Sparse Nonlinear Regression: Parameter Estimation under Nonconvexity

Zhuoran Yang ZY6@PRINCETON.EDU
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

Zhaoran Wang ZHAORAN@PRINCETON.EDU
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

Han Liu HANLIU@PRINCETON.EDU
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

Yonina C. Eldar YONINA@EE.TECHNION.AC.IL
Department of EE Technion, Israel Institute of Technology, Haifa 32000, Israel

Tong Zhang TZHANG@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

Abstract

We study parameter estimation for sparse nonlinear regression. More specifically, we assume the data are given by $y = f(\mathbf{x}^\top \boldsymbol{\beta}^*) + \epsilon$, where f is nonlinear. To recover $\boldsymbol{\beta}^*$, we propose an ℓ_1-regularized least-squares estimator. Unlike classical linear regression, the corresponding optimization problem is nonconvex because of the nonlinearity of f. In spite of the nonconvexity, we prove that under mild conditions, every stationary point of the objective enjoys an optimal statistical rate of convergence. Detailed numerical results are provided to back up our theory.

1. Introduction

We study a family of sparse nonlinear regression models. Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^\top \in \mathbb{R}^d$ be the sparse parameter vector of interest. We consider the model

$$y = f(\mathbf{x}^\top \boldsymbol{\beta}^*) + \epsilon, \tag{1.1}$$

where $y \in \mathbb{R}$ is a response variable, $\mathbf{x} \in \mathbb{R}^d$ is the covariate and $\epsilon \in \mathbb{R}$ is the exogenous noise. When f is the identity function, model (1.1) reduces to the well studied linear model. Given independent and identically distributed observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$, our goal is to estimate $\boldsymbol{\beta}^*$ even when $d \gg n$.

We can view (1.1) as a perceptron with noise, which is the basic building block of a feed forward neural network [41]. Establishing the theoretical guarantees of the estimation in (1.1) may provide insight on more complicated neural networks. Our model is also inspired by the nonlinear sparse recovery problems [2, 5, 7] which aim to recover a sparse parameter from a nonlinear system.

1.1. Main Results

Assuming f is monotonic, a straightforward way to estimate $\boldsymbol{\beta}^*$ is to solve a sparse linear regression problem [18] using the transformed data $\{f^{-1}(y_i), \mathbf{x}_i\}_{i=1}^n$. However, this approach works well only in the noiseless case with $\epsilon = 0$. Otherwise, it results in inaccurate parameter estimation and high prediction error due to the inverse operation. In this paper, we propose estimating the parameter $\boldsymbol{\beta}^*$ by solving the following ℓ_1-regularized least-squares problem:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\text{minimize}} \, \frac{1}{n} \sum_{i=1}^n \big[y_i - f(\mathbf{x}_i^\top \boldsymbol{\beta})\big]^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{1.2}$$

where λ is a regularization parameter and $\|\cdot\|_1$ is the vector ℓ_1-norm. Unlike the linear model for which (1.2) is a convex optimization problem, in general settings (1.2) could be highly nonconvex due to the nonlinearity of f, which prevents us from obtaining the global optimum. The existence of f also prevents us from having the restricted strongly convex property of the loss function.

In spite of the challenge of nonconvexity, we prove that any stationary point $\widehat{\boldsymbol{\beta}}$ of (1.2) enjoys optimal statistical rates of convergence under suitable conditions, i.e., with high probability

$$\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\big\|_2 \le C_1 \cdot \sqrt{s^* \log d / n} \text{ and}$$
$$\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\big\|_1 \le C_2 \cdot s^* \sqrt{\log d / n},$$

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

where s^* is the number of nonzero entries of $\boldsymbol{\beta}^*$ and C_1, C_2 are some absolute constants which do not depend on n, d or s^*. The statistical rates of convergence cannot be improved even when f is the identity function. In addition, we require a scaling of $n = O(s^* \log d)$ samples to obtain a vanishing error, which is also needed for linear sparse recovery problems [18]. Next, we provide an efficient gradient-based algorithm that provably converges to a stationary point. Our method is iterative and consist of soft-thresholding after a gradient descent step. This approach can be viewed as a generalization of the ISTA algorithm [6] to the nonlinear setting.

1.2. Related Work

The model in (1.1) is closely related to the single index model, which assumes $(y, \mathbf{x})$ satisfy $y = f(\mathbf{x}^\top \boldsymbol{\beta}^*) + \epsilon$ with an unknown f. The single index model is well studied in low dimensional settings where $d \ll n$. See, e.g., [16, 17, 21–23, 34, 44, 52, 53] and references therein. They mostly consider M-estimators that simultaneously estimate f and $\boldsymbol{\beta}^*$. However, these M-estimators are defined as the global optima of nonconvex minimization problems which are intractable to obtain. In high-dimensional settings where $\boldsymbol{\beta}^*$ is sparse, [3] establish PAC-Bayesian analysis for sparse single index models. [37, 38] propose marginal regression and generalized Lasso estimators which attain fast statistical rates of convergence. Nevertheless, the flexibility of the unknown link function f comes at a price. In detail, [37, 38] require $\mathbf{x}$ to be exactly Gaussian for their methods to succeed, even if f is known a priori. Also, unknown f raises identifiability issues, since the magnitude of $\boldsymbol{\beta}^*$ can be incorporated into f. As a result, these methods only estimate the direction of $\boldsymbol{\beta}^*$.

Another related line of work is sufficient dimension reduction, for which we aim to recover a subspace $\mathcal{U}$ such that y only depends on the projection of $\mathbf{x}$ onto $\mathcal{U}$. Both single index model and our problem can be viewed as special cases of the framework in which $\mathcal{U}$ is a one-dimensional subspace. See [14, 15, 28–30] and the references therein. Most works in this direction use spectral methods, which also rely on the Gaussian assumption and can only estimate the direction of $\boldsymbol{\beta}^*$. In comparison, we assume f is known. In this setting, we allow $\mathbf{x}$ to follow more general distributions and can directly estimate $\boldsymbol{\beta}^*$. [26, 27] propose iterative algorithms that alternatively estimate f and $\boldsymbol{\beta}^*$ based on the isotonic regression in the setting with $d \ll n$. However, their analysis focuses on generalization error instead of estimation error, which is the primary goal in this paper.

Our work is also related to problems of phase retrieval where the goal is to recover a signal $\boldsymbol{\beta}^* \in \mathbb{C}^d$ from the magnitude of its linear measurements contaminated by random noise. More specifically, the model of phase retrieval is given by $y = |\mathbf{x}^\top \boldsymbol{\beta}|^2 + \epsilon$. For high-dimensional settings, this problem is extensively studied under noisy or noiseless settings. See, e.g., [8, 11, 12, 19, 20, 24, 31, 35, 36, 42, 43, 47, 50]. These works show that a high dimensional signal can be accurately estimated up to global phase under restrictive assumptions on $\mathbf{x}$, e.g., $\mathbf{x}$ is Gaussian or certain classes of measurements. However, our work considers general measurements. Note that phase retrieval does not fall in the model under (1.1) because it uses a quadratic function, which is not monotonic. See §4 for a more detailed discussion.

1.3. Main Contribution

Our contribution is twofold. First, we propose an ℓ_1-regularized least-squares estimator for parameter estimation. We prove that every stationary point of the optimization problem in (1.2) converges to the true parameter, which explains the empirical success of regularized least-squares in the presence of nonlinear transforms. In the noiseless setting, as long as the number of samples is proportional to $s^* \log d$, we are able to exactly recover $\boldsymbol{\beta}^*$. To the best of our knowledge, this is the first parameter estimation result for the model (1.1) in high dimensional settings that does not rely on the normality of $\mathbf{x}$, and recovers both the magnitude and direction of $\boldsymbol{\beta}^*$. Our analysis for the stationary points of nonconvex optimization problems is of independent interest. Second, we establish the minimax rate of parameter estimation for the model (1.1), which establishes the minimax optimality of the stationary points of the proposed optimization problem in (1.2).

Organization of the rest of this paper In §2 we present our method for parameter estimation. We lay out the theory in §3. We discuss the connection to prior work with more details in §4. We corroborate our theoretical results with thorough numerical results in §5. In addition, we sketch the proof the statistical rates in §6. We conclude the paper in §7.

2. High-dimensional Estimation

In this section, we introduce the proposed methods for parameter estimation. In addition, we present the intuition behind our methods and compare our estimation procedures with the one that inverts the nonlinear function f directly.

Recall that we observe $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ satisfying $y_i = f(\mathbf{x}_i^\top \boldsymbol{\beta}^*) + \epsilon_i$. We assume the function f is monotonic and continuously differentiable. We define the least-square loss function as

$$L(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i^\top \boldsymbol{\beta})]^2. \tag{2.1}$$

We assume $\boldsymbol{\beta}^*$ is sparse and estimate it by solving the ℓ_1-regularized optimization problem in (1.2).

Due to the nonlinearity of f, $L(\boldsymbol{\beta})$ can be nonconvex. As a result, we can only find a stationary point $\widehat{\boldsymbol{\beta}}$ satisfying $\nabla L(\widehat{\boldsymbol{\beta}}) + \lambda \cdot \boldsymbol{\xi} = \mathbf{0}$, where $\boldsymbol{\xi} \in \partial \|\widehat{\boldsymbol{\beta}}\|_1$ and $\nabla L(\boldsymbol{\beta})$ is the gradient of $L(\boldsymbol{\beta})$. To obtain a stationary point, we apply the proximal gradient method, which generates an iterative sequence $\{\boldsymbol{\beta}^{(t)}, t \geq 0\}$ satisfying

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \big\{ \langle \nabla L(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)} \rangle + \alpha_t/2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \big\}, \quad (2.2)$$

where $1/\alpha_t > 0$ is the stepsize at the t-th iteration. In our setting, $\nabla L(\boldsymbol{\beta}^{(t)})$ is given by

$$\nabla L(\boldsymbol{\beta}^{(t)}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i^\top \boldsymbol{\beta}^{(t)})] f'(\mathbf{x}_i^\top \boldsymbol{\beta}^{(t)}) \mathbf{x}_i.$$

We denote $\mathbf{u}^{(t)}$ is given by $\mathbf{u}^{(t)} := \boldsymbol{\beta}^{(t)} - 1/\alpha_t \cdot \nabla L(\boldsymbol{\beta}^{(t)})$. then (2.2) has an explicit solution given by

$$\beta_i^{(t+1)} = \text{soft}(u_i^{(t)}, \lambda/\alpha_t) \quad \text{for } 1 \leq i \leq d, \quad (2.3)$$

where $\text{soft}(u, a) := \text{sign}(u) \max\{|u| - a, 0\}$ is the soft-thresholding operator.

The resulting algorithm is given in Algorithm 1, which is an application of the SpaRSA method proposed by [51] to our nonconvex problem. The main step is given in (2.3), which performs a soft-thresholding step on a gradient-descent update. This algorithm reduces to ISTA [6] when f is the identity. For nonlinear sparse recovery problems, this technique is also similar to the thresholded Wirtinger flow algorithm proposed for phase retrieval [8, 12].

To pick a suitable α_t, we use the line search procedure described in Algorithm 2. It iteratively increases α_t by a factor of η to ensure that $\boldsymbol{\beta}^{(t+1)}$ satisfies the acceptance criterion, which guarantees sufficient decrease of the objective function. To choose the initial α_t at the beginning of each line search iteration, we use the Barzilai-Borwein (BB) spectral method [4] in Algorithm 2, which guarantees that the initial value of each stepsize α_t lies in the interval $[\alpha_{\min}, \alpha_{\max}]$. Using the theory of [51], we establish the numerical convergence of the iterative sequence to a stationary point of (1.2) . However, it is challenging to establish the statistical properties of the stationary points. Our theory in §3 shows that, surprisingly, any stationary point enjoys satisfactory statistical guarantees. Consequently, Algorithm 1 yields a stationary point that is desired for parameter estimation.

When f is known, it seems tempting to apply linear compressed sensing procedures to the inverted data $\{z_i, \mathbf{x}_i\}$ where $z_i = f^{-1}(y_i)$. If f is linear, say $f(u) = au + b$, then $f^{-1}(u) = a^{-1}(u - b)$. In this case we have $z = f^{-1}(y) = \mathbf{x}^\top \boldsymbol{\beta}^* + a^{-1}\epsilon$, which is exactly a linear model. However, this method does not work well for general nonlinear f.

Algorithm 1 Proximal gradient algorithm for solving the ℓ_1-regularized problem in (1.2).

1: **Input:** regularization parameter $\lambda > 0$, update factor $\eta > 1$, constants $\zeta > 0$, $\alpha_{\min}, \alpha_{\max}$ with $0 < \alpha_{\min} < 1 < \alpha_{\max}$, integer $M > 0$, and $\phi(\boldsymbol{\beta}) := L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$
2: **Initialization:** set the iteration counter $t \leftarrow 0$ and choose $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^d$
3: **Repeat**
4: Choose stepsize α_t according to Algorithm 2
5: **Repeat**
6: $\mathbf{u}^{(t)} \leftarrow \boldsymbol{\beta}^{(t)} + \frac{1}{n\alpha_t} \cdot \sum_{i=1}^{n} \big[y_i - f\big(\mathbf{x}_i^\top \boldsymbol{\beta}^{(t)}\big)\big] f'\big(\mathbf{x}_i^\top \boldsymbol{\beta}^{(t)}\big) \mathbf{x}_i$.
7: $\beta_i^{(t+1)} \leftarrow \text{soft}(u_i^{(t)}, \lambda/\alpha_t)$ for $1 \leq i \leq d$.
8: $\alpha_t \leftarrow \eta \cdot \alpha_t$
9: **Until** $\boldsymbol{\beta}^{(t+1)}$ satisfies the acceptance criterion:
10: $\phi(\boldsymbol{\beta}^{(t+1)}) \leq \max\big\{\phi(\boldsymbol{\beta}^{(j)}) - \zeta \cdot \alpha_t/2 \cdot \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \colon \max(t - M, 0) \leq j \leq t\big\}$
11: Update the iteration counter $t \leftarrow t + 1$
12: **Until** $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2 / \|\boldsymbol{\beta}^{(t)}\|_2$ is sufficiently small
13: **Output:** $\widehat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(t)}$

Algorithm 2 The Barzilai-Borwein (BB) spectral approach for choosing α_t in Line 1 of Algorithm 1.

1: **Input:** the iteration counter t, $\boldsymbol{\delta}^{(t)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}$ and $\boldsymbol{g}^{(t)} = \nabla L(\boldsymbol{\beta}^{(t)}) - \nabla L(\boldsymbol{\beta}^{(t-1)})$
2: **if** $t = 0$ **then**
3: **Output:** $\alpha_t = 1$
4: **else**
5: **Output:** $\alpha_t = \langle \boldsymbol{\delta}^{(t)}, \boldsymbol{g}^{(t)} \rangle / \langle \boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t)} \rangle$ or $\alpha_t = \langle \boldsymbol{g}^{(t)}, \boldsymbol{g}^{(t)} \rangle / \langle \boldsymbol{\delta}^{(t)}, \boldsymbol{g}^{(t)} \rangle$
6: **end if**

To see this, denote $z = f^{-1}(y) = f^{-1}[f(\mathbf{x}^\top \boldsymbol{\beta}^*) + \epsilon]$ and $\mu = \mathbb{E}[z|\mathbf{x}] - \mathbf{x}^\top \boldsymbol{\beta}^*$. Then we have model

$$z = \mathbf{x}^\top \boldsymbol{\beta}^* + \mu + \xi, \quad (2.4)$$

where ξ is the remaining term that satisfies $\mathbb{E}[\xi|\mathbf{x}] = 0$. Note that both μ and ξ depend on $\boldsymbol{\beta}^*$ implicitly. When treating (2.4) as a sparse linear model with intercept, we discard such dependency and thus incur large estimation error. We numerically compare the proposed method with the linear approach that inverts f in §5 and show that our approach outperforms the linear framework.

3. Theoretical Results

In this section, we present the main theoretical results. The statistical model is defined in (1.1). Hereafter we assume that ϵ is sub-Gaussian with variance proxy σ^2. By saying that a random vector $\mathbf{z} \in \mathbb{R}^k$ is sub-Gaussian with zero mean and variance proxy $\tau^2 \geq 0$, we mean that $\mathbb{E}[\mathbf{z}] = \mathbf{0}$

and

$$\mathbb{E}[\exp(\boldsymbol{\theta}^\top \mathbf{z})] \leq \exp(\|\boldsymbol{\theta}\|_2^2 \tau^2/2) \text{ for all } \boldsymbol{\theta} \in \mathbb{R}^k.$$

3.1. Theory of Parameter Estimation

Before presenting the main results for parameter estimation, we first state the following assumptions on $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, which are standard for sparse linear regression problems with fixed design.

Assumption 1. *Sparse-Eigenvalue*(s^*, k^*). For $k \in \{1, \ldots, d\}$, we denote the k-sparse eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ as $\rho_-(k)$ and $\rho_+(k)$ respectively, which are defined as

$$\rho_-(k) \coloneqq \inf\big\{\mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}\mathbf{v} \colon \|\mathbf{v}\|_0 \leq k, \|\mathbf{v}\|_2 = 1\big\} \text{ and}$$
$$\rho_+(k) \coloneqq \sup\big\{\mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}\mathbf{v} \colon \|\mathbf{v}\|_0 \leq k, \|\mathbf{v}\|_2 = 1\big\}.$$

We assume that, for $s^* = \|\boldsymbol{\beta}^*\|_0$, there exists a $k^* \in \mathbb{N}$ such that $k^* \geq 2s^*$ and

$$\rho_+(k^*)/\rho_-(2k^* + s^*) \leq 1 + 0.5k^*/s^*. \tag{3.1}$$

The condition $\rho_+(k^*)/\rho_-(2k^* + s^*) \leq 1 + 0.5k^*/s^*$ requires that the eigenvalue ratio $\rho_+(k)/\rho_-(2k + s^*)$ grows sub-linearly in k. This condition, commonly referred to as *sparse eigenvalue condition*, is standard in sparse estimation problems and has been studied by [56]. This condition is weaker than the well-known restricted isometry property (RIP) in compressed sensing [10], which states that there exists a constant $\delta \in (0, 1)$ and integer $s \in \{1, \ldots, d\}$ such that for all s-sparse $\mathbf{v} \in \mathbb{R}^d$, we have

$$(1 - \delta)\|\mathbf{v}\|_2^2 \leq \mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{v} \leq (1 + \delta)\|\mathbf{v}\|_2^2. \tag{3.2}$$

Comparing (3.1) and (3.2), we see that (3.1) holds with $k^* = (s - s^*)/2$ if the RIP condition holds with $s \geq 5s^*$ and $\delta = 1/3$. As is shown in [49], RIP holds with high probability for sub-Gaussian random matrices. Therefore Assumption 1 holds at least when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. sub-Gaussian, which contains many well-known distributions as special cases.

We note that although Assumption 3.1 holds since it does not depend on the nonlinear transfmration f, the *restricted strong convexity* (RSC) condition defined in [32, 33] on the loss function $L(\boldsymbol{\beta})$ does not directly hold in general in our setting since $L(\boldsymbol{\beta})$ depends on the nonlinear transformation f.

In addition to the sparse eigenvalue assumption, we need a regularity condition, which states that the elements of $\widehat{\boldsymbol{\Sigma}}$ are uniformly bounded.

Assumption 2. *Bounded-Design*(D). We assume there exists an absolute constant D that does not depends on n, d, or s^* such that $\|\widehat{\boldsymbol{\Sigma}}\|_\infty \leq D$, where $\|\cdot\|_\infty$ is the matrix elementwise ℓ_∞-norm.

If the population version of $\widehat{\boldsymbol{\Sigma}}$, i.e., $\boldsymbol{\Sigma} \coloneqq \mathbb{E}(\mathbf{x}\mathbf{x}^\top)$, has bounded elements and $\mathbf{x}$ has sub-Gaussian or sub-exponential tails, then by concentration inequalities we can prove that Assumption 2 holds with high probability with $D = 2\|\boldsymbol{\Sigma}\|_\infty$. We verify this assumption for sub-Gaussian $\mathbf{x}$ in the appendix. This assumption is generally unnecessary for high dimensional linear regression. However, it is required in our setting where it is used to control the effect of the nonlinear transform.

We note that we do not make any further assumptions except Assumptions 1 and 2 on the distribution of $\mathbf{x}$ for the theory of parameter estimation to hold. These two assumptions are shown to be true when $\mathbf{x}$ is sub-Gaussian.

We are now ready to present our main theorem for parameter estimation, which states that any stationary point of the ℓ_1-regularized optimization problem enjoys optimal statistical rates of convergence and that Algorithm 1 successfully converges to a stationary point.

Theorem 1. We assume that the univariate function f satisfies $f(0) = 0$ and is continuously differentiable with $f'(x) \in [a, b], \forall x \in \mathbb{R}$ for some $0 < a < b$. We further assume that Assumptions 1 and 2 hold. Then there exists a constant B such that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \leq B\sigma \cdot \sqrt{\log d/n}$ with probability tending to one. Suppose we choose the regularization parameter λ in (1.2) as

$$\lambda = C\sigma\sqrt{\log d/n} \text{ with } C \geq \max\{L_1 B, L_2\}, \tag{3.3}$$

where L_1 and L_2 satisfy $L_1^{-1} + 3b\sqrt{D}L_2^{-1} \leq 0.1$. Then for any stationary point $\widehat{\boldsymbol{\beta}}$ satisfying $\nabla L\big(\widehat{\boldsymbol{\beta}}\big) + \lambda \cdot \boldsymbol{\xi} = \mathbf{0}$ with $\boldsymbol{\xi} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$, it holds with probability at least $1 - d^{-1}$ that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 25/\rho_-(k^* + s^*) \cdot a^{-2}\sqrt{s^*}\lambda; \tag{3.4}$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 25/\rho_-(k^* + s^*) \cdot a^{-2}s^*\lambda. \tag{3.5}$$

Furthermore, Algorithm 1 attains a stationary point with the statistical rates in (3.4).

By our discussion under Assumption 1, we can take $k^* = Cs^*$ for some constant $C > 0$. Then plugging (3.3) into (3.4), we obtain the rate of $\sqrt{s^* \log d/n}$ in ℓ_2-norm and the rate of $s^*\sqrt{\log d/n}$ in ℓ_1-norm. Similar results are also established for sparse linear regression, and more generally, high-dimensional generalized linear models [9, 25, 55]. These rates are optimal in the sense that they cannot be improved even if f equals to the identity. Note that the lower bound a of f' shows up in the statistical rates of convergence in (3.4). If a is close to zero, we obtain a large statistical error. To see the intuition, we consider a worst case where f is constant, i.e., $a = 0$. Then it is impossible to consistently estimate $\boldsymbol{\beta}^*$, since in this case the observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$ provide no information on $\boldsymbol{\beta}^*$.

The statistical rates of convergence are proportional to the noise level σ, which implies that the proposed method exactly recovers $\boldsymbol{\beta}^*$ in the noiseless setting. In the noisy case, by (3.4), to get ϵ accuracy of estimating $\boldsymbol{\beta}^*$ in ℓ_2-norm with high probability, the sample complexity is $n = O(\epsilon^{-2} s^* \log d)$, which is of the same order as that of high-dimensional linear models.

3.2. Minimax Lower Bound

To understand the optimality of the estimation result, we study the minimax lower bound of parameter estimation in our model, which reveals the fundamental limits of the estimation problem. We define the minimax risk as

$$\mathcal{R}_f^*(s, n, d) = \inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in \mathbb{B}_0(s)} \mathbb{E}_{\boldsymbol{\beta}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2, \tag{3.6}$$

where the expectation is taken over the probability model in (1.1) with parameter $\boldsymbol{\beta}$ and $\mathbb{B}_0(s) \coloneqq \big\{\boldsymbol{\beta} \in \mathbb{R}^d \colon \|\boldsymbol{\beta}\|_0 \leq s\big\}$. Here the supremum is taken over all s-sparse parameters and the infimum is taken over all estimators $\widehat{\boldsymbol{\beta}}$ based on samples $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$. We assume f is continuously differentiable with $f'(u) \in [a, b], \forall u \in \mathbb{R}$. The following theorem gives a lower bound on the minimax risk $\mathcal{R}_f(s, n, d)$, which implies the optimality of the proposed estimator.

Theorem 2. For integer s and d satisfying $1 \leq s \leq d/8$, the minimax risk defined in (3.6) has the following lower bound

$$\mathcal{R}_f^*(s, n, d) \geq \frac{\sigma^2}{192 b^2 \rho_+(2s)} \frac{s \log[1 + d/(2s)]}{n}. \tag{3.7}$$

By Theorem 2, if we consider a, b as constants and assume that k^*/s^* is bounded, then the ℓ_2-statistical rate of convergence of $\widehat{\boldsymbol{\beta}}$ in (3.4) matches the minimax lower bound in (3.7) in terms of order. This establishes the optimality of the proposed estimator.

4. Connection to Prior Work

The model we consider is closely related to the single index model where the function f is unknown. Both of these two models fall in the framework of sufficient dimension reduction with a one-dimensional subspace $\mathcal{U}$ [14, 15, 28–30]. In low dimensional settings, most works in this direction use spectral methods, which rely on the Gaussian assumption and can only estimate $\boldsymbol{\theta}^* = \boldsymbol{\beta}^* \|\boldsymbol{\beta}^*\|_2^{-1}$ because the norm of $\boldsymbol{\beta}^*$ is not identifiable when f is unknown. As introduced in [30], many moment based sufficient dimension reduction methods can be stated as a generalized eigenvalue problem $\mathbf{M}_n \boldsymbol{\theta}_i = \lambda_i \mathbf{N}_n \boldsymbol{\theta}_i$ for $i = 1, \ldots, d$, where $\mathbf{M}_n$ and $\mathbf{N}_n$ are symmetric matrices computed from the data; $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d$ are generalized eigenvectors such that $\boldsymbol{\theta}_i^\top \mathbf{N}_n \boldsymbol{\theta}_j = \mathbb{1}_{\{i=j\}}$ and $\lambda_1 \geq \cdots \geq \lambda_d$ are the generalized eigenvalues. In addition, it is required that $\mathbf{M}_n$ and $\mathbf{N}_n$ are positive semidefinite and positive definite, respectively. Here $\mathbf{M}_n$ and $\mathbf{N}_n$ are the sample versions of the corresponding population quantities $\mathbf{M}$ and $\mathbf{N}$. For example, in sliced inverse regression [28], we have $\mathbf{M} = \mathrm{Cov}\{\mathbb{E}[\mathbf{x} - \mathbb{E}(\mathbf{x})|y]\}$ and $\mathbf{N} = \mathrm{Cov}(\mathbf{x})$ and $\mathbf{M}_n$ and $\mathbf{N}_n$ are their population analogs. When $\mathcal{U}$ is one-dimensional, $\boldsymbol{\theta}^*$ corresponds to the generalized eigenvector with the largest eigenvalue. In low dimensional settings, [30] showed that $\boldsymbol{\theta}^*$ can be estimated by the following optimization problem:

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{maximize}} \ \boldsymbol{\theta}^\top \mathbf{M}_n \boldsymbol{\theta} \quad \text{subject to } \boldsymbol{\theta}^\top \mathbf{N}_n \boldsymbol{\theta}. \tag{4.1}$$

Since the works in this direction all require the matrix $\mathbf{N}_n$, which is the sample covariance matrix of $\mathbf{x}$ in most cases, to be invertible, such methods cannot be generalized to high-dimensional settings where $\mathbf{N}_n$ is not invertible.

For high-dimensional single index models, [38] proposes an estimator by projecting $n^{-1} \sum_{i=1}^n y_i \mathbf{x}_i$ onto a fixed star-shaped closed subset K of $\mathbb{R}^d$. Similarly, [37] propose a least-squares estimator with a geometric constraint:

$$\text{minimize} \sum_{i=1}^n (\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 \ \text{ subject to } \ \boldsymbol{\theta} \in K. \tag{4.2}$$

Both of these methods rely on the assumption that $\mathbf{x}_i$ is Gaussian to have good estimation of $\mathbb{E}(y \cdot \mathbf{x})$. Under the Gaussian assumption, we achieve the same statistical rate, which is optimal. When $\mathbf{x}$ is not Gaussian, as shown in [1], their methods will have some extra terms in the error bound that may or may not tend to zero. Our method, however, works when $\mathbf{x}$ has a general distribution with optimal statistical rates of convergence.

5. Numerical Experiments

In this section, we evaluate the finite sample performance of parameter estimation on both simulated data and a real-world dataset.

For parameter estimation, we compute the ℓ_2-error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, where $\widehat{\boldsymbol{\beta}}$ is the solution of Algorithm 1. In addition, we compare our method with the linear approach that inverts the nonlinear function. For the linear framework we apply the ℓ_1-regularized regresion (Lasso) [45].

5.1. Simulated Data

Throughout this section, we sample independent data from model (1.1) with $\epsilon \sim N(0, 1)$ and $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a Toeplitz matrix with $\Sigma_{jk} = 0.95^{|j-k|}$. The sparse parameter vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is set to have nonzero values in the first s^* entries. That is, $\beta_j^* \neq 0$ for $1 \leq j \leq s^*$ and $\beta_j^* = 0$ otherwise. In addition, we consider the nonlinear

function $f(x) = 2x + \cos(x)$. In this case the derivative $f'(\cdot)$ is bounded by $a = 1$ and $b = 4$.

For parameter estimation, we compare the ℓ_2-error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ with $\sqrt{s^* \log d/n}$ under two settings: (i) we fix $d = 256$, $s^* = 6, 8$, or 10, and vary n, and (2) fix $s^* = 10$, $d = 128, 256$ or 512, and vary n. For the parameter $\boldsymbol{\beta}^*$, the first s^* entries are sampled independently from the uniform distribution on the interval $[0, 2]$. That is, $\beta_j^* \sim \mathrm{U}(0, 2)$ for $1 \le j \le s^*$ and $\beta_j^* = 0$ for $j > s^*$. We set the regularization parameter $\lambda = 3\sigma \cdot \sqrt{\log d/n}$. The parameters of Algorithm 1 are chosen as $\alpha_{\min} = 1/\alpha_{\max} = 10^{30}$, $\eta = 2$, $M = 5$, and $\zeta = \mathtt{tol} = 10^{-5}$. The ℓ_2-errors reported are based on 100 independent experiments. We plot the ℓ_2-errors against the effective sample size $\sqrt{s^* \log d/n}$ in Figure 1. The figure illustrates that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ grows sublinearly with $\sqrt{s^* \log d/n}$, which corroborates with our argument that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le C\sqrt{s^* \log d/n}$ for some absolute constant C.

To compare Algorithm 1 with inverting f, we consider the settings where $d = 256$, $s^* = 8$. We then apply Lasso to the inverted data $\{f^{-1}(y_i), \mathbf{x}_i\}_{i=1}^n$ where the regularization parameter of Lasso is selected via 5-fold cross-validation. The optimization problem of Lasso is also solved using Algorithm 1. We plot the ℓ_2-errors of these two techniques against the effective sample size in Figure 1-(c), which shows that the proposed method outperforms the linear approach.

5.2. Real Data Analysis

To show the effectiveness of the proposed method, we study the *Computer Audition Lab 500-Song (CAL500)* dataset [48], which can be obtained from the publicly available *Mulan* data library [46]. The *CAL500* dataset consists of music annotations of 502 popular music tracks. The attributes of this dataset consist of both continuous and binary subsets. The continuous features are obtained from the coefficients of short time Fourier transforms on each music track. In specific, there are four types of continuous features: *spectral centroids*, *spectral flux*, *zero crossings* and a time series of *Mel-frequency cepstral coefficient (MFCC)*. In addition, for each music track, the values of the binary features are assigned by human listeners to give semantic descriptions. For accuracy, each music track is annotated by at least three human listeners. See [48] for a more detailed introduction of the *CAL500* dataset. This dataset is previously analyzed in [13, 54], where they study the conditional independence of the attributes by fitting graphical models. Similar to [13], we study model (1.1) only using the continuous features. In specific, we use n random subsamples of the 502 instances of $d = 68$ continuous attributes, where n is an integer that will be specified later. We generate the response according model (1.1) with $\sigma = 1$, $f(x) = 4x + \cos(x)$. Moreover, we choose support of $\boldsymbol{\beta}^*$ uniformly over $\{1, \ldots, d\}$.

Given the response and the design matrix, we study the performance of the proposed estimator. Specifically, we compare the ℓ_2-error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ with $\sqrt{s^* \log d/n}$ under the setting where we fix $d = 68$, $s^* = 4, 6$, or 8, and vary n. In this setting, the nonzero entries of $\boldsymbol{\beta}^*$ are sampled independently from the uniform distribution over $[0, 2]$. We set the regularization parameter to be $\lambda = 2\sigma\sqrt{\log d/n}$ and the parameters of Algorithm 1 the same as those in the simulation studies in §5.1. We plot the ℓ_2-errors against the effective sample size $\sqrt{s^* \log d/n}$ in Figure 2-(a) based on 100 random experiments . The figure also shows that the estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ grows sublinearly with $\sqrt{s^* \log d/n}$.

In addition, we also study the setting where the nonzero entries of $\boldsymbol{\beta}^*$ are set to a constant $\beta_0 > 0$. In addition, we fix $d = 68$, $n = 50$, and $s^* = 4, 6$, or 8. The regularization parameter λ and the parameters of Algorithm 1 remain the same. In this case, the value of β_0 corresponds to the magnitude of the signal parameter. Thus, estimation is easier for large β_0 whereas the error is large for small β_0. For presentation, we plot the ℓ_2-error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ against β_0 based on 100 independent trials. As show in in Figure 2-(b), as β_0 grows, the estimation error gradually decreases, which coincides with the intuition.

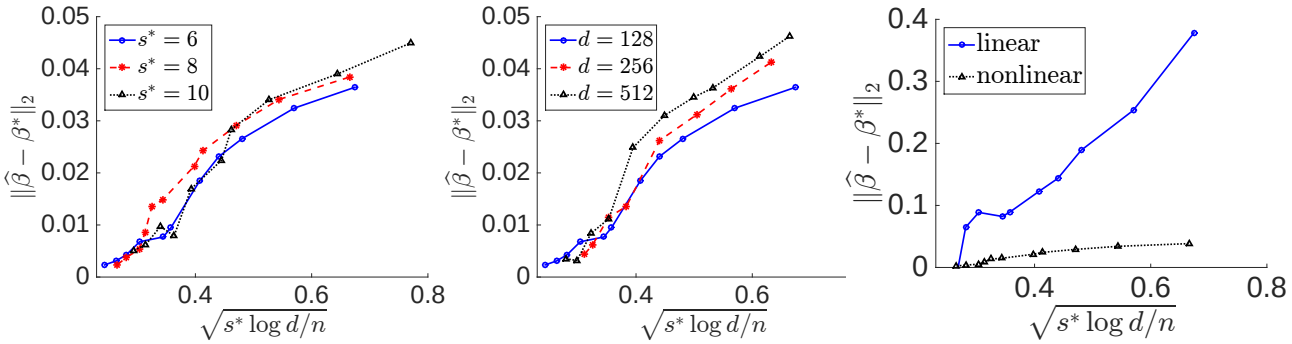
Moreover, similar to the simulation studies, we also compare the proposed method with the linear framework which inverts f. In particular, we fix $d = 68$, $s^* = 4$ and vary n. The support of $\boldsymbol{\beta}^*$ is chosen uniformly with the nonzero entries sampled independently from $\mathrm{U}(0, 2)$. We compute the ℓ_2-error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ of the Lasso estimator obtained using 5-fold cross-validation. In addition, for the proposed method, the regularization parameter λ and the parameters of Algorithm 1 is the same as in the previous setting. In Figure 2-(c) we plot the ℓ_2-errors of these two estimators against the effective sample size $\sqrt{s^* \log d/n}$ based on 100 independent experiments. It clear that the error of the estimator constructed by the linear framework is much larger, which shows the superiority of the proposed method.

6. Proof of the Statistical Rates

In this section we sketch the proof of the statistical rates of convergence for the proposed estimator. We defer a more detailed proofs the main results in the appendix.

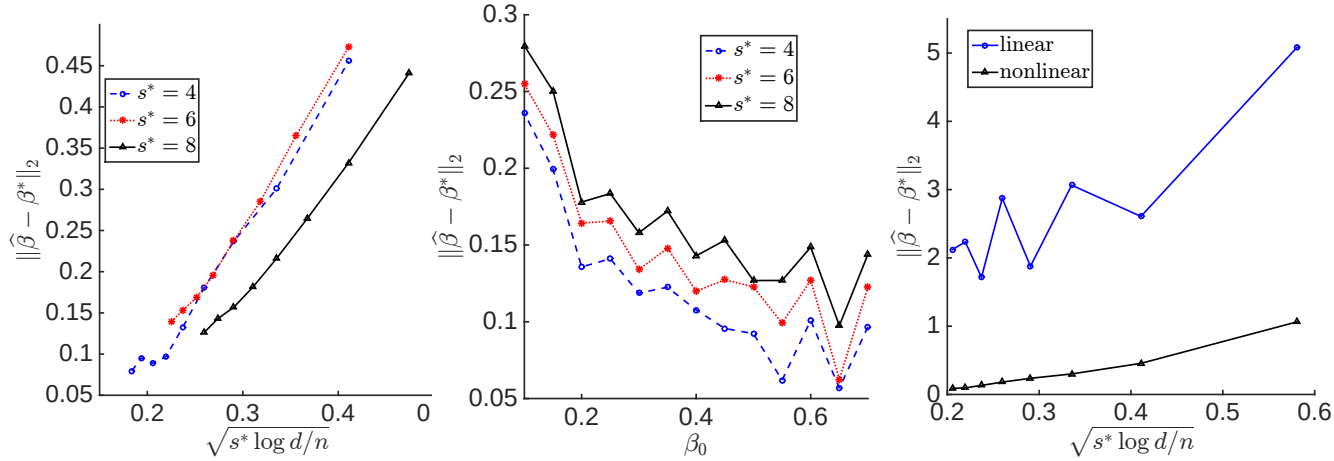
Similar to the analysis of Lasso estimator, a main step of the proof is to show that the error vector lies in an ℓ_1-cone. In specific, for any stationary point $\widehat{\boldsymbol{\beta}}$ of the optimization problem in (1.2), we denote $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Let $\mathcal{S}$ be the support of $\boldsymbol{\beta}^*$, we show that

$$\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \le \gamma \cdot \|\boldsymbol{\delta}_{\mathcal{S}}\|_1$$



(a). Fix $d = 256$. (b). Fix $s^* = 10$. (c). Linear vs. Nonlinear.

Figure 1: Statistical errors $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ plotted against the effective sample size $\sqrt{s^* \log d/n}$ with d or s^* fixed and n varied are shown in (a) and (b), respectively. The comparison between the method of inverting f and the proposed method is shown in (c).



(a). Compare ℓ_2-errors against $\sqrt{s^* \log d/n}$. (b). Compare ℓ_2-errors against β_0. (c). Linear vs. Nonlinear.

Figure 2: Statistical errors $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ plotted against the effective sample size $\sqrt{s^* \log d/n}$ and the magnitude of signal parameter β_0 are shown in (a) and (b), respectively. We fix s^* and vary n in (a) and fix $s^* = 4, n = 50$ in (b). The comparison between the method of inverting f and the proposed method with $s^* = 4$ and n varied is shown in (c).

for some constant $\gamma > 0$. This is established by combining a upper and an lower bound for

$$\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle. \tag{6.1}$$

For an upper bound, by the optimality of $\widehat{\boldsymbol{\beta}}$ we have $\nabla L(\widehat{\boldsymbol{\beta}}) + \lambda \cdot \boldsymbol{\xi} = \mathbf{0}$, where $\boldsymbol{\xi} \in \partial \|\widehat{\boldsymbol{\beta}}\|_1$. Note that the support of $\boldsymbol{\beta}^*$ is $\mathcal{S}$, that is, $\mathcal{S} = \{j \colon \beta_j^* \neq 0\}$. Also note that the optimality of $\widehat{\boldsymbol{\beta}}$ implies that

$$\langle \boldsymbol{\xi}_{\mathcal{S}^c}, \widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} \rangle = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 = \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1.$$

By Hölder's inequality, since $\|\boldsymbol{\xi}\|_\infty \leq 1$ and $\boldsymbol{\beta}_{\mathcal{S}}^* = \mathbf{0}$, (6.1) is bounded by

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \|\nabla L(\boldsymbol{\beta}^*)\|_\infty \|\boldsymbol{\delta}\|_1. \end{aligned} \tag{6.2}$$

Moreover, by calculation, we have

$$\begin{aligned} \nabla L(\boldsymbol{\beta}^*) &= -\frac{1}{n} \sum_{i=1}^n [y_i - f(\mathbf{x}_i^\top \boldsymbol{\beta}^*)] \cdot f'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) \cdot \mathbf{x}_i \\ &= -\frac{1}{n} \sum_{i=1}^n f'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) \cdot \mathbf{x}_i \cdot \epsilon_i. \end{aligned}$$

Since f' is bounded and that ϵ_i's are i.i.d. sub-Gaussian random variables, conditioning on $\{\mathbf{x}_i\}_{i=1}^n$, $\nabla L(\boldsymbol{\beta}^*)$ is the mean of i.i.d. sub-Gaussian random variables [49]. Concentration of measure guarantees that $\nabla L(\boldsymbol{\beta}^*)$ is not far away from its mean, which is $\mathbf{0}$. The following lemma shows that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty$ is of order $\sigma \cdot \sqrt{\log d/n}$ with high probability.

Lemma 3. Let $L(\boldsymbol{\beta})$ be the least-square loss function defined in (2.1), there exist an absolute constant $B > 0$

that does not depend on n, d or s^* and $\delta = \delta(n,d)$ that tends to 0 as $n \to \infty$ such that $\delta \le (2d)^{-1}$ and that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \le B\sigma \cdot \sqrt{\log d/n}$ with probability $1-\delta$.

In what follows, we condition on the event that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \le B\sigma \cdot \sqrt{\log d/n}$, which holds with probability at least $1-(2d)^{-1}$ by Lemma 3. By the definition of λ and (6.2) we have

$$\begin{aligned}&\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\le -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + L_1^{-1}\lambda \|\boldsymbol{\delta}\|_1. \end{aligned} \tag{6.3}$$

Thus we derive an upper bound for (6.1). Moreover, the lemma establishes a lower bound (6.1).

Lemma 4. Recall that $\widehat{\boldsymbol{\Sigma}} \coloneqq n^{-1}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. Under the Assumption *Bounded-Design*(D), it holds with probability at least $1-(2d)^{-1}$ that, for any $\boldsymbol{\beta} \in \mathbb{R}^d$,

$$\begin{aligned}\langle \nabla L(\boldsymbol{\beta}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \ge a^2 (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \widehat{\boldsymbol{\Sigma}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ - 3b\sigma\sqrt{D \log d/n} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1.\end{aligned}$$

Thus combining the upper bound and the lower bound for (6.1), we obtain that

$$a^2 \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} \le -\lambda(1-\mu)\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda(1+\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1, \tag{6.4}$$

where $\mu = L_1^{-1} + 3b\sqrt{D}L_2^{-1} \le 0.1$. Hence it follows that $\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \le (1+\mu)/(1-\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \le 1.23\|\boldsymbol{\delta}_{\mathcal{S}}\|_1$. This shows that the error vector lies in the ℓ_1-cone

$$\{\mathbf{v} \in \mathbb{R}^d \colon \|\mathbf{v}_{\mathcal{S}^c}\|_1 \le 1.23\|\mathbf{v}_{\mathcal{S}}\|_1\}.$$

Note that by (6.4) we have $a^2\boldsymbol{\delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta} \le \lambda(1+\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1$. The final part of the proof is to compare this upper bound with a bound of $\boldsymbol{\delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}$ from below, which is given in the following lemma to bound $\boldsymbol{\delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}$ from below.

Lemma 5. For any $\boldsymbol{\eta} \in \mathbb{R}^d$ and any index set $\mathcal{S}$ with $|\mathcal{S}| = s^*$, let $\mathcal{J}$ be the set of indices of the largest k^* entries of $\boldsymbol{\eta}_{\mathcal{S}^c}$ in absolute value and let $\mathcal{I} = \mathcal{J} \cup \mathcal{S}$. Here s^* and k^* are the same as those in Assumption *Sparse-Eigenvalue*(s^*, k^*). Assume that $\|\boldsymbol{\eta}_{\mathcal{S}^c}\|_1 \le \gamma\|\boldsymbol{\eta}_{\mathcal{S}}\|_1$ for some $\gamma > 0$. Then we obtain that $\|\boldsymbol{\eta}\|_2 \le (1+\gamma)\|\boldsymbol{\eta}_{\mathcal{I}}\|_2$ and that

$$\begin{aligned}\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta} \ge \rho_-(s^*+k^*) \cdot \big[\|\boldsymbol{\eta}_{\mathcal{I}}\|_2 - \gamma\sqrt{s^*/k^*} \cdot \\ \sqrt{\rho_+(k^*)/\rho_-(s^*+2k^*) - 1} \cdot \|\boldsymbol{\eta}_{\mathcal{S}}\|_2\big] \cdot \|\boldsymbol{\eta}_{\mathcal{I}}\|_2.\end{aligned} \tag{6.5}$$

Under Assumption 1, we have $\rho_+(k^*)/\rho_-(s^*+2k^*) \le 1 + 0.5k^*/s^*$. By Lemma 5 we obtain that $\|\boldsymbol{\delta}\|_2 \le 2.23\|\boldsymbol{\delta}_{\mathcal{I}}\|_2$ and that

$$\begin{aligned}\boldsymbol{\delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta} &\ge \big(1 - 1.23\sqrt{0.5}\big) \cdot \rho_-(s^*+k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2 \\ &\ge 0.1 \cdot \rho_-(s^*+k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2,\end{aligned} \tag{6.6}$$

where $\mathcal{J}$ is the set of indices of the largest k^* entries of $\boldsymbol{\delta}_{\mathcal{S}^c}$ in absolute value and $\mathcal{I} = \mathcal{J} \cup \mathcal{S}$. Combining the upper and lower bounds for $\boldsymbol{\delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\delta}$ we obtain

$$\|\boldsymbol{\delta}_{\mathcal{I}}\|_2 \le 11/\rho_-(s^*+k^*) \cdot a^{-2}\sqrt{s^*}\lambda.$$

Therefore we have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &= \|\boldsymbol{\delta}\|_1 \le 2.23\|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \le 2.23\sqrt{s^*}\|\boldsymbol{\delta}_{\mathcal{S}}\|_2 \\ &\le 25/\rho_-(s^*+k^*) \cdot a^{-2}s^*\lambda; \\ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &= \|\boldsymbol{\delta}\|_2 \le 2.23\|\boldsymbol{\delta}_{\mathcal{I}}\|_2 \\ &\le 25/\rho_-(s^*+k^*) \cdot a^{-2}\sqrt{s^*}\lambda.\end{aligned}$$

Thus we establish the statistical rates of convergence for the proposed estimator.

7. Conclusion

We study parameter estimation for high dimensional regression under known nonlinear transform. We propose an ℓ_1-regularized least-square estimator for estimation. Although the optimization problem is non-convex, we show that every stationary point converges to the true signal with the optimal statistical rate of convergence. We establish the optimality by deriving a minimax lower bound for the regression model. In addition, we propose an efficient algorithm that successfully converges to a stationary point. Both simulation experiments and real data analysis are provided to back up the developed theory.

References

[1] Ai, A., Lapanowski, A., Plan, Y., and Vershynin, R. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441: 222–239, 2014.

[2] Aksoylar, C. and Saligrama, V. Sparse recovery with linear and nonlinear observations: Dependent and noisy data. *arXiv preprint arXiv:1403.3109*, 2014.

[3] Alquier, P. and Biau, G. Sparse single-index model. *The Journal of Machine Learning Research*, 14(1): 243–280, 2013.

[4] Barzilai, J. and Borwein, J. M. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[5] Beck, A. and Eldar, Y. C. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[6] Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[7] Beck, A. and Eldar, Y. C. Sparse signal recovery from nonlinear measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5464–5468. IEEE, 2013.

[8] Cai, T. T., Li, X., and Ma, Z. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *arXiv preprint arXiv:1506.03382*, 2015.

[9] Candès, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[10] Candès, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51 (12):4203–4215, 2005.

[11] Candès, E. J., Strohmer, T., and Voroninski, V. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8): 1241–1274, 2013.

[12] Candès, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251, 2015.

[13] Cheng, J., Levina, E., and Zhu, J. High-dimensional mixed graphical models. *arXiv preprint arXiv:1304.2810*, 2013.

[14] Cook, R. D. Principal Hessian directions revisited. *Journal of the American Statistical Association*, 93 (441):84–94, 1998.

[15] Cook, R. D. and Lee, H. Dimension reduction in binary response regression. *Journal of the American Statistical Association*, 94(448):1187–1200, 1999.

[16] Delecroix, M., Hristache, M., and Patilea, V. Optimal smoothing in semiparametric index approximation of regression functions. Technical report, Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, 2000.

[17] Delecroix, M., Hristache, M., and Patilea, V. On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3): 730–769, 2006.

[18] Eldar, Y. C. and Kutyniok, G. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[19] Eldar, Y. C. and Mendelson, S. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[20] Eldar, Y. C., Sidorenko, P., Mixon, D. G., Barel, S., and Cohen, O. Sparse phase retrieval from short-time Fourier measurements. *Signal Processing Letters, IEEE*, 22(5):638–642, 2015.

[21] Härdle, W., Hall, P., and Ichimura, H. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178, 1993.

[22] Horowitz, J. L. *Semiparametric and Nonparametric Methods in Econometrics*, volume 692. Springer, 2000.

[23] Ichimura, H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

[24] Jaganathan, K., Oymak, S., and Hassibi, B. On robust phase retrieval for sparse signals. In *50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 794–799. IEEE, 2012.

[25] Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *International Conference on Artificial Intelligence and Statistics*, pp. 381–388, 2010.

[26] Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pp. 927–935, 2011.

[27] Kalai, A. T. and Sastry, R. The isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory*, 2009.

[28] Li, K.-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[29] Li, K.-C. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[30] Li, L. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

[31] Li, X. and Voroninski, V. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5): 3019–3033, 2013.

[32] Loh, P.-L. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*, 2014.

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

[34] McCullagh, P., Nelder, J. A., and McCullagh, P. *Generalized linear models*, volume 2. Chapman and Hall London, 1989.

[35] Ohlsson, H. and Eldar, Y. C. On conditions for uniqueness in sparse phase retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1841–1845. IEEE, 2014.

[36] Ohlsson, H., Yang, A., Dong, R., and Sastry, S. Compressive phase retrieval from squared output measurements via semidefinite programming. In *16th IFAC Symposium on System Identification, Brussels, Belgium, 11-13 July, 2012*, pp. 89–94, 2012.

[37] Plan, Y. and Vershynin, R. The generalized Lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*, 2015.

[38] Plan, Y., Vershynin, R., and Yudovina, E. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[39] Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 10(57):6976–6994, 2011.

[40] Rigollet, P., Tsybakov, A., et al. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[41] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[42] Shechtman, Y., Beck, A., and Eldar, Y. C. Gespar: Efficient phase retrieval of sparse signals. *IEEE Transactions on Signal Processing*, 62(4):928–938, 2014.

[43] Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. Phase retrieval with application to optical imaging: a contemporary overview. *Signal Processing Magazine, IEEE*, 32(3):87–109, 2015.

[44] Sherman, R. P. U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric theory*, 10(02):372–395, 1994.

[45] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[46] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414, 2011.

[47] Tu, S., Boczar, R., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[48] Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[49] Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[50] Waldspurger, I., dAspremont, A., and Mallat, S. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[51] Wright, S. J., Nowak, R. D., and Figueiredo, M. A. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*.

[52] Xia, Y. and Li, W. On single-index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285, 1999.

[53] Xia, Y., Tong, H., and Li, W. On extended partially linear single-index models. *Biometrika*, 86(4):831–842, 1999.

[54] Yang, Z., Ning, Y., and Liu, H. On semiparametric exponential family graphical models. *arXiv preprint arXiv:1412.8697*, 2014.

[55] Zhang, C.-H. and Huang, J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pp. 1567–1594, 2008.

[56] Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010.

A. Proof of the Main Results

In this section, we lay out the proofs of the main results presented in §3. We first establish the statistical rate of convergence for the proposed estimator, and then show the optimality of our procedure by deriving the minimax lower bound.

A.1. Proof of Theorem 1

Proof. For any stationary point $\widehat{\boldsymbol{\beta}}$ of the optimization problem in (1.2), by definition we have

$$\nabla L(\widehat{\boldsymbol{\beta}}) + \lambda \cdot \boldsymbol{\xi} = \mathbf{0}, \;\; \text{where } \boldsymbol{\xi} \in \partial \|\widehat{\boldsymbol{\beta}}\|_1.$$

For notational simplicity, we denote $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ as $\boldsymbol{\delta}$. By definition, we have

$$\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle = \langle -\lambda \cdot \boldsymbol{\xi} - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle.$$

We denote the support of $\boldsymbol{\beta}^*$ as $\mathcal{S}$, that is, $\mathcal{S} = \{j \colon \beta_j^* \neq 0\}$. By writing $\boldsymbol{\xi} = \boldsymbol{\xi}_{\mathcal{S}} + \boldsymbol{\xi}_{\mathcal{S}^c}$ we have

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &= \langle -\lambda \cdot \boldsymbol{\xi}_{\mathcal{S}^c} - \lambda \cdot \boldsymbol{\xi}_{\mathcal{S}} - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle. \end{aligned} \tag{A.1}$$

Note that $\boldsymbol{\beta}^*_{\mathcal{S}} = \mathbf{0}$ and $\langle \boldsymbol{\xi}_{\mathcal{S}^c}, \widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} \rangle = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 = \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1$. By Hölder's inequality, since $\|\boldsymbol{\xi}\|_\infty \leq 1$, the right-hand side of (A.1) can be bounded by

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \|\nabla L(\boldsymbol{\beta})\|_\infty \|\boldsymbol{\delta}\|_1. \end{aligned} \tag{A.2}$$

Now we invoke Lemma 3 to bound the right hand side of (A.2). In what follows, we condition on the event that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \leq B\sigma \cdot \sqrt{\log d / n}$, which holds with probability at least $1 - \delta$, where $\delta \geq (2d)^{-1}$. By the definition of λ, we have $\lambda \geq L_1 \cdot \|\nabla L(\boldsymbol{\beta})\|_\infty$ with probability at least $1 - \delta$. By (A.2) we have

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + L_1^{-1} \lambda \|\boldsymbol{\delta}\|_1. \end{aligned} \tag{A.3}$$

Now we invoke Lemma 4 to establish a lower bound of the left-hand side of (A.3). Combining (3.3), (A.3) and Lemma 4 we obtain that

$$\begin{aligned} 0 &\leq a^2 \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} \leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \mu \lambda \|\boldsymbol{\delta}\|_1 \\ &= -\lambda(1-\mu)\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda(1+\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1. \end{aligned} \tag{A.4}$$

where $\mu = L_1^{-1} + 3b\sqrt{D} L_2^{-1} \leq 0.1$. Hence it follows that $\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leq (1+\mu)/(1-\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \leq 1.23 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1$.

Now we invoke the Lemma 5 to bound $\boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta}$ from below. Under Assumption 1, we have

$$\rho_+(k^*) / \rho_-(s^* + 2k^*) \leq 1 + 0.5 k^* / s^*.$$

Combining this inequality with Lemma 5 we obtain that

$$\begin{aligned} \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} &\geq \big(1 - 1.23\sqrt{0.5}\big) \cdot \rho_-(s^* + k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2 \\ &\geq 0.1 \cdot \rho_-(s^* + k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2, \end{aligned} \tag{A.5}$$

where $\mathcal{J}$ is the set of indices of the largest k^* entries of $\boldsymbol{\delta}_{\mathcal{S}^c}$ in absolute value and $\mathcal{I} = \mathcal{J} \cup \mathcal{S}$. Here the first inequality of (A.5) follows from Lemma 5 and that $\mathcal{S} \subset \mathcal{I}$. Combining (A.4) and (A.5) we obtain that

$$\begin{aligned} 0.1 \cdot \rho_-(s^* + k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2 &\leq \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} \leq a^{-2} \lambda (1+\mu) \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \\ &\leq 1.1 \cdot a^{-2} \sqrt{s^*} \lambda \|\boldsymbol{\delta}_{\mathcal{I}}\|_2, \end{aligned}$$

which implies that $\|\boldsymbol{\delta}_{\mathcal{I}}\|_2 \leq 11 / \rho_-(s^* + k^*) \cdot a^{-2} \sqrt{s^*} \lambda$. Note that by Lemma 5 we also have $\|\boldsymbol{\delta}\|_2 \leq 2.23 \|\boldsymbol{\delta}_{\mathcal{I}}\|_2$. Combining this inequality with the fact that $\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leq 1.23 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1$, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &= \|\boldsymbol{\delta}\|_1 \leq 2.23 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \leq 2.23 \sqrt{s^*} \|\boldsymbol{\delta}_{\mathcal{S}}\|_2 \\ &\leq 25 / \rho_-(s^* + k^*) \cdot a^{-2} s^* \lambda; \\ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &= \|\boldsymbol{\delta}\|_2 \leq 2.23 \|\boldsymbol{\delta}_{\mathcal{I}}\|_2 \\ &\leq 25 / \rho_-(s^* + k^*) \cdot a^{-2} \sqrt{s^*} \lambda. \end{aligned}$$

Finally, to show that Algorithm 1 indeed catches a stationary point, we note that the acceptance criterion of the Algorithm (Line 1) implies that $\phi(\boldsymbol{\beta}^{(1)}) \leq \phi(\boldsymbol{\beta}^{(0)})$ where $\phi(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$. Moreover, for $t = 2$, we also have $\phi(\boldsymbol{\beta}^{(2)}) \leq \max\{\phi(\boldsymbol{\beta}^{(0)}), \phi(\boldsymbol{\beta}^{(1)})\}$. By induction, we conclude that for all $t \geq 1$, $\phi(\boldsymbol{\beta}^{(t)}) \leq \phi(\boldsymbol{\beta}^{(0)})$. Therefore we have $\boldsymbol{\beta}^{(t)} \in \mathcal{C} := \big\{\boldsymbol{\beta} \in \mathbb{R}^d \colon \|\boldsymbol{\beta}\|_1 \leq \lambda^{-1} \cdot L(\boldsymbol{\beta}^{(0)}) + \|\boldsymbol{\beta}^{(0)}\|_1\big\}$. Since set $\mathcal{C}$ is compact and the loss function L is continuously differentiable, it is also Lipschitz on $\mathcal{C}$. Therefore, by the convergence result of in Theorem 1 of [51], we conclude that every accumulation point of Algorithm 1 is a stationary point of optimization problem (1.2). □

A.2. Proof of Theorem 2

In what follows, inspired by [39], we apply Fano's method to derive the minimax risk of estimation for the nonlinear regression model defined in (1.1).

Proof. Let $M = M(\delta_n)$ be the cardinality of a $2\delta_n$-packing set of $B_0(s)$ with respect to the ℓ_2-metric where δ_n will be specified later. We denote the elements of this packing set as $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^M\}$. For any estimator $\widehat{\boldsymbol{\beta}}$, let $\psi = \operatorname{argmin}_{i \leq M} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^i\|_2$, triangle inequality implies that

$$\begin{aligned} 2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^i\|_2 &\geq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^i\|_2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\psi\|_2 \\ &\geq \|\boldsymbol{\beta}^i - \boldsymbol{\beta}^\psi\|_2 \geq 2\delta_n \;\; \text{for } \; i \neq \psi. \end{aligned}$$

Thus we conclude that

$$\begin{aligned}\mathcal{R}_f^*(s,n,d) &\geq \inf_{\psi} \sup_{1\leq i\leq M} \delta_n^2 \cdot \mathbb{P}_{\boldsymbol{\beta}^i}(\psi \neq i) \\ &\geq \inf_{\psi} \delta_n^2 \cdot \mathbb{P}_{\boldsymbol{\beta}^U}(\psi \neq U),\end{aligned}$$

where U is uniform distributed over $\{1,\ldots,N\}$. We consider the following data-generating process: For a continuously differentiable function f with $f'(u) \in [a,b], \forall u \in \mathbb{R}$, we first sample a random variable U uniformly over $1,\ldots M$, then generate data $y_i = f(\mathbf{x}_i^\top \boldsymbol{\beta}^U) + \epsilon_i$. Fano's inequality implies that

$$\mathbb{P}(\psi \neq U) \geq 1 - \big[I(U; y_1,\ldots,y_n) + \log 2\big] \big/ \log N.$$

In what follows, we establish an upper bound for the mutual information $I(U; y_1,\ldots,y_n)$. For $s \in \{1,\ldots,d\}$, we define the high-dimensional sparse hypercube as $\mathcal{C}_o(s) \coloneqq \big\{\mathbf{v} \in \{0,1\}^d, \|\mathbf{v}\|_0 = s\big\}$. We define the Hamming distance on $\mathcal{C}_0(s)$ as $\rho_H(\mathbf{v},\mathbf{v}') = \sum_{i=1}^d \mathbb{1}\{v_i \neq v_i'\}$. The following lemma, obtained from [40], is an extension of the Varhsamov-Gilbert lemma to $\mathcal{C}_0(s)$.

Lemma 6 (Sparse Varshamov-Gilbert lemma)**.** For any two integers s and d satisfying $1 \leq s \leq d/8$, there exist $\mathbf{v}_1,\ldots,\mathbf{v}_M \in \{0,1\}^d$ with $\|\mathbf{v}_i\|_0 = s$ for $1 \leq i \leq M$ such that

$$\begin{aligned}\rho_H(\mathbf{v}_i,\mathbf{v}_j) \geq s/2 \text{ for all } i \neq j, \text{ and} \\ \log(M) \geq s/8 \cdot \log[1 + d/(2s)].\end{aligned}$$

By Lemma 6 there exist $\mathcal{C}' \subset \mathcal{C}_0$ with $|\mathcal{C}'| \geq \exp\big\{s/8 \cdot \log[1+d/(2s)]\big\}$ such that $\rho_H(\mathbf{v},\mathbf{v}') \geq s/2$ for all $\mathbf{v},\mathbf{v}' \in \mathcal{C}'$. Then for $\boldsymbol{\beta},\boldsymbol{\beta}' \in \mathcal{C} \coloneqq \delta_n \cdot \sqrt{2/s} \cdot \mathcal{C}'$, we have

$$\begin{aligned}\delta_n^2 \cdot 2/s \cdot \rho_H(\boldsymbol{\beta},\boldsymbol{\beta}') &\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 \\ &\leq 2(\|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\beta}'\|_2^2) \leq 8\delta_n^2,\end{aligned}$$

which implies that $\delta_n^2 \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 \leq 8\delta_n^2$ for all $\boldsymbol{\beta},\boldsymbol{\beta}' \in \mathcal{C}$. By the convexity of mutual information, we have $I(U; y_1,\ldots,y_n) \leq M^{-2} \sum_{1\leq m,m'\leq M} D_{KL}(\boldsymbol{\beta}^m, \boldsymbol{\beta}^{m'})$. Since given $\boldsymbol{\beta}$ and f, $y_i \sim N\big(f(\mathbf{x}_i^\top \boldsymbol{\beta}), \sigma^2\big)$, direct computation yields that

$$\begin{aligned}&D_{KL}(\boldsymbol{\beta}^m, \boldsymbol{\beta}^{m'}) \\ &= 1/(2\sigma^2) \sum_{i=1}^n [f(\mathbf{x}_i^\top \boldsymbol{\beta}^m) - f(\mathbf{x}_i^\top \boldsymbol{\beta}^{m'})]. \qquad \text{(A.6)}\end{aligned}$$

By mean-value theorem, (A.6) can be bounded by

$$\begin{aligned}&D_{KL}(\boldsymbol{\beta}^m, \boldsymbol{\beta}^{m'}) \\ &\leq n \cdot b^2/(2\sigma^2)(\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'})^\top \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'}) \\ &\leq n \cdot b^2 \cdot \rho_+(2s)/(2\sigma^2)\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'}\|_2^2 \\ &\leq 4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2,\end{aligned}$$

where the second inequality follows from $\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'}\|_0 \leq 2s$. Therefore we conclude that $I(U; y_1,\ldots,y_n) \leq 4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2$, which yields that

$$\begin{aligned}\inf_{\psi} \mathbb{P}_{\boldsymbol{\beta}}(\psi \neq U) &\geq 1 - \frac{4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2 + \log 2}{\log M} \\ &\geq 1 - \frac{4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2 + \log 2}{s/8 \cdot \log[1 + d/(2s)]}.\end{aligned}$$

Setting $\delta_n^2 = \frac{\sigma^2 s \log[1+d/(2s)]}{96nb^2\rho_+(2s)}$, since $s \geq 4$ and $d \geq 8s$, we conclude that the right-hand side is no less than $1/2$. Now we obtain the following minimax lower bound

$$\mathcal{R}_f^*(s,n,d) \geq \frac{\sigma^2}{192b^2\rho_+(2s)} \frac{s\log[1+d/(2s)]}{n}.$$

This concludes the proof of Theorem 2. □

B. Proof of Auxiliary Results

In this appendix, we provide the proofs of the auxiliary lemmas appearing in the proof of the main results.

Proof of Lemma 3. By the definition of loss function L, for $j = 1,\ldots,d$, the j-th entry of $\nabla L(\boldsymbol{\beta}^*)$ can be written as $\nabla_j L(\boldsymbol{\beta}^*) = 1/n \cdot \sum_{i=1}^n \epsilon_i f'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) x_{ij}$. Recall that ϵ_i's are i.i.d. centered sub-Gaussian random variables with variance proxy σ^2. Thus conditioning on $\mathbf{x}_1,\ldots,\mathbf{x}_n$, $\nabla_j L(\boldsymbol{\beta}^*)$ is a centered sub-Gaussian random variable with variance proxy bounded by

$$\sigma^2 \cdot \frac{1}{n^2} \sum_{i=1}^n f'(\mathbf{x}_i^\top \boldsymbol{\beta}^*)^2 x_{ij}^2 \leq \sigma^2 \cdot b^2 \cdot \widehat{\Sigma}_{j,j}/n,$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$. Under Assumption *Bounded-Design*(D), the variance proxy of $\nabla_j L(\boldsymbol{\beta}^*)$ is bounded by $\sigma^2 \cdot b^2 \cdot D/n$. By the definition of variance proxy of sub-Gaussian random variables, we have

$$\begin{aligned}&\mathbb{P}\big(\big|\nabla_i \mathcal{L}(\boldsymbol{\beta}^*)\big| > \sigma \cdot b \cdot t \cdot \sqrt{D/n} \big| \mathbf{x}_1,\ldots,\mathbf{x}_n\big) \\ &\leq 2\exp(-t^2/2), \ \ \forall t > 0. \qquad \text{(B.1)}\end{aligned}$$

Taking a union bound over $j = 1,2,\ldots,d$ in for the left-hand side of (B.1) we obtain that

$$\begin{aligned}&\mathbb{P}\big(\big\|\nabla L(\boldsymbol{\beta}^*)\big\|_\infty > \sigma \cdot b \cdot t\sqrt{D/n} \big| \mathbf{x}_1,\ldots,\mathbf{x}_n\big) \\ &\leq 2\exp(-t^2/2 + \log d), \ \ \forall t > 0. \qquad \text{(B.2)}\end{aligned}$$

By choosing $t = C\sqrt{\log d}$ in (B.2) for a sufficiently large C, we conclude that there exist a constant $B = C \cdot b \cdot \sqrt{D} > 0$ such that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \leq B\sigma\sqrt{\log d/n}$ with probability at least $1 - \delta$, where we have $\delta \leq (2d)^{-1}$. □

Proof of Lemma 4. By the definition of $L(\boldsymbol{\beta})$, the gradient $\nabla L(\boldsymbol{\beta})$ is given by

$$\nabla L(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i^\top\boldsymbol{\beta})]f'(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i. \quad \text{(B.3)}$$

Hence for $\nabla L(\boldsymbol{\beta}^*)$, (B.3) can be reduced to

$$\nabla L(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f'(\mathbf{x}_i^\top\boldsymbol{\beta}^*)\mathbf{x}_i, \quad \text{(B.4)}$$

where $\epsilon_1, \ldots, \epsilon_n$ are n i.i.d. realizations of the random noise ϵ in (1.1). For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we denote $\boldsymbol{\eta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. Recalling that $y_i = f(\mathbf{x}_i^\top\boldsymbol{\beta}^*) + \epsilon_i$, Taylor expansion of (B.3) implies that

$$\begin{aligned}\nabla L(\boldsymbol{\beta}) = &-\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f'(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\ &+ \frac{1}{n}\sum_{i=1}^{n} f'(\mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}})f'(\mathbf{x}_i^\top\boldsymbol{\beta})(\mathbf{x}_i^\top\boldsymbol{\eta})\mathbf{x}_i, \quad \text{(B.5)}\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}$ lies on the line segment between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$. Combining (B.4) and (B.5) we have

$$\langle \nabla L(\boldsymbol{\beta}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle = A_1 + A_2, \quad \text{(B.6)}$$

where A_1 and A_2 are defined respectively as

$$\begin{aligned}A_1 &= \frac{1}{n}\sum_{i=1}^{n} f'(\mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}})f'(\mathbf{x}_i^\top\boldsymbol{\beta})(\mathbf{x}_i^\top\boldsymbol{\eta})^2 \text{ and} \\ A_2 &= \frac{1}{n}\sum_{i=1}^{n}\big\{f'(\mathbf{x}_i^\top\boldsymbol{\beta}^*) - f'(\mathbf{x}_i^\top\boldsymbol{\beta})\big\}(\mathbf{x}_i^\top\boldsymbol{\eta})\epsilon_i.\end{aligned}$$

By the boundedness of f', we can lower bound A_1 by

$$A_1 \geq a^2\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\boldsymbol{\eta})^2 = a^2\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}. \quad \text{(B.7)}$$

For the second part A_2, by the sub-Gaussianity of the random noise ϵ_i's,

$$\big\{f'(\mathbf{x}_i^\top\boldsymbol{\beta}^*) - f'(\mathbf{x}_i^\top\boldsymbol{\beta})\big\}\cdot(\mathbf{x}_i^\top\boldsymbol{\eta})\cdot\epsilon_i$$

is a centered sub-Gaussian random variable with variance proxy

$$\sigma^2[f'\big(\mathbf{x}_i^\top\boldsymbol{\beta}^*\big) - f'\big(\mathbf{x}_i^\top\boldsymbol{\beta}\big)]^2\cdot(\mathbf{x}_i^\top\boldsymbol{\eta})^2 \leq 4\sigma^2b^2(\mathbf{x}_i^\top\boldsymbol{\eta})^2.$$

Therefore we conclude that A_2 is centered and sub-Gaussian with variance proxy bounded by

$$4b^2n^{-2}\sigma^2\sum_{i=1}^{n}(\mathbf{x}_i^\top\boldsymbol{\eta})^2 = 4b^2\sigma^2n^{-1}\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}.$$

By the tail bound for sub-Gaussian random variables, we obtain that for any $x > 0$,

$$\mathbb{P}\big(|A_2| \geq x\big) \leq 2\exp(-x^2/C),$$

where $C = 8b^2\sigma^2n^{-1}\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}$. With probability at least $1 - (2d)^{-1}$, it holds that

$$\begin{aligned}A_2 &\geq \sqrt{C\cdot\log(4d)} \geq -3b\sigma\sqrt{\log d/n}\sqrt{\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}} \\ &\geq -3b\sigma\sqrt{D\log d/n}\|\boldsymbol{\eta}\|_1, \quad \text{(B.8)}\end{aligned}$$

where the last inequality is derived from Hölder's inequality $\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta} \leq \|\widehat{\boldsymbol{\Sigma}}\|_\infty\|\boldsymbol{\eta}\|_1^2 \leq D\|\boldsymbol{\eta}\|_1^2$. Therefore combining (B.6), (B.7) and (B.8) with probability at least $1 - (2d)^{-1}$, we have

$$\begin{aligned}&\langle \nabla L(\boldsymbol{\beta}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \\ &\geq a^2\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta} - 3b\sigma\sqrt{D\log d/n}\|\boldsymbol{\eta}\|_1.\end{aligned}$$

This concludes the proof of Lemma 4. □

Proof of Lemma 5. Recall that $\mathcal{J}$ is the set of indices of the largest k^* entries of $\boldsymbol{\eta}_{\mathcal{S}^c}$ in absolute value and let $\mathcal{I} = \mathcal{J}\cup\mathcal{S}$. The following Lemma establishes a lower-bound on $\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}$.

Lemma 7. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix and $\rho_-(k)$ and $\rho_+(k)$ be its k-sparse eigenvalues. Suppose that for some integer s and k, $\rho_-(s+2k) > 0$. For any $\mathbf{v} \in \mathbb{R}^d$, let $\mathcal{F}$ be any index set of size $d - s$, that is, $|\mathcal{F}^c| = s$. We let $\mathcal{J}$ be the set of indices of the k largest component of $\mathbf{v}_{\mathcal{F}^c}$ in absolute value and let $\mathcal{I} = \mathcal{F}^c\cup\mathcal{J}$. Then we have

$$\begin{aligned}\mathbf{v}^\top\boldsymbol{\Sigma}\mathbf{v} \geq &\rho_-(s+k)\cdot\Big[\|\mathbf{v}_{\mathcal{I}}\|_2 - \\ &\sqrt{\rho_+(k)/\rho_-(s+2k) - 1}\cdot\|\mathbf{v}_{\mathcal{F}}\|_1/\sqrt{k}\Big]\cdot\|\mathbf{v}_{\mathcal{I}}\|_2.\end{aligned}$$

By Assumption 1, $\rho_-(s^* + 2k^*) > 0$. Combining Lemma 7 with $\mathcal{F} = \mathcal{S}^c$ and that $\|\boldsymbol{\eta}_{\mathcal{S}^c}\|_1 \leq \gamma\|\boldsymbol{\eta}_{\mathcal{S}}\|_1 \leq \gamma\sqrt{s}\|\boldsymbol{\eta}_{\mathcal{S}}\|_2$ together yield inequality (6.5).

For the second part of the lemma, by the definition of $\mathcal{J}$ we obtain that

$$\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_\infty \leq \|\boldsymbol{\eta}_{\mathcal{J}}\|_1/k^* \leq \|\boldsymbol{\eta}_{\mathcal{S}^c}\|_1/k^* \leq \gamma/k^*\|\boldsymbol{\eta}_{\mathcal{S}}\|_1,$$

hence by Hölder's inequality we have

$$\begin{aligned}\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_2 &\leq \|\boldsymbol{\eta}_{\mathcal{I}^c}\|_1^{1/2}\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_\infty^{1/2} \\ &\leq (\gamma/k^*)^{1/2}\|\boldsymbol{\eta}_{\mathcal{S}}\|_1^{1/2}\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_1^{1/2} \\ &\leq \gamma k^{*-1/2}\cdot\|\boldsymbol{\eta}_{\mathcal{S}}\|_1,\end{aligned}$$

where we use the fact that $\mathcal{I}^c \subset \mathcal{S}^c$. Thus it holds that

$$\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_2 \leq \gamma\sqrt{s^*/k^*}\cdot\|\boldsymbol{\eta}_{\mathcal{S}}\|_2 \leq \gamma\cdot\|\boldsymbol{\eta}_{\mathcal{I}}\|_2 \text{ and} \quad \text{(B.9)}$$

$$\big\|\boldsymbol{\eta}\big\|_2 \leq (1+\gamma)\cdot\|\boldsymbol{\eta}_{\mathcal{I}}\|_2. \quad \text{(B.10)}$$

Thus we conclude the proof of Lemma 5. □

Proof of Lemma 7. Without loss of generality, we assume that $\mathcal{F}^c = \{1, \ldots, s\}$. We also assume that for $\mathbf{v} \in \mathbb{R}^d$,when $j > s$, v_j is arranged in descending order of $|v_j|$. That is, we rearrange the components of $\mathbf{v}$ such that $|v_j| \geq |v_{j+1}|$ for all j greater than s. Let $\mathcal{J}_0 = \{1, \ldots, s\}$ and $\mathcal{J}_i = \{s + (i-1)k + 1, \ldots, \min(s + ik, d)\}$ for $i \geq 1$. By definition, we have $\mathcal{J} = \mathcal{J}_1$ and $\mathcal{I} = \mathcal{J}_0 \cup J_1$. Moreover, we have $\|\mathbf{v}_{\mathcal{J}_i}\|_\infty \leq \|\mathbf{v}_{\mathcal{J}_{i-1}}\|_1/k$ when $i \geq 2$ because of the descending order of $|v_j|$ for $j > s$. Then we further have $\sum_{i \geq 2} \|\mathbf{v}_{\mathcal{J}_i}\|_\infty \leq \|\mathbf{v}_{\mathcal{F}}\|_1/k$.

We define the restricted correlation coefficients of $\mathbf{\Sigma}$ as

$$\pi(s,k) \coloneqq \sup\Bigg\{ \frac{\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{J}} \|\mathbf{v}_{\mathcal{I}}\|_2}{\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}} \|\mathbf{v}_{\mathcal{J}}\|_\infty} : I \cap \mathcal{J} = \emptyset, \\ |\mathcal{I}| \leq s, |\mathcal{J}| \leq k, \mathbf{v} \in \mathbb{R}^d \Bigg\}.$$

As shown in [56], if $\rho_-(s+k) > 0$ we have

$$\pi(s,k) \leq \frac{\sqrt{k}}{2} \cdot \sqrt{\rho_+(k)/\rho_-(s+k) - 1}. \tag{B.11}$$

Then by the definition of $\pi(s+k,k)$ we obtain

$$\big|\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{J}_i}\big| \leq \pi(s+k,k) \cdot \big(\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \cdot \|\mathbf{v}_{\mathcal{J}_i}\|_\infty / \|\mathbf{v}_{\mathcal{I}}\|_2.$$

Thus we have the following upper bound for $\big|\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}^c}\big|$:

$$\begin{aligned} \big|\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}^c}\big| &\leq \sum_{i \geq 2} \big|\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{J}_i}\big| \\ &\leq \pi(s+k,k) \cdot \|\mathbf{v}_{\mathcal{I}}\|_2^{-1} \big(\mathbf{v}_I^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \sum_{i \geq 2} \|\mathbf{v}_{\mathcal{J}_i}\|_\infty \\ &\leq \pi(s+k,k) \cdot \|\mathbf{v}_{\mathcal{I}}\|_2^{-1} \big(\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \|\mathbf{v}_{\mathcal{F}}\|_1/k. \end{aligned} \tag{B.12}$$

Because $\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} \geq \mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}} + 2\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}^c}$, by (B.12) we have

$$\begin{aligned} &\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} \\ &\quad \geq \mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}} - 2\pi(s+k,k) \|\mathbf{v}_I\|_2^{-1} \big(\mathbf{v}_I^\top \mathbf{\Sigma} \mathbf{v}_I\big) \|\mathbf{v}_{\mathcal{F}}\|_1/k \\ &\quad = \big(\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \big[1 - 2\pi(s+k,k) \|\mathbf{v}_{\mathcal{I}}\|_2^{-1} \|\mathbf{v}_{\mathcal{F}}\|_1/k\big]. \end{aligned} \tag{B.13}$$

Combining (B.13), the fact that $\mathbf{v}_{\mathcal{I}}^\top \mathbf{\Sigma} \mathbf{v}_{\mathcal{I}} \geq \rho_-(s+k) \cdot \|\mathbf{v}_{\mathcal{I}}\|_2^2$ and (B.11) for $\pi(s+k,k)$, we conclude the proof of Lemma 7. □