Convergence guarantee for the sparse monotone single index model

Ran Dai*, Hyebin Song†, Rina Foygel Barber‡, Garvesh Raskutti§

May 18, 2021

Abstract

We consider a high-dimensional monotone single index model (hdSIM), which is a semiparametric extension of a high-dimensional generalize linear model (hdGLM), where the link function is unknown, but constrained with monotone and non-decreasing shape. We develop a scalable projection-based iterative approach, the "Sparse Orthogonal Descent Single-Index Model" (SOD-SIM), which alternates between sparse-thresholded orthogonalized "gradient-like" steps and isotonic regression steps to recover the coefficient vector. Our main contribution is that we provide finite sample estimation bounds for both the coefficient vector and the link function in high-dimensional settings under very mild assumptions on the design matrix $\mathbf{X}$, the error term ϵ, and their dependence. The convergence rate for the link function matched the low-dimensional isotonic regression minimax rate up to some poly-log terms ($n^{-1/3}$). The convergence rate for the coefficients is also $n^{-1/3}$ up to some poly-log terms. This method can be applied to many real data problems, including GLMs with misspecified link, classification with mislabeled data, and classification with positive-unlabeled (PU) data. We study the performance of this method via both numerical studies and also an application on a rocker protein sequence data.

1 Introduction

Single index models (SIMs) provide a semi-parametric extension of linear models, where a scalar response variable $Y \in \mathbb{R}$ is related to a predictor vector $X \in \mathbb{R}^p$ via

$$\mathbb{E}[Y|X] = g_\star(X^\top u_\star) \tag{1}$$

for some unknown parameter vector $u_\star$ and unknown link function $g_\star$.

*Department of Biostatistics, University of Nebraska Medical Center
†Department of Statistics, The Pennsylvania State University
‡Department of Statistics, University of Chicago
§Department of Statistics, University of Wisconsin-Madison

In this paper, we consider a shape-constrained single index model (1) where $g_\star$ belongs to a class of monotonic (but not necessarily smooth) functions and the parameter $u_\star$ is high-dimensional and sparse (with sparsity level $s_\star$). We develop a scalable algorithm and provide theoretical guarantees for high-dimensional single index models. Prior theoretical guarantees for the high-dimensional single index model rely on the distribution of X being known or symmetric (see e.g. [28, 38]). In a number of settings, such assumptions on X are not satisfied (in section 1.1.3 we show X follows a non-symmetric mixture distribution) and prior theoretical guarantees do not apply. In this paper, we provide a scalable algorithm with theoretical guarantees for the high-dimensional single index model under more general design assumptions that include non-symmetric distributions.

We consider a general case where mild assumptions are made on the design matrix X, the error $\epsilon := Y - \mathbb{E}[Y \mid X]$, and their interaction. In particular, X can be deterministic or random. If X is random, the distribution of X can be asymmetric. The distribution of ϵ may depend on X. In addition we assume that the unknown vector $u_\star$ is sparse and high-dimensional. As $g_\star$ is not assumed to be known a priori, the model (1) provides a more flexible model specification than a parametric model, while still circumventing the curse-of-dimensionality by avoiding a p-dimensional non-parametric function estimation. Two estimation problems exist in single index model framework: estimation of the unknown vector $u_\star$ and of the 1-dimensional regression function $g_\star$. Both are addressed in this paper.

To begin, we provide three motivating examples for the hdSIMs in (1). We see that each example reduces to the estimation problem of $u_\star$ or $g_\star$, or both under the model (1).

1.1 Examples

1.1.1 Example 1: Mis-specified generalized linear models (GLMs)

Generalized linear models (GLMs) are parametric extensions of linear models, which includes many popular models such as logistic and Poisson regression. GLMs relate the conditional expectation of Y given X via a link function ψ, such that $\mathbb{E}[Y \mid X] = \psi^{-1}(X^\top u_\star)$, where a link function ψ is assumed to be known. The parameter of interest $u_\star$ is usually estimated via an iterative procedure using the knowledge of ψ; therefore, when ψ is misspecified, an output is inevitably biased for $u_\star$, even in an asymptotic sense. Using the single index model framework, we produce an estimate of $u_\star$ without particular link function specification.

1.1.2 Example 2: Classification with corrupted labels

Another interesting example involves classification with corrupted labels. In particular, instead of observing a sample $(X, \tilde{Y}) \in \mathbb{R}^p \times \{0, 1\}$ where $\tilde{Y}$ may follow, e.g., a logistic model, we observe the pair $(X, Y) \in \mathbb{R}^p \times \{0, 1\}$ where Y is a systematically corrupted version of $\tilde{Y}$. In particular $\tilde{Y}$ is related to X via $\mathbb{E}[\tilde{Y}|X] = \tilde{g}(X^\top u_\star)$ for some unknown parameter vector $u_\star \in \mathbb{R}^p$ and a monotonic function $\tilde{g}$, which is potentially unknown.

We assume the corrupted sample Y is independent of X given $\tilde{Y}$, which corresponds to a missing-completely-at-random assumption. The corruption structure is completely specified by corruption probabilities $\rho_1 = \mathbb{P}(Y = 0|\tilde{Y} = 1)$ (the probability that a

positive label is corrupted) and $\rho_0 = \mathbb{P}(Y = 1|\tilde{Y} = 0)$ (the probability that a negative label is corrupted), assuming $\rho_0 + \rho_1 < 1$ for identifiability of the model. Under these assumptions, $\mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$ is given by

$$\begin{aligned}\mathbb{P}(Y = 1|X) &= \tilde{g}(X^\top u_\star)\mathbb{P}(Y = 1|\tilde{Y} = 1) + (1 - \tilde{g}(X^\top u_\star))\mathbb{P}(Y = 1|\tilde{Y} = 0) \\ &= \tilde{g}(X^\top u_\star)(1 - \rho_1) + (1 - \tilde{g}(X^\top u_\star))\rho_0 \\ &= g_\star(X^\top u_\star) \end{aligned} \tag{2}$$

where we let $g_\star(s) := \tilde{g}(s)(1 - \rho_1) + (1 - \tilde{g}(s))\rho_0$. (Note $g_\star$ is monotonic since $\tilde{g}$ is monotonic and $\rho_0 + \rho_1 < 1$ by assumption.)

In many practical situations the corruption probabilities ρ_0 and ρ_1 may be unknown. In such cases, even with the assumption of a specific $\tilde{g}$ (such as a sigmoid function $\tilde{g}(u) = 1/(1 + \exp(-u))$ in the case of a logistic model for $\tilde{Y}$), with the unknown parameters ρ_0 and ρ_1, the new link function $g_\star$ is still unknown, which motivates moving beyond the parametric model setting.

1.1.3 Example 3: Classification with positive-unlabeled (PU) data and unknown prevalence π

In various applications such as ecology, text-mining, and biochemistry, negative samples are often hard or even impossible to be obtained. In positive-unlabeled (PU) learning, we use positive and unlabeled samples, where positive samples are taken from the sub-population where responses are known to be positive, and unlabeled samples are random samples from the population.

To make this concrete, we begin with a distribution $\mathcal{D}$ on data $(\tilde{X}, \tilde{Y}) \in \mathbb{R}^p \times \{0, 1\}$. Here $\tilde{X}$ is a feature vector while $\tilde{Y} \in \{0, 1\}$ is a true label, "positive" (1) or "negative" (0). In some applications, it is not possible to observe (negatively) labeled data; instead, we observe unlabeled data (i.e., a feature vector X, drawn from the marginal distribution of $\tilde{X}$ under the joint distribution $\mathcal{D}$), and a subsample of positively labeled data (i.e., a feature vector X drawn from the conditional distribution of $\tilde{X} \mid \tilde{Y} = 1$, derived from the joint distribution $\mathcal{D}$). The new "label" $Y \in \{0, 1\}$ then specifies whether X was drawn as an unlabeled sample ($Y = 0$) or as a positive-only draw ($Y = 1$). Writing γ as the ratio of positive-only samples to unlabeled samples, the conditional distribution $Y \mid X$ can be calculated as

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{\gamma\pi^{-1}}{\frac{1}{\mathbb{P}(\tilde{Y}=1|\tilde{X}=x)} + \gamma\pi^{-1}},$$

where $\pi = \mathbb{P}(\tilde{Y} = 1)$ is the proportion of positive labels under the original distribution $\mathcal{D}$ on (X, Y). Therefore, we can see that if the distribution of $\tilde{Y} \mid \tilde{X}$, in the underlying population, satisfies the monotone single-index model, then the same is true for the new PU data—specifically, if $\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X} = x) = \tilde{g}(x^\top u_\star)$ for some monotone function $\tilde{g}$ and some vector $u_\star$, then

$$\mathbb{P}(Y = 1 \mid X = x) = g_\star(x^\top u_\star)$$

for

$$g_\star(v) = \frac{\gamma\pi^{-1}}{\frac{1}{\tilde{g}(v)} + \gamma\pi^{-1}},$$

which is also a monotone function, inheriting this property from $\tilde{g}$.

In [33], this problem is addressed with a parametric approach, by assuming that $\tilde{g}$ is known (commonly the sigmoid function, i.e., assuming a logistic model for $\tilde{Y} \mid \tilde{X}$), and assuming that the proportion $\pi = \mathbb{P}(\tilde{Y} = 1)$ of positive labels in the underlying population is known as well. In settings where π is unknown, however, this approach can no longer be applied, even if $\tilde{g}$ is known; reliable estimation of such proportion is very challenging and heavily relies on model specification [16] and can lead to unreliable inference. In contrast, assuming only that $g_\star$ is a monotone function allows for both unknown π and unknown (monotone) $\tilde{g}$.

Another interesting aspect of this PU problem is that the distribution of X is, in general, not symmetric. Even in a setting where the distribution of features in the population (i.e., the marginal distribution of $\tilde{X}$) can plausibly be assumed to be symmetric, this will no longer be the case for X—this is because the marginal distribution of X is a mixture between the marginal distribution of $\tilde{X}$, and the conditional distribution of $\tilde{X} \mid \tilde{Y} = 1$. Therefore, this setting cannot be addressed with existing methods that rely on a symmetric design assumption.

1.2 Prior work

The single index model has been an active area in statistics for a number of decades, since its introduction in econometrics and statistics [19, 17, 23]. There is extensive literature about estimation of the index vector $u_\star$, which can be broadly classified into two categories: methods which directly estimate the index vector while avoiding estimation of a "nuisance" infinite-dimensional function, and methods which involve estimation of both the infinite-dimensional function and the index vector.

Methods of the first type usually require strong design assumption such as Gaussian or an elliptically symmetric distribution. Duan and Li [13] proposed an estimator of $u_\star$ using sliced inverse regression in the context of sufficient dimension reduction. $\sqrt{n}$ consistency and asymptotically normality of the estimator are established in the low-dimensional regime, under the elliptically symmetrical design assumption. This result is extended in Lin et al. [25] in high-dimensional regime under similar design assumptions. Plan et al. [28] studied sparse recovery of the index vector $u_\star$ when X is Gaussian, and established a non-asymptotic $\mathcal{O}((s \log(2p/s)/n)^{1/2})$ mean-squared error bound, where s is a sparsity level, i.e. $s := \|u_\star\|_0$. Ai et al. [1] studied the error bound of the same estimator when the design is not Gaussian and showed that the estimator is biased and the bias depends on the size of $\|u_\star\|_\infty$. The proposed estimator in Plan et al. [28] is generalized in several directions; Yang et al. [37] propose an estimator based on the score function of the covariate, which relies on prior knowledge of covariate distribution. Chen and Banerjee [8] proposed estimators based on U-statistics which also include Plan et al. [28] as a special case with standard Gaussian design assumption.

Methods of the second type usually do not require strong design assumption, and our method also falls into this category. One popular approach is via M-estimation or solving estimation equations. Under smoothness assumptions of $g_\star$, methods for a link function estimation were proposed via Kernel estimation [19, 15, 23, 12, 10], local polynomials [6, 9], or splines [39, 24]. An estimator for the index vector is then obtained by minimizing certain criterion function such as squared loss [19, 15, 39] or solving an estimating equation [9, 10]. $\sqrt{n}$ consistency and asymptotic normality of those proposed estimators were established for those estimators in the low-dimensional regime.

Another approach is the average derivative method [29, 18], which takes advantage of the fact $\frac{d}{dx}\mathbb{E}[Y \mid X = x] = u_\star g'(x^\top u_\star)$. An estimator is then obtained through a non-parametric estimation of $\frac{d}{dx}\mathbb{E}[Y \mid X = x]$, also under smoothness assumption of $g_\star$. There are also studies for single index model in high-dimensional regime, using PAC-Bayesian approach [2] or incorporating penalties [35, 30].

There have been an increasing number of studies where the link function is restricted to have a particular shape, such as a monotone shape, but is not necessarily a smooth function. Kalai and Sastry [22] and Kakade et al. [21] investigated single index problem in low-dimensional regime under monotonic and Lipschitz continuous link assumption, and proposed iterative perceptron-type algorithms with prediction error guarantees. In particular, an isotonic regression is used for the estimation of $g_\star$ then update $u_\star$ based on estimated $g_\star$. Ganti et al. [14] extended [22, 21] to the high-dimensional setting via incorporating a projection step onto a sparse subset of the parameter space. They empirically demonstrated good predictive performance of their algorithms, but no theoretical guarantee were provided. Balabdaoui et al. [3] study the least square estimators under monotonicity constraint and establish $n^{-1/3}$ consistency of the least-square estimator. Also in Balabdaoui et al. [4], assuming a smooth link function $g_\star$, a $\sqrt{n}$ consistent estimator for the index vector $u_\star$, based on solving a score function, is proposed.

Our work focuses on estimation of the index vector in high-dimensional regime where an unknown function is assumed to be Lipschitz-continuous and monotonic. We provide an efficient algorithm via iterative hard-thresholding and a non-asymptotic ℓ_2 and mean function error bound. Unlike Ganti et al. [14] and Kakade et al. [21], our algorithm does not require to input a Lipschitz constant of the unknown function.

1.3 Our contributions

Our major contributions can be summarized as follows:

- Develop a scalable projection-based iterative approach, the "Sparse Orthogonal Descent Single-Index Model" (SOD-SIM) algorithm, which alternates between sparse-thresholded orthogonalized "gradient-like" steps and isotonic regression steps to recover the coefficient vector.
- Provide finite sample convergence guarantees for the SOD-SIM algorithm, both in estimating the parameter vector $u_\star$ and the mean function $g_\star(X^\top u_\star)$ that the error scales as $\mathcal{O}\big(\frac{s_\star^{1/2}\log np}{n^{1/3}}\big)$, where $s_\star$ is the sparsity level of $u_\star$. Note that the minimax rate for low-dimensional isotonic regression is $n^{-1/3}$ and our algorithm achieves this rate up to log factors. This rate also matches the $n^{-1/3}$ rate shown for least square estimator in [3]. However, the algorithm to obtain the least square estimator is computationally intensive even in low dimensional cases. Whereas in our work, we simultaneously showed the statistical and algorithmic convergence of a computationally efficient algorithm, which can also handle high dimensional case. In estimating $u_\star$, with X being elliptically symmetrically distributed, the link-free slicing regression estimates by Duan and Li [13] has been shown to be $\sqrt{n}$ consistent with asymptotic normal distribution. When $g_\star$ is bounded and continuously differentiable, the score estimates proposed by Balabdaoui et al. [4] has been shown to be $\sqrt{n}$ consistent with asymptotic normal distribution. In

comparison, our convergence result is not asymptotic and has milder assumptions on X and $g_\star$; our algorithm is also much more computationally efficient than the score estimator.

- Finally we provide empirical study results based on simulated and real data, which support our theoretical findings and also shows that our algorithm performs well compared with existing approaches.

2 Main results

Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be the feature vectors and $Y_1, \ldots, Y_n \in \mathbb{R}$ the response values,[1] which we model as

$$Y_i = g_\star(X_i^\top u_\star) + Z_i,$$

where $g_\star$ is monotone non-decreasing; $u_\star$ is a $s_\star$-sparse unit vector; and $Z_i \in \mathbb{R}$ denotes the noise term.

We write $\mathbf{X} \in \mathbb{R}^{n\times p}$ to denote the matrix with rows $X_i, i \in \{1, \cdots, n\}$, and $\mathbf{Y} \in \mathbb{R}^n$ and $\mathbf{Z} \in \mathbb{R}^n$ to denote the vectors with entries Y_i and Z_i, respectively. For any function $g : \mathbb{R} \to \mathbb{R}$ and any $v \in \mathbb{R}^p$, we write $g(\mathbf{X}v)$ to mean that g is applied elementwise to the vector $\mathbf{X}v$, i.e. $g(\mathbf{X}v) = \big(g(X_1^\top v), \ldots, g(X_n^\top v)\big)$.

In the remaining of this section, we first present our SOD-SIM algorithm. Then we present the finite sample convergence results for the estimation of $u_\star$ and $g_\star$.

2.1 Algorithm

Before defining our algorithm, we first define notations for the hard-thresholding operator $\Psi_s(\cdot)$ and the isotonic regression operator $\mathrm{iso}(\cdot)$. First let $\Psi_s : \mathbb{R}^p \to \mathbb{R}^p$ be the "hard-thresholding operator" at sparsity level s, i.e.,

$$\Psi_s(v)_i = \begin{cases} v_i & i \in S, \\ 0 & i \notin S, \end{cases}$$

where $S \subseteq \{1, \cdots, p\}$ indexes the s largest-magnitude entries of v (ties may be broken with any mechanism). Next, for any vectors $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^n$, define

$$\mathrm{iso}_{\mathbf{X}u}(v) = \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{ \|v - w\|_2^2 : w_i \le w_j \text{ whenever } (\mathbf{X}u)_i \le (\mathbf{X}u)_j \right\},$$

i.e. the isotonic regression of v onto $\mathbf{X}u$. This convex problem can be solved efficiently using pool-adjacent-violators algorithm (PAVA) [11].

With these definitions in place, we are ready to define our algorithm for solving the sparse single index model.[2] Given a target sparsity level $s \ge 1$ and a step size $\eta > 0$, the algorithm alternates between taking an orthogonalized gradient-like step (with step size η), and hard-thresholding to enforce s-sparsity. The steps of the algorithm are shown in Algorithm 1.

[1] We denote Y_i, $i = 1, \cdots, n$ as continuous variables for simplicity. In practice, they can be in other data types (continuous/categorical/mixed).

[2] Implicitly, the initialization step in our algorithm assumes that $\mathbf{X}^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n) \neq 0$. In our proofs, we will verify that this holds with high probability.

Algorithm 1 Sparse Orthogonal Descent Single-Index Model (SOD-SIM)

Initialize:

$$u_0 = \frac{\Psi_s\big(\mathbf{X}^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n)\big)}{\|\Psi_s\big(\mathbf{X}^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n)\big)\|_2}, \tag{3}$$

where $\bar{Y} = \frac{1}{n}\sum_i Y_i$ and $\mathbf{1}_n$ is the vector of 1's.

for $t = 1, 2, \ldots$ **do**

- Compute $\mathrm{iso}_{\mathbf{X}u_{t-1}}(\mathbf{Y})$, the isotonic regression of $\mathbf{Y}$ onto $\mathbf{X}u_{t-1}$, using PAVA.
- Take an orthogonal gradient-like step,

$$\tilde{u}_t = u_{t-1} + \eta \cdot \mathcal{P}^{\perp}_{u_{t-1}}\left(n^{-1}\mathbf{X}^\top\big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u_{t-1}}(\mathbf{Y})\big)\right).$$

- Enforce sparsity and unit norm,

$$u_t = \frac{\Psi_s(\tilde{u}_t)}{\|\Psi_s(\tilde{u}_t)\|_2}.$$

until some convergence criterion is reached.

For our theoretical guarantee to hold, we will see that the target sparsity level s needs to be sufficiently large relative to the true sparsity $s_\star$ of $u_\star$. Under mild assumptions, our theory will guarantee that the iterations of this algorithm converge to $u_\star$ up to an error level that is $\mathcal{O}(s^{1/2}\log(np)/n^{1/3})$.

The iterative framework of this algorithm is similar to the CSI method of Ganti et al. [14]. A key difference is that, in the $g_\star$ estimation step, the CSI method enforces a Lipschitz constraint in addition to the monotonicity constraint, which requires choosing the Lipschitz constant in advance; whereas our method only uses isotonic regression to enforce monotonicity. In addition, no theoretical guarantee is given for CSI. Our gradient update step also has some similarities to the least squares method of Balabdaoui et al. [3]; however, their algorithm works explicitly to minimize an objective function $\|\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\|_2^2$, while our SOD-SIM algorithm recovers $u_\star$ through an iterative procedure. The descent direction is not the gradient of least square objective exactly, but a modified version for computational simplicity. The orthogonal projection step enforce a steady "gradient-like" decending rate. Empirically in many cases the algorithm performance is similar without the orthogonal projection step; but in some cases when the initialization $\langle u_0, u_\star\rangle$ is close to 0, the orthogonal projected algorithm has better performance. The orthogonal projection step is also important in solving the technical issue from the normalization step in the proof.

2.2 Convergence guarantees for fixed design X

We first provide upper bounds of convergence rates for the estimation of $u_\star$ and $g_\star$ when $\mathbf{X}$ is fixed. We begin with our assumptions.

2.2.1 Assumptions

We assume that

$$\mathbf{Y} = g_\star(\mathbf{X}u_\star) + \mathbf{Z}, \text{ where } \mathbf{X} \in \mathbb{R}^{n\times p}, \mathbf{Y} \in \mathbb{R}^n, \mathbf{Z} \in \mathbb{R}^n, \tag{4}$$

and we place the following assumptions on the underlying function $g_\star$, parameters $u_\star$, design matrix $\mathbf{X}$ and noise $\mathbf{Z}$.

Assumptions on the signal We assume $g_\star$ is a bounded, Lipschitz, and monotone non-decreasing function, i.e. for $t_1 > t_0$,

$$0 \le g_\star(t_1) - g_\star(t_0) \le L(t_1 - t_0), \text{ and } g_\star(\cdot) : \mathbb{R} \to [-B, B], \tag{5}$$

while the true parameter vector $u_\star$ is a sparse unit vector,

$$\|u_\star\|_2 = 1 \text{ and } |\text{Support}(u_\star)| \le s_\star. \tag{6}$$

For the design matrix $\mathbf{X}$, we assume upper-bounded sparse eigenvalues,

$$\frac{1}{n}\|\mathbf{X}u\|_2^2 \le \beta\|u\|_2^2 \text{ for all } (2s + s_\star)\text{-sparse } u \in \mathbb{R}^p. \tag{7}$$

For the corresponding lower bound, we require an identifiability condition, which is slightly stronger than the usual restricted eigenvalue type condition. This is because we are working with a substantially larger class of models than the usual parametric regression setting. In particular, there exists $\epsilon_n > 0$ such that,

$$\begin{aligned}&\text{for any } s\text{-sparse unit vector } u \text{ with } \|u - u_\star\|_2 \ge \epsilon_n \\ &\text{and any monotone non-decreasing } g, \quad \frac{1}{L^2 n}\|g(\mathbf{X}u) - g_\star(\mathbf{X}u_\star)\|_2^2 \ge \alpha\|u - u_\star\|_2^2.\end{aligned} \tag{8}$$

This condition effectively ensures that we cannot reproduce the true regression function $g_\star(\mathbf{X}u_\star)$ with some other sparse vector $u \neq u_\star$ without any loss in accuracy, unless $\|u - u_\star\|_2$ is very small. In particular, comparing assumptions (5), (7), and (8), we can see that we must have $\alpha \le \beta$. To help interpret the parameters in this assumption, we should think of α and β as constants (expressing properties of the function $g_\star$ that are constant regardless of sample size), while ϵ_n is vanishing as $n \to \infty$ (i.e., the identifiability condition will hold for vectors u increasingly close to $u_\star$, as sample size n increases). Later in Section 2.3, we will show that (7) and (8) are satisfied with high probability when X is a mixture of Gaussian distributions.

Assumptions on the noise For $\mathbf{Z}$, we assume for all $1 \le i \le j \le n$, Z_i's are independent, with

$$\begin{aligned}&\mathbb{E}[Z_i] = 0 \text{ and } \mathbb{E}\left[e^{tZ_i}\right] \le e^{t^2\sigma^2/2} \text{ for all } t \in \mathbb{R}, \\ &\text{and } \sigma_{\max}^2 \ge \text{Var}(Z_i) = \sigma_i^2 \ge \sigma_{\min}^2 > 0 \text{ and } |\sigma_i - \sigma_j| \le L_\sigma \cdot \frac{j-i}{n}.\end{aligned} \tag{9}$$

2.2.2 Convergence guarantee for the estimation of $u_\star$

Theorem 1. *Assume the conditions (4), (5), (6), (7), and (8) hold for the signal, and condition (9) holds for the noise. Assume we run the algorithm with step size $\eta > 0$ and working sparsity level $s \geq 1$ satisfying*

$$\eta \leq \frac{1}{L\beta}, \quad \frac{s}{s_\star} \geq c_s > \left(\frac{1}{\alpha\eta L}\right)^2. \tag{10}$$

For $\delta > 0$, assume furthermore that

$$\frac{s_\star \log(4p/\delta)}{n} \leq \frac{\sigma\alpha^2 L^2}{2\beta}. \tag{11}$$

Then, with probability at least $1-\delta$, it holds for all $t \geq 0$ that

$$\|u_t - u_\star\|_2 \leq \sqrt{2} \cdot r^t + C\left(\epsilon_n + \frac{s^{1/2}\log(np/\delta)}{n^{1/3}}\right), \tag{12}$$

where

$$r = \left(\frac{1-\alpha\eta L}{1-\sqrt{s_\star/s}}\right)^{1/2} < 1,$$

and where the constant C (specified in the proof) depends on $\alpha, \beta, L, B, \sigma, \eta, c_s$, but not on n, p, s, δ and ϵ_n.

This theorem provides information about both computation and statistical convergence in estimating $u_\star$. Define

$$\Delta = C\left(\epsilon_n + \frac{s^{1/2}\log(np/\delta)}{n^{1/3}}\right).$$

Then, from a computation perspective, the algorithm converges linearly up to the time when it reaches the statistical error Δ. In particular, for any tolerance $\tau > \Delta$, running the algorithm for $t \geq \frac{\log\left(\frac{\tau-\Delta}{\sqrt{2}}\right)}{\log(r)}$ many iterations will lead to $\|u_t - u_\star\|_2^2 \leq \tau$ (with probability at least $1-\delta$).

Next, we consider the dependence on the sample size n. Suppose we assume $\epsilon_n \leq \mathcal{O}(n^{-1/3})$ (as we will see in Proposition 2 below, this holds with high probability under a random model for $\mathbf{X}$). Then Δ scales with sample size n as $n^{-1/3}$ (up to log factors). Balabdaoui et al. [3] also obtained $n^{-1/3}$ consistency of a least square estimator of $u_\star$ under similar assumptions. With $g_\star$ being monotone increasing and twice continuously differentiable, and some further assumptions on the $u_\star$ and $\mathbf{X}$, Balabdaoui et al. [4] proposed a score based estimator which is $\sqrt{n}$-consistent, i.e., error scales with sample size as $n^{-1/2}$.

Comparing these results, an open question remains regarding the convergence rate of our algorithm in settings where the true $g_\star$ is smooth — the $n^{-1/2}$ rate achieved by Balabdaoui et al. [4] with slightly stronger assumptions suggests that perhaps our $n^{-1/3}$ rate can be improved with an additional smoothness assumption. In Section 3.1.2 below, we examine this open question empirically, and will see that while the $n^{-1/3}$ rate appears to be tight for a non-smooth $g_\star$, simulation results with a smooth $g_\star$ suggest that an improved rate might be possible in that setting.

2.2.3 Convergence guarantee for the estimation of $g_\star$

We show that with the SOD-SIM algorithm in 2.1, the estimation of $g_\star$ has convergence rate of $\mathcal{O}(\epsilon_n + \frac{s^{1/2}\log(np)}{n^{1/3}})$. In particular, assuming $\epsilon_n \leq O(n^{-1/3})$, the estimation of $g_\star$ has $\mathcal{O}(n^{-1/3})$ convergence rate omitting the log terms.

Proposition 1. *Assume the conditions (4), (5), (6), (7), and (8) hold for the signal, and condition (9) holds for the noise. For $\delta > 0$, assume (10) and (11) hold for the step size η, sparsity levels s and $s_\star$, and n, p. Then as n is sufficiently large, with probability at least $1-\delta$, for all $t \geq 1$, the prediction error is bounded by:*

$$n^{-1/2}\|\mathrm{iso}_{\mathbf{X}u_t}\mathbf{Y} - g_\star(\mathbf{X}u_\star)\|_2 \quad \leq \quad L\sqrt{2\beta}\ \cdot\ r^t\ +\ C'\left(\epsilon_n + \frac{s^{1/2}\log(np/\delta)}{n^{1/3}}\right), \quad (13)$$

where r is defined in Theorem 1, and where the constant C' (specified in the proof) depends on $\alpha, \beta, L, B, \sigma, \eta, c_s$, but not on n, p, s, δ and ϵ_n.

The proof for Proposition 1 is deferred to section A.4. We show our estimator $\mathrm{iso}_{\mathbf{X}u_t}\mathbf{Y}$ of $g_\star$ converges at a $n^{-1/3}$ rate in ℓ_2 norm. The upper bound of the convergence rate for estimating $g_\star$ matches its lower bound up to the log factors, as we know the minimax rate for isotonic regression is $n^{-1/3}$ [7]. This result matches the convergence results shown in [3] and [4].

2.3 Convergence guarantees with random design X

In this section we verify that the assumptions in section 2.2.1 are likely to hold for random design matrices $\mathbf{X}$. In particular, we show the convergence rate for random design $\mathbf{X}$ with normal mixture distribution, which is a much milder condition than the symmetric elliptical $\mathbf{X}$.

2.3.1 Assumptions

We assume the same model (4) as in section 2.2.1. We have the following assumptions on $g_\star(\cdot)$, $\mathbf{X}$, $u_\star$ and $\mathbf{Z}$ under random design matrix $\mathbf{X}$.

Assumptions on the signal Assume $g_\star(\cdot)$ is bounded Lipschitz and monotone nondecreasing (5), and $u_\star$ is a sparse unit vector (6). The rows of $\mathbf{X}$ are i.i.d. draws from a distribution on $X \in \mathbb{R}^p$, and $g_\star(\cdot)$, X and $u_\star$ satisfy that

$$X \sim \sum_{k=1}^{K} a_k \cdot \mathcal{N}(\mu_k, \Sigma_k), \text{ i.e. a mixture-of-normals distribution,}$$
$$\text{with } \sum_{k=1}^{K} a_k = 1,\ \|\mu_k\|_2 \leq V \text{ and } c_0\mathbf{I}_p \preceq \Sigma_k \preceq c_1\mathbf{I}_p \text{ for all } k = 1, \ldots, K. \quad (14)$$

and that

$$\mathrm{Var}\left(g_\star(X^\top u_\star)\right) \geq \nu^2 > 0, \quad (15)$$

Assumptions on the noise Conditional on $\mathbf{X}$, the noise $\mathbf{Z}$ is subgaussian with scale σ,

$$\text{i.e. } \mathbb{E}\left[e^{\langle v,\mathbf{Z}\rangle} \;\middle|\; \mathbf{X}\right] \le e^{\sigma^2\|v\|_2^2/2} \text{ holds almost surely over } \mathbf{X}, \text{ for any fixed } v \in \mathbb{R}^n. \quad (16)$$

2.3.2 Convergence guarantees

Proposition 2. *Under assumptions (4), (5), (6), (14), (15), and (16). For $\delta > 0$, there exists $\epsilon_n \ge C_\alpha \cdot \sqrt{\frac{s\log(np/\delta)}{n}}$ and constants C_α, α, β depending on L, c_0, c_1, B, ν, V but not on n, p, s, δ; as n is sufficiently large, with probability at least $1-\delta$, conditions (7) and (8) hold.*

Proposition 2 verifies that under the random X setting proposed in section 2.3.1, conditions (7) and (8) are satisfied with high probability. This implies that when $\mathbf{X}$ has a normal mixture distribution, with high probability, the results of Theorem 1 still apply, that our proposed SOD-SIM algorithm leads to $n^{-1/3}$ convergence rate in ℓ_2 norm for the estimation of the index $u_\star$. The proof is deferred in section A.5.

2.4 Proof of Theorem 1

We will now prove our convergence theorem. The detailed proofs for the lemmas are presented later in Section A.

First, we give two deterministic results, that will use our assumptions (4), (5), (6), (7), and (8) on the signal (i.e., on $\mathbf{X}$, $g_\star$, and $u_\star$), but hold for any *fixed* noise vector $\mathbf{Z}$. Define

$$\mathrm{Err}_\infty(\mathbf{Z}) = n^{-1}\|\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty,$$

where $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$, and

$$\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z}) = \sup_{u\in\mathbb{S}_s^{p-1}} \left\{ n^{-1/2}\|\mathrm{iso}_{\mathbf{X}u}(g_\star(\mathbf{X}u_\star) + \mathbf{Z}) - \mathrm{iso}_{\mathbf{X}u}(g_\star(\mathbf{X}u_\star))\|_2 \right\},$$

where

$$\mathbb{S}_s^{p-1} = \{u \in \mathbb{R}^p : \|u\|_2 = 1,\ u \text{ is } s\text{-sparse}\}$$

is the set of s-sparse unit vectors in $\mathbb{R}^p$.

First, we verify that the initialization u_0 defined in (3) is well-defined (i.e., $\mathbf{X}^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n) \neq 0$), and is not worse than a random guess, meaning that we have $\langle u_0, u_\star\rangle \ge 0$.

Lemma 1. *Assume the conditions (4), (5), (6), (7), and (8) hold. If $s \ge s_\star \cdot \frac{\beta}{\alpha}$ and*

$$\mathrm{Err}_\infty(\mathbf{Z}) \le \frac{\alpha L}{\sqrt{s_\star}},$$

then the initialization u_0 defined in (3) is well-defined, and satisfies

$$\langle u_0, u_\star\rangle \ge 0.$$

Next, we prove a bound on the iterative update step of the algorithm.

Lemma 2. *Assume the conditions (4), (5), (6), (7), and (8) hold. Fix any $u \in \mathbb{S}_s^{p-1}$ satisfying $\langle u, u_\star \rangle \geq 0$, and fix any step size $\eta \in [0, \frac{1}{L\beta}]$ and any sparsity level $s > s_\star$. Define a hard-thresholded update step as*

$$\check{u} = \frac{\Psi_s(\tilde{u})}{\|\Psi_s(\tilde{u})\|_2} \text{ where } \tilde{u} = u + \eta \cdot \mathcal{P}_u^\perp \left(\tfrac{1}{n} \mathbf{X}^\top \left(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y}) \right) \right).$$

Then

$$\|\check{u} - u_\star\|_2 \leq \left(\frac{1 - \alpha\eta L}{1 - \sqrt{s_\star/s}} \right)^{1/2} \cdot \|u - u_\star\|_2 + \mathrm{Remainder}(\mathbf{Z}), \tag{17}$$

where

$$\mathrm{Remainder}(\mathbf{Z}) = \epsilon_n \left(\frac{1}{1 - \sqrt{s_\star/s}} \right)^{1/2} + \frac{2\eta \left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z}) \right)}{1 - \sqrt{s_\star/s}}.$$

Combining these two lemmas proves the following deterministic convergence result:

Lemma 3. *Assume the conditions (4), (5), (6), (7), and (8) hold. Fix a step size $\eta \geq 0$ and a sparsity level $s \geq 1$ satisfying*

$$\eta \leq \frac{1}{L\beta}, \quad s > s_\star \cdot \left(\frac{1}{\alpha\eta L} \right)^2.$$

Assume also that

$$\mathrm{Err}_\infty(\mathbf{Z}) \leq \frac{\alpha L}{\sqrt{s_\star}}.$$

Then for all $t \geq 0$, it holds that

$$\|u_t - u_\star\|_2 \leq \sqrt{2} \cdot r^t + \frac{\mathrm{Remainder}(\mathbf{Z})}{1 - r},$$

where

$$r = \left(\frac{1 - \alpha\eta L}{1 - \sqrt{s_\star/s}} \right)^{1/2}$$

and where $\mathrm{Remainder}(\mathbf{Z})$ *is defined as in Lemma 2.*

Proof of Lemma 3. We will prove the lemma by induction. We will show that, for each $t \geq 0$, it holds that

$$\|u_t - u_\star\|_2 \leq \sqrt{2} \cdot r^t + \mathrm{Remainder}(\mathbf{Z}) \cdot \sum_{s=1}^{t} r^{s-1}. \tag{18}$$

At $t = 0$, we have

$$\|u_0 - u_\star\|_2^2 = \|u_0\|_2^2 + \|u_\star\|_2^2 - 2\langle u_0, u_\star \rangle \leq 2,$$

since by Lemma 1 we know that $u_0, u_\star$ are unit vectors satisfying $\langle u_0, u_\star \rangle \geq 0$. Therefore (18) holds at $t = 0$. Now suppose (18) holds at $t = T \geq 0$. We then have

$$\begin{aligned} \|u_{T+1} - u_\star\|_2 &\leq r \cdot \|u_T - u_\star\|_2 + \mathrm{Remainder}(\mathbf{Z}) \\ &\leq r \cdot \left(\sqrt{2} \cdot r^T + \mathrm{Remainder}(\mathbf{Z}) \cdot \sum_{s=1}^{T} r^{s-1} \right) + \mathrm{Remainder}(\mathbf{Z}), \end{aligned}$$

where the first step holds by Lemma 2 and the second step applies (18) with $t = T$. This proves that (18) holds with $t = T + 1$, completing the proof. □

With these deterministic results in place, we now need to bound $\mathrm{Err}_\infty(\mathbf{Z})$ and $\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})$, under our assumptions on the noise $\mathbf{Z}$.

Lemma 4. *Assume the conditions (4), (5), (6), (7), and (8) on the signal, and (9) on the noise. For any $\delta > 0$, with probability at least $1 - \delta$ it holds that*

$$\mathrm{Err}_\infty(\mathbf{Z}) \le \sigma \sqrt{\frac{2\beta \log(4p/\delta)}{n}}$$

and

$$\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z}) \le n^{-1/3} \left(2B + \sigma \sqrt{8 \log n} \cdot \sqrt{\log \left(\frac{3n^{2s+1}p^s}{\delta} \right)} \right).$$

Combining Lemma 3 with Lemma 4, we have proved Theorem 1 with the constant

$$C = \max \left\{ \frac{r/(1-r)}{\sqrt{1-\alpha\eta L}}, \frac{r^2/(1-r)}{1-\alpha\eta L} \cdot \eta\sqrt{\beta}\,(7 + 4B + 13\sigma) \right\}.$$

3 Experiments

In this section, we first use simulation experiments to demonstrate the performance of the SOD-SIM algorithm in estimating $u_\star$, and explore the lower bound for the convergence rate; then we apply the SOD-SIM algorithm to a classification problem with PU rocker protein data.

3.1 Simulation studies

3.1.1 Performance in estimating $u_\star$ for PU data

In this section, we study the performance of the SOD-SIM algorithm in estimating $u_\star$ for PU data without knowing the prevalence π.

Methods We simulate PU data with 400 positive data and 400 unlabeled data. Samples in the population are independent and for individual i, $X_i \in \mathbb{R}^p \sim \mathcal{N}(0, \Sigma_\rho)$, where Σ_ρ is the autoregressive matrix with its $i,j-$th entry being $\rho^{|i-j|}$, where we let $\rho = 0.2$. Next, let $Y_i | X_i \sim \text{Bernoulli}\left(g(X_i^\top u_\star)\right)$, where $g(t) = \frac{e^t}{1+e^t}$ is the expit function, $u_\star = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0, \cdots, 0) \in \mathbb{R}^p$ is with sparsity level $s_\star = 2$ and we vary $p = 100, 400, 800, 1600$.

We compare the performance of our proposed SOD-SIM in estimating $u_\star$ with the logistic model with ℓ_1 penalized maximum likelihood (sparseLR) method. For the sparseLR, we choose the tuning parameter using cross validation. For the SOD-SIM algorithm, we let the working sparsity $s = 10$ and set the learning rate $\eta = 0.1$. We simulate for $M = 100$ times.

Results The performance is shown in Figure 1. Since $u_\star$ is only identifiable up to direction, both $u_\star$ and the estimations $\widehat{u}$ are rescaled to have unit norms. Besides bias, standard deviation (SD) and rooted mean square error (RMSE), we also characterize the mean inner product of $u_\star$ and $\widehat{u}$ as their correlation. We can see that for all settings with different p, the SOD-SIM method has smaller bias and SD than the sparseLR

method and the correlation from SOD-SIM with $u_\star$ is closer to 1. With the dimension of the covariates p increases, the performance of the sparseLR method becomes worse, whereas the SOD-SIM method remains a good performance. The results shows that the mis-specification of the link function and the non-symmetric covariate distribution of the PU data affects the estimation of $u_\star$ using the parametric sparseLR method. The proposed SOD-SIM method performs well for the PU data without specifying the prevalence π.

3.1.2 Lower bound exploration

In this section we study whether our convergence rate for $\|u_t - u_\star\|_2$ is tight in terms of its dependence on the sample size n. We are interested in two questions: (1) Is the upper bound from Theorem 1 tight for the SOD-SIM algorithm? (2) Does the SOD-SIM algorithm attain the optimal convergence rate of the global minimizer?

Methods We vary the sample size $n = 100, 150, 200, 300, 400, 600, 800, 1200$; and construct two low-dimensional examples with $s_\star = 2$, $p = 2$, $u^\star = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$. We let $\mathbf{Y} = g_n(\mathbf{X}u^\star) + \mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I)$. In Example 1, $\sigma = 1$ and in Example 2, $\sigma = 0.8$.

We first construct the $g_n(x)$ function. In Example 1, we construct $g_n(x)$ of which the second derivative is unbounded as $n \to \infty$. Specifically, we have

$$g_n(x) = x - \frac{n^{-1/3} f(n^{1/3}x)}{2+\epsilon} \text{ where } f(x) = 2(x - \lfloor x \rfloor) - 4(x - \lfloor x \rfloor - 1/2)_+, \tag{19}$$

where we set $\epsilon = 0.1$. Notice that in this construction, $g_n''(x) \propto n^{1/3} f''(n^{1/3}x) \propto n^{1/3}$.

In Example 2, we construct a smoother version of $g_n(x)$ by letting

$$g_n(x) = x - \frac{n^{-2/3} f(n^{1/3}x)}{2+\epsilon} \tag{20}$$

where we set $\epsilon = 0.1$. For this example, $g_n(x)$ has bounded limiting second derivative as $n \to \infty$. These two example functions are plotted in Figure 2.

To construct the covariates, for each sample, we first generate $t_i \sim \text{Unif}[0, 1]$ and we define X_i as

$$X_{1i} = \frac{\sqrt{2}}{\text{sd}_1} n^{-1/3} f(n^{1/3} t_i),$$
$$X_{2i} = \frac{\sqrt{2}}{\text{sd}_2} (t_i - n^{-1/3} f(n^{1/3} t_i)) + \sqrt{2}\varepsilon_i,$$

for $i = 1, \cdots, n$, where $\varepsilon_i \sim \text{Unif}[-0.5, 0.5]$ are i.i.d. and independent of X_{1i}. The constants sd_1 and sd_2 are used to make X_1 and X_2 to have roughly equal variances. In particular, in Example 1, we let $\text{sd}_1 = 0.025$ and $\text{sd}_2 = 0.4$, and in Example 2, we let $\text{sd}_1 = 0.0014$ and $\text{sd}_2 = 0.4$.

We compute the global minimizer of the ℓ_2 loss. For $\theta \in [0, \pi/2]$, with increment 0.001, we compute $\|\mathbf{Y} - \text{iso}_{\mathbf{X}u_\theta}(\mathbf{Y})\|_2$, where $u_\theta = (\cos\theta, \sin\theta)$, to obtain the minimizer $\hat{u}_g$ of the ℓ_2 loss.

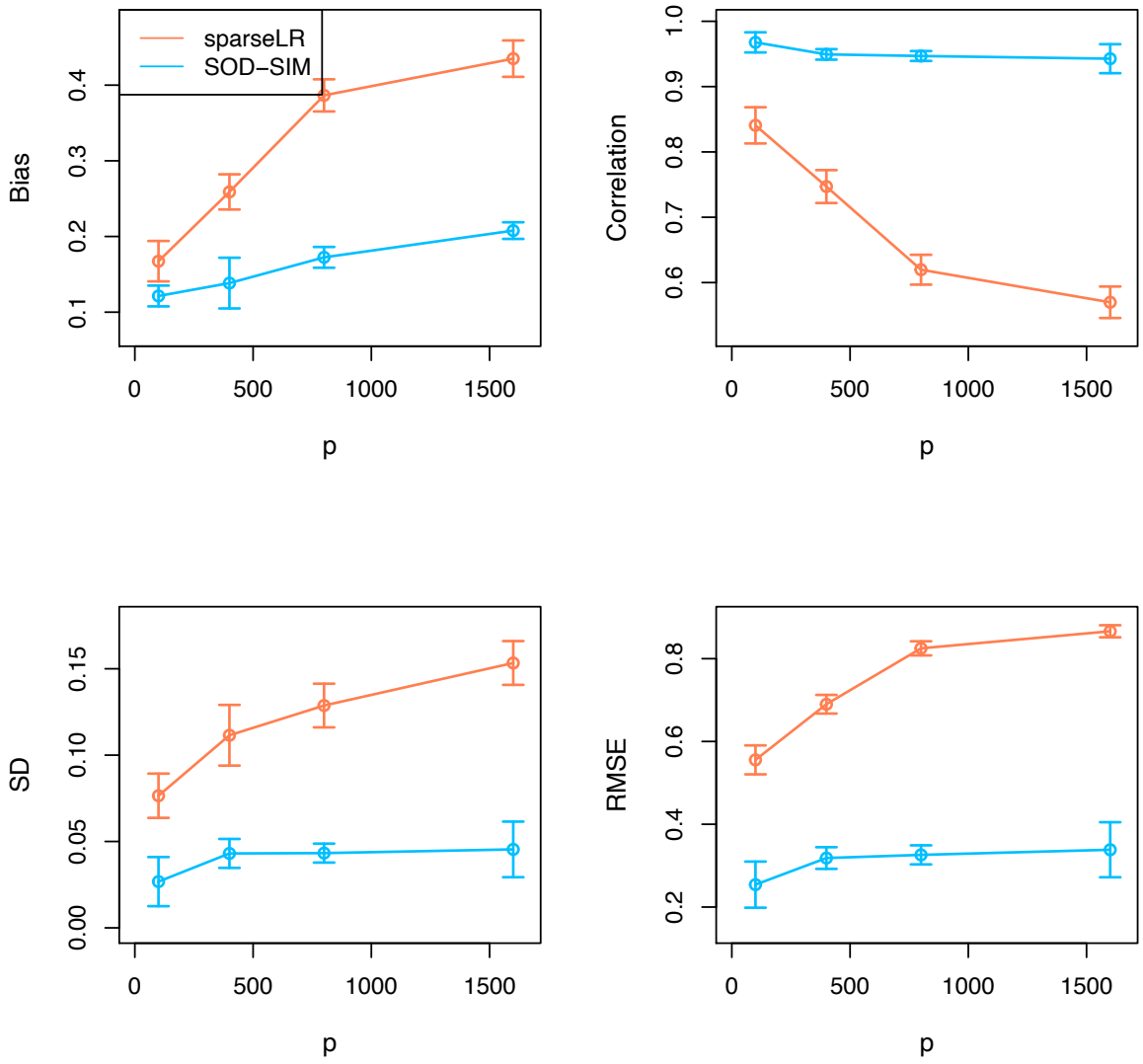

Figure 1: Empirical performance of the SOD-SIM and sparseLR algorithms on PU data in terms of bias, SD, correlation and RMSE with their 95% confidence intervals.

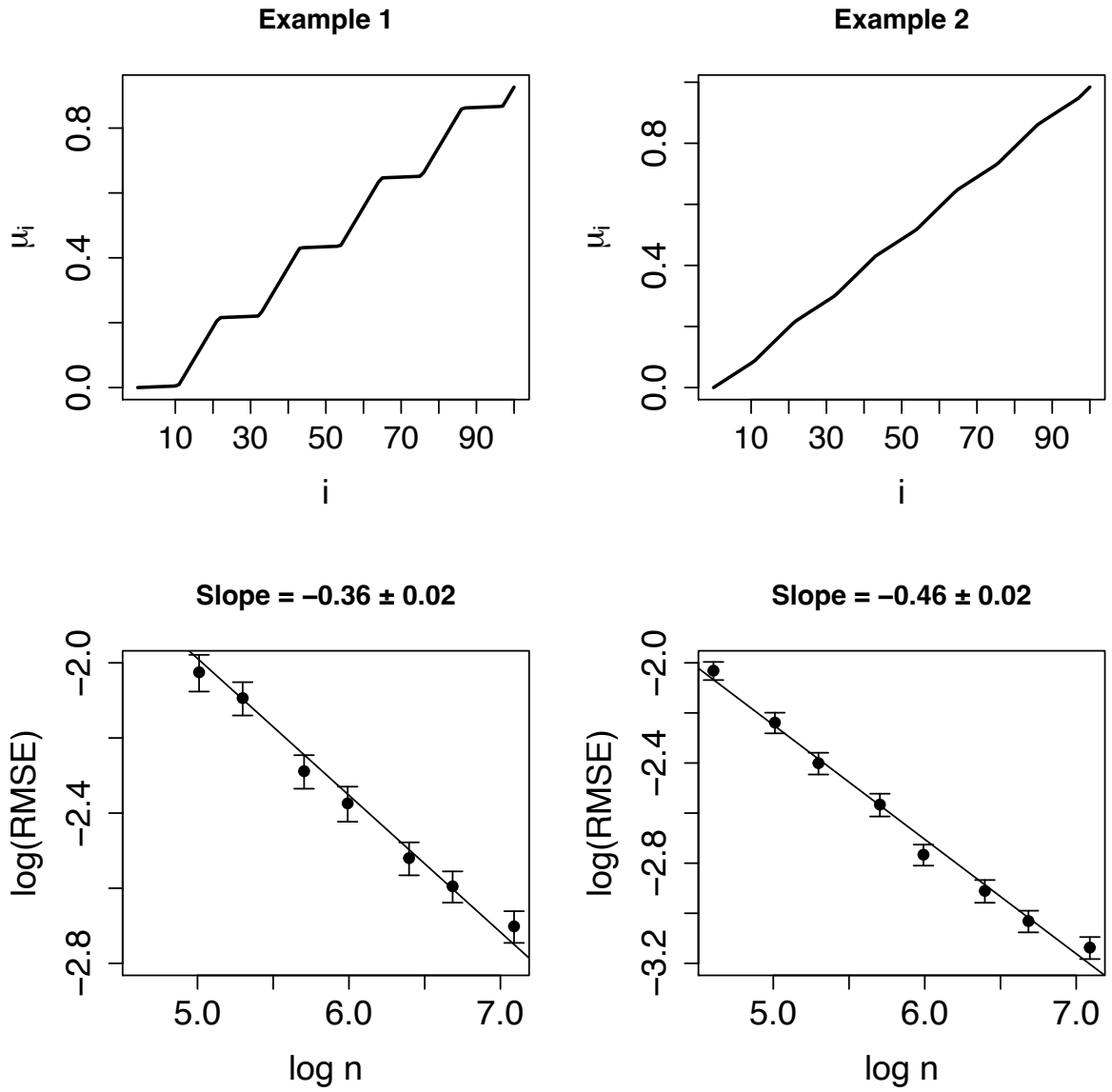

Figure 2: Upper: g_n from (19) (left) and (20) (right). Lower: log(RMSE) (with 95% confidence interval) vs $\log(n)$ for the global estimator for Example 1 (19) (left) and Example 2 (20) (right).

Results We plot the logarithm of rooted mean square error (log(RMSE)) against $\log(n)$ to show the rate of convergence. Example 1 (non-smooth example) has a convergence rate close to $n^{-1/3}$, which indicates our proved convergence rate for the SOD-SIM algorithm is as tight as the global minimizer. The second example (with bounded limiting second derivative) has a faster convergence rate, which is close to $n^{-1/2}$.

Our exploratory simulation results match with the reported convergence rate of the score based estimator in [4]. We do not have a definitive answer for the convergence rate of our algorithm when the function $g_\star$ is smooth. A better rate could potentially be achieved with alternative proof techniques, and is beyond the scope of this work.

3.2 Application to rocker protein data

In this section, we study an application of our SOD-SIM algorithm to the rocker protein sequence data. The rocker data is composed of sequences with confirmed function ($Y = 1$) and sequences with unknown functionality ($Y = 0$).

To give some biological context to this dataset, Rocker is a *de novo* designed protein which recreates the biological function of substrate transportation [20]. A functional protein successfully transports ions across cell membranes. To study how a mutation at each site affects the membrane transport ability of proteins, mutations are introduced to the original Rocker sequence (wild-type) by double site saturation mutagenesis. Mutated sequences are screened by fluorescence-activated cell sorting (FACS), which sort out *functional* protein variants. Due to an experimental challenge of obtaining negative sequences, an additional set of sequences are obtained from the initial library whose associated functionality is unknown. Therefore, the resulting data set is a Positive-Unlabeled (PU) dataset, because the first set consists of functional sequence variants and the second set consists unlabeled examples.

Methods A sequence consists of 25 positions taking one of the 21 discrete values, which correspond to 20 amino acid letter codes plus an additional letter for the alignment gap. Each sequence contains at most two mutations. The data consists of $n_\ell = 703030$ functional (positive label) and $n_u = 1287155$ unlabeled sequences. To compare different algorithms reliably, we split the data into 10 subsets, and use half of each split for training and remaining half for testing. Each training dataset contains 35K labeled and 64K unlabeled examples on average. We generate features with main (site-wise) and pairwise (interaction between two sites) effects using one-hot encoding of the sequences. Removing columns with zero counts, we obtain 27K features on average, where 490 of the generated features correspond to the main effects. We take the amino acid levels in the WT sequence as the baseline levels to generate a sparse design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ where $(n, p) \approx (100\text{K}, 27\text{K})$. The number of unique sequences is around 26K (i.e., rank of row space of $\mathbf{X} \approx 26\text{K}$), which makes the problem high-dimensional. The response vector $\mathbf{Y} \in \mathbb{R}^n$ represents whether each sequence i is labeled ($\mathbf{Y}_i = 1$) or unlabeled ($\mathbf{Y}_i = 0$).

We applied four methods to each train dataset which we denote as follows:

- SOD-SIM: our proposed algorithm
- sparseLR: the logistic regression with ℓ_1-penalty

- PV1: the proposed method in [26]. We solve the following objective:

$$\hat{u}^{PV1} \in \arg\max_{u \in K(s)} \langle y, \mathbf{X}u \rangle$$

  where $K(s)$ is a s-approximate sparse set, defined as $K(s) := \{x \in \mathbb{R} : \|x\|_2 \leq 1, \|x\|_1 \leq \sqrt{s}\}$

- PV2: the proposed method in [27].

$$\hat{u}^{PV2} \in \arg\min_{u \in K(s)} \|y - \mathbf{X}u\|_2.$$

We used the **glmnet** package to solve the logistic regression objective with ℓ_1-penalty. To solve objectives in PV1 and PV2, we implemented the projected gradient methods, where we iteratively projected the gradients onto the intersection of ℓ_1 and ℓ_2 balls. We used the Dykstra's projection algorithm to obtain the projection onto the intersection of the two balls [5]. Note both objectives in PV1 and PV2 are convex, and therefore the projected gradient descent algorithm guarantees to find a global minimum. For PV1 and PV2, we input the standardized $\mathbf{X}$ where each column is centered and scaled. Since sparseLR, PV1, and PV2 are convex problems, we run until the algorithms the models converge. For SOD-SIM, we let the learning rate $\eta = 1$ and run the algorithm until changes of estimated parameters is small (< 0.0005) or the iteration number t reaches the pre-defined maximum number of iterations (≤ 1000). In addition, PV1 and PV2 methods do not provide estimates of the link function. We run an isotonic regression after obtaining $\hat{u}^{PV1}$ and $\hat{u}^{PV2}$ to estimate the link function, and used such estimates to perform downstream prediction tasks.

We used four metrics (Accuracy, F1 score, Brier Score, and AUC value) to evaluate predictive performance of the four methods. Accuracy is the proportion of correctly classified examples. F1 score is the harmonic mean of the precision and recall, whose value lies between 0 and 1. Higher number corresponds to a better performance. Brier Score is an average squared ℓ_2 loss, i.e., Brier Score $:= \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2$, and therefore, lower numbers correspond to better performances. AUC measures the area under the ROC curve. The perfect classification corresponds to the AUC value of 1, and a random classification corresponds to 0.5. For the choice of hyperparameters (s for SOD-SIM, PV1, and PV2, and λ for the ℓ_1 logistic regression), we use a grid of 100 working sparsity s values from 1 to p, interpolated in a square root scale, and a grid of 100 lambda values obtained from the **glmnet** package by default. We picked the hyperparameters that gave the best results for the most of the metrics on the test datasets.

Results Figure 3.2 demonstrates the empirical performances of the four methods. SOD-SIM performed the best in all metrics on average. The predictive performance of the sparse logistic regression was slightly worse than the SOD-SIM. It is likely due to the mis-specification error, since we are forcing the link function to be a sigmoid function. Both PV1 and PV2 seem to suffer from the deviation $\mathbf{X}$ from the Gaussian design; however, PV2 seems to be more robust to such deviation.

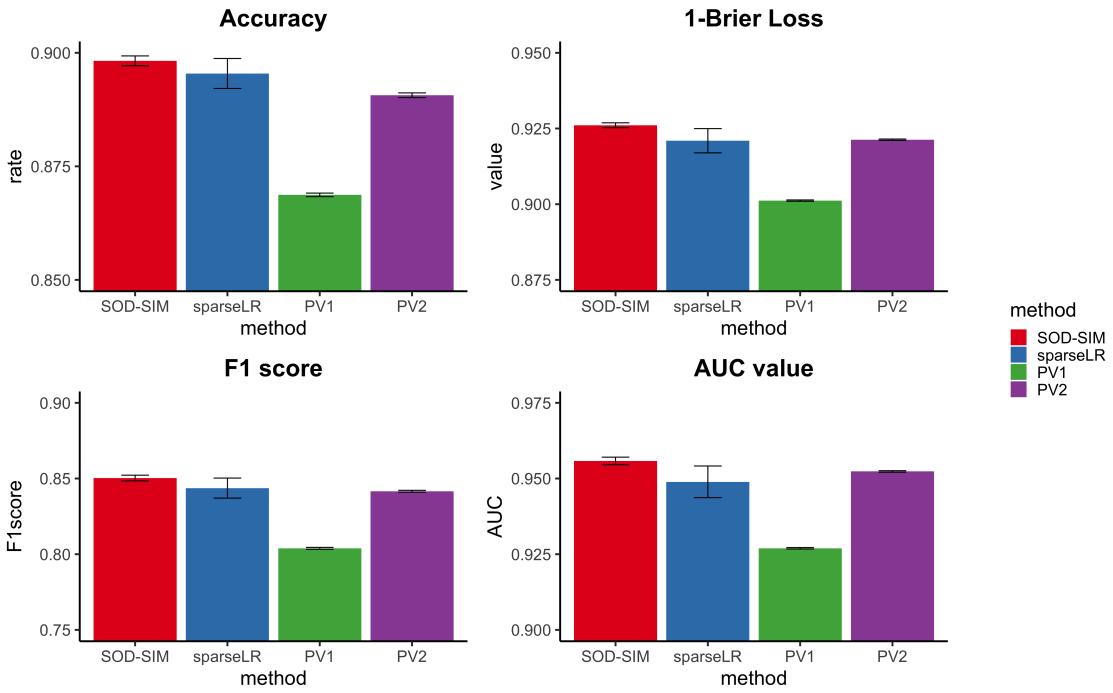

Figure 3: Average accuracy rates, F1 scores, 1-Brier Score, and AUC values of the four methods on 10 test datasets. Higher values correspond to better performance in all four plots. The error bars represent one standard error.

4 Discussion

In this paper, we propose a scalable iterative projection-based algorithm (SOD-SIM) for the estimation of hdSIMs, which we show the statistical convergence guarantees for the estimation of both the coefficient $u_\star$ and the unknown function $g_\star$. From the minimax results of isotonic regression perspective, our convergence rate for $g_\star$ is tight with respect to n up to some poly log terms. Our simulation results suggest that under the mild model assumptions, the convergence rate for the estimation of $u_\star$ is also tight. However, our simulation results for the smooth $g_\star$ case suggests that the convergence rate might potentially decay faster with sample size n, under stronger model assumptions—while the theoretical upper bounds proved here scale as $n^{-1/3}$ (ignoring log terms), the simulations suggest that the lower bound for the settings with additional smoothness conditions might potentially be improved. The gap between these two rates in the smooth case remains an open question.

Our statistical guarantee requires very mild model assumptions, especially for the covariates $\mathbf{X}$. This allows our theoretical guarantee to cover many useful examples such as the PU data. From our simulation studies, we have shown that even if at population level the covariates $\mathbf{X}$ is normally distributed, the positive-only data has asymmetric distribution and parametric models such as sparseLR suffer from estimation bias caused by both misspecification of the link function and the asymmetric distribution of the covariates; whereas our SOD-SIM algorithm has good performance in estimating $u_\star$ in high-dimensional settings. In our real data analysis, we have shown that our method (SOD-SIM) also outperforms methods based on the slicing regression idea (PV1, PV2).

Acknowledgements

R.F.B. was partially supported by the National Science Foundation via grants DMS-1654076 and DMS-2023109, and by the Office of Naval Research via grant N00014-20-1-2337. G.R. was partially supported by the National Science Foundation via grant DMS-1811767 and by the National Institute of Health via grant R01 GM131381-01. H. S. was partially supported by the National Institute of Health via grant R01 GM131381-01. The authors thank Sabyasachi Chatterjee for helpful discussions.

References

[1] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222 – 239, 2014. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2013.04.002. Special Issue on Sparse Approximate Solution of Linear Systems.

[2] P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 01 2013.

[3] F. Balabdaoui, C. Durot, and H. Jankowski. Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B):3276–3310, 11 2019. doi: 10.3150/18-BEJ1090.

[4] F. Balabdaoui, P. Groeneboom, and K. Hendrickx. Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46(2):517–544, 2019. doi: 10.1111/sjos.12361.

[5] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in Order Restricted Statistical Inference*, pages 28–47. Springer New York, 1986.

[6] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997. doi: 10.1080/01621459.1997.10474001.

[7] S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43:1774–1800, 08 2015. doi: 10.1214/15-AOS1324.

[8] S. Chen and A. Banerjee. Robust structured estimation with single-index models. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 712–721, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[9] J.-M. Chiou and H.-G. Müller. Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association*, 93(444):1376–1387, 1998. doi: 10.1080/01621459.1998.10473799.

[10] X. Cui, W. K. Härdle, and L. Zhu. The efm approach for single-index models. *Ann. Statist.*, 39(3):1658–1688, 06 2011. doi: 10.1214/10-AOS871.

[11] J. de Leeuw, K. Hornik, and P. Mair. Isotone optimization in r: Pool-adjacent-violators (pava) and active set methods. *Journal of Statistical Software*, 32:1–24, 2009.

[12] M. Delecroix, W. Härdle, and M. Hristache. Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86(2):213 – 226, 2003. ISSN 0047-259X. doi: https://doi.org/10.1016/S0047-259X(02)00046-5.

[13] N. Duan and K.-C. Li. Slicing regression: A link-free regression method. *Ann. Statist.*, 19(2):505–530, 06 1991. doi: 10.1214/aos/1176348109.

[14] R. Ganti, N. S. Rao, L. Balzano, R. Willett, and R. D. Nowak. On learning high dimensional structured single index models. In *AAAI Conference on Artificial Intelligence*, pages 1898–1904. AAAI Press, 2017.

[15] W. Hardle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *Ann. Statist.*, 21(1):157–178, 03 1993. doi: 10.1214/aos/1176349020.

[16] T. J. Hastie and W. Fithian. Inference from presence-only data; the ongoing controversy. *Ecography*, 36 8:864–867, 2013.

[17] J. L. Horowitz and W. Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996. doi: 10.1080/01621459.1996.10476732.

[18] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):593–623, 06 2001. doi: 10.1214/aos/1009210682.

[19] H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71 – 120, 1993. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(93)90114-K.

[20] N. H. Joh, T. Wang, M. P. Bhate, R. Acharya, Y. Wu, M. Grabe, M. Hong, G. Grigoryan, and W. F. DeGrado. De novo design of a transmembrane zn^{2+}-transporting four-helix bundle. *Science*, 346(6216):1520–1524, Dec. 2014.

[21] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 927–935. Curran Associates, Inc., 2011.

[22] A. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. 01 2009.

[23] R. W. Klein and R. H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421, 1993. ISSN 00129682, 14680262.

[24] A. K. Kuchibhotla and R. K. Patra. Efficient estimation in single index models through smoothing splines. *ArXiv*, 2016.

[25] Q. Lin, Z. Zhao, and J. S. Liu. On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Statist.*, 46(2):580–610, 04 2018. doi: 10.1214/17-AOS1561.

[26] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inf. Theory*, 59(1): 482–494, Jan. 2013.

[27] Y. Plan and R. Vershynin. The generalized lasso with Non-Linear observations. *IEEE Trans. Inf. Theory*, 62(3):1528–1537, Mar. 2016.

[28] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017. doi: 10.1093/imaiai/iaw015.

[29] J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989. ISSN 00129682, 14680262.

[30] P. Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266 – 282, 2015. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2015.02.007.

[31] C. Sammut and G. I. Webb, editors. *McDiarmid's Inequality*, pages 651–652. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_521. URL https://doi.org/10.1007/978-0-387-30164-8_521.

[32] C. Sammut and G. I. Webb, editors. *Symmetrization Lemma*, pages 954–954. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_808. URL https://doi.org/10.1007/978-0-387-30164-8_808.

[33] H. Song and G. Raskutti. Pulasso: High-dimensional variable selection with presence-only data. *Journal of the American Statistical Association*, 115(529): 334–347, 2020. doi: 10.1080/01621459.2018.1546587.

[34] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

[35] T. Wang, P.-R. Xu, and L.-X. Zhu. Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis*, 109:221 – 235, 2012. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2012.03.009.

[36] F. Yang and R. F. Barber. Contraction and uniform convergence of isotonic regression. *Electron. J. Statist.*, 13(1):646–677, 2019. doi: 10.1214/18-EJS1520.

[37] Z. Yang, K. Balasubramanian, and H. Liu. High-dimensional non-Gaussian single index models via thresholded score function estimation. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3851–3860, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[38] Z. Yang, K. Balasubramanian, P. Zhaoran Wang, and H. Liu. Estimating high-dimensional non-gaussian multiple index models via stein's lemma. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6097–6106. Curran Associates, Inc., 2017.

[39] Y. Yu and D. Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002. doi: 10.1198/016214502388618861.

A Additional proofs

A.1 Proof of Lemma 2

First, writing $\bar{u} = \Psi_s(\tilde{u})$, since u is a s-sparse unit vector we can calculate

$$1 = \|u\|_2^2 = \langle \tilde{u}, u\rangle \le \|\tilde{u}_{\mathrm{support}(u)}\|_2 \le \max_{S\subset[d],|S|=s} \|\tilde{u}_S\|_2 = \|\bar{u}\|_2,$$

where the second equality holds since $\tilde{u}$ is equal to u plus an orthogonal vector. Since $u_\star$ is a unit vector, we have

$$\|\check{u} - u_\star\|_2 \le \|\bar{u} - u_\star\|_2,$$

since $\check{u} = \bar{u}/\|\bar{u}\|_2$ is the projection of $\bar{u}$ onto the unit ball, which is a convex set containing $u_\star$.

A trivial calculation shows that

$$\|\bar{u} - u_\star\|_2^2 = \|u - u_\star\|_2^2 - \|u - \bar{u}\|_2^2 + 2\langle u - \bar{u}, u_\star - \bar{u}\rangle. \tag{21}$$

Now we work with this inner product. By definition of $\tilde{u}$, we can write

$$\begin{aligned}\langle u - \bar{u}, u_\star - \bar{u}\rangle &= \langle \tilde{u} - \bar{u}, u_\star - \bar{u}\rangle + \frac{\eta}{n}\langle \mathcal{P}_u^\perp \mathbf{X}^\top\big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big), \bar{u} - u_\star\rangle \\ &\le \frac{\sqrt{s_\star}}{2\sqrt{s}}\|u_\star - \bar{u}\|_2^2 + \frac{\eta}{n}\langle \mathcal{P}_u^\perp \mathbf{X}^\top\big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big), \bar{u} - u_\star\rangle,\end{aligned}$$

where the last step applies Lemma 5 (presented later on). Plugging this back into (21) and rearranging terms, we obtain

$$\left(1 - \sqrt{\frac{s_\star}{s}}\right)\|\bar{u} - u_\star\|_2^2 \le \|u - u_\star\|_2^2 - \|u - \bar{u}\|_2^2 + \frac{2\eta}{n}\langle \mathcal{P}_u^\perp \mathbf{X}^\top\big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big), \bar{u} - u_\star\rangle.$$

Moreover, noting that $\mathbf{1}_n^\top(y - \mathrm{iso}_v(y)) = 0$ for any $y, v \in \mathbb{R}^n$ by properties of isotonic regression, we can equivalently write this as

$$\left(1 - \sqrt{\frac{s_\star}{s}}\right) \|\bar{u} - u_\star\|_2^2 \\ \leq \|u - u_\star\|_2^2 - \|u - \bar{u}\|_2^2 + \frac{2\eta}{n}\langle \mathcal{P}_u^\perp \mathbf{X}^\top \mathcal{P}_{\mathbf{1}_n}^\perp \big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big), \bar{u} - u_\star\rangle. \quad (22)$$

Next, working with this last term, denote $\mu_\star = g_\star(\mathbf{X}u_\star)$, we can write

$$\mathcal{P}_{\mathbf{1}_n}^\perp \big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big) \\ = \mathcal{P}_{\mathbf{1}_n}^\perp \left(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\right) + \mathcal{P}_{\mathbf{1}_n}^\perp \left(\mathbf{Y} - \mu_\star\right) + \mathcal{P}_{\mathbf{1}_n}^\perp \left(\mathrm{iso}_{\mathbf{X}u}(\mu_\star) - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\right) \\ = \left(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\right) + \big(\mathbf{Z} - \bar{Z}\mathbf{1}_n\big) + \mathcal{P}_{\mathbf{1}_n}^\perp \left(\mathrm{iso}_{\mathbf{X}u}(\mu_\star) - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\right),$$

and since

$$\langle \mathcal{P}_u^\perp \mathbf{X}^\top \mathcal{P}_{\mathbf{1}_n}^\perp \big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big), \bar{u} - u_\star\rangle \leq \langle \mathcal{P}_u^\perp \mathbf{X}^\top \big(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\big), \bar{u} - u_\star\rangle \\ + \left\|\mathbf{X}^\top \left(\mathbf{Z} - \bar{Z}\mathbf{1}_n\right)\right\|_\infty \|\mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_1 + \|\mathrm{iso}_{\mathbf{X}u}(\mathbf{Y}) - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2 \|\mathcal{P}_{\mathbf{1}_n}^\perp \mathbf{X} \mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_2,$$

and therefore plugging in our definitions of $\mathrm{Err}_\infty(\mathbf{Z})$ and $\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})$ from above,

$$\langle \mathcal{P}_u^\perp \mathbf{X}^\top \big(\mathbf{Y} - \mathrm{iso}_{\mathbf{X}u}(\mathbf{Y})\big), \bar{u} - u_\star\rangle \leq \langle \mathcal{P}_u^\perp \mathbf{X}^\top \big(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\big), \bar{u} - u_\star\rangle \\ + n\mathrm{Err}_\infty(\mathbf{Z}) \cdot \|\mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_1 + n^{1/2}\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z}) \cdot \|\mathcal{P}_{\mathbf{1}_n}^\perp \mathbf{X} \mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_2.$$

Since $u, \bar{u}$ are s-sparse and $u_\star$ is $s_\star$-sparse, we can calculate that $\mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)$ is $(2s + s_\star)$-sparse and therefore

$$\|\mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_1 \leq \sqrt{2s + s_\star}\|\mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_2 \leq \sqrt{2s + s_\star}\|\bar{u} - u_\star\|_2.$$

By our condition (7) bounding the sparse eigenvalues of $\mathbf{X}$, we also have

$$\|\mathcal{P}_{\mathbf{1}_n}^\perp \mathbf{X} \mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_2 \leq \|\mathbf{X} \mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_2 \leq \sqrt{\beta n}\|\mathcal{P}_u^\perp \big(\bar{u} - u_\star\big)\|_2 \leq \sqrt{\beta n}\|\bar{u} - u_\star\|_2.$$

Combining everything and returning to (22), then,

$$\left(1 - \sqrt{\frac{s_\star}{s}}\right) \|\bar{u} - u_\star\|_2^2 \leq \|u - u_\star\|_2^2 - \|u - \bar{u}\|_2^2 \\ + 2\eta\|\bar{u} - u_\star\|_2 \left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})\right) \\ + \frac{2\eta}{n}\langle \mathcal{P}_u^\perp \mathbf{X}^\top \big(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\big), \bar{u} - u_\star\rangle. \quad (23)$$

Next we work with the remaining inner product. We have

$$\begin{aligned}
&\langle \mathcal{P}_u^\perp \mathbf{X}^\top \big(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\big), \bar{u} - u_\star\rangle \\
&= \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}\mathcal{P}_u^\perp(\bar{u} - u_\star))\rangle \\
&= \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}\mathcal{P}_u^\perp \bar{u}\rangle - \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}\mathcal{P}_u^\perp u_\star\rangle \\
&= \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}\mathcal{P}_u^\perp(\bar{u} - u)\rangle - \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}(u_\star - u \cdot u^\top u_\star)\rangle \\
&\leq \|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2 \cdot \|\mathbf{X}\mathcal{P}_u^\perp(\bar{u} - u)\|_2 - \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}u_\star\rangle \\
&\qquad + \langle u, u_\star\rangle \cdot \langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}u\rangle \\
&\leq \|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2 \cdot \sqrt{\beta n}\|\bar{u} - u\|_2 - \frac{1}{L}\|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2^2,
\end{aligned}$$

where to prove the last step, we apply (7) to the first term, apply Lemma 6 (presented later) to the second term, and for the third term we observe that it is ≤ 0, because $\langle u, u_\star \rangle \geq 0$ by assumption while $\langle \mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star), \mathbf{X}u \rangle \leq 0$ by definition of isotonic regression (i.e., since isotonic regression is projection onto a convex cone). Finally, by Cauchy–Schwarz we have

$$\|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2 \cdot \sqrt{\beta n}\|\bar{u} - u\|_2 \leq \frac{L\beta n}{2}\|\bar{u} - u\|_2^2 + \frac{1}{2L}\|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2^2.$$

Therefore, we have calculated

$$\langle \mathcal{P}_u^\perp \mathbf{X}^\top \big(\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\big), \bar{u} - u_\star \rangle \leq \frac{L\beta n}{2}\|\bar{u} - u\|_2^2 - \frac{1}{2L}\|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2^2.$$

Plugging these calculations back into (23), and applying the assumption that $\eta \leq \frac{1}{L\beta}$, we obtain

$$\begin{aligned}\left(1 - \sqrt{\frac{s_\star}{s}}\right)\|\bar{u} - u_\star\|_2^2 \leq \|u - u_\star\|_2^2 - \frac{\eta}{Ln}\|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2^2 \\ + 2\eta\|\bar{u} - u_\star\|_2 \left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})\right). \quad (24)\end{aligned}$$

Next, we split into cases. If $\|u - u_\star\| \leq \epsilon_n$, then (24) implies that

$$\left(1 - \sqrt{\frac{s_\star}{s}}\right)\|\bar{u} - u_\star\|_2^2 \leq \epsilon_n^2 + 2\eta\|\bar{u} - u_\star\|_2 \left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})\right)$$

which can be relaxed to

$$\|\bar{u} - u_\star\|_2 \leq \frac{\epsilon_n}{\left(1 - \sqrt{\frac{s_\star}{s}}\right)^{1/2}} + \frac{2\eta\left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})\right)}{1 - \sqrt{\frac{s_\star}{s}}}.$$

If instead $\|u - u_\star\|_2 \geq \epsilon_n$, then by the identifiability assumption (8), noting that by the definition of isotonic regression, we can write $\mathrm{iso}_{\mathbf{X}u}(\mu_\star) = g(\mathbf{X}u)$ for some monotone non-decreasing function g, we have

$$\frac{1}{L^2 n}\|\mu_\star - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2^2 = \frac{1}{L^2 n}\|g_\star(\mathbf{X}u_\star) - g(\mathbf{X}u)\|_2^2 \geq \alpha\|u - u_\star\|_2^2.$$

In this case, (24) implies that

$$\begin{aligned}&\left(1 - \sqrt{\frac{s_\star}{s}}\right)\|\bar{u} - u_\star\|_2^2 \\ &\quad \leq (1 - \alpha\eta L)\|u - u_\star\|_2^2 + 2\eta\|\bar{u} - u_\star\|_2 \left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})\right)\end{aligned}$$

which can be relaxed to

$$\|\bar{u} - u_\star\|_2 \leq \left(\frac{1 - \alpha\eta L}{1 - \sqrt{\frac{s_\star}{s}}}\right)^{1/2}\|u - u_\star\|_2 + \frac{2\eta\left(\sqrt{2s + s_\star} \cdot \mathrm{Err}_\infty(\mathbf{Z}) + \sqrt{\beta} \cdot \mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})\right)}{1 - \sqrt{\frac{s_\star}{s}}}.$$

Finally, since we proved above that $\|\check{u} - u_\star\|_2 \leq \|\bar{u} - u_\star\|_2$, this completes the proof.

A.1.1 Supporting lemmas

Lemma 5 (Liu & Barber 2018, Lemma 1)**.** *For any $v \in \mathbb{R}^p$ and $s_\star$-sparse $w \in \mathbb{R}^p$, if $s \geq s_\star$, then*

$$\langle v - \Psi_s(v), w - \Psi_s(v) \rangle \leq \frac{\sqrt{s_\star}}{2\sqrt{s}} \|w - \Psi_s(v)\|_2^2.$$

Lemma 6. *For any vectors $v, w \in \mathbb{R}^n$ and any L-Lipschitz monotone nondecreasing function g,*

$$\langle v, g(v) - \mathrm{iso}_w(g(v)) \rangle \geq L^{-1} \|g(v) - \mathrm{iso}_w(g(v))\|_2^2.$$

Proof of Lemma 6. Let $a_1 < \cdots < a_K$ be the unique values of the vector $\mathrm{iso}_w(g(v))$, and for each $k = 1, \ldots, K$ let

$$I_k = \Big\{ i : \big(\mathrm{iso}_w(g(v))\big)_i = a_k \Big\}.$$

By definition of isotonic regression, we have

$$\frac{1}{|I_k|} \sum_{i \in I_k} g(v) = a_k. \tag{25}$$

For each k, let $b_k \in \mathbb{R}$ satisfy

$$g(b_k) = a_k,$$

which must exist since $g : \mathbb{R} \to \mathbb{R}$ is continuous (due to being Lipschitz) and a_k is a convex combination of values taken by g. We then have

$$\begin{aligned}
\langle v, g(v) - \mathrm{iso}_w(g(v)) \rangle &= \sum_{i=1}^n v_i \cdot \big(g(v_i) - \mathrm{iso}_w(g(v))_i\big) \\
&= \sum_{k=1}^K \sum_{i \in I_k} v_i \cdot \big(g(v_i) - a_k\big) \\
&= \sum_{k=1}^K \left(\sum_{i \in I_k} v_i \cdot \big(g(v_i) - a_k\big) - b_k \cdot \sum_{i \in I_k} \big(g(v_i) - a_k\big) \right) \text{ by (25)} \\
&= \sum_{k=1}^K \sum_{i \in I_k} (v_i - b_k) \cdot \big(g(v_i) - g(b_k)\big) \text{ by definition of } b_k \\
&\geq \sum_{k=1}^K \sum_{i \in I_k} L^{-1} \big(g(v_i) - a_k\big)^2 \text{ since } g \text{ is } L\text{-Lipschitz and monotone} \\
&= \sum_{k=1}^K \sum_{i \in I_k} L^{-1} \Big(g(v_i) - \big(\mathrm{iso}_w(g(v))\big)_i\Big)^2 \text{ by definition of } a_k \\
&= L^{-1} \|g(v) - \mathrm{iso}_w(g(v))\|_2^2.
\end{aligned}$$

□

A.2 Proof of Lemma 4

First we bound $\mathrm{Err}_\infty(\mathbf{Z})$. By the sparse eigenvalue bound (7) on $\mathbf{X}$, we have $\|\mathbf{X}_j\|_2 = \|\mathbf{X}\mathbf{e}_j\|_2 \leq \sqrt{\beta n}$ for all $j = 1, \ldots, p$. Since $\mathbf{Z} = \mathbf{Y} - \mu_\star$, $\mathbf{Z} - \bar{Z}\mathbf{1}_n$ is zero-mean and σ-subgaussian, we therefore see that $\mathbf{X}_j^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)$ is also zero-mean and is $\sigma\sqrt{\beta n}$-subgaussian. Therefore, for each j,

$$\mathbb{P}\left\{n^{-1}\left|\mathbf{X}_j^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\right| > \sigma\sqrt{\frac{2\beta\log(4p/\delta)}{n}}\right\} \leq \frac{\delta}{2p},$$

and so taking a union bound, we have shown that

$$\mathbb{P}\left\{\mathrm{Err}_\infty(\mathbf{Z}) \leq \sigma\sqrt{\frac{2\beta\log(4p/\delta)}{n}}\right\} \geq 1 - \frac{\delta}{2}.$$

Next we turn to $\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z})$. First, for any vector $v \in \mathbb{R}^n$, define π_v to be a permutation of $\{1, \ldots, n\}$ such that $v_{\pi_v(1)} \leq \cdots \leq v_{\pi_v(n)}$, i.e., we rearrange the indices into the ordering of v. (If v has some repeated entries then π_v will not be unique but we can arbitrarily choose one such permutation for each v.)

First, for any fixed v, we will compute a deterministic bound on $\|\mathrm{iso}_v(\mathbf{Y}) - \mathrm{iso}_v(\mu_\star)\|_2$. We know that $\mu_\star \in [-B, B]^n$ and therefore $\mathrm{iso}_v(\mu_\star) \in [-B, B]^n$, and is monotone non-decreasing by definition. By Lemma 11.1 in [7], we can find some partition

$$\{1, \ldots, n\} = \underbrace{\{\pi_v(1), \ldots, \pi_v(k_1)\}}_{=:S_1} \cup \underbrace{\{\pi_v(k_1+1), \ldots, \pi_v(k_2)\}}_{=:S_2} \cup \cdots \cup \underbrace{\{\pi_v(k_{M-1}+1), \ldots, \pi_v(n)\}}_{=:S_M},$$

with $M \leq \lceil n^{1/3} \rceil$, such that

$$\max_{i \in S_m} \mathrm{iso}_v(\mu_\star) - \min_{i \in S_m} \mathrm{iso}_v(\mu_\star) \leq 2Bn^{-1/3} \text{ for all } m = 1, \ldots, M.$$

In other words, this condition ensures $\mathrm{iso}_v(\mu_\star)$ has little variation within each index subset S_m. We define $k_0 = 0$ and $k_M = n$ below, so that we can write $S_m = \{\pi_v(k_{m-1}+1), \ldots, \pi_v(k_m)\}$ for each $m = 1, \ldots, M$.

Next for any permutation π of $\{1, \ldots, n\}$, define a seminorm

$$\|w\|_{\mathrm{SW},\pi} = \max_{1 \leq j \leq k \leq n} \frac{\left|\sum_{\ell=j}^k w_{\pi(\ell)}\right|}{\sqrt{k-j+1}}.$$

This is the "sliding window" norm of [36], defined according to the ordering of the vector v. By Theorem 1 and Lemma 2 of [36], it holds that

$$\|\mathrm{iso}_v(x) - \mathrm{iso}_v(x')\|_{\mathrm{SW},\pi_v} \leq \|x - x'\|_{\mathrm{SW},\pi_v}$$

for all $x, x' \in \mathbb{R}^n$. Therefore, for all $1 \leq j \leq k \leq n$,

$$\frac{\left|\sum_{\ell=j}^k \left(\mathrm{iso}_v(\mathbf{Y})_{\pi_v(\ell)} - \mathrm{iso}_v(\mu_\star)_{\pi_v(\ell)}\right)\right|}{\sqrt{k-j+1}} \leq \|\mathrm{iso}_v(\mathbf{Y}) - \mathrm{iso}_v(\mu_\star)\|_{\mathrm{SW},\pi_v} \leq \|\mathbf{Y} - \mu_\star\|_{\mathrm{SW},\pi_v} = \|\mathbf{Z}\|_{\mathrm{SW},\pi_v}.$$

Now fix any $\ell \in \{1, \dots, n\}$ and let m_ℓ be the index such that $\ell \in S_{m_\ell}$. We have

$$\begin{aligned}
\mathrm{iso}_v(\mathbf{Y})_{\pi_v(\ell)} &\le \frac{\sum_{j=\ell}^{k_{m_\ell}} \mathrm{iso}_v(\mathbf{Y})_{\pi_v(j)}}{k_{m_\ell} - \ell + 1} \quad \text{since } \mathrm{iso}_v(\mathbf{Y}) \text{ is monotone nondecreasing} \\
&\le \frac{\sum_{j=\ell}^{k_{m_\ell}} \mathrm{iso}_v(\mu_\star)_{\pi_v(j)}}{k_{m_\ell} - \ell + 1} + \frac{\|\mathbf{Z}\|_{\mathrm{SW},\pi_v}}{\sqrt{k_m - \ell + 1}} \quad \text{by the bound above} \\
&\le \mathrm{iso}_v(\mu_\star)_{\pi_v(\ell)} + 2Bn^{-1/3} + \frac{\|\mathbf{Z}\|_{\mathrm{SW},\pi_v}}{\sqrt{k_{m_\ell} - \ell + 1}} \quad \text{by construction of the partition.}
\end{aligned}$$

Similarly,

$$\mathrm{iso}_v(\mathbf{Y})_{\pi_v(\ell)} \ge \mathrm{iso}_v(\mu_\star)_{\pi_v(\ell)} - 2Bn^{-1/3} - \frac{\|\mathbf{Z}\|_{\mathrm{SW},\pi_v}}{\sqrt{\ell - k_{m_\ell - 1}}}.$$

Therefore,

$$\left|\mathrm{iso}_v(\mathbf{Y})_{\pi_v(\ell)} - \mathrm{iso}_v(\mu_\star)_{\pi_v(\ell)}\right| \le 2Bn^{-1/3} + \frac{\|\mathbf{Z}\|_{\mathrm{SW},\pi_v}}{\sqrt{\min\{\ell - k_{m_\ell - 1}, k_{m_\ell} - \ell + 1\}}},$$

and so

$$\begin{aligned}
\|\mathrm{iso}_v(\mathbf{Y}) - \mathrm{iso}_v(\mu_\star)\|_2 &\le \sqrt{n} \cdot 2Bn^{-1/3} + \|\mathbf{Z}\|_{\mathrm{SW},\pi_v} \cdot \sqrt{\sum_{\ell=1}^{n} \frac{1}{\min\{\ell - k_{m_\ell - 1}, k_{m_\ell} - \ell + 1\}}} \\
&= 2Bn^{1/6} + \|\mathbf{Z}\|_{\mathrm{SW},\pi_v} \cdot \sqrt{\sum_{m=1}^{M} \sum_{\ell=1}^{k_m - k_{m-1}} \frac{1}{\min\{\ell, k_m - k_{m-1} - \ell + 1\}}} \\
&\le 2Bn^{1/6} + \|\mathbf{Z}\|_{\mathrm{SW},\pi_v} \sqrt{\sum_{m=1}^{M} 2\log\left(2(k_m - k_{m-1})\right)} \\
&= 2Bn^{1/6} + \|\mathbf{Z}\|_{\mathrm{SW},\pi_v} \sqrt{2\log\left(2^M (k_1 - k_0)(k_2 - k_1)\dots(k_M - k_{M-1})\right)}.
\end{aligned}$$

Since $0 = k_0 < k_1 < \cdots < k_M = n$, it holds that $(k_1 - k_0)(k_2 - k_1)\dots(k_M - k_{M-1}) \le (n/M)^M$, and therefore,

$$\|\mathrm{iso}_v(\mathbf{Y}) - \mathrm{iso}_v(\mu_\star)\|_2 \le 2Bn^{1/6} + \|\mathbf{Z}\|_{\mathrm{SW},\pi_v} \sqrt{2M\log(2n/M)}.$$

Since $M = \lceil n^{1/3} \rceil$, as long as $n \ge 2$ we can relax this to

$$\|\mathrm{iso}_v(\mathbf{Y}) - \mathrm{iso}_v(\mu_\star)\|_2 \le 2Bn^{1/6} + \|\mathbf{Z}\|_{\mathrm{SW},\pi_v} 2n^{1/6} \sqrt{\log n}.$$

Since this holds deterministically for any v, by considering $v = \mathbf{X}u$ we see that

$$\begin{aligned}
\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z}) = \sup_{u \in \mathbb{S}_s^{p-1}} \left\{ n^{-1/2} \|\mathrm{iso}_{\mathbf{X}u}(\mathbf{Y}) - \mathrm{iso}_{\mathbf{X}u}(\mu_\star)\|_2 \right\} \\
\le 2Bn^{-1/3} + 2n^{-1/3}\sqrt{\log n} \cdot \sup_{u \in \mathbb{S}_s^{p-1}} \left\{ \|\mathbf{Z}\|_{\mathrm{SW},\pi_{\mathbf{X}u}} \right\}.
\end{aligned}$$

Next define

$$\mathcal{S}_{n,p}^{s\text{-sparse}}(X_1, \dots, X_n) = \left\{ \pi \in \mathcal{S}_n : \pi = \pi_{\mathbf{X}u} \text{ for some } u \in \mathbb{S}_s^{p-1} \right\}.$$

In other words, this is the set of possible orderings of the entries of $\mathbf{X}u$, under any $u \in \mathbb{S}_s^{p-1}$. Write $\mathcal{S}_{n,p}^{s\text{-sparse}}(X_1, \ldots, X_n) = \{\pi_1, \ldots, \pi_K\}$ where $K = |\mathcal{S}_{n,p}^{s\text{-sparse}}(X_1, \ldots, X_n)|$. This means that for every s-sparse u,

$$\|\mathbf{Z}\|_{\mathrm{SW},\pi_{\mathbf{X}u}} = \|\mathbf{Z}\|_{\mathrm{SW},\pi_k} \text{ for some } k \in \{1, \ldots, K\},$$

and so we can write

$$\mathrm{Err}_{\mathrm{iso}}(\mathbf{Z}) \leq 2Bn^{-1/3} + 2n^{-1/3}\sqrt{\log n} \cdot \max_{k=1,\ldots,K} \|\mathbf{Z}\|_{\mathrm{SW},\pi_k}.$$

Finally, for σ-subgaussian zero-mean $\mathbf{Z}$ and any fixed permutation π, Lemma 3 of [36] proves that

$$\|\mathbf{Z}\|_{\mathrm{SW},\pi} \leq \sqrt{2\sigma^2 \log\left(\frac{n^2+n}{\delta'}\right)}$$

with probability at least $1 - \delta'$, for any $\delta' > 0$. Setting $\delta' = \frac{\delta}{2K}$, then,

$$\max_{k=1,\ldots,K} \|\mathbf{Z}\|_{\mathrm{SW},\pi_k} \leq \sqrt{2\sigma^2 \log\left(\frac{(n^2+n)\cdot 2K}{\delta}\right)}$$

with probability at least $1 - \frac{\delta}{2}$. Finally, Lemma 7 below proves that, deterministically,

$$K = \left|\mathcal{S}_{n,p}^{s\text{-sparse}}(X_1, \ldots, X_n)\right| \leq n^{2s-1}p^s.$$

Therefore, with probability at least $1 - \frac{\delta}{2}$,

$$\max_{k=1,\ldots,K} \|\mathbf{Z}\|_{\mathrm{SW},\pi_k} \leq \sqrt{2\sigma^2 \log\left(\frac{(n^2+n)\cdot 2n^{2s-1}p^s}{\delta}\right)},$$

which completes the proof.

A.2.1 Supporting lemma

Lemma 7. *For any sample size n, dimension p, sparsity level s, and points $x_1, \ldots, x_n \in \mathbb{R}^p$, define the set*

$$\mathcal{S}_{n,p}^{s\text{-sparse}}(x_1, \ldots, x_n) = \left\{\pi \in \mathcal{S}_n : x_{\pi(1)}^\top u \leq x_{\pi(2)}^\top u \leq \cdots \leq x_{\pi(n)}^\top u \text{ for some } s\text{-sparse } u \in \mathbb{R}^p\right\}.$$

Then we have

$$\left|\mathcal{S}_{n,p}^{x\text{-sparse}}(x_1, \ldots, x_n)\right| \leq n^{2s-1}p^s.$$

Proof. Define

$$\mathcal{S}_{n,s}(x_1, \ldots, x_n) = \left\{\pi \in \mathcal{S}_n : x_{\pi(1)}^\top u \leq x_{\pi(2)}^\top u \leq \cdots \leq x_{\pi(n)}^\top u \text{ for some } u \in \mathbb{R}^s\right\}.$$

First, for any s-sparse $u \in \mathbb{R}^p$, let $A \subseteq [p]$ be some subset of size $|A| = s$ containing the support of u. Denote the permutation π such that $x_{\pi(1)}^\top u \leq x_{\pi(2)}^\top u \leq \cdots \leq x_{\pi(n)}^\top u$ as $\pi_u(x_1, \ldots, x_n)$. Then it's clear that

$$\pi_u(x_1, \ldots, x_n) \in \mathcal{S}_{n,s}(x_{1A}, \ldots, x_{nA}).$$

where we write $x_{iA} \in \mathbb{R}^s$ as the subvector of x_i with entries indexed by A. Therefore we have

$$\left|\mathcal{S}_{n,p}^{s\text{-sparse}}(x_1,\ldots,x_n)\right| \le \sum_{A\subset[p],|A|\le s} |\mathcal{S}_{n,s}(x_{1A},\ldots,x_{nA})| \le p^s \cdot \max_{y_1,\ldots,y_n\in\mathbb{R}^s} |\mathcal{S}_{n,s}(y_1,\ldots,y_n)|,$$

since the number of subsets A is bounded by $\binom{p}{s} \le p^s$. Therefore, it suffices to compute a bound for the non-sparse case, i.e. to bound $|\mathcal{S}_{n,s}(y_1,\ldots,y_n)|$. Cover (1967) proves that for points $y_1,\ldots,y_n$ in generic position, this is exactly equal to $Q(n,s)$ which is defined by the recursive relation

$$Q(n+1,s) = \begin{cases} 1, & n = 0, \\ Q(n,s) + nQ(n,s-1), & s \ge 2, \text{ and } n \ge 1 \\ 2, & s = 1 \text{ and } n \ge 1. \end{cases}$$

If $y_1,\ldots,y_n$ are not in generic position, then the cardinality is instead bounded by $Q(n,s)$. Finally, using this recursive relation, we check that for the case $n = 1$, trivially $Q(1,s) = 1 \le 1^{2s-1}$ for any $s \ge 1$, and for the case $s = 1$ and $n \ge 2$, $Q(n,1) = 2 \le n^{2\cdot 1-1}$. Then for $n \ge 2$ and $s \ge 2$, we inductively have

$$Q(n+1,s) = Q(n,s)+nQ(n,s-1) \le n^{2s-1}+n\cdot n^{2(s-1)-1} = (n+1)\cdot n^{2s-2} \le (n+1)^{2s-1},$$

proving the desired bound. □

A.3 Proof of Lemma 1

Define $\tilde{u}_0 = \frac{1}{n}\mathbf{X}^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n)$. First, writing $\bar{g}_\star(\mathbf{X}u_\star) = \frac{1}{n}\sum_{i=1}^n g_\star(X_i^\top u_\star)$, we have

$$\begin{aligned}
\langle \tilde{u}_0, u_\star\rangle &= \frac{1}{n}\langle u_\star, \mathbf{X}^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n)\rangle \\
&= \frac{1}{n}\langle \mathbf{X}u_\star, g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\rangle + \frac{1}{n}\langle u_\star, \mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\rangle \\
&\ge \frac{1}{n}\langle \mathbf{X}u_\star, g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\rangle - \sqrt{s_\star}\|n^{-1}\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty \\
&\ge \frac{1}{nL}\|g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2 - \sqrt{s_\star}\|n^{-1}\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty,
\end{aligned}$$

where the first inequality holds since

$$\langle u_\star, \mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\rangle \le \|u_\star\|_1\|\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty$$

and $\|u_\star\|_1 \le \sqrt{s_\star}\|u_\star\|_2 = \sqrt{s_\star}$ since $u_\star$ is a $s_\star$-sparse unit vector, while the second inequality can be seen by applying Lemma 6 with $w = \mathbf{1}_n$ and $g = g_\star$. Next, we have

$$\begin{aligned}
\frac{1}{L^2 n}\|g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2 &\ge \frac{1}{L^2 n}\|g_\star(\mathbf{X}u_\star) - \mathrm{iso}_{-\mathbf{X}u}(g_\star(\mathbf{X}u_\star))\|_2^2 \\
&\ge \alpha\|u_\star - (-u_\star)\|_2^2 = 4\alpha,
\end{aligned}$$

where the first step holds since trivially, the constant vector $\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n$ agrees with the ordering of $-\mathbf{X}u_\star$, while the second step holds by assumption (8) applied with

$u = -u_\star$. In particular, by the assumption of the lemma, we see that $\langle \tilde{u}_0, u_\star \rangle > 0$, verifying that $\tilde{u}_0 \neq 0$. Note that, by definition, we have

$$u_0 = \frac{\Psi_s(\tilde{u}_0)}{\|\Psi_s(\tilde{u}_0)\|_2}.$$

Next, we apply Lemma 5 with $v = \frac{\tilde{u}_0}{\|\Psi_s(\tilde{u}_0)\|_2}$ and $w = u_\star$. Noting that $\Psi_s(v) = \frac{\Psi_s(\tilde{u}_0)}{\|\Psi_s(\tilde{u}_0)\|_2} = u_0$, this lemma yields

$$\left\langle \frac{\tilde{u}_0}{\|\Psi_s(\tilde{u}_0)\|_2} - u_0, u_\star - u_0 \right\rangle \leq \frac{\sqrt{s_\star}}{2\sqrt{s}} \|u_\star - u_0\|_2^2 = \sqrt{\frac{s_\star}{s}} \left(1 - \langle u_0, u_\star \rangle\right),$$

where the last step holds since u_0 and $u_\star$ are both unit vectors. We also have

$$\left\langle \frac{\tilde{u}_0}{\|\Psi_s(\tilde{u}_0)\|_2} - u_0, u_0 \right\rangle = 0,$$

since these two vectors have disjoint support by definition. Therefore,

$$\left\langle \frac{\tilde{u}_0}{\|\Psi_s(\tilde{u}_0)\|_2} - u_0, u_\star \right\rangle \leq \sqrt{\frac{s_\star}{s}} \left(1 - \langle u_0, u_\star \rangle\right),$$

and after rearranging terms, we have

$$\frac{1}{\|\Psi_s(\tilde{u}_0)\|_2} \langle \tilde{u}_0, u_\star \rangle \leq \sqrt{\frac{s_\star}{s}} + \left(1 - \sqrt{\frac{s_\star}{s}}\right) \langle u_0, u_\star \rangle.$$

Combining this with the calculations above, we have proved that

$$\langle u_0, u_\star \rangle \geq \left(1 - \sqrt{\frac{s_\star}{s}}\right)^{-1} \left(\frac{1}{\|\Psi_s(\tilde{u}_0)\|_2} \langle \tilde{u}_0, u_\star \rangle - \sqrt{\frac{s_\star}{s}}\right).$$

Combining this with our lower bound on $\langle \tilde{u}_0, u_\star \rangle$ calculated above, we now have

$$\begin{aligned}\langle u_0, u_\star \rangle \geq \frac{1}{\|\Psi_s(\tilde{u}_0)\|_2} \cdot \left(1 - \sqrt{\frac{s_\star}{s}}\right)^{-1} \cdot \\ \left(\frac{1}{nL} \|g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2 - \sqrt{s_\star} \|n^{-1}\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty - \sqrt{\frac{s_\star}{s}} \cdot \|\Psi_s(\tilde{u}_0)\|_2\right).\end{aligned}$$

Next, writing $S \subseteq \{1, \ldots, p\}$ to denote the support of u_0 (with $|S| \leq s$), we calculate

$$\begin{aligned}\|\Psi_s(\tilde{u}_0)\|_2 &= \|n^{-1}\mathbf{X}_S^\top(\mathbf{Y} - \bar{Y}\mathbf{1}_n)\|_2 \\ &\leq \|n^{-1}\mathbf{X}_S^\top(g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star))\|_2 + \|n^{-1}\mathbf{X}_S^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_2 \\ &\leq \|n^{-1}\mathbf{X}_S^\top(g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star))\|_2 + \sqrt{s} \cdot \|n^{-1}\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty,\end{aligned}$$

and so we now have

$$\begin{aligned}\langle u_0, u_\star \rangle \geq \frac{1}{\|\Psi_s(\tilde{u}_0)\|_2} \cdot \left(1 - \sqrt{\frac{s_\star}{s}}\right)^{-1} \cdot \\ \left(\frac{1}{nL} \|g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2 - 2\sqrt{s_\star} \|n^{-1}\mathbf{X}^\top(\mathbf{Z} - \bar{Z}\mathbf{1}_n)\|_\infty \right. \\ \left. - \sqrt{\frac{s_\star}{s}} \cdot \|n^{-1}\mathbf{X}_S^\top(g_\star(\mathbf{X}u_\star) - \bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n)\|_2\right).\end{aligned}$$

Furthermore,

$$\begin{aligned}\|n^{-1}\mathbf{X}_S^\top(g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n)\|_2 &= \sup_{u\in\mathbb{R}^p:\|u\|_2\le 1,\text{support}(u)\subseteq S}\left|n^{-1}u^\top\mathbf{X}^\top(g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n)\right| \\ &\le \sup_{u\in\mathbb{R}^p:\|u\|_2\le 1,\text{support}(u)\subseteq S}\|n^{-1}\mathbf{X}u\|_2\|(g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n)\|_2 \\ &\le \sqrt{\frac{\beta}{n}}\cdot\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2,\end{aligned}$$

where the last step holds by assumption (7). Combining everything, we have shown that

$$\begin{aligned}\langle u_0,u_\star\rangle \ge \frac{1}{\|\Psi_s(\tilde{u}_0)\|_2}\cdot\left(1-\sqrt{\frac{s_\star}{s}}\right)^{-1}\cdot \\ \left(\frac{1}{nL}\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2-2\sqrt{s_\star}\|n^{-1}\mathbf{X}^\top(\mathbf{Z}-\bar{Z}\cdot\mathbf{1})\|_\infty\right. \\ \left.-\sqrt{\frac{\beta s_\star}{sn}}\cdot\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2\right).\end{aligned}$$

Next, recalling that we have shown $\frac{1}{L^2n}\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2\ge 4\alpha$, since we've assumed that $s\ge s_\star\cdot\frac{\beta}{\alpha}$ we have

$$\begin{aligned}&\frac{1}{nL}\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2-\sqrt{\frac{\beta s_\star}{sn}}\cdot\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2 \\ &\quad\ge \frac{1}{nL}\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2^2-\sqrt{\frac{\alpha}{n}}\cdot\|g_\star(\mathbf{X}u_\star)-\bar{g}_\star(\mathbf{X}u_\star)\mathbf{1}_n\|_2\ge 2\alpha L.\end{aligned}$$

Therefore,

$$\langle u_0,u_\star\rangle \ge \frac{1}{\|\Psi_s(\tilde{u}_0)\|_2}\cdot\left(1-\sqrt{\frac{s_\star}{s}}\right)^{-1}\cdot\left(2\alpha L-2\sqrt{s_\star}\|n^{-1}\mathbf{X}^\top(\mathbf{Z}-\bar{Z}\mathbf{1}_n)\|_\infty\right).$$

This completes the proof.

A.4 Proofs for Proposition 1

By definition of isotonic regression,

$$\begin{aligned}& n^{-1/2}\|\text{iso}_{\mathbf{X}u_t}\mathbf{Y}-g_\star(\mathbf{X}u_\star)\|_2 \\ \le\ & n^{-1/2}\|\text{iso}_{\mathbf{X}u_t}\mathbf{Y}-\text{iso}_{\mathbf{X}u_t}g_\star(\mathbf{X}u_\star)\|_2+n^{-1/2}\|\text{iso}_{\mathbf{X}u_t}g_\star(\mathbf{X}u_\star)-g_\star(\mathbf{X}u_\star)\|_2 \\ \le\ & n^{-1/2}\|\text{iso}_{\mathbf{X}u_t}\mathbf{Y}-\text{iso}_{\mathbf{X}u_t}g_\star(\mathbf{X}u_\star)\|_2+n^{-1/2}\|g_\star(\mathbf{X}u_t)-g_\star(\mathbf{X}u_\star)\|_2 \\ \le\ & \text{Err}_{\text{iso}}(\mathbf{Z})+L\sqrt{\beta}\|u_t-u_\star\|_2,\end{aligned}$$

where the second inequality is from the contractive property of isotonic regression in ℓ_2-norm (see Yang and Barber [36]), and $\text{Err}_{\text{iso}}(\mathbf{Z})$ is defined as in the proof of Theorem 1. We finish the proof by plugging in the results from Lemma 4 and Theorem 1:

$$\begin{aligned}& n^{-1/2}\|\text{iso}_{\mathbf{X}u_t}\mathbf{Y}-g_\star(\mathbf{X}u_\star)\|_2 \\ \le\ & L\sqrt{2\beta}\cdot r^t+C(L\sqrt{\beta}+4\sigma)\left(\epsilon_n+\frac{s^{1/2}\log(np/\delta)}{n^{1/3}}\right)\ \text{with probability at least } 1-\delta.\end{aligned}$$

A.5 Proofs for the normal mixture X (Proposition 2)

In this section we provide the proof for Proposition 2. The main body of the proof is to verify that under the assumptions in section 2.3.1, conditions (7) and (8) are satisfied with high probability. We show this in the following two lemmas.

For the upper bound on sparse eigenvalue condition, we have the following result:

Lemma 8. *Assuming that the rows of* $\mathbf{X}$ *are i.i.d. draws from a distribution on* $X \in \mathbb{R}^p$, *with second moment* $\Sigma = \mathbb{E}\left[XX^\top\right]$ *satisfying*

$$c_0\mathbf{I}_p \preceq \Sigma \preceq c_1\mathbf{I}_p \text{ for some constants } c_1 \geq c_0 > 0,$$

then for all $\delta > 0$, *as* n *is large enough,* (7) *is satisfied with probability at least* $1 - \delta/2$ *with* $\beta = 2c_1$.

For the lower bound on sparse monotone single index eigenvalues, we have the following lemma:

Lemma 9. *Under assumptions* (5), (14), (15), *for all* $\delta > 0$, *as* n *is sufficiently large, there exists* $\epsilon_n \geq C_\alpha \cdot \sqrt{\frac{s\log(np/\delta)}{n}}$ *that* (8) *is satisfied with probability at least* $1 - \delta/2$, *with* $\alpha > 0$ *and* $C_\alpha > 0$ *being constants depending on* L, c_0, c_1, B, ν, V *but not on* n, p, s, δ.

Combining Lemmas 8 and 9 we finish the proof of Proposition 2. The proof of Lemmas 8 and 9 are deferred in section A.5.1.

A.5.1 Proof of Lemmas 8 and 9

Proof of Lemma 8. For any fixed support $S \subset [p]$ with $|S| = 2s + s_\star$, by 5.40 of [34],

$$\mathbb{P}\left\{\|\frac{1}{n}\mathbf{X}_S^\top\mathbf{X}_S - \mathrm{Cov}(X_S)\| \leq \sigma^2 C\left(\sqrt{\frac{2s + s_\star + \log(1/\delta)}{n}} + \frac{2s + s_\star + \log(1/\delta)}{n}\right)\right\} \geq 1 - \delta/2,$$

for a universal constant C. Taking a union bound over all $\binom{p}{2s+s_\star}$ possible supports S of size $|S| = 2s + s_\star$,

$$\mathbb{P}\left\{\sup_S \|\frac{1}{n}\mathbf{X}_S^\top\mathbf{X}_S - \mathrm{Cov}(X_S)\| \right.$$
$$\left. \leq \sigma^2 C\left(\sqrt{\frac{2s + s_\star + (2s + s_\star)\log(p/\delta)}{n}} + \frac{2s + s_\star + (2s + s_\star)\log(p/\delta)}{n}\right)\right\} \geq 1 - \delta/2.$$

Simplifying,

$$\mathbb{P}\left\{\sup_S \|\frac{1}{n}\mathbf{X}_S^\top\mathbf{X}_S - \mathrm{Cov}(X_S)\| \leq 6\sigma^2 C\left(\sqrt{\frac{s\log(p/\delta)}{n}} + \frac{s\log(p/\delta)}{n}\right)\right\} \geq 1 - \delta/2.$$

We can rewrite this as

$$\mathbb{P}\left\{\sup_{u \in \mathbb{S}_s^{p-1}} \frac{1}{n}\|\mathbf{X}u\|_2^2 \leq c_1 + 6\sigma^2 C\left(\sqrt{\frac{s\log(p/\delta)}{n}} + \frac{s\log(p/\delta)}{n}\right)\right\} \geq 1 - \delta/2,$$

where

$$\mathbb{S}_s^{p-1} = \{u \in \mathbb{R}^p : \|u\|_2 = 1,\ u \text{ is } s\text{-sparse}\}$$

is the set of s-sparse unit vectors in $\mathbb{R}^p$. Let $\beta = 2c_1$, then with n sufficiently large,

$$\mathbb{P}\left\{\sup_{u\in\mathbb{S}_s^{p-1}} \frac{1}{n}\|\mathbf{X}u\|_2^2 \le \beta\right\} \ge 1 - \delta/2.$$

This concludes the proof of Lemma 8. □

Proof of Lemma 9. We would like to place a lower bound on $\frac{1}{L^2 n}\|g(\mathbf{X}u) - g_\star(\mathbf{X}u_\star)\|_2^2$ for all $u \in \mathbb{S}_s^{p-1}$ with $\|u - u_\star\|_2 \ge \epsilon_n$ and all monotone non-decreasing functions $g : \mathbb{R} \to [-B, B]$, where $[-B, B]$ is the range of $g_\star$. For $\alpha > 0$ (we will specify α later in the proof), we have

$$\begin{aligned}
&\inf_{\substack{u\in\mathbb{S}_s^{p-1},\|u-u_\star\|_2\ge\epsilon_n \\ \text{non-decr. } g:\mathbb{R}\to[-B,B]}} \left(\sqrt{\frac{1}{L^2 n}\|g(\mathbf{X}u) - g_\star(\mathbf{X}u_\star)\|_2^2} - \sqrt{\alpha}\|u - u_\star\|_2\right) \\
&= \inf_{\substack{u\in\mathbb{S}_s^{p-1},\|u-u_\star\|_2\ge\epsilon_n \\ \text{non-decr. } g:\mathbb{R}\to[-B,B]}} \left(\sqrt{\frac{1}{L^2 n}\sum_i \left(g(X_i^\top u) - g_\star(X_i^\top u_\star)\right)^2} - \sqrt{\alpha}\|u - u_\star\|_2\right) \\
&\ge \inf_{\substack{u\in\mathbb{S}_s^{p-1},\|u-u_\star\|_2\ge\epsilon_n \\ \text{non-decr. } g:\mathbb{R}\to[-B,B]}} \left(\frac{1}{Ln}\sum_i \left|g(X_i^\top u) - g_\star(X_i^\top u_\star)\right| - \sqrt{\alpha}\|u - u_\star\|_2\right),
\end{aligned}$$

where the last step holds by the standard inequality relating the ℓ_1 and ℓ_2 norms. We can further decompose this as

$$\inf_{\substack{u\in\mathbb{S}_s^{p-1},\|u-u_\star\|_2\ge\epsilon_n \\ \text{non-decr. } g:\mathbb{R}\to[-B,B]}} \left(\sqrt{\frac{1}{L^2 n}\|g(\mathbf{X}u) - g_\star(\mathbf{X}u_\star)\|_2^2} - \sqrt{\alpha}\|u - u_\star\|_2\right) \ge \text{Term 1} - \text{Term 2}$$

where

$$\text{Term 1} = \inf_{\substack{u\in\mathbb{S}_s^{p-1},\|u-u_\star\|_2\ge\epsilon_n \\ \text{non-decr. } g:\mathbb{R}\to[-B,B]}} \left(\frac{1}{L}\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right] - \sqrt{\alpha}\|u - u_\star\|_2\right)$$

and

$$\text{Term 2} = \sup_{\substack{u\in\mathbb{S}_s^{p-1},\|u-u_\star\|_2\ge\epsilon_n \\ \text{non-decr. } g:\mathbb{R}\to[-B,B]}} \left(\frac{1}{L}\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right] - \frac{1}{Ln}\sum_i \left|g(X_i^\top u) - g_\star(X_i^\top u_\star)\right|\right).$$

For Term 1, by Lemma 10 (26), which states that for all s-sparse unit u and any non-decreasing function $g : \mathbb{R} \to [-B, B]$ we have

$$\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right]^2 \ge \alpha_\star\|u - u_\star\|_2^2,$$

where $\alpha_\star > 0$ is a constant which does not depend on n, p, s, ϵ_n. Therefore, let $\alpha = \frac{\alpha_\star}{2L^2}$,

$$\text{Term 1} \geq \sqrt{\alpha} \cdot \epsilon_n.$$

For Term 2, we will use concentration bounds to control Term 2. First, since both g and $g_\star$ take values in $[-B, B]$, and therefore $\left|g(X_i^\top u) - g_\star(X_i^\top u_\star)\right| \in [0, 2B]$ for any X_i and any non-decreasing $g : \mathbb{R} \to [-B, B]$ and $u \in \mathbb{S}_s^{p-1}$, we can see that replacing a single X_i with a different vector X_i' can perturb the value of Term 2 by at most $\frac{2B}{n}$. Therefore by McDiarmid's inequality [31], for any $\delta_1 > 0$,

$$\mathbb{P}\left\{\text{Term 2} > \mathbb{E}\left[\text{Term 2}\right] + 2B \cdot \sqrt{\frac{\log(1/\delta_1)}{2n}}\right\} \leq \delta_1.$$

Next, by the symmetrization lemma [32], we have

$$\mathbb{E}\left[\text{Term 2}\right] \leq 2\mathbb{E}\left[\sup_{\substack{s \in \mathbb{S}_s^{p-1} \\ \text{non-decr. } g : \mathbb{R} \to [-B,B]}} \frac{1}{n}\sum_i \xi_i \left|g(X_i^\top u) - g_\star(X_i^\top u_\star)\right|\right],$$

where $\xi_i \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$ are i.i.d. Rademacher variables, drawn independently from the X_i's. Next, consider the composition

$$X \mapsto \left(g(X_i^\top u) - g_\star(X_i^\top u_\star)\right) \mapsto \left|g(X_i^\top u) - g_\star(X_i^\top u_\star)\right|.$$

The second map is the absolute value function, which is 1-Lipschitz. By the Lipschitz inequality for Rademacher complexity, therefore,

$$\mathbb{E}\left[\text{Term 2}\right] \leq 4\mathbb{E}\left[\sup_{\substack{s \in \mathbb{S}_s^{p-1} \\ \text{non-decr. } g : \mathbb{R} \to [-B,B]}} \frac{1}{n}\sum_i \xi_i \left(g(X_i^\top u) - g_\star(X_i^\top u_\star)\right)\right] \leq 16B\sqrt{\frac{s\log(np)}{n}}.$$

The last inequality is because of Lemma 11.

Therefore, combining everything, with probability at least $1 - \delta/2$,

$$\text{Term 2} \leq B\left(16\sqrt{\frac{s\log(np)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}\right).$$

With n sufficiently large, let $\epsilon_n \geq \frac{16B}{\sqrt{\alpha}}\sqrt{\frac{s\log(np/\delta)}{n}}$, Term 2 $\leq 16B\sqrt{\frac{s\log(np/\delta)}{n}} \leq \sqrt{\alpha}\cdot\epsilon_n$. This concludes the proof of Lemma 9. □

A.5.2 Supporting lemmas

Lemma 10. *Suppose assumptions (5), (14), and (15) hold, then*

$$\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right]^2 \geq \alpha_\star \|u - u_\star\|_2^2 \tag{26}$$

holds for all s-sparse unit vectors u, and any non-decreasing function $g : \mathbb{R} \to [-B, B]$. where $\alpha_\star > 0$ is a constant depending on L, c_0, c_1, B, ν, V but not on n, p, s.

Proof. Let $A \in \{1, \ldots, K\}$ be the latent variable indicating which component X was drawn from. Fix any $u \in \mathbb{S}_s^{p-1}$. Then for any g,

$$\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right] = \sum_{k=1}^K a_k \mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right| \,\middle|\, A = k\right] = \sum_{k=1}^K a_k \mathbb{E}\left[\left|g(R_k) - g_\star(S_k)\right|\right],$$

where (R_k, S_k) is bivariate normal with mean

$$\begin{pmatrix} \mu_k^\top u \\ \mu_k^\top u_\star \end{pmatrix} =: m,$$

and covariance

$$\begin{pmatrix} u^\top \Sigma_k u & u^\top \Sigma_k u_\star \\ u_\star^\top \Sigma_k u & u_\star^\top \Sigma_k u_\star \end{pmatrix}.$$

Here we can calculate

$$\rho = \frac{u_\star^\top \Sigma_k u}{\sqrt{u^\top \Sigma_k u \cdot u_\star^\top \Sigma_k u_\star}} = \frac{\frac{c}{2} u^\top \Sigma_k u + \frac{1}{2c} u_\star^\top \Sigma_k u_\star - \frac{1}{2}(c^{1/2} u - c^{-1/2} u_\star)^\top \Sigma_k (c^{1/2} u - c^{-1/2} u_\star)}{\sqrt{u^\top \Sigma_k u \cdot u_\star^\top \Sigma_k u_\star}},$$

for any $c > 0$. Choosing $c = \sqrt{\frac{u_\star^\top \Sigma_k u_\star}{u^\top \Sigma_k u}}$, we obtain

$$\rho = 1 - \frac{(c^{1/2} u - c^{-1/2} u_\star)^\top \Sigma_k (c^{1/2} u - c^{-1/2} u_\star)}{2\sqrt{u^\top \Sigma_k u \cdot u_\star^\top \Sigma_k u_\star}} \leq 1 - \frac{c_0}{2c_1} \|c^{1/2} u - c^{-1/2} u_\star\|_2^2,$$

where c_0 and c_1 are the smallest and largest restricted eigenvalues of Σ_k. Next,

$$\|c^{1/2} u - c^{-1/2} u_\star\|_2^2 = c\|u\|_2^2 + c^{-1}\|u_\star\|_2^2 - 2\langle u, u_\star\rangle = c + c^{-1} - 2\langle u, u_\star\rangle \geq 2 - 2\langle u, u_\star\rangle = \|u - u_\star\|_2^2,$$

since u and $u_\star$ are both unit vectors. Combining everything,

$$\rho \leq 1 - \frac{c_0}{2c_1} \|u - u_\star\|_2^2.$$

Note that $|m_2| \leq V$ since $\|\mu_k\|_2 \leq V$, and $u^\top \Sigma_k u, u_\star^\top \Sigma_k u_\star \geq c_0$ since $\Sigma_k \succeq c_0 \mathbf{I}_p$. Now suppose for the moment that $g_\star(T) - g_\star(-T) \geq C$ for some $T > 0$ and $C > 0$. Applying Lemma 12 in the Appendix, we see that

$$\mathbb{E}\left[\left|g(R_k) - g_\star(S_k)\right|\right] \geq C_0 \sqrt{1 - (\rho \vee 0)^2},$$

where $C_0 > 0$ depends only on L, C, T, c_0, V, for each $k = 1, \ldots, K$. Plugging in our bound on ρ, we see that

$$\rho \vee 0 \leq 1 - \frac{c_0}{4c_1} \|u - u_\star\|_2^2$$

must hold, since $c_0 \leq c_1$ and $\|u - u_\star\|_2^2 \leq 4$ (since u and $u_\star$ are unit vectors).

$$\mathbb{E}\left[\left|g(R_k) - g_\star(S_k)\right|\right] \geq C_0 \sqrt{1 - \left(1 - \frac{c_0}{4c_1} \|u - u_\star\|_2^2\right)^2} \geq C_0 \sqrt{\frac{c_0}{2c_1}} \|u - u_\star\|_2.$$

Since this is true for each k, therefore we have

$$\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right] \geq C_0\sqrt{\frac{c_0}{2c_1}}\|u - u_\star\|_2.$$

Finally, we just need to find $T, C > 0$ such that $g_\star(T) - g_\star(-T) \geq C$. Recall that $\mathrm{Var}\left(g_\star(X^\top u_\star)\right) \geq \nu^2 > 0$ by assumption. Recall also that $g_\star$ takes values in $[-B, B]$. Define

$$T = V + \sqrt{c_1} \cdot \Phi^{-1}\left(1 - \frac{\nu^2}{16B^2}\right).$$

Then

$$\begin{aligned}\mathbb{P}\left\{|X^\top u_\star| > T\right\} &= \max_k \mathbb{P}\left\{\left|\mathcal{N}(u_\star^\top \mu_k, u_\star^\top \Sigma_k u_\star)\right| > T\right\} \\ &\leq \max_k \mathbb{P}\left\{|\mathcal{N}(0,1)| > \frac{T - |u_\star^\top \mu_k|}{\sqrt{u_\star^\top \Sigma_k u_\star}}\right\} \leq \frac{\nu^2}{8B^2},\end{aligned}$$

by our bounds on μ_k and Σ_k and our definition of T. Now let $C = g_\star(T) - g_\star(-T)$. Then

$$\begin{aligned}\nu^2 &= \mathrm{Var}\left(g_\star(X^\top u_\star)\right) \\ &\leq \mathbb{E}\left[\left(g_\star(X^\top u_\star) - g_\star(0)\right)^2\right] \\ &= \mathbb{E}\left[\left(g_\star(X^\top u_\star) - g_\star(0)\right)^2 \cdot \mathbb{1}\left\{|X^\top u_\star| > T\right\}\right] + \mathbb{E}\left[\left(g_\star(X^\top u_\star) - g_\star(0)\right)^2 \cdot \mathbb{1}\left\{|X^\top u_\star| \leq T\right\}\right] \\ &\leq \mathbb{E}\left[4B^2 \cdot \mathbb{1}\left\{|X^\top u_\star| > T\right\}\right] + \mathbb{E}\left[C^2\right] \\ &= 4B^2\mathbb{P}\left\{|X^\top u_\star| > T\right\} + C^2 \\ &\leq 4B^2 \cdot \frac{\nu^2}{8B^2} + C^2 \\ &= \nu^2/2 + C^2,\end{aligned}$$

and therefore, we must have $C \geq \nu/\sqrt{2}$. Combining everything,

$$\mathbb{E}\left[\left|g(X^\top u) - g_\star(X^\top u_\star)\right|\right] \geq \sqrt{\alpha_\star}\|u - u_\star\|_2$$

for all $u \in \mathbb{S}_s^{p-1}$, where $\alpha_\star > 0$ is a constant depending on L, c_0, c_1, B, ν, V. □

Lemma 11. *With $g_\star : \mathbb{R} \to [a, b]$, and $b - a \leq 2B$,*

$$\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u \\ \text{non-decr. } g:\mathbb{R}\to[a,b]}} \frac{1}{n}\sum_i \xi_i\big(g(X_i^\top u) - g_\star(X_i^\top u_\star)\big)\right] \leq 4B\sqrt{\frac{s\log(np)}{n}}$$

Proof of Lemma 11. The expectation can be simplified as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ \text{non-decr. } g:\mathbb{R}\to[a,b]} \frac{1}{n}\sum_i \xi_i\big(g(X_i^\top u) - g_\star(X_i^\top u_\star)\big)\right]\\
=&\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ \text{non-decr. } g:\mathbb{R}\to[a,b]} \frac{1}{n}\sum_i \xi_i\cdot\big(g(X_i^\top u) - a\big)\right] - 4\mathbb{E}\left[\frac{1}{n}\sum_i \xi_i\cdot\big(g_\star(X_i^\top u_\star) - a\big)\right]\\
=&\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ \text{non-decr. } g:\mathbb{R}\to[a,b]} \frac{1}{n}\sum_i \xi_i\cdot\big(g(X_i^\top u) - a\big)\right]\\
=&2B\cdot\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ \text{non-decr. } g:\mathbb{R}\to[0,1]} \frac{1}{n}\sum_i \xi_i\cdot g(X_i^\top u)\right].
\end{aligned}$$

where the next-to-last step holds since $\mathbb{E}[\xi_i] = 0$. Next, any non-decreasing function $g : \mathbb{R} \to [0,1]$ can be written as a (possibly infinite) convex combination of step functions. This means that for any u, the supremum over g is attained at some step function, in other words, we can write

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ \text{non-decr. } g:\mathbb{R}\to[a,b]} \frac{1}{n}\sum_i \xi_i\big(g(X_i^\top u) - g_\star(X_i^\top u_\star)\big)\right]\\
&\qquad\qquad\le 2B\cdot\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ t\in\mathbb{R}}} \frac{1}{n}\sum_i \xi_i\cdot\mathbb{1}\left\{X_i^\top u\ge t\right\}\right].
\end{aligned}$$

Now let $\pi_{u;X_1,\dots,X_n}$ be the permutation of $\{1,\dots,n\}$ induced by the ordering of the $X_i^\top u$'s, that is, the permutation $\pi \in \mathcal{S}_n$ that satisfies

$$X_{\pi(1)}^\top u \le X_{\pi(2)}^\top u \le \cdots \le X_{\pi(n)}^\top u.$$

To handle the possibility that this might not define a unique permutation, i.e. since it might be the case that $\langle X_i, u\rangle = \langle X_{i'}, u\rangle$ for some $i \ne i'$, we place an arbitrary total ordering on $\mathcal{S}_n$ and then, in the case of ties, we define $\pi_{u;X_1,\dots,X_n}$ as the minimal permutation satisfying the ordering above.

Then the above can be rewritten as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u\\ \text{non-decr. } g:\mathbb{R}\to[a,b]} \frac{1}{n}\sum_i \xi_i\big(g(X_i^\top u) - g_\star(X_i^\top u_\star)\big)\right]\\
&\qquad\qquad\le 2B\cdot\mathbb{E}\left[\max\left\{0, \sup_{\substack{s\text{-sparse unit } u\\ k\in[n]}} \frac{1}{n}\sum_{i\ge k}\xi_{\pi_{u;X_1,\dots,X_n}(i)}\right\}\right].
\end{aligned}$$

Now define

$$\Pi_{p,s}(X_1,\dots,X_n) = \left\{\pi_{u;X_1,\dots,X_n} : s\text{-sparse unit } u \in \mathbb{R}^p\right\} \subseteq \mathcal{S}_n,$$

i.e. the set of all permutations attained by any sparse vector u (using the given feature vectors $X_1,\dots,X_n$). The above can therefore be rewritten as

$$\mathbb{E}\left[\sup_{\substack{s\text{-sparse unit } u \\ \text{non-decr. } g:\mathbb{R}\to[a,b]}} \frac{1}{n}\sum_i \xi_i\big(g(X_i^\top u) - g_\star(X_i^\top u_\star)\big)\right] \leq 2B\cdot\mathbb{E}\left[\max\left\{0, \sup_{\substack{\pi\in\Pi_{p,s}(X_1,\dots,X_n) \\ k\in[n]}} \frac{1}{n}\sum_{i\geq k}\xi_{\pi(i)}\right\}\right].$$

By Lemma 7 in the Appendix, for any $X_1,\dots,X_n$ we have $|\Pi_{p,s}(X_1,\dots,X_n)| \leq n^{2s-1}p^s$. Now, for each π and each k, $\frac{1}{n}\sum_{i\geq k}\xi_{\pi(i)}$ is a zero-mean subgaussian random variable at the scale $\frac{\sqrt{n-k+1}}{n} \leq n^{-1/2}$, and we are taking a supremum over $n^{2s-1}p^s\cdot n = n^{2s}p^s$ such variables. This expected value is therefore bounded as $\sqrt{\frac{2\log(2n^{2s}p^s)}{n}} \leq 2\sqrt{\frac{s\log(np)}{n}}$ (where the last step holds as long as $p\geq 2$). □

Lemma 12. *Let*

$$\begin{pmatrix} R \\ S \end{pmatrix} \sim \mathcal{N}\left(m, \begin{pmatrix} r^2 & rs\cdot\rho \\ rs\cdot\rho & s^2 \end{pmatrix}\right),$$

and let $g, g_\star : \mathbb{R}\to\mathbb{R}$ be any non-decreasing functions, such that $g_\star$ is L-Lipschitz and satisfies $g_\star(T) - g_\star(-T) \geq C > 0$ for some $T > 0$. Then

$$\mathbb{E}\left[\left|g(R) - g_\star(S)\right|\right] \geq C_0\sqrt{1-(\rho\vee 0)^2},$$

where $C_0 > 0$ is a function of $L, T, C, s, |m_2|$ (not dependent on ρ), and is defined in the proof.

Proof of Lemma 12. First, we will replace R with $\tilde{R} = \frac{R-m_1}{r}$ and S with $\tilde{S} = \frac{S-m_2}{s}$, to obtain

$$\begin{pmatrix} \tilde{R} \\ \tilde{S} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Define non-decreasing functions $\tilde{g}(t) = g(m_1 + r\cdot t)$ and $\tilde{g}_\star(t) = g_\star(m_2 + s\cdot t)$. Then we are looking to lower bound

$$\mathbb{E}\left[\left|g(R) - g_\star(S)\right|\right] = \mathbb{E}\left[\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\tilde{S})\right|\right].$$

Note that $\tilde{g}_\star$ is $\tilde{L}$-Lipschitz for $\tilde{L} = Ls$, and

$$\tilde{g}_\star\left(\frac{T-m_2}{s}\right) - \tilde{g}_\star\left(\frac{-T-m_2}{s}\right) = g_\star(T) - g_\star(-T) \geq C,$$

and since $\tilde{g}_\star$ is nondecreasing, we can weaken this bound to

$$\tilde{g}_\star(\tilde{T}) - \tilde{g}_\star(-\tilde{T}) \geq C \text{ where } \tilde{T} = \frac{T+|m_2|}{s}.$$

First consider the case that $\rho \le \rho_0$, where $\rho_0 \in [0,1)$ is defined to satisfy $2Ls\sqrt{1-\rho_0^2} < \frac{1}{2}C\sqrt{2/\pi}$. In this case, define $c = \frac{\tilde{g}_\star(\tilde{T})+\tilde{g}_\star(-\tilde{T})}{2}$, then $\tilde{g}_\star(\tilde{T})-c \ge \frac{C}{2}$ and $c-\tilde{g}_\star(-\tilde{T}) \ge \frac{C}{2}$. First suppose that $\tilde{g}(0) \le c$. Then $\tilde{g}(r) \le c$ for any $r \le 0$ since $\tilde{g}$ is non-decreasing. Now, if $\tilde{S} \ge T$ and $\tilde{R} \le 0$, this means that

$$\tilde{g}_\star(\tilde{S}) \ge \tilde{g}_\star(T) \ge \frac{C}{2} + c \ge \frac{C}{2} + \tilde{g}(\tilde{R}),$$

and so $\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\tilde{S})\right| \ge \frac{C}{2}$. In other words,

$$\mathbb{E}\left[\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\tilde{S})\right|\right] \ge \frac{C}{2} \cdot \mathbb{P}\left\{\tilde{S} \ge T, \tilde{R} \le 0\right\}.$$

Finally,

$$\mathbb{P}\left\{\tilde{S} \ge T, \tilde{R} \le 0\right\} \ge \mathbb{P}\left\{\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}\right) \in [T,\infty) \times (-\infty, 0]\right\} =: P_0(\rho_0, T) > 0,$$

since for the first inequality, the probability is minimized at $\rho = \rho_0$ (under the assumption $\rho \le \rho_0$), and is therefore lower bounded by some positive value $P_0 = P_0(T, \rho_0)$. Therefore, the conclusion of the lemma holds in this case, as long as we set $C_0 \le \frac{C}{2} \cdot P_0(T, \rho_0)$.

Next we will work on the case that $\rho \ge \rho_0$. Let

$$\check{S} = \rho \cdot \tilde{R} - \left(\tilde{S} - \rho \cdot \tilde{R}\right),$$

so that $(\tilde{R}, \check{S})$ is equal in distribution to $(\tilde{R}, \tilde{S})$. Then

$$\mathbb{E}\left[\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\tilde{S})\right|\right] = \mathbb{E}\left[\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\check{S})\right|\right],$$

and so

$$\begin{aligned}
\mathbb{E}\left[\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\tilde{S})\right|\right] &= \frac{1}{2}\mathbb{E}\left[\left(\left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\tilde{S})\right| + \left|\tilde{g}(\tilde{R}) - \tilde{g}_\star(\check{S})\right|\right)\right] \\
&\ge \frac{1}{2}\mathbb{E}\left[\left|\tilde{g}_\star(\tilde{S}) - \tilde{g}_\star(\check{S})\right|\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left|\tilde{g}_\star(\tilde{S}) - \tilde{g}_\star\left(\frac{\tilde{S}+\check{S}}{2}\right)\right| + \left|\tilde{g}_\star\left(\frac{\tilde{S}+\check{S}}{2}\right) - \tilde{g}_\star(\check{S})\right|\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left|\tilde{g}_\star(\tilde{S}) - \tilde{g}_\star(\rho\cdot\tilde{R})\right| + \left|\tilde{g}_\star(\rho\cdot\tilde{R}) - \tilde{g}_\star(\check{S})\right|\right],
\end{aligned}$$

where the third step holds since since $g_\star$ is non-decreasing. Now let $W = \tilde{S} - \rho \cdot \tilde{R} \sim$

$\mathcal{N}(0, 1-\rho^2)$, and note $W \perp\!\!\!\perp \tilde{R}$. Then we can rewrite this as

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\left|\tilde{g}_\star(\tilde{S}) - \tilde{g}_\star(\rho\cdot\tilde{R})\right| + \left|\tilde{g}_\star(\rho\cdot\tilde{R}) - \tilde{g}_\star(\check{S})\right|\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left|\tilde{g}_\star(\rho\cdot\tilde{R}+W) - \tilde{g}_\star(\rho\cdot\tilde{R}-W)\right|\right] \\
&= \frac{1}{2}\mathbb{E}_W\left[\int_{t=-\infty}^{\infty}\phi(t)\cdot\left|\tilde{g}_\star\big(\rho\cdot t+W\big) - \tilde{g}_\star\big(\rho\cdot t-W\big)\right|dt\right] \text{ since } \tilde{R}\sim\mathcal{N}(0,1) \text{ is independent of } W \\
&\geq \frac{1}{2}\mathbb{E}_W\left[\int_{t=-\tilde{T}/\rho}^{\tilde{T}/\rho}\phi(t)\cdot\left|\tilde{g}_\star\big(\rho\cdot t+W\big) - \tilde{g}_\star\big(\rho\cdot t-W\big)\right|dt\right] \text{ since the integrand is nonnegative} \\
&= \frac{1}{2}\mathbb{E}_W\left[\int_{t=-\tilde{T}/\rho}^{\tilde{T}/\rho}\phi(t)\cdot\Big(\tilde{g}_\star\big(\rho\cdot t+|W|\big) - \tilde{g}_\star\big(\rho\cdot t-|W|\big)\Big)dt\right] \text{ since } \tilde{g}_\star \text{ is non-decreasing} \\
&= \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\int_{t=-\tilde{T}/\rho}^{\tilde{T}/\rho}\Big(\tilde{g}_\star\big(\rho\cdot t+|W|\big) - \tilde{g}_\star\big(\rho\cdot t-|W|\big)\Big)dt\right] \text{ since } \phi(t)\geq\phi(\tilde{T}/\rho) \text{ for all } |t|\leq\tilde{T}/\rho \\
&= \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\int_{t=-\tilde{T}/\rho}^{\tilde{T}/\rho}\tilde{g}_\star\big(\rho\cdot t+|W|\big)\,\mathsf{d}t - \int_{t=-\tilde{T}/\rho}^{\tilde{T}/\rho}\tilde{g}_\star\big(\rho\cdot t-|W|\big)dt\right] \\
&= \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\int_{t=-\tilde{T}/\rho+|W|/\rho}^{\tilde{T}/\rho+|W|/\rho}\tilde{g}_\star\big(\rho\cdot t\big)\,\mathsf{d}t - \int_{t=-\tilde{T}/\rho-|W|/\rho}^{\tilde{T}/\rho-|W|/\rho}\tilde{g}_\star\big(\rho\cdot t\big)dt\right] \\
&\text{by a change of variables in each integral} \\
&= \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\int_{t=\tilde{T}/\rho-|W|/\rho}^{\tilde{T}/\rho+|W|/\rho}\tilde{g}_\star\big(\rho\cdot t\big)\,\mathsf{d}t - \int_{t=-\tilde{T}/\rho-|W|/\rho}^{-\tilde{T}/\rho+|W|/\rho}\tilde{g}_\star\big(\rho\cdot t\big)dt\right] \\
&\text{by cancellation of the overlapping regions of integration} \\
&\geq \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\frac{2|W|}{\rho}\Big(\tilde{g}_\star(\tilde{T}-|W|) - \tilde{g}_\star(-\tilde{T}+|W|)\Big)\right] \text{ since } \tilde{g}_\star \text{ is non-decreasing} \\
&\geq \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\frac{2|W|}{\rho}\Big(\tilde{g}_\star(\tilde{T}) - \tilde{g}_\star(-\tilde{T}) - 2|W|\tilde{L}\Big)\right] \text{ since } \tilde{g}_\star \text{ is } \tilde{L}\text{-Lipschitz} \\
&\geq \frac{\phi(\tilde{T}/\rho)}{2}\mathbb{E}_W\left[\frac{2|W|}{\rho}\Big(C - 2|W|\tilde{L}\Big)\right] \text{ since } \tilde{g}_\star(\tilde{T}) - \tilde{g}_\star(-\tilde{T}) \geq C \\
&= \frac{1}{\rho}\phi(\tilde{T}/\rho)\left(C\mathbb{E}\left[|W|\right] - 2\tilde{L}\mathbb{E}\left[W^2\right]\right) \\
&= \frac{1}{\rho}\phi(\tilde{T}/\rho)\left(C\sqrt{2/\pi}\sqrt{1-\rho^2} - 2\tilde{L}(1-\rho^2)\right) \text{ since } W\sim\mathcal{N}(0,1-\rho^2).
\end{aligned}
$$

Plugging in our definitions of $\tilde{T}$ and $\tilde{L}$ and combining everything,

$$\mathbb{E}\left[\left|g(R) - g_\star(S)\right|\right] \geq \frac{1}{\rho}\cdot\phi\left(\frac{T+|m_2|}{s\rho}\right)\cdot\left(C\sqrt{2/\pi}\sqrt{1-\rho^2} - 2Ls(1-\rho^2)\right).$$

Now, recall that ρ_0 satisfies $2Ls\sqrt{1-\rho_0^2} < \frac{1}{2}C\sqrt{2/\pi}$. Then since $\frac{1}{\rho}\geq 1\geq\rho_0$,

$$\mathbb{E}\left[\left|g(R) - g_\star(S)\right|\right] \geq \sqrt{1-\rho^2}\cdot\rho_0\cdot\phi\left(\frac{T+|m_2|}{s\rho_0}\right)\cdot\left(C\sqrt{2/\pi} - 2Ls\sqrt{1-\rho_0^2}\right).$$

Therefore, as long as $C_0 \leq \rho_0 \cdot \phi\left(\frac{T+|m_2|}{s\rho_0}\right) \cdot \left(C\sqrt{2/\pi} - 2Ls\sqrt{1-\rho_0^2}\right)$, in both cases ($\rho \geq \rho_0$ and $\rho < \rho_0$), we have proved that $\mathbb{E}\left[\left|g(R) - g_\star(S)\right|\right] \geq C_0\sqrt{1-\rho^2}$ in the case that $\rho \geq \rho_0$, as desired. $\square$