High-dimensional Non-Gaussian Single Index Models via Thresholded Score Function Estimation

Zhuoran Yang [1] **Krishnakumar Balasubramanian** [1] **Han Liu** [1]

Abstract

We consider estimating the parametric component of single index models in high dimensions. Compared with existing work, we do not require the covariate to be normally distributed. Utilizing Stein's Lemma, we propose estimators based on the score function of the covariate. Moreover, to handle score function and response variables that are heavy-tailed, our estimators are constructed via carefully thresholding their empirical counterparts. Under a bounded fourth moment condition, we establish optimal statistical rates of convergence for the proposed estimators. Extensive numerical experiments are provided to back up our theory.

1. Introduction

Estimators for high-dimensional parametric (linear) models have been developed and analyzed extensively in the last two decades (see for example (Bühlmann & van de Geer, 2011; Vershynin, 2015) for comprehensive overviews). While being a useful testbed for illustrating conceptual phenomenon, they often suffer from a lack of flexibility in modeling real-world situations. On the other hand, completely nonparametric models, although flexible, suffer from the curse of dimensionality unless restrictive additive sparsity or smoothness assumptions are imposed (Ravikumar et al., 2009; Yuan et al., 2016). An interesting compromise between the parametric and nonparametric models is provided by the so-called semiparametric index models (Horowitz, 2009). Here, the response and the covariate are linked through a low-dimensional nonparametric function that takes in as input a linear transformation of the covariate. The nonparametric component is also called as the link function and the linear components are called as the indices.

In this work, we focus on the simplest family of such models, the single index models (SIMs), which assume that the response Y and the covariate X satisfy $Y = f(\langle X, \beta^* \rangle) + \epsilon$, where β^* is the true signal, ϵ is the mean-zero random noise, and f is a univariate link function. (see §2 for the precise definition). They form the basis of more complicated models such as Multiple Index Models (MIMs) (Diaconis & Shahshahani, 1984) and Deep Neural Networks (DNNs) (LeCun et al., 2015), which are cascades of MIMs. Moreover, we focus on the task of estimating the parametric (linear) component β^* without the knowledge of the nonparametric part f in the high-dimensional setting, where the number of samples is much smaller than the dimensionality of β^*.

Estimating the parametric component without depending on the specific form of the nonparametric part appears naturally in several situations. For example, in one-bit compressed sensing (Boufounos & Baraniuk, 2008) and sparse generalized linear models (Loh & Wainwright, 2015), we are interested in recovering the true signal vector based on nonlinear measurements. Furthermore, in a DNN, the activation function is pre-specified and the task is to estimate the linear components, which are used for prediction in the test stage. Performing nonlinear least-squares in this setting, leads to nonconvex optimization problems that are invariably sub-optimal without further assumptions. Hence, developing estimators for the linear component that are both statistically accurate and computationally efficient for a class of activation functions provide a compelling alternative. Understanding such estimators for SIMs is hence crucial for understanding the more complicated DNNs.

Although SIMs appear to be a simple extension of the standard linear models, most existing work in the high-dimensional setting assume X follows a Gaussian distribution for estimating β^* without the knowledge of the nonparametric part. It is not clear whether those estimation methods are still valid and optimal when X is drawn from a more general class of distributions. To relax the Gaussian assumption, we study the setting where the distribution of X is non-Gaussian but known a priori.

[1] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. Correspondence to: Han Liu <hanliu@princeton.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

1.1. Challenges of the Single Index Models

There are significant challenges that appear when we are dealing with estimators for SIMs. They can be summarized as assumptions on either the link function or the data distribution (for example, non-Gaussian assumption).

1. **Knowledge of link function**: Suppose the link function is known, for example, $f(u) = u^2$ which corresponds to the phase retrieval model (see (Jaganathan et al., 2015) for a survey and history of this model). Then using an M-estimator to estimate β^* is a natural procedure (Jaganathan et al., 2015). But computationally the problem becomes nonconvex and one need to resort to either SDP based convex relaxations that are computationally expensive or do non-convex alternating minimization that require Gaussian assumptions on the data for successful initialization in the high-dimensional setting (Cai et al., 2015). Furthermore, if the link function is changed, it might become challenging or impossible to obtain provably computable estimators.

2. **Knowledge of data distribution**: Now suppose we want to be agnostic about the link function, i.e., we want to estimate the linear component for a general class of link functions. Then it becomes necessary to make assumptions about the distribution from which the covariates are sampled from. In particular, assuming the covariate has Gaussian and symmetric elliptical distributions respectively, (Plan & Vershynin, 2016) and (Goldstein et al., 2016) propose estimators in the high-dimensional setting for a large class of unknown link functions.

As mentioned previously, our estimators are based on Stein's Lemma for non-Gaussian distributions, which utilizes the score function. Estimating with the score function is challenging due to their heavy tails. In order to illustrate that, consider the univariate histograms provided in Figure-1. The dark shaded, more concentrated one corresponds to the histogram of 10000 i.i.d. samples from Gamma distribution with scale and shape parameters set to 5 and 0.2 respectively. The transparent histogram corresponds to the distribution of the score function of the same Gamma distribution. Note that even when the actual Gamma distribution is well concentrated, the distribution of the corresponding score function is well-spread and heavy-tailed. In the high dimensional setting, in order to estimate with the score functions, we require certain vectors or matrices based on the score functions to be well-concentrated in appropriate norms. In order to achieve that, we construct robust estimators via careful truncation arguments to balance the bias (due to thresholding)-variance (of the estimator) tradeoff and achieve the required concentration.

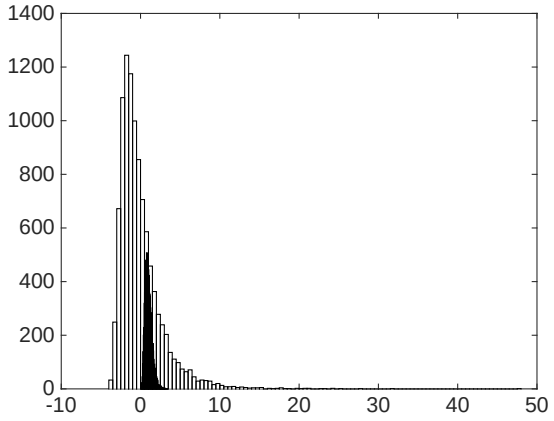

Figure 1: Histogram of Score Function based on 10000 independent samples from the Gamma distribution with shape 5 and scale 0.2. The dark histogram (we recommend the reader to zoom in to notice it) concentrated around zero corresponds to the Gamma distribution and the transparent histogram corresponds to the distribution of the score of the same Gamma distribution.

1.2. Related Work

There is a significant body of work on SIMs in the low-dimensional setting. They are based on assumptions on either the distribution of the covariate or the link functions. Assuming a monotonic link function, (Han, 1987; Sherman, 1993) propose the maximum rank correlation estimator exploiting the relationship between monotonic functions and rank-correlations. Furthermore, (Li & Duan, 1989) propose an estimator for a wide class of unknown link functions under the assumption that the covariate follows a symmetric elliptical distribution. This assumption is restrictive as often times the covariates are not from a symmetric distribution. For example, in several economic applications where the covariates are usually highly skewed and heavy-tailed (Horowitz, 2009). A line of work for estimation in SIMs is proposed by Ker-Chau Li which is based on sliced inverse regression (Li, 1991) and principal Hessian directions (Li, 1992) . These estimators are based on similar symmetry assumptions and involve computing second-order (conditional and unconditional) moments which are difficult to estimate in high-dimensions without restrictive assumptions.

The success of Lasso and related linear estimators in high-dimensions (Bühlmann & van de Geer, 2011), also enabled the exploration of high-dimensional SIMs. Although, this is very much work in progress. As mentioned previously, (Plan & Vershynin, 2016) show that the Lasso estimator works for the SIMs in high dimensions when the data is Gaussian. A more tighter albeit an asymptotic results under the same setting was proved in (Thrampoulidis et al., 2015). Very recently (Goldstein et al., 2016) extend

the results of (Li & Duan, 1989) to the high dimensional setting but it suffers from similar problems as mentioned in the low-dimensional setting. For the case of monotone nonparametric component, (Yang et al., 2015) analyze a non-convex least squares approach under the assumption that the data is sub-Gaussian. However, the success of their method hinges on the knowledge of the link function. Furthermore, (Jiang & Liu, 2014; Lin et al., 2015; Zhu et al., 2006) analyze the sliced inverse regression estimator in the high-dimensional setting concentrating mainly on support recovery and consistency properties. Similar to the low-dimensional case, the assumptions made on the covariate distribution restrict them from several real-world applications involving non-Gaussian or non-symmetric covariate, for example high-dimensional problems in economics (Fan et al., 2011). Furthermore, several results are established on a case-by-case basis for fixed link function. Specifically (Boufounos & Baraniuk, 2008; Ai et al., 2014) and (Davenport et al., 2014) consider 1-bit compressed sensing and matrix completion respectively, where the link is assumed to be the sign function. Also, (Waldspurger et al., 2015) and (Cai et al., 2015) propose and analyze convex and non-convex estimators for phase retrieval respectively, in which the link is the square function. All the above works, except (Ai et al., 2014) make Gaussian assumptions on the data and are specialized for the specific link functions. The non-asymptotic result obtained in (Ai et al., 2014) is under sub-Gaussian assumptions, but the estimator is not consistent. Finally, there is a line of work focussing on estimating both the parametric and the nonparametric component (Kalai & Sastry, 2009; Kakade et al., 2011; Alquier & Biau, 2013; Radchenko, 2015). We do not focus on this situation in this paper as mentioned before.

To summarize, all the above works require restrictive assumption on either the data distribution or on the link function. We propose and analyze an estimator for a class of (unknown) link functions for the case when the covariates are drawn from a non-Gaussian distribution – under the assumption that we know the distribution *a priori*. Note that in several situations, one could fit specialized distributions, to real-world data that is often times skewed and heavy-tailed, so that it provides a good generative model of the data. Also, mixture of Gaussian distribution, with the number of components selected appropriately, approximates the set of all square integrable distributions to arbitrary accuracy (see for example (McLachlan & Peel, 2004)). Furthermore, since this is a density estimation problem it is unlabeled and there is no issue of label scarcity. Hence it is possible to get accurate estimate of the distribution in most situations of interest. Thus our work is complementary to the existing literature and provides an estimator for a class of models that is not addressed in the previous works. We conclude this section with a summary of our main contributions in this paper:

- We propose estimators for the parametric component of a sparse SIM and low-rank SIM for a class of unknown link function under the assumption that the covariate distribution is non-Gaussian but known a priori.
- We show that it is possible to recover a s-sparse d-dimensional vector and a rank-r, $d_1 \times d_2$ dimensional matrix with number of samples of the order of $s \log d$ and $r(d_1 + d_2) \log(d_1 + d_2)$ respectively under significantly mild moment assumptions in the SIM setting.
- We provide numerical simulation results that confirm our theoretical predictions.

2. Single Index Models

In this section, we introduce the notation and define the single index models. Throughout this work, we use $[n]$ to denote the set $\{1, \dots, n\}$. In addition, for a vector $v \in \mathbb{R}^d$, we denote by $\|v\|_p$ the ℓ_p-norm of v for any $p \geq 1$. We use $\mathcal{S}^{d-1}$ to denote the unit sphere in $\mathbb{R}^d$, which is defined as $\mathcal{S}^{d-1} = \{v \in \mathbb{R}^d \colon \|v\|_2 = 1\}$. In addition, we define the support of $v \in \mathbb{R}^d$ as $\text{supp}(v) = \{j \in [d], v_j \neq 0\}$. Moreover, we denote the nuclear norm, operator norm, and Frobenius norm of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ by $\|\cdot\|_\star$, $\|\cdot\|_{\text{op}}$, and $\|\cdot\|_{\text{fro}}$, respectively. We denote by $\text{vec}(A)$ the vectorization of matrix A, which is a vector in $\mathbb{R}^{d_1 \cdot d_2}$. For two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ we define the trace inner product as $\langle A, B \rangle = \text{Trace}(A^\top B)$. Note that it can be viewed as the standard inner product between $\text{vec}(A)$ and $\text{vec}(B)$. In addition, for an univariate function $g \colon \mathbb{R} \to \mathbb{R}$, we denote by $g \circ (v)$ and $g \circ (A)$ the output of applying g to each element of a vector v and a matrix A, respectively. Finally, for a random variable $X \in \mathbb{R}$ with density p, we use $p^{\otimes d} \colon \mathbb{R}^d \to \mathbb{R}$ to denote the joint density of $\{X_1, \cdots, X_d\}$, which are d identical copies of X.

Now we are ready to define the statistical model. Let $f \colon \mathbb{R} \to \mathbb{R}$ be an univariate function and β^* be the parameter of interest, which is a structured vector or a matrix. The single index model in general is formulated as

$$Y = f(\langle X, \beta^* \rangle) + \epsilon, \tag{2.1}$$

where X is the covariate, $Y \in \mathbb{R}$ is the response, and ϵ is the exogenous noise that is independent of X. We assume that ϵ *is centered and has bounded fourth moment, i.e.,* $\mathbb{E}_{p_0}(\epsilon) = 0$ *and* $\mathbb{E}(\epsilon^4) \leq C$ *for an absolute constant* $C > 0$. Note in particular that this allows for heavy-tailed noise as well. In addition, we assume that the entries of X are i.i.d. random variables with density p_0. This assumption could be further relaxed using more sophisticated concentration arguments; here we focus on the i.i.d. setting to clearly present the main message of this paper.

Let $\{(Y_i, X_i)\}_{i=1}^n$ be n i.i.d. observations of the SIM. Our goal is to consistently estimate β^* without the knowledge of f. In particular, we focus on the case when β^* is either sparse or low-rank, which are defined as follows.

Sparse single index model: In this setting, we assume that $\beta^* = (\beta_1^*, \cdots, \beta_d^*)^\top$ is a sparse vector in $\mathbb{R}^d$ with s^* nonzero entries, such that $s^* \ll n \ll d$. Moreover, for the model in (2.1) to be identifiable, we further assume β^* lies on the unit sphere $\mathcal{S}^{d-1}$ as the norm of β^* can always be absorbed in the unknown link function f.

Low-rank single index model: In this setting, we assume that $\beta^* \in \mathbb{R}^{d_1 \times d_2}$ has rank $r^* \ll \min\{d_1, d_2\}$. In this scenario, $X \in \mathbb{R}^{d_1 \times d_2}$ and the inner product in (2.1) is $\langle X, \beta^* \rangle = \text{Trace}(X^\top \beta^*)$. For model identifiability, we further assume that $\|\beta^*\|_F = 1$, similar to the sparse case.

3. Estimation via Score Functions

Our estimator is primarily motivated by an interesting phenomenon illustrated in (Plan & Vershynin, 2016) for the Gaussian setting. Below, we first briefly summarize the result from (Plan & Vershynin, 2016) and then provide our *alternative justification* for the same result via Stein's Lemma. We mainly leverage this alternative justification and propose our estimators for the more general setting we consider. Assuming for simplicity, we work in the one-dimensional setting and are given n i.i.d. samples from the SIM. Consider the least-squares estimator

$$\widehat{\beta}_{LS} = \underset{\beta \in \mathbb{R}}{\text{argmin}} \ \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \beta)^2 .$$

Note that the above estimator is the standard least-squares estimator assuming a linear model (i.e., identity link function). The surprising observation from (Plan & Vershynin, 2016) is that, under the *crucial* assumption that X is standard Gaussian, $\widehat{\beta}_{LS}$ is a good estimator of β^* (up to a scaling) even when the data is generated from the nonlinear SIM. The same holds true for the high-dimensional setting when the minimization is performed in an appropriately constrained norm-ball (for example, the ℓ_1-ball). Hence the theory developed for the linear setting could be leveraged to understand the performance in the SIM setting. Below, we give an alternative justification for the above estimator as an implication of Stein's Lemma in the Gaussian case, which is summarized as follows.

Proposition 3.1 (Gaussian Stein's Lemma (Stein, 1972))**.** Let $X \sim N(0, 1)$ and $g : \mathbb{R} \to \mathbb{R}$ be a continuos function such that $\mathbb{E}|g'(X)| \leq \infty$. Then we have $\mathbb{E}[g(X)X] = \mathbb{E}[g'(X)]$.

Note that in our context for SIMs, we have $\mathbb{E}[f'(X)] \propto \beta^*$ and $\mathbb{E}[f(X)X] = \mathbb{E}[Y \cdot X]$. Now consider the following estimator, which is based on performing least-squares on the sample version of the above proposition:

$$\widehat{\beta}_{SL} = \underset{\beta \in \mathbb{R}}{\text{argmin}} \ \frac{1}{n} \sum_{i=1}^n (Y_i X_i - \beta)^2$$

Note that $\widehat{\beta}_{LS}$ and $\widehat{\beta}_{SL}$ are the same estimators assuming $X \sim N(0, 1)$, as $n \to \infty$. This observation leads to an alternative interpretation of the estimator proposed by (Plan & Vershynin, 2016) via Stein's Lemma for Gaussian random variables. Thus it provides an alternative justification for why the linear least-squares estimator should work in the SIM setting. This observation naturally leads to leveraging non-Gaussian versions of Stein's Lemma for dealing with non-Gaussian covariates.

We now describe our estimator for the non-Gaussian setting based on the above observation. We first define the score function associate to a density. Let $p \colon \mathbb{R}^d \to \mathbb{R}$ be a probability density function defined on $\mathbb{R}^d$. The score function $S_p \colon \mathbb{R}^d \to \mathbb{R}$ associated to p is defined as

$$S_p(x) = -\nabla_x [\log p(x)] = -\nabla_x p(x)/p(x).$$

Note that in the above definition, the derivative is taken with respect to x. This is different from the more traditional definition of the score function where the density belongs to a parametrized family and the derivative is taken with respect to the parameters. In the rest of the paper to simplify the notation, we omit the subscript x from ∇_x. We also omit the subscript p from S_p when the underlying density p is clear from the context.

We now describe a version of Stein's Lemma that is applicable for non-Gaussian random variables. Note from the motivating example for the Gaussian case that while utilizing the Stein's Lemma for SIM estimation, assumptions on the function in Stein's Lemma translate directly to those on the link function in SIM. We now introduce a version of Stein's Lemma that applies to non-Gaussian random variables and for continuously differentiable functions from (Stein et al., 2004). A more general version of the Stein's Lemma that applies to a class of regular functions is available in (Stein et al., 2004). We assume continuously differentiable functions in the Stein's Lemma below as they cover a wide range of practical SIM such as generalized linear models and single-layer neural networks.

Lemma 3.2 (Non-Gaussian Stein's Lemma (Stein et al., 2004))**.** Let $g \colon \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable function and $X \in \mathbb{R}^d$ be a random vector with density $p \colon \mathbb{R}^d \to \mathbb{R}$, which is also continuously differentiable. Under the assumption that the expectations $\mathbb{E}[g(X) \cdot S(X)]$ and $\mathbb{E}[\nabla g(X)]$ are both well-defined, we have the follow-

ing generalized Stein's identity

$$\mathbb{E}[g(X)\cdot S(X)] = -\int_{\mathbb{R}^d} g(x)\cdot\nabla p(x)\mathrm{d}x$$
$$= \int_{\mathbb{R}^d} \nabla g(x)\cdot p(x)\mathrm{d}x = \mathbb{E}[\nabla g(X)]. \quad (3.1)$$

Recall that in the two single index models introduced in §2, X in (2.1) has i.i.d. entries with density p_0. To unify both the vector and matrix settings, in the low-rank SIM, we identify X with $\mathrm{vec}(X)\in\mathbb{R}^d$ where $d = d_1\cdot d_2$. In this case, X has density $p = p_0^{\otimes d}$ and the corresponding score function $S\colon \mathbb{R}^d\to\mathbb{R}^d$ is given by

$$S(x) = -\nabla\log p(x) = -\nabla p(x)/p(x) = s_0\circ(x), \quad (3.2)$$

where the univariate function $s_0 = p_0'/p_0$ is applied to each entry of x. Thus $S(X)$ has i.i.d. entries. In addition, by Lemma 3.2, we have $\mathbb{E}[S(X)] = 0$ by setting g to be a constant function in (3.1). Moreover, in the context of SIMs specified in (2.1), we have

$$\mathbb{E}[Y\cdot S(X)] = \mathbb{E}\big[f(\langle X,\beta^*\rangle)\cdot S(X)\big]$$
$$= \mathbb{E}[f'(\langle X,\beta^*\rangle)]\cdot\beta^*,$$

as long as the density and the link function satisfy the conditions stated in Lemma 3.2. This implies that optimization problem

$$\underset{\beta\in\mathbb{R}^d}{\mathrm{minimize}}\big\{\langle\beta,\beta\rangle - 2\mathbb{E}[Y\cdot\langle S(X),\beta\rangle]\big\} \quad (3.3)$$

has solution $\beta = \mu\cdot\beta^*$, where $\mu = \mathbb{E}[f'(\langle X,\beta^*\rangle)]$. Hence the above program could be used to obtain the unknown β^* as long as $\mu\neq 0$. Before we proceed to describe the sample version of the above program, we make the following brief remark. The requirement $\mu\neq 0$ rules out in particular the use of our approach for non-Gaussian phase retrieval (where $f(u) = u^2$) as in that case we have $\mu = 0$ when X is centered. But we emphasize that the same holds true in the Gaussian and elliptical setting as well, as noted in (Plan & Vershynin, 2016) and (Goldstein et al., 2016). Their methods also fail to recover the true β^* when the SIM model corresponds to phase retrieval. We refer the reader to §6 for a discussion on overcoming this limitation.

Finally, we use a sample version of the above program as an estimator for the unknown β^*. In order to deal with the high-dimensional setting, we consider a regularized version of the above formulation. More specifically, we use the ℓ_1-norm and nuclear norm regularization in the vector and matrix settings respectively. However, a major difficulty in the sample setting for this procedure is that $\mathbb{E}[Y\cdot S(X)]$ and its empirical counterpart may not be close enough due to a lack of concentration. Recall our discussion from §1.1 that even if the random variable X is light-tailed, its score-function $S(x)$ might be arbitrarily heavy-tailed. Furthermore, bounded-fourth moment assumption on the noise, Y too can be heavy-tailed. Thus the naive method of using the sample version of (3.3) to estimate β^* leads to sub-optimal statistical rates of convergence.

To improve concentration and obtain optimal rates of convergence, we replace $Y\cdot S(X)$ with a transformed random variable $\mathcal{T}(Y,X)$, which will be defined precisely in §4 for the sparse and low-rank cases. In particular, $\mathcal{T}(Y,X)$ is a carefully truncated version of $Y\cdot S(X)$, introduced and analyzed in (Catoni et al., 2012; Fan et al., 2016) for related problems, that enables us to obtain well-concentrated estimators. Thus our final estimator $\widehat{\beta}$ is defined as the solution to the following regularized optimization problem

$$\underset{\beta\in\mathbb{R}^d}{\mathrm{minimize}}\ L(\beta) + \lambda\cdot R(\beta), \quad (3.4)$$
$$L(\beta) = \langle\beta,\beta\rangle - \frac{2}{n}\sum_{i=1}^{n}\langle\mathcal{T}(Y_i,X_i),\beta\rangle,$$

where $\lambda>0$ is the regularization parameter which will be specified later and $R(\cdot)$ is the ℓ_1-norm in the vector case and the nuclear norm in the matrix case.

4. Theoretical Results

In this section, we state our main results in Theorem 4.2 and Theorem 4.3,which establish the statistical rates of convergence of the estimator defined in §3. The proof for both Theorems is presented in the supplementary material. Before doing so, we introduce our main moment assumption for the single index model. This assumption is made apart from the assumptions made on the noise and the link function in §2 and §3 respectively. Recall that each entry of the score function defined in (3.2) is equal to $s_0(u) = -p_0'(u)/p_0(u)$. We first state the assumption and make a few remarks about it.

Assumption 4.1. There exists an absolute constant $M > 0$ such that $\mathbb{E}(Y^4)\le M$ and $\mathbb{E}_{p_0}[s_0^4(U)]\le M$, where random variable $U\in\mathbb{R}$ has density p_0.

Consider the assumption $\mathbb{E}(Y^4)\le M$. By Cauchy-Schwarz inequality we have $\mathbb{E}(Y^4)\le 4\mathbb{E}(\epsilon^4) + 4\mathbb{E}[f^4(\langle X,\beta^*\rangle)]$. Note that we assum ϵ to be centered, independent of X and has bounded fourth moment (see §2). If the covariate X has bounded fourth moment along the direction of true parameter, since $f(\cdot)$ is continuously differentiable, $f(\langle X,\beta^*\rangle)$ has bounded fourth moment as well if $f(\cdot)$ is defined on a compact subset of $\mathbb{R}$. Hence the condition $\mathbb{E}(Y^4)\le M$ is relatively easy to satisfy and significantly milder than assuming that Y is bounded or has lighter tails. Furthermore, $\mathbb{E}_{p_0}[s_0^4(U)]\le M$ is relatively mild and satisfied by a wide class of random variables. Specifically random variables that are non-symmetric and non-Gaussian satisfy this property thereby allowing our approach to work with covariates not previously possible.

We believe it is highly non-trivial to weaken this condition without losing significantly in the rates of convergence that we discuss below.

4.1. Sparse Single Index Model

Under the above assumptions, we first state our theorem on the sparse SIM. As discussed in §3, $Y \cdot S(X)$ can by heavy-tailed and hence we apply truncation to achieve concentration. Denote the j-th entry of the score function S in (3.2) as $S_j \colon \mathbb{R}^d \to \mathbb{R}$, $j \in [d]$. We define the truncated response and score function as

$$\widetilde{Y} = \text{sign}(Y) \cdot (|Y| \wedge \tau),$$
$$\widetilde{S}_j(x) = \text{sign}[S_j(x)] \cdot \big[|S_j(x)| \wedge \tau\big], \tag{4.1}$$

where $\tau > 0$ is a predetermined threshold value. We define $\widetilde{Y}_i$ similarly for all Y_i, $i \in [n]$. Then we define the estimator $\widehat{\beta}$ as the solution to the optimization problem in (3.4) with $\mathcal{T}(Y_i, X_i) = \widetilde{Y}_i \cdot \widetilde{S}(X_i)$ and $R(\beta) = \|\beta\|_1$. Here we apply elementwise truncation in $\mathcal{T}$ to ensure the sample average of $\mathcal{T}$ converges to $\mathbb{E}[Y \cdot S(X)]$ in the ℓ_∞-norm for an appropriately chosen τ. Note that the ℓ_∞-norm is the dual norm of the ℓ_1-norm. Such a convergence requirement in the dual norm is standard in the analysis of regularized M-estimators (Negahban et al., 2012) to achieve optimal rates. The following theorem characterizes the convergence rates of $\widehat{\beta}$.

Theorem 4.2 (Signal recovery for the sparse single index model)**.** For the sparse SIM defined in §2, we assume that $\beta^* \in \mathbb{R}^d$ has s^* nonzero entries. Under Assumption 4.1, we let $\tau = 2(M \cdot \log d/n)^{1/4}$ in (4.1) and set the regularization parameter λ in (3.4) as $C\sqrt{M \cdot \log d/n}$, where $C > 0$ is an absolute constant. Then with probability at least $1 - d^{-2}$, the ℓ_1-regularized estimator $\widehat{\beta}$ defined in (3.4) satisfies

$$\|\widehat{\beta} - \mu\beta^*\|_2 \le \sqrt{s^*} \cdot \lambda, \;\; \|\widehat{\beta} - \mu\beta^*\|_1 \le 4s^* \cdot \lambda.$$

From this theorem, the ℓ_1- and ℓ_2-convergence rates of $\widehat{\beta}$ are $\|\widehat{\beta} - \mu\beta^*\|_1 = \mathcal{O}(s^*\sqrt{\log d/n})$ and $\|\widehat{\beta} - \mu\beta^*\|_2 = \mathcal{O}(\sqrt{s^* \log d/n})$, respectively. These rates match the convergence rates of sparse generalized linear models (Loh & Wainwright, 2015) and sparse single index models with Gaussian and symmetric elliptical covariates (Plan & Vershynin, 2016; Goldstein et al., 2016) which are known to be minimax-optimal for this problem via matching lower bounds.

4.2. Low-rank Single Index Model

We next state our theorem for the low-rank SIM. In this case, we apply the nuclear norm regularization to promote low-rankness. Note that by definition, $\mathcal{T}$ is matrix-valued. Since the dual norm of the nuclear norm is the operator norm, we need the sample average of $\mathcal{T}$ to converge to $\mathbb{E}[Y \cdot S(X)]$ in the operator norm rapidly to achieve optimal rates of convergence. To achieve such a goal, we leverage the truncation argument from (Catoni et al., 2012; Minsker, 2016) to construct $\mathcal{T}(Y, X)$.

Let $\phi \colon \mathbb{R} \to \mathbb{R}$ be a non-decreasing function such that

$$-\log(1 - x + x^2/2) \le \phi(x) \le \log(1 + x + x^2/2), \;\; \forall x \in \mathbb{R}.$$

Based on ϕ, we define a linear mapping $\psi \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ as follows. For any $A \in \mathbb{R}^{d_1 \times d_2}$, let

$$\widetilde{A} = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$$

and let $\Upsilon \Lambda \Upsilon^\top$ be the eigenvalue decomposition of $\widetilde{A}$. In addition, let $B = \Upsilon[\phi \circ (\Lambda)]\Upsilon^\top$, where ψ is applied elementwisely on Λ. Then we write B in block from as

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

and define $\psi(A) = B_{12}$. Finally, we define $\mathcal{T}(Y, X) = 1/\kappa \cdot \psi\big[\kappa \cdot Y \cdot S(X)\big]$, where $\kappa > 0$ will be specified later. Therefore, our final estimator $\widehat{\beta} \in \mathbb{R}^{d_1 \times d_2}$ is defined as the solution to the optimization problem in (3.4) with $R(\beta) = \|\beta\|_\star$. We note here the minimization in (3.4) is taken over $\mathbb{R}^{d_1 \times d_2}$. The following theorem quantifies the convergence rates of the proposed estimator.

Theorem 4.3 (Signal recovery for the low-rank single index model)**.** For the low-rank single index model defined in §2, we assume that $\text{rank}(\beta^*) = r^*$. Under Assumption 4.1, we let

$$\kappa = \frac{2\sqrt{n \cdot \log(d_1 + d_2)}}{\sqrt{(d_1 + d_2)M}}$$

in $\mathcal{T}(Y, X)$. Moreover, the regularization parameter λ in (3.4) is set to $C\sqrt{M \cdot (d_1 + d_2) \cdot \log(d_1 + d_2)/n}$, where $C > 0$ is an absolute constant. Then with probability at least $1 - (d_1 + d_2)^{-2}$, the nuclear norm regularized estimator $\widehat{\beta}$ satisfies

$$\|\widehat{\beta} - \mu\beta^*\|_{\text{fro}} \le 3\sqrt{r^*} \cdot \lambda, \;\; \|\widehat{\beta} - \mu\beta^*\|_\star \le 12r^* \cdot \lambda.$$

By this theorem, we have $\|\widehat{\beta} - \mu\beta^*\|_{\text{fro}} = \mathcal{O}(\sqrt{r^*(d_1 + d_2) \cdot \log(d_1 + d_2)/n})$ and $\|\widehat{\beta} - \mu\beta^*\|_\star = \mathcal{O}(r^* \cdot \sqrt{(d_1 + d_2) \cdot \log(d_1 + d_2)/n})$. Note that the rate obtained is minimax-optimal up to a logarithmic factor. Furthermore, it matches the rates for low-rank single index models with Gaussian and symmetric elliptical distributions up to a logarithmic factor (Plan & Vershynin, 2016; Goldstein et al., 2016).

5. Numerical Experiments

We assess the finite sample performance of the proposed estimators on simulated data. Throughout this section, we let $\epsilon \sim N(0,1)$ and set the link function in (2.1) as one of $f_1(u) = 3u + 10\sin(u)$ and $f_2(u) = \sqrt{2}u + 4\exp(-2u^2)$, which are plotted in Figure 2. We set p_0 to be one of (i) Gamma distribution with shape parameter 5 and scale parameter 1, (ii) Student's t-distribution with 5 degrees of freedom, and (iii) Rayleigh distribution with scale parameter 2. To measure the estimation accuracy, we use the cosine distance

$$\cos\theta(\widehat{\beta}, \beta^*) = 1 - \|\widehat{\beta}\|_\bullet^{-1} |\langle \widehat{\beta}, \beta^* \rangle|,$$

where $\bullet$ stands for the Euclidean norm in the vector case and the Frobenius norm when β^* is a matrix. Here we report the cosine distance rather than $\|\widehat{\beta} - \mu\beta^*\|_\bullet$ to compare the performances for X having different distributions, where μ may have different values.

For the vector case, we fix $d = 2000$, $s^* = 5$ and vary n. The support of β^* is chosen uniformly random among all subsets of $\{1, \ldots, d\}$. For each $j \in \text{supp}(\beta^*)$, we set $\beta_j^* = 1/\sqrt{s^*} \cdot \gamma_j$, where each γ_j is an i.i.d. Rademacher random variable. In addition, the regularization parameter λ is set to $4\sqrt{\log d/n}$. We plot the cosine distance against the signal strength $\sqrt{s^* \log d/n}$ in Figure 4-(a) and (b) for f_1 and f_2 respectively, based on 200 independent trials for each n. As shown in this figure, the estimation error grows sublinearly as a function of the signal strength.

As for the matrix case, we fix $d_1 = d_2 = 20$, $r^* = 3$ and let n vary. The signal parameter β^* is equal to $USV^\top$, where $U, V \in \mathbb{R}^{d \times d}$ are random orthogonal matrices and S is a diagonal matrix with r^* nonzero entries. Moreover, we set the nonzero diagonal entries of S as $1/\sqrt{r^*}$, which implies $\|\beta^*\|_{\text{fro}} = 1$. We set the regularization parameter as $\lambda = 2\sqrt{(d_1 + d_2)\log(d_1 + d_2)/n}$. Furthermore, we use the proximal gradient descent algorithm (with the learning rate fixed to 0.05) to solve the nuclear norm regularization problem in (3.4). To present the result, we plot the cosine distant against the signal strength $\sqrt{r^*(d_1 + d_2)\log(d_1 + d_2)/n}$ in Figure 4-(b) based on 200 independent trials. As shown in this figure, the error is bounded by a linear function of the signal strength, which corroborates Theorem 4.3.

6. Conclusion

In this paper, we consider SIMs in the high-dimensional non-Gaussian setting and proposed estimators based on Stein's Lemma for a wider class of unknown link functions and covariate distributions. We consider both sparse and low-rank models and propose minimax rate-optimal estimators under fairly mild assumptions. An interesting avenue of future work is the problem of phase retrieval with non-Gaussian data. Our current approach requires that $\mu \neq 0$ which is not applicable. The main reason this happens is we use a first-order version of Stein's Lemma. Such a problem could overcome by second-order Stein's Lemma (Janzamin et al., 2014). Obtaining rate-optimal estimators based on second-order score functions require addressing several challenges. Concentrating on phase retrieval (and sparse phase retrieval) we plan to report our results for the above problem in the near future.

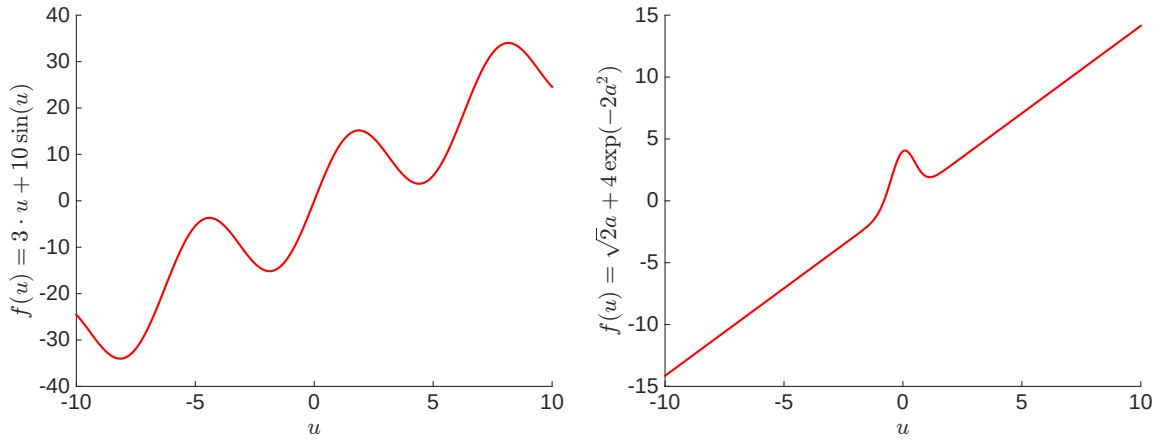

Figure 2: Plot of the link functions $f_1(u) = 3u + 10 \cdot \sin(u)$ (left) and $f_2(u) = \sqrt{2}u + 4\exp(-2u^2)$ (right). Both functions are nonlinear and not monotone.

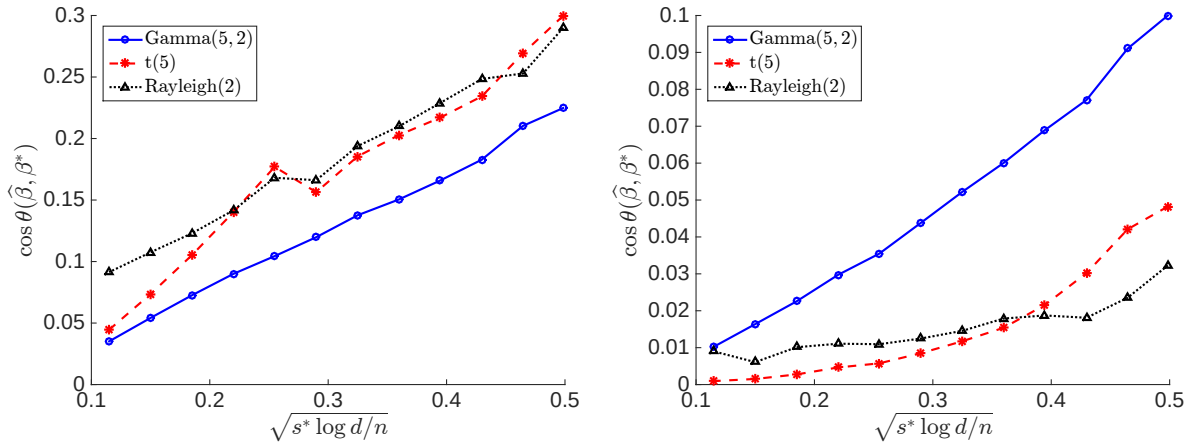

Figure 3: Cosine distances between the true parameter and the estimated parameter in the sparse SIM with the link function in 2.1 set to f_1 (left) and f_2 (right). Here we set $d_{=}2000$. $s^* = 5$ and vary n.

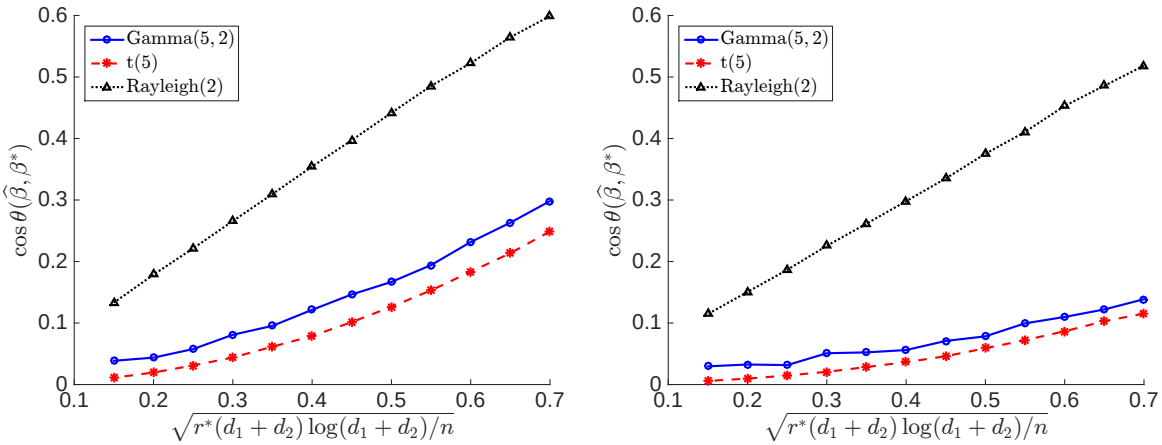

Figure 4: Cosine distances between the true parameter and the estimated parameter in the low-rank SIM with link function in 2.1 set to f_1 (left) and f_2 (right). Here we set $d_1 = d_2 = 20$. $r^* = 3$ and vary n.

References

Ai, Albert, Lapanowski, Alex, Plan, Yaniv, and Vershynin, Roman. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

Alquier, Pierre and Biau, Gérard. Sparse single-index model. *The Journal of Machine Learning Research*, 14 (1):243–280, 2013.

Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Boufounos, Petros T and Baraniuk, Richard G. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pp. 16–21. IEEE, 2008.

Bühlmann, Peter and van de Geer, Sara. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Cai, T Tony, Li, Xiaodong, and Ma, Zongming. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *arXiv preprint arXiv:1506.03382*, 2015.

Catoni, Olivier et al. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48 (4):1148–1185, 2012.

Davenport, Mark A, Plan, Yaniv, van den Berg, Ewout, and Wootters, Mary. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

Diaconis, P. and Shahshahani, M. On nonlinear functions of linear combinations. *SIAM Journal on Scientific and Statistical Computing*, 5(1):175–191, 1984.

Fan, J., Lv, J., and Qi, L. Sparse high-dimensional models in economics. *Annual review of economics*, 3(1):291–317, 2011.

Fan, Jianqing, Wang, Weichen, and Zhu, Ziwei. Robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*, 2016.

Goldstein, Larry, Minsker, Stanislav, and Wei, Xiaohan. Structured signal recovery from non-linear and heavy-tailed measurements. *arXiv preprint arXiv:1609.01025*, 2016.

Han, Aaron K. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

Horowitz, Joel L. *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer, 2009.

Jaganathan, Kishore, Eldar, Yonina C, and Hassibi, Babak. Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*, 2015.

Janzamin, Majid, Sedghi, Hanie, and Anandkumar, Anima. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

Jiang, B. and Liu, J. S. Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5):1751–1786, 2014.

Kakade, Sham M, Kanade, Varun, Shamir, Ohad, and Kalai, Adam. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pp. 927–935, 2011.

Kalai, Adam Tauman and Sastry, Ravi. The isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory*, 2009.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, Ker-Chau. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Li, Ker-Chau. 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.

Li, Ker-Chau and Duan, Naihua. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989.

Lin, Q., Zhao, Z., and Liu, J. S. On consistency and sparsity for sliced inverse regression in high dimensions. *arXiv preprint arXiv:1507.03895*, 2015.

Loh, Po-Ling and Wainwright, Martin J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

McLachlan, Geoffrey and Peel, David. *Finite mixture models*. John Wiley & Sons, 2004.

Minsker, Stanislav. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *arXiv preprint arXiv:1605.07129*, 2016.

Negahban, Sahand N., Ravikumar, Pradeep, Wainwright, Martin J., and Yu, Bin. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012.

Plan, Yaniv and Vershynin, Roman. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Radchenko, Peter. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

Ravikumar, Pradeep, Lafferty, John, Liu, Han, and Wasserman, Larry. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Sherman, Robert P. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, 61(1):123–137, 1993.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

Stein, Charles, Diaconis, Persi, Holmes, Susan, Reinert, Gesine, et al. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*. Institute of Mathematical Statistics, 2004.

Thrampoulidis, Christos, Abbasi, Ehsan, and Hassibi, Babak. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 2015.

Vershynin, Roman. Estimation in high dimensions: a geometric perspective. In *Sampling theory, a renaissance*, pp. 3–66. Springer, 2015.

Waldspurger, Irène, dAspremont, Alexandre, and Mallat, Stéphane. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

Yang, Zhuoran, Wang, Zhaoran, Liu, Han, Eldar, Yonina C, and Zhang, Tong. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *International Conference on Machine Learning*, 2015.

Yuan, Ming, Zhou, Ding-Xuan, et al. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

Zhu, Lixing, Miao, Baiqi, and Peng, Heng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474): 630–643, 2006.

A. Proofs of the Main Results

In this section, we lay out the proofs of the two theorems in §4, which establish the statistical rates of convergence of our estimators.

A.1. Proof of Theorem 4.2

Proof. Since $\widehat{\beta}$ is the solution of the optimization problem in (3.4), the first-order optimality condition states that

$$\nabla L(\widehat{\beta}) + \lambda\xi = 0, \;\; \text{where} \;\; \xi \in \partial\|\widehat{\beta}\|_1. \tag{A.1}$$

Then the entries of $\xi \in \mathbb{R}^d$ are given by

$$\xi_j = \text{sign}(\widehat{\beta}_j), \;\; \forall j \in \text{supp}(\widehat{\beta}); \;\; \xi_j \in [-1, 1], \;\; \forall j \notin \text{supp}(\widehat{\beta}).$$

For any index set $\mathcal{A} \subseteq [d]$ and $z \in \mathbb{R}^d$, we define the restriction of z to $\mathcal{A}$, $z_{\mathcal{A}} \in \mathbb{R}^d$, by letting

$$[z_{\mathcal{A}}]_j = z_j \;\; \text{if} \;\; j \in \mathcal{A}, \;\; [z_{\mathcal{A}}]_j = 0 \;\; \text{otherwise}.$$

Here $[z_{\mathcal{A}}]_j$ is the j-th entry of $z_{\mathcal{A}}$. Let $\mathcal{S} = \text{supp}(\beta^*)$, then we can write $\xi = \xi_{\mathcal{S}} + \xi_{\mathcal{S}^c}$. For notational simplicity, in the sequel, we define $\theta = \widehat{\beta} - \mu \cdot \beta^*$. Thus by (A.1) it holds that

$$\begin{aligned} \langle \nabla L(\widehat{\beta}) - \nabla L(\mu\beta^*), \theta \rangle &= \langle -\lambda \cdot \xi - \nabla L(\mu\beta^*), \theta \rangle \\ &\le \langle -\lambda \cdot \xi_{\mathcal{S}} - \lambda \cdot \xi_{\mathcal{S}^c}, \theta \rangle + \|\nabla L(\mu\beta^*)\|_\infty \cdot \|\theta\|_1. \end{aligned} \tag{A.2}$$

By the definition of ξ, we have

$$\langle -\lambda \cdot \xi_{\mathcal{S}^c}, \widehat{\beta} - \mu\beta^* \rangle = -\lambda \cdot \|\widehat{\beta}_{\mathcal{S}}\|_1. \tag{A.3}$$

Moreover, since $\|\xi\|_\infty \le 1$, Hölder's inequality implies that

$$\langle -\lambda \cdot \xi_{\mathcal{S}}, \theta \rangle \le \|\theta_{\mathcal{S}}\|_1. \tag{A.4}$$

Note that $\nabla^2 L(\beta) = 2I_d$. Combining (A.9), (A.3), and (A.4), we obtain

$$2\|\theta\|_2^2 = \langle \nabla L(\widehat{\beta}) - \nabla L(\mu\beta^*), \theta \rangle \le -\lambda\|\theta_{\mathcal{S}^c}\|_1 + \lambda\|\theta_{\mathcal{S}}\|_1 + \|\nabla L(\mu\beta^*)\|_\infty \cdot \|\theta\|_1. \tag{A.5}$$

For an upper bound of the right-hand side of (A.5), we apply the following lemma to obtain an upper bound on $\|\nabla L(\mu\beta^*)\|_\infty$.

Lemma A.1 (Bound on $\|\nabla L(\mu\beta^*)\|_\infty$)**.** We set the truncation level in (4.1) as $\tau = 2(M \cdot n/\log d)^{1/4}$. Then we have

$$\mathbb{P}\Big[\|\nabla L(\mu\beta^*)\|_\infty > 7\sqrt{M \cdot \log d/n}\Big] \le d^{-2}.$$

Proof. See §B.1 for a detailed proof. □

Thus by Lemma A.1 and the choice of λ, we have $\lambda > 2\|\nabla L(\mu\beta^*)\|_\infty$ with probability at least $1 - d^{-2}$. This implies that

$$2\|\theta\|_2^2 \le -\lambda/2 \cdot \|\theta_{\mathcal{S}^c}\|_1 + 3\lambda/2 \cdot \|\theta_{\mathcal{S}}\|_1 \le 2\lambda \cdot \|\theta_{\mathcal{S}}\|_1. \tag{A.6}$$

Since the leftmost term in (A.6) is nonnegative, we obtain $\|\theta_{\mathcal{S}^c}\|_1 \le 3 \cdot \|\theta_{\mathcal{S}}\|_1$. In addition, since $\mathcal{S}| = s^*$, $\|\theta_{\mathcal{S}}\|_1 \le \sqrt{s^*} \cdot \|\theta_{\mathcal{S}}\|_2$. Thus by (A.6) we have $\|\theta\|_2 \le \sqrt{s^*} \cdot \lambda$. Moreover, we also have $\|\theta_{\mathcal{S}}\|_1 \le s^*\lambda$, which further implies that

$$\|\theta\|_1 = \|\theta_{\mathcal{S}}\|_1 + \|\theta_{\mathcal{S}^c}\|_1 \le 4 \cdot \|\theta_{\mathcal{S}}\|_1 \le 4s^*\lambda.$$

Therefore, we conclude the proof. □

A.2. Proof of Theorem 4.3

Proof. The proof of Theorem 4.3 is parallel to that of Theorem 4.2. Here the difference is to handle the nuclear norm regularization, instead of the ℓ_1-penalty. Since $\widehat{\beta}$ is the solution of the optimization problem in (3.4), the first order optimality condition states that

$$L(\widehat{\beta}) + \lambda\|\widehat{\beta}\|_\star \le L(\mu\beta^*) + \lambda\|\mu\beta^*\|_\star. \tag{A.7}$$

To simplify the notation, we define $\Theta = \widehat{\beta} - \mu\cdot\beta^*$. Since L is quadratic,

$$L(\widehat{\beta}) - L(\mu\beta^*) = \langle \nabla L(\mu\beta^*), \Theta\rangle + 2\|\Theta\|_{\text{fro}}^2, \tag{A.8}$$

where ∇L takes values in $\mathbb{R}^{d_1\times d_2}$. Then combining (A.7), (A.8), and Hölder's inequality, we have

$$\|\Theta\|_{\text{fro}}^2 \le -\langle \nabla L(\mu\beta^*), \Theta\rangle + \lambda\|\mu\beta^*\|_\star - \lambda\|\widehat{\beta}\|_\star \le \big\|\nabla L(\mu\beta^*)\big\|_{\text{op}}\cdot\|\Theta\|_\star + \lambda\|\mu\beta^*\|_\star - \lambda\|\widehat{\beta}\|_\star. \tag{A.9}$$

In the following, we focus on the term $\|\mu\beta^*\|_\star - \|\widehat{\beta}\|_\star$ in (A.9). Let $U\Lambda^*V^\top$ be the singular value decomposition of $\mu\beta^*$, where $U\in\mathbb{R}^{d_1\times d_1}$ and $V\in\mathbb{R}^{d_2\times d_2}$ are orthogonal matrices, and $\Lambda^*\in\mathbb{R}^{d_1\times d_2}$ be formed by the singular values of $\mu\beta^*$. Moreover, since $\mathrm{rank}(\beta^*) = r^*$, Λ^* can be written in block form as

$$\Lambda^* = \begin{bmatrix}\Lambda_{11}^* & 0\\ 0 & 0\end{bmatrix}, \tag{A.10}$$

where $\Lambda_{11}^*\in\mathbb{R}^{r^*\times r^*}$ is a diagonal matrix whose diagonal elements are the nonzero singular values of $\mu\beta^*$. We define $\Gamma = U^\top\Theta V$, which can be written in block form as

$$\Gamma = \begin{bmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & \Gamma_{22}\end{bmatrix},$$

where $\Gamma_{11}\in\mathbb{R}^{r^*\times r^*}$. In addition, we define matrices

$$\Gamma^{(1)} = \begin{bmatrix}0 & 0\\ 0 & \Gamma_{22}\end{bmatrix} \text{ and } \Gamma^{(2)} = \begin{bmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & 0\end{bmatrix}.$$

Then by (A.10) and triangle inequality of the nuclear norm, we have

$$\begin{aligned}\|\widehat{\beta}\|_\star &= \|\mu\beta^* + \Theta\|_\star = \|U(\Lambda^* + \Gamma)V^\top\|_\star = \|\Lambda^* + \Gamma\|_\star\\ &\ge \|\Lambda^* + \Gamma^{(1)}\|_\star - \|\Gamma^{(2)}\|_\star = \|\Lambda^*\|_\star + \|\Gamma^{(1)}\|_\star - \|\Gamma^{(2)}\|_\star,\end{aligned} \tag{A.11}$$

where the last equality follows from the fact that $\Lambda^* + \Gamma^{(1)}$ is block diagonal. Since $\|\mu\beta^*\|_\star = \|\Lambda^*\|_\star$, by (A.11) we obtain

$$\|\mu\beta^*\|_\star - \|\widehat{\beta}\|_\star \le \|\Gamma^{(2)}\|_\star - \|\Gamma^{(1)}\|_\star. \tag{A.12}$$

In addition, triangle inequality implies that

$$\|\Theta\|_\star = \|U\Gamma V^\top\|_\star \le \|\Gamma^{(1)}\|_\star + \|\Gamma^{(2)}\|_\star. \tag{A.13}$$

Thus combining (A.11), (A.12), (A.13), we have

$$\|\Theta\|_{\text{fro}}^2 \le \big(\big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} + \lambda\big)\cdot\|\Gamma^{(2)}\|_\star + \big(\big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} - \lambda\big)\cdot\|\Gamma^{(1)}\|_\star, \tag{A.14}$$

We utilize the following lemma to obtain an upper bound of $\|\nabla L(\mu\beta^*)\|_{\text{op}}$.

Lemma A.2 (Upper bound of $\|\nabla L(\mu\beta^*)\|_{\text{op}}$)**.** Let loss function $L\colon \mathbb{R}^{d_1\times d_2}\to\mathbb{R}$ be defined in (3.4) for the matrix setting. Setting $\kappa = 2\sqrt{n\cdot\log(d_1+d_2)}/\sqrt{(d_1+d_2)M}$, then it holds that

$$\mathbb{P}\Big[\|\nabla L(\mu\beta^*)\|_{\text{op}} > 6\sqrt{(d_1+d_2)/n}\Big] \le (d_1+d_2)^{-2}.$$

Proof. See B.2 for a detailed proof. □

By Lemma A.2 and the choice of λ, we conclude that $\lambda > 2 \cdot \|\nabla L(\mu\beta^*)\|_{\text{op}}$ with probability at least $1 - (d_1 + d_2)^{-3}$. Thus by (A.14) we have

$$\|\Theta\|_{\text{fro}}^2 \leq 3\lambda/2 \cdot \|\Gamma^{(2)}\|_\star - \lambda/2 \cdot \|\Gamma^{(1)}\|_\star \tag{A.15}$$

which implies that $\|\Gamma^{(1)}\|_\star \leq 3 \cdot \|\Gamma^{(2)}\|_\star$. Moreover, by the subadditivity of rank, we obtain

$$\operatorname{rank}(\Gamma^{(2)}) \leq \operatorname{rank}\left(\begin{bmatrix} \Gamma_{11}/2 & \Gamma_{12} \\ 0 & 0 \end{bmatrix}\right) + \operatorname{rank}\left(\begin{bmatrix} \Gamma_{11}/2 & 0 \\ \Gamma_{21} & 0 \end{bmatrix}\right) = 2r^*,$$

which implies that $\|\Gamma^{(2)}\|_\star \leq \sqrt{2r^*} \cdot \|\Gamma^{(2)}\|_{\text{fro}}$ Then by (A.15) we obtain that $\|\Theta\|_{\text{fro}} \leq 3/\sqrt{2} \cdot \sqrt{r^*} \cdot \lambda$. Finally, by triangle inequality for the nuclear norm,

$$\|\Theta\|_\star = \|\Gamma\|_\star \leq \|\Gamma^{(1)}\|_\star + \|\Gamma^{(2)}\|_\star \leq 4 \cdot \|\Gamma^{(2)}\|_\star \leq 4\sqrt{2r^*}\|\Gamma^{(2)}\|_{\text{fro}} = 12r^*\lambda.$$

Thus we conclude the proof of Theorem 4.3.

□

B. Proof of Auxiliary Results

B.1. Proof of Lemma A.1

Proof. By definition of the loss function L in (3.4), we have

$$\nabla L(\mu\beta^*) = 2\mu\beta^* - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i) = \mathbb{E}\big[2Y_i \cdot S(X_i)\big] - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i).$$

By triangle inequality,

$$\|\nabla L(\mu\beta^*)\|_\infty \leq \Big\|\mathbb{E}\big[2Y \cdot S(X)\big] - \mathbb{E}\big[2\widetilde{Y} \cdot \widetilde{S}(X)\big]\Big\|_\infty + \Big\|\mathbb{E}\big[2\widetilde{Y} \cdot \widetilde{S}(X)\big] - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i)\Big\|_\infty. \tag{B.1}$$

For any $j \in [d]$, by the definition of the truncated response $\widetilde{Y}$ and truncated score $\widetilde{S}$, we obtain

$$\begin{aligned}
\big|\mathbb{E}\big[\widetilde{Y} \cdot \widetilde{S}_j(X)\big] - \mathbb{E}\big[Y \cdot S_j(X)\big]\big| &\leq \Big|\mathbb{E}\Big\{\widetilde{Y} \cdot \big[\widetilde{S}_j(X) - S_j(X)\big]\Big\}\Big| + \big|\mathbb{E}\big[(\widetilde{Y} - Y) \cdot S_j(X)\big]\big| \\
&= \underbrace{\big|\mathbb{E}\big[\widetilde{Y} \cdot S_j(X) \cdot \mathbb{1}\{|S_j(X)| > \tau\}\big]\big|}_{a_1} + \underbrace{\big|\mathbb{E}\big[Y \cdot S_j(X) \cdot \mathbb{1}\{|Y| > \tau\}\big]\big|}_{a_2}.
\end{aligned} \tag{B.2}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
a_1^2 &\leq \mathbb{E}\big[\widetilde{Y}^2 S_j^2(X)\big] \cdot \mathbb{P}\big[|S_j(X)| \geq \tau\big] \\
&\leq \sqrt{\mathbb{E}(\widetilde{Y}^4) \cdot \mathbb{E}\big[S_j^4(X)\big]} \cdot \mathbb{E}\big[S_j^4(X)\big] \cdot \tau^{-4} \\
&= M^2 \cdot \tau^{-4},
\end{aligned} \tag{B.3}$$

where the second inequality follows from Chebyshev's inequality. Similarly, for a_2 we have

$$\begin{aligned}
a_2^2 &\leq \mathbb{E}\big[Y^2 S_j^2(X)\big] \cdot \mathbb{P}\big(|Y| \geq \tau\big) \\
&\leq \sqrt{\mathbb{E}(\widetilde{Y}^4) \cdot \mathbb{E}\big[S_j^4(X)\big]} \cdot \mathbb{E}(Y^4) \cdot \tau^{-4} \\
&\leq M^2 \cdot \tau^{-4}.
\end{aligned} \tag{B.4}$$

Thus combining (B.2), (B.3), and (B.4), we conclude that

$$\Big|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\Big|\le a_1+a_2\le 2M\cdot\tau^{-2}$$

for all $j\in[d]$. Thus choosing $\tau=2(M\cdot n/\log d)^{1/4}$, we have

$$\Big\|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\Big\|_\infty\le 1/2\cdot\sqrt{M\cdot\log d/n}. \tag{B.5}$$

Furthermore, under Assumption 4.1, the variance of $\widetilde{Y}\cdot\cdot\widetilde{S}_j(X)$ is bounded by

$$\mathrm{Var}[\widetilde{Y}\cdot\widetilde{S}_j(X)]\le\mathbb{E}[\widetilde{Y}^2\cdot\widetilde{S}_j^2(X)]\le\mathbb{E}[Y^2\cdot S_j^2(X)]\le\sqrt{\mathbb{E}(Y^4)\cdot\mathbb{E}[S_j^4(X)]}\le M.$$

Thus for the second term in (B.1), since $|\widetilde{Y}\cdot\widetilde{S}_j(X)|\le\tau^2$, by the Bernstein inequality in (Boucheron et al., 2013) (Theorem 2.10), for any $j\in[d]$ and any $t>0$, we have

$$\mathbb{P}\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big|\ge\sqrt{\frac{2M\cdot t}{n}}+\frac{\tau^2\cdot t}{3n}\bigg\}\le\exp(-t). \tag{B.6}$$

Taking union bound over $j\in[t]$ in (B.6) yields

$$\mathbb{P}\bigg\{\Big\|\frac{1}{n}\sum_{i=1}^n\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big\|_\infty\ge\sqrt{\frac{2M\cdot t}{n}}+\frac{\tau^2\cdot t}{3n}\bigg\}\le\exp(-t+\log d). \tag{B.7}$$

Finally, we plug in $\tau=2(M\cdot n/\log d)^{1/4}$ and set $t=3\log d$ in (B.7) to obtain that

$$\Big\|\frac{1}{n}\sum_{i=1}^n\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big\|_\infty\le(4+\sqrt{6})\sqrt{\frac{M\cdot\log d}{n}} \tag{B.8}$$

with probability at least $1-d^{-2}$. Finally, combining (B.1), (B.5), and (B.8), we conclude the proof. □

B.2. Proof of Lemma A.2

Proof. For loss function L defined in (3.4) in the matrix setting, we have

$$\nabla L(\mu\beta^*)=2\mu\beta^*-\frac{2}{\kappa\cdot n}\sum_{i=1}^n\psi\big[\kappa\cdot Y_i\cdot S(X_i)\big]=2\mathbb{E}[Y\cdot S(X)]-\frac{2}{\kappa\cdot n}\sum_{i=1}^n\psi\big[\kappa\cdot Y_i\cdot S(X_i)\big]. \tag{B.9}$$

Here the last equality follows from the generalized Stein's identity. In the sequel, we apply results in (Minsker, 2016) to bound $\|\nabla L(\mu\beta^*)\|_{\mathrm{op}}$. To begin with, we first consider the operator norm of $\mathbb{E}[Y^2\cdot S(X)S(X)^\top]\in\mathbb{R}^{d_1\times d_2}$ and $\mathbb{E}[Y^2\cdot S(X)^\top S(X)]\in\mathbb{R}^{d_2\times d_2}$. For notational simplicity, we denote by $S_{j,\cdot}(\cdot)\in\mathbb{R}^{d_2}$ $S_{\cdot,k}(\cdot)\in\mathbb{R}^{d_1}$ the j-th row and k-the column of the score function $S(\cdot)$, respectively. For any $u\in\mathcal{S}^{d_1-1}$, by Cauchy-Schwarz inequality we have

$$\mathbb{E}[Y^2\cdot u^\top S(X)S(X)^\top u]=\sum_{k=1}^{d_2}\mathbb{E}\big\{[Y^2\cdot S_{\cdot,k}(X)^\top u]^2\big\}\le d_2\cdot\sqrt{\mathbb{E}(Y^4)\cdot\mathbb{E}\big\{[S_{\cdot,1}(X)^\top u]^4\big\}}, \tag{B.10}$$

where we use the fact that the entries of $S(X)$ are i.i.d. Since $\mathbb{E}[S_{ij}(X)]=0$ and $\mathbb{E}[S_{ij}^4(X)]\le M$, by Cauchy-Schwarz inequality we obtain that

$$\begin{aligned}\mathbb{E}\big\{[S_{\cdot,1}(X)^\top u]^4\big\}&=\sum_{j_1=1}^d\sum_{j_2=1}^d\mathbb{E}[S_{j_1,1}(X)^2\cdot S_{j_2,1}^2(X)]\cdot u_{j_i}^2u_{j_2}^2\\&\le\sum_{j_1=1}^d\sum_{j_2=1}^d\sqrt{\mathbb{E}[S_{j_1,1}^4(X)]\cdot\mathbb{E}[S_{j_2,1}^4(X)]}\cdot u_{j_i}^2u_{j_2}^2\le M\sum_{j_1=1}^d\sum_{j_2=1}^d u_{j_i}^2u_{j_2}^2=M.\end{aligned} \tag{B.11}$$

Thus combining (B.10) and (B.11) we obtain that

$$\mathbb{E}[Y^2 \cdot u^\top S(X)S(X)^\top u] \le d_2 \cdot M,$$

which implies that $\|\mathbb{E}[Y^2 \cdot S(X)S(X)^\top]\|_{\mathrm{op}} \le d_2 \cdot M$. Similarly, we obtain $\|\mathbb{E}[Y^2 \cdot S(X)^\top S(X)]\|_{\mathrm{op}} \le d_1 \cdot M$. Thus by Corollary 3.1 in (Minsker, 2016), we have

$$\mathbb{P}\Big\{\Big\|\frac{1}{\kappa \cdot n}\sum_{i=1}^n \psi\big[\kappa \cdot Y_i \cdot S(X_i)\big] - \mathbb{E}[Y \cdot S(X)]\Big\|_{\mathrm{op}} \ge \frac{t}{\sqrt{n}}\Big\} \le 2(d_1+d_2)\exp\big[-\kappa t\sqrt{n} + \kappa^2(d_1+d_2)M/2\big] \quad \text{(B.12)}$$

for any $t > 0$ and $\kappa > 0$. We set $\kappa = 2\sqrt{n \cdot \log(d_1+d_2)}/\sqrt{(d_1+d_2)M}$ and $t = \sqrt{(d_1+d_2)M} \cdot s$ in (B.12), which implies that

$$\begin{aligned}&\mathbb{P}\Big\{\Big\|\frac{1}{\kappa \cdot n}\sum_{i=1}^n \psi\big[\kappa \cdot Y_i \cdot S(X_i)\big] - \mathbb{E}[Y \cdot S(X)]\Big\|_{\mathrm{op}} \ge \sqrt{\frac{(d_1+d_2)M}{n}} \cdot s\Big\} \\ &\quad \le 2(d_1+d_2)\exp\big[-2\sqrt{\log(d_1+d_2)} \cdot s + 2 \cdot \log(d_1+d_2)\big]. \quad \text{(B.13)}\end{aligned}$$

Now we set $s = 3 \cdot \sqrt{\log(d_1+d_2)}$, which implies that the right-hand side of (B.13) is less than

$$2(d_1+d_2)\exp\big[-6\log(d_1+d_2) + 2 \cdot \log(d_1+d_2)\big] \le (d_1+d_2)^2 \cdot \exp\big[-4 \cdot \log(d_1+d_2)\big] = (d_1+d_2)^{-2}.$$

Therefore, combining (B.9) and (B.13) we conclude that

$$\|\nabla L(\mu\beta^*)\|_{\mathrm{op}} \le 6\sqrt{(d_1+d_2) \cdot M/n} \cdot \sqrt{\log(d_1+d_2)}$$

with probability at least $1 - (d_1+d_2)^{-2}$, which concludes the proof.

□