma 5.3,

all $\boldsymbol{\beta}_t^k$'s within the tth path-following stage have nice recovery performance, that is,

$$\|\boldsymbol{\beta}_t^k - \boldsymbol{\beta}^*\|_2 \le \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*} \qquad \text{for } k = 1, 2, \ldots.$$

Furthermore, by Lemma 5.1 the sparsity of $\boldsymbol{\beta}_t^k$'s implies the restricted strong convexity and smoothness of $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$, which enable us to establish the geometric rate of convergence within the tth path-following stage. These results are formally presented in Theorem 5.5.

THEOREM 5.5. *Suppose within the tth path-following stage, the proximal-gradient method in Algorithm 3 is initialized by $\boldsymbol{\beta}_t^0$ and L_t^0, which satisfy*

$$\|(\boldsymbol{\beta}_t^0)_{\overline{S^*}}\|_0 \le \widetilde{s}, \qquad \omega_{\lambda_t}(\boldsymbol{\beta}_t^0) \le \lambda_t/2 \quad \text{and} \quad L_t^0 \le 2(\rho_+ - \zeta_+).$$

For logistic loss we further assume $\|\boldsymbol{\beta}_t^0\|_2 \le R$ with R specified in Definition 4.3. Then we have the following results:

- *For $k = 1, 2, \ldots$, we have*

$$\|(\boldsymbol{\beta}_t^k)_{\overline{S^*}}\|_0 \le \widetilde{s}, \qquad \|\boldsymbol{\beta}_t^k - \boldsymbol{\beta}^*\|_2 \le \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*}, \qquad L_t^k \le 2(\rho_+ - \zeta_+). \tag{5.2}$$

- *The iterative sequence $\{\boldsymbol{\beta}_t^k\}_{k=0}^{\infty}$ converges toward a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$, which satisfies $\|(\widehat{\boldsymbol{\beta}}_{\lambda_t})_{\overline{S^*}}\|_0 \le \widetilde{s}$ and the exact optimality condition that $\omega_{\lambda_t}(\boldsymbol{\beta}_t^k) \le 0$.*
- *To achieve an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ that satisfies $\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t) \le \lambda_t/4$, we need no more than $C'\log(4C\sqrt{s^*})$ proximal-gradient iterations defined in lines 5–9 of Algorithm 3. Here*

$$C = 2\sqrt{21}\cdot\sqrt{\kappa}(1+\kappa), \qquad C' = 2/\log\biggl(\frac{1}{1 - 1/(8\kappa)}\biggr). \tag{5.3}$$

- *To obtain an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ such that $\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t) \le \varepsilon_{\mathrm{opt}}$, we need no more than $C'\log(C\lambda_t\sqrt{s^*}/\varepsilon_{\mathrm{opt}})$ proximal-gradient iterations. Here C and C' are defined in (5.3).*

PROOF. See Section D.6 of the supplementary material [Wang, Liu and Zhang (2014)] for a detailed proof. □

To prove the geometric rate of convergence and desired statistical recovery results hold within all path-following stages, that is, $t = 0, \ldots, N$, we need to verify that the conditions of Theorem 5.5 hold within each stage. We prove by induction. We assume the initialization of $(t-1)$th path-following stage satisfies

$$\|(\boldsymbol{\beta}_{t-1}^0)_{\overline{S^*}}\|_0 \le \widetilde{s}, \qquad \omega_{\lambda}(\boldsymbol{\beta}_{t-1}^0) \le \lambda_t/2 \quad \text{and} \quad L_{t-1}^0 \le 2(\rho_+ - \zeta_+). \tag{5.4}$$

Applying Theorem 5.5, we obtain

$$\|(\boldsymbol{\beta}_{t-1}^k)_{\overline{S^*}}\|_0 \le \widetilde{s}, \qquad L_{t-1}^k \le 2(\rho_+ - \zeta_+) \qquad \text{for } k = 1, 2, \ldots.$$

Consequently, the approximate solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ produced by the $(t-1)$th stage satisfies $\|(\widetilde{\boldsymbol{\beta}}_{t-1})_{\overline{S^*}}\|_0 \le \widetilde{s}$, while L_{t-1} satisfies $L_{t-1} \le 2(\rho_+ - \zeta_+)$. Since we warm start the tth path-following stage with $\boldsymbol{\beta}_t^0 = \widetilde{\boldsymbol{\beta}}_{t-1}$ and $L_t^0 = L_{t-1}$ (line 8 of Algorithm 1), we have

$$\|(\boldsymbol{\beta}_t^0)_{\overline{S^*}}\|_0 \le \widetilde{s}, \qquad L_t^0 \le 2(\rho_+ - \zeta_+). \tag{5.5}$$

Moreover, note that the stopping criterion of the proximal-gradient method ensures $\omega_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) \le \lambda_{t-1}/4$ (line 9 of Algorithm 3), which implies $\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) \le \lambda_t/2$ according to Lemma D.4 of the supplementary material [Wang, Liu and Zhang (2014)]. Thus we have

$$\omega_{\lambda_t}(\boldsymbol{\beta}_t^0) \le \lambda_t/2. \tag{5.6}$$

Therefore, we know that (5.4) implies (5.5) and (5.6). We will verify (5.5) and (5.6) hold for $t = 0$ in the proof of Theorem 4.5 in the supplementary material [Wang, Liu and Zhang (2014)]. By induction, we have that (5.5) and (5.6) hold for $t = 0, \ldots, N$. As a consequence of Theorem 5.5, all path-following stages have geometric rates of convergence along the solution path, which implies the global geometric rate of convergence in Theorem 4.5. See the supplementary material [Wang, Liu and Zhang (2014)] for a detail proof. Meanwhile, all $\boldsymbol{\beta}_t^k$'s have desired statistical properties, that is,

$$\|\boldsymbol{\beta}_t^k - \boldsymbol{\beta}^*\|_2 \le \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*} \qquad \text{for } t = 1, \ldots, N \text{ and } k = 0, 1, \ldots,$$

which leads to the statistical rates of convergence of the approximate local solutions $\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N$ in Theorem 4.7, the more refined rates of convergence of the exact local solutions $\{\widehat{\boldsymbol{\beta}}_{\lambda_t}\}_{t=1}^N$ in Theorem 4.8 and the support recovery results in Theorem 4.10. See Sections D.8–D.10 of the supplementary material [Wang, Liu and Zhang (2014)] for detailed proofs, respectively.

6. Discussion. Our work is related to recent works on understanding nonconvex regularization in the context of least squares regression. Zhang (2010a) proposed an MC+ procedure for MCP penalized least squares regression. However, the computation of MC+ might be inefficient because there can be exponentially many switching points on its solution path. To remedy this issue, Zhang (2010b, 2013) proposed the multi-stage convex relaxation method, which iteratively solves

$$\widehat{\boldsymbol{\beta}}^k \leftarrow \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\mathcal{L}(\boldsymbol{\beta}) + \sum_{j=1}^d p'_\lambda(|\widehat{\beta}_j^{k-1}|)|\beta_j|\right\}, \qquad k = 1, 2, \ldots, \tag{6.1}$$

where $p_\lambda(\beta_j)$ is defined in Section 2, and the initialization $\widehat{\boldsymbol{\beta}}^0$ is set to be the Lasso estimator corresponding to λ. For k sufficiently large, $\widehat{\boldsymbol{\beta}}^k$ has the same oracle properties as in Theorems 4.8 and 4.10. However, for each k we need to solve the minimization problem in (6.1) exactly, which is not realistic in practice, since practical optimization methods only attain finite numerical precision in finite iterations. In contrast, we provide simultaneous statistical and computational analysis by explicitly taking the numerical precision into account, and establish the global geometric rate of convergence in terms of iteration complexity for calculating the full regularization path.

This multi-stage convex relaxation method was previously referred to as local linear approximation (LLA), and was analyzed on fixed dimensional models by Zou and Li (2008). Fan, Xue and Zou (2014) recently provided nonasymptotic analysis of LLA and proved that LLA finds the oracle estimator in two iterations. However, their results rely on that the Lasso initialization satisfies $\|\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}^*\|_\infty \le C\lambda$ with high probability, which requires λ to take the value of $C'\sqrt{s^* \log d/n}$. Consequently, their requirement on the minimum signal strength is of the order of $\sqrt{s^* \log d/n}$, which is suboptimal. In contrast, we only require a minimum signal strength of the order of $\sqrt{\log d/n}$, which is optimal [Zhang and Zhang (2012)]. Also, they did not analyze the iteration complexity for computing each step of LLA, that is, solving (6.1).

Very recently, Wang, Kim and Li (2013) considered a two-step approach similar to the two-step LLA procedure, named the calibrated CCCP. It differs from the two-step LLA in that its Lasso initialization $\widehat{\boldsymbol{\beta}}^0$ is obtained using the regularization parameter $\tau\lambda$, where $\tau = o(1)$ and $\lambda = \sqrt{\log d/n}$. It attains the oracle estimator under the restricted eigenvalue (RE) condition [Bickel, Ritov and Tsybakov (2009)], but requires the minimum signal strength to be larger than $Cs^*\sqrt{\log d/n}$. Under a stronger assumption than the RE condition, namely the relaxed sparse Riesz condition, a minimum signal strength of the order of $\sqrt{\log d/n}/\tau$ is required. Such requirement is still suboptimal, but is close to the optimal scaling of $\sqrt{\log d/n}$ in our results, since τ can take $1/\log n$. They proposed a novel high-dimensional BIC criterion, which can be used to choose the best λ_{tgt} in our procedure. Also, they provided extensions to logistic regression.

The iterative hard thresholding (IHT) algorithm [Blumensath and Davies (2009)] can also achieve a local solution with desired statistical recovery performance at a global geometric rate of convergence. However, the theoretical results of IHT are not directly comparable with ours because of the usage of different noise models. If we have to cast the theoretical results of IHT into our model, their results are much weaker than ours. In detail, IHT attains an approximate local solution $\widetilde{\boldsymbol{\beta}}$, which satisfies

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le 6\|\mathbf{e}\|_2 \tag{6.2}$$

with high probability. Here $\mathbf{e} \in \mathbb{R}^n$ is the noise vector in their setting, which is often considered to be perturbation noise. Note that a proper normalization gives

$\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)/\sqrt{n}$, where $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*$ is considered to be the sub-Gaussian noise with zero mean and variance proxy σ^2 in our setting. Then (6.2) gives

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le 6\|\mathbf{e}\|_2 = 6\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2/\sqrt{n} \le 6 \cdot 4\sigma\sqrt{n}/\sqrt{n} = 24\sigma \tag{6.3}$$

with high probability. Note that the upper bound on the right-hand side of (6.3) does not depend on s^* and d, and fails to converge to zero as $n \to \infty$. In summary, casting the results of IHT into our setting of sub-Gaussian noise yields a rather weak result. Also, IHT requires prior knowledge on the true sparsity level s^* to achieve fast global convergence, while our method does not.

In addition, the difference between our work and the independent work by Loh and Wainwright (2013) has been discussed in Section 1 and Section 4 with details.

7. Numerical results. We provide numerical results illustrating the computational efficiency and statistical accuracy of the proposed method. In detail, first we illustrate the effectiveness of our method on a problem with both nonconvex loss and penalty functions. Then we conduct comparison between our method and existing nonconvex procedures.

In the first experiment, we consider $\mathcal{L}(\boldsymbol{\beta})$ to be semiparametric elliptical random design loss defined in (2.6) and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ to be the MCP penalty defined in (2.2). We test on a synthetic dataset with $n = 500$ samples and $d = 2500$ dimensions. See the supplementary material [Wang, Liu and Zhang (2014)] for the detailed settings.

As shown in Figure 3(a), the objective function value $\phi_\lambda(\boldsymbol{\beta}_t^k)$ is monotone decreasing along the regularization path, as characterized by our theory (see Lemma D.1 of the supplementary material [Wang, Liu and Zhang (2014)]) and converges eventually.

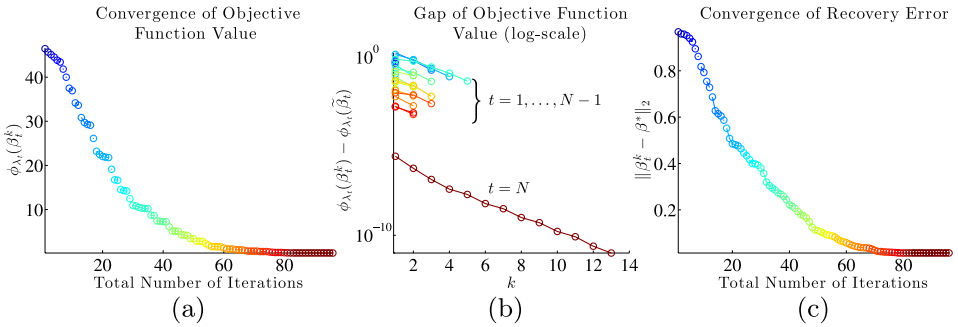

FIG. 3. *Semiparametric elliptical design regression with MCP*: (a) *plot of the objective function value* $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k)$ *along the regularization path*; (b) *plot of* $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k) - \phi_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t)$ *(log-scale) within each path-following stage*; (c) *plot of the recovery error* $\|\boldsymbol{\beta}_t^k - \boldsymbol{\beta}^*\|_2$. *Here we illustrate each path-following stage* $(t = 1, \ldots, N)$ *with a different color. Note that each point in the figure denotes* $\boldsymbol{\beta}_t^k$, *which corresponds to the* k*th iteration of the proximal-gradient method (Algorithm 3) within the* t*th path-following stage.*

Figure 3(b) illustrates the geometric rate of convergence within each path-following stage. In detail, each line denotes a path-following stage. It shows the objective function value gap, that is, $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k) - \phi_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t)$ decays exponentially with k within each stage. Note that

$$\begin{aligned}\phi_{\lambda_t}(\boldsymbol{\beta}_t^{k-1}) - \phi_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t) &\geq \phi_{\lambda_t}(\boldsymbol{\beta}_t^{k-1}) - \phi_{\lambda_t}(\boldsymbol{\beta}_t^k) \\ &\geq \frac{L_t^k}{2}\|\boldsymbol{\beta}_t^k - \boldsymbol{\beta}_t^{k-1}\|_2^2 \geq \frac{L_t^k}{2}\frac{\omega_{\lambda_t}^2(\boldsymbol{\beta}_t^k)}{(L_t^k + \rho_+ - \zeta_-)^2}.\end{aligned} \tag{7.1}$$

Here the first inequality is because the objective function $\phi_{\lambda_t}(\boldsymbol{\beta}_t^k)$ is monotone decreasing, while the second and third inequalities follow from Lemmas D.1 and D.2 of the supplementary material [Wang, Liu and Zhang (2014)], respectively. Therefore, $\omega_{\lambda_t}(\boldsymbol{\beta}_t^k)$ also decays exponentially within each stage, which implies that we only need a logarithmic number of iterations to attain the desired approximate local solution within each path-following stage, as characterized by Theorem 4.5.

Figure 3(b) illustrates the success of the path-following scheme in Figure 1: The $k = 1$ point on each line denotes the initialization of the corresponding path-following stage, for example, the tth stage. Recall that such initialization is set to be the approximate local solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ obtained within the $(t-1)$th stage, which falls into the region of optimization precision in Figure 1. Meanwhile, the fast convergence within the tth stage suggests that $\widetilde{\boldsymbol{\beta}}_{t-1}$ also falls into the region of fast convergence in Figure 1. Thus the path-following scheme works exactly as we have described in Figure 1 empirically.

Figure 3(c) shows that the ℓ_2 recovery error decays toward a small value as the optimization method proceeds, which implies that the attained approximate local solution has desired statistical properties, as predicted by Theorem 4.7.

In the second experiment, we compare our method with several existing nonconvex procedures on statistical performance, including LLA [Zou and Li (2008)], the calibrated CCCP [Wang, Kim and Li (2013)], SparseNet [Mazumder, Friedman and Hastie (2011)] and the multi-stage convex relaxation method [Zhang (2010b, 2013)]. We consider an example of least squares regression with MCP, where $n = 200$, $d = 2000$ and $\|\boldsymbol{\beta}^*\|_0 = 10$. See the supplementary material [Wang, Liu and Zhang (2014)] for the detailed settings.

We compare the support recovery performance and ℓ_2 recovery error of the estimators obtained from these procedures in Table 1, where we use the Lasso estimator and the oracle estimator defined in (4.19) as references. For support recovery, we are interested in the cardinality of true positive sets (TPS) and false positive set (FPS), both of which are defined in Table 1.

Ideally, the cardinality of TPS should be as large as $\|\boldsymbol{\beta}^*\|_0$ (which is 10 in this example), since a good procedure should exactly identify $S^* = \operatorname{supp}(\boldsymbol{\beta}^*)$. Meanwhile, the cardinality of FPS should be close to zero; that is, few of the coordinates in $\overline{S^*}$ are wrongly identified as nonzero. Table 1 shows that all nonconvex procedures significantly outperform the Lasso, which produces a less sparse estimator

TABLE 1

Comparing statistical performance of nonconvex procedures: TPS/FPS denote the true/false positive sets, which are defined as $\{j \in S^* : \widehat{\beta}_j \neq 0\}$ *and* $\{j \in \overline{S^*} : \widehat{\beta}_j \neq 0\}$, *respectively, and* $|\cdot|$ *denotes their cardinality. The* ℓ_2 *recovery error is defined as* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, *where* $\widehat{\boldsymbol{\beta}}$ *is the estimator. Standard deviations are present in the parentheses*

Method	**|TPS|**	**|FPS|**	$\boldsymbol{\ell_2}$ **error**
Approximate path following	**10** (0)	**0.180** (0.0411)	**0.702** (0.0278)
SparseNet	10 (0)	0.950 (0.108)	0.848 (0.0230)
Multi-stage convex relaxation	10 (0)	2.21 (0.146)	1.28 (0.0753)
LLA	10 (0)	2.98 (0.304)	1.28 (0.0996)
Calibrated CCCP	9.99 (0.01)	3.28 (0.308)	1.40 (0.122)
Lasso	9.98 (0.0141)	31.15 (0.799)	2.63 (0.0460)
Oracle estimator	10 (0)	0 (0)	0.484 (0.0221)

with larger ℓ_2 recovery error. In this specific example, our method outperforms the existing nonconvex procedures. Moreover, our method almost recovers S^* exactly and achieves a small ℓ_2 recovery error that is very close to the ℓ_2 error of the oracle estimator, as characterized by Theorem 4.8.

8. Conclusion. In this paper, we provide a unified theory for penalized M-estimators with possibly nonconvex loss and penalty functions. These problems are motivated by generalized linear models with nonconvex penalties and semiparametric elliptical design regression, as well as a broad range of other applications. Because it is intractable to compute the global solutions of these problems due to the nonconvex formulation, we need to establish theory that characterizes both the computational and statistical properties of the local solutions obtained by specific algorithms. For this purpose, we propose an approximate regularization path-following method, which serves as a unified framework for solving a variety of high-dimensional sparse learning problems with nonconvexity. Computationally, our method enjoys a fast global geometric rate of convergence for calculating the entire regularization path; statistically, all the approximate and exact local solutions attained by our method along the regularization path possess sharp statistical rate of convergence in both estimation and support recovery. In particular, we provide sharp theoretical analysis that demonstrates the advantage of using nonconvex penalties. This paper shows that, under suitable conditions, we can efficiently obtain the entire regularization path of a broad class of nonconvex sparse learning problems.

Our work can be extended in many directions: Our method and theory for least squares loss and logistic loss can be easily extended to other generalized linear models (see Section C.2 of the supplementary material [Wang, Liu and Zhang (2014)] for details); for inverse covariance matrix estimation, our work is directly applicable to the sparse column inverse operator (SCIO) [Liu and Luo (2012)];

meanwhile, it might need more effort than verifying Assumptions 4.1 and 4.4 to adapt the graphical Lasso into our framework; for example, the optimization algorithm also has to be modified to enforce the positive semidefinite constraint; it is also interesting to consider other loss functions, for example, quantile regression [Wang, Wu and Li (2012)], for which Assumption 4.4 may no longer hold.

Acknowledgments. We sincerely thank Po-Ling Loh, Martin Wainwright and Yiyuan She for their helpful personal communications. We are grateful to the Editor, Associate Editor and referees for their insightful comments.

SUPPLEMENTARY MATERIAL

Optimal computational and statistical rates of convergence for sparse nonconvex learning problems (DOI: 10.1214/14-AOS1238SUPP; .pdf). We provide the detailed proof in the supplement [Wang, Liu and Zhang (2014)].

REFERENCES

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

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

BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726

BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396

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

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

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

FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

HASTIE, T., ROSSET, S., TIBSHIRANI, R. and ZHU, J. (2004). The entire regularization path for the support vector machine. *J. Mach. Learn. Res.* **5** 1391–1415. MR2248021

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

KIM, Y., CHOI, H. and OH, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103** 1665–1673. MR2510294

KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** 7–57. MR2500227

LIU, W. and LUO, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. Preprint. Available at arXiv:1203.3896.

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Preprint. Available at arXiv:1305.2436.

MAIRAL, J. and YU, B. (2012). Complexity analysis of the lasso regularization path. Preprint. Available at arXiv:1205.0079.

MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769

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

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization* **87**. Springer, New York.

NESTEROV, YU. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. MR3071865

PARK, M. Y. and HASTIE, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 659–677. MR2370074

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

ROSSET, S. and ZHU, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012–1030. MR2341696

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.* **3** 384–415. MR2501318

SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comput. Statist. Data Anal.* **56** 2976–2990. MR2929353

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation* **45**. Cambridge Univ. Press, Cambridge.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

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

WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. MR3127873

WANG, Z., LIU, H. and ZHANG, T. (2014). Supplement to "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems." DOI:10.1214/14-AOS1238SUPP.

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222. MR2949353

WRIGHT, S. J., NOWAK, R. D. and FIGUEIREDO, M. A. T. (2009). Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.* **57** 2479–2493. MR2650165

XIAO, L. and ZHANG, T. (2013). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM J. Optim.* **23** 1062–1091. MR3057332

ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109–2144. MR2543687

ZHANG, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

ZHANG, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli* **19** 2277–2293. MR3160554

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

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

ZHAO, P. and YU, B. (2007). Stagewise lasso. *J. Mach. Learn. Res.* **8** 2701–2726. MR2383572

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

Z. WANG
H. LIU
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: zhaoran@princeton.edu
hanliu@princeton.edu

T. ZHANG
DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: tzhang@stat.rutgers.edu