Electronic Journal of Statistics
Vol. 13 (2019) 710–746
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1537

Isotonic regression meets LASSO

Matey Neykov

Department of Statistics & Data Science,
Carnegie Mellon University,
Pittsburgh, PA 15213, USA
e-mail: mneykov@stat.cmu.edu

Abstract: This paper studies a two step procedure for monotone increasing additive single index models with Gaussian designs. The proposed procedure is simple, easy to implement with existing software, and consists of consecutively applying LASSO and isotonic regression. Aside from formalizing this procedure, we provide theoretical guarantees regarding its performance: 1) we show that our procedure controls the in-sample squared error; 2) we demonstrate that one can use the procedure for predicting new observations, by showing that the absolute prediction error can be controlled with high-probability. Our bounds show a tradeoff of two rates: the minimax rate for estimating high dimensional quadratic loss, and the minimax nonparametric rate for estimating a monotone increasing function.

Keywords and phrases: Monotone single index models, isotonic regression, LASSO, sparsity, high dimensional statistics.

Received February 2018.

1. Introduction

Linear regression modeling and least squares estimation are two closely related topics with pervasive applications in the field of statistics and beyond. Often times however, linear models are simply approximations to the true data generating mechanism. Nonparametric and semiparametric regressions present more flexible alternatives to the linear model, and have been studied extensively in the statistics literature in low dimensional regimes [36, 40, 21].

This paper focuses on the *high dimensional, sparse, monotone increasing, additive, semiparametric* generalization of the linear regression model:

$$Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}^*) + \varepsilon, \tag{1.1}$$

where the noise ε is independent of the design $\boldsymbol{X}$. Here, the terms high dimensional and sparse, refer to the fact that the vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ for a large ambient dimension p, but it is assumed to be sparse, i.e., its effective support (the set of its non-zero elements) is small; the terms semiparametric and monotone increasing refer to the "link" function f, which is assumed to be unspecified (unknown) and monotone increasing, while the term additive refers to the fact that the noise term ε is additive. In addition to assuming that f is increasing, in this paper we further suppose that f is a Lipschitz function, which is in general not necessary when studying models of the type (1.1), but it eases the theoretical analysis in the high dimensional regime. On an important note, model (1.1) is not fully identifiable, as multiplying $\boldsymbol{\beta}^*$ by a constant and dividing f by the same constant generates the same outcome; similarly adding and subtracting a constant from f and ε yields the same outcome. However (1.1) can be easily identified by assuming that f is increasing*, fixing any function of $\boldsymbol{\beta}^*$ which attains different values when scaling $\boldsymbol{\beta}^*$ proportionally, such as any vector norm, and further assuming that the noise ε has a zero mean†. Notice that the linear regression model is a special case of the model (1.1) where f is a linear function with positive slope. Model (1.1) is also known as monotone single index model (SIM), and is a special case of the general SIM which imposes less restrictions on the outcome generation. Specifically, the general SIM framework assumes

$$Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}^*, \varepsilon). \tag{1.2}$$

Note that unlike (1.1), (1.2) makes no monotonicity assumptions on f, and the error term need not be additive.

Assuming a general SIM (1.2) with Gaussian design $\boldsymbol{X}$, and that the error term ε is independent of the design $\boldsymbol{X}$, [32, 38] showed that a constrained variant of the LASSO [39] can successfully estimate $\boldsymbol{\beta}^*$ (up to a scalar), while [30] showed that LASSO can recover the support of $\boldsymbol{\beta}^*$, given that

$$\mathrm{Cov}(Y, \boldsymbol{X}^\top \boldsymbol{\beta}^*) \neq 0. \tag{1.3}$$

A nice feature of model (1.1) is that condition (1.3) holds true by default. Indeed, it is not hard to check that in (1.1) with a non-constant increasing link function f‡, Gaussian design $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, and noise ε independent of the design, one has

$$\mathrm{Cov}(Y, \boldsymbol{X}^\top \boldsymbol{\beta}^*) = \mathrm{Cov}(f(\boldsymbol{X}^\top \boldsymbol{\beta}^*), \boldsymbol{X}^\top \boldsymbol{\beta}^*) > 0.\text{§}$$

Therefore, under the above assumptions, given existing results, model (1.1) is amenable to a LASSO application for obtaining proportional estimates of the vector $\boldsymbol{\beta}^*$.

Isotonic regression is a line of work which is highly relevant to model (1.1) yet very distinct from the aforementioned LASSO developments. Specifically, isotonic regression aims to estimate the increasing function f after observing n samples from the model

$$Y = f(X) + \varepsilon,$$

where the design $X \in \mathbb{R}$ is considered fixed, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise. The model fitting is done via the following optimization procedure

$$\underset{f \text{ is increasing}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^n (Y_i - f(X_i))^2$$

[*] To fix multiplication by ±1.

[†] Note that since ε and $\boldsymbol{X}$ are independent this is equivalent to assuming that $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$.

[‡] f needs to assume at least two distinct values on sets of non-zero Lebesgue measure.

[§] The last inequality follows from the Chebyshev's association inequality [cf. Theorem 2.14 6].

There has been a multitude of advances on isotonic regression [12, 45, 13, 8, 7, 4], where the authors showed estimation rates for general increasing, as well as for piecewise constant f.

This manuscript puts the two ideas above into one procedure to tackle problems of the type (1.1). Specifically, we study a two step estimator, which starts with applying LASSO to obtain an initial estimate $\overline{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, then "plugs-in" $\overline{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}^*$ effectively reducing the dimension from p to 1. In the second step the procedure uses isotonic regression to obtain an estimate of f. Finally, predicting a new observation is done on the basis of how close this observation is to a data point along the direction $\overline{\boldsymbol{\beta}}$. A detailed description of the procedure along with more in-depth motivation can be found in Section 2.

1.1. Contributions

This manuscript contains two main contributions, which are proved under the assumption that $\boldsymbol{X}_i^\top \boldsymbol{\beta}^* \sim \mathcal{N}(0,1)$ and that the link function f is increasing and Lipschitz with Lipschitz constant L. In order to informally summarize our findings, recall that $\overline{\boldsymbol{\beta}}$ denotes the LASSO estimate of the direction $\boldsymbol{\beta}^*$, and let $\widehat{f}$ denote the isotonic regression estimate of f, given a sample of n observations $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ of model (1.1). An informal version of our first main finding is that applying sequentially LASSO and isotonic regression gives the following bound on the in-sample squared ℓ_2 prediction error

$$\frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}))^2 \le r_1 \vee r_2, \tag{1.4}$$

with high probability. In the above inequality we observe the tradeoff of two rates r_1 and r_2 which omitting constants for simplification are

$$r_1 \asymp \sigma^2 \left[\frac{\sigma + |f(\max \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}) - f(\min \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}})|}{n\sigma}\right]^{\frac{2}{3}}, \; r_2 \asymp \frac{L^2 s \log p}{n}.$$

These two rates have natural interpretations: r_1 is the general nonparametric rate of isotonic regression for increasing functions as shown in [8, 4]; r_2 is the standard estimation rate for LASSO of the ℓ_2^2 norm [5].

Our second main contribution is to derive guarantees for the conditional mean integrated absolute error (CMIAE), which is a measurement of the prediction capability of our estimate. The CMIAE is defined as

$$\text{CMIAE} := \mathbb{E}\big[|f(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\min \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}, \;\max \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}]\big],$$

and measures the absolute integrated error of prediction for a new observation $\boldsymbol{X}$, conditional on the new observation lying within the data estimated range. Informally, we can show that up to constant factors with high probability the CMIAE satisfies:

$$\text{CMIAE} \le [r_1 \vee r_2]^{\frac{1}{2}}.$$

In other words, conditional on a new observation falling within the range $[\min \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}, \ \max \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}]$, the squared mean absolute integrated error obeys a similar inequality to the one in (1.4). The difference between (1.4) and the inequality in the preceding display is that in (1.4) we evaluate the performance of the estimator only on the dataset, while the CMIAE inequality shows that the estimator performs well on new observations. This is a non-trivial distinction, both in practice and in theory.

The major difficulty in showing the above two results stems from the fact that one uses the data once when estimating $\overline{\boldsymbol{\beta}}$ and therefore in the second step when one applies isotonic regression complicated data dependencies may occur.

1.2. Related work

Estimation for SIMs in the case when p is fixed, has been studied extensively in the literature [see e.g., 42, 20, 31, 28, among others]. Recently there have been active developments for high dimensional SIMs. The first work on high dimensional SIM was [2] where the authors proposed a PAC-Bayesian approach. [32] and later [38] demonstrated that under condition (1.3) running the least squares with ℓ_1 regularization can obtain a consistent estimate of the direction of $\boldsymbol{\beta}^*$, while [30] showed that this procedure also recovers the signed support of the direction. Even more recently [16, 43] extended some of those ideas to cases where the design has a known elliptical and not necessarily Gaussian distribution, yet non of these works offer estimation of the unknown function f. It is noteworthy to mention that condition (1.3) which is instrumental in the aforementioned papers traces roots to the following seminal works [26, 25, 10] in the area of sufficient dimension reduction.

Some of the recent literature focuses not only on recovering the direction $\boldsymbol{\beta}^*$ but also aims at estimating f, which is also what the main goal of the present paper is. [33], e.g., proposed a nonparametric least squares with an equality ℓ_1 constraint to handle simultaneous estimation of $\boldsymbol{\beta}^*$ and f. This procedure is computationally challenging and in order to establish estimation guarantees, [33] assumes that f is smooth; the rates obtained by [33] are suboptimal for monotone links f. Furthermore, these rates do not hold in the "ultra-high" dimensional setting $p \gg n$, while our procedure provably works even in such extreme situations. [17] considered a smoothed-out U-process type of loss function with ℓ_1 regularization, and proved their approach works for a model class which encompasses the monotone SIM (1.1). In contrast to [17] however, our nonparametric procedure requires only one tuning parameter, less computation, and has better estimation guarantees (albeit within a smaller model class). The algorithm proposed by [17] was in part inspired by one of the seminal works studying models with monotone link functions in the low dimensional setting – [27]. [27] proposed the maximum score estimator of the multinomial choice model, which is the minimizer to a non-smooth loss function. [27] also established the consistency for the maximum score estimator. Later [18, 19] suggested a smoothed version of [27]'s loss function and showed that the smoothed out version can have computational and estimation benefits [see 37, for an overview].

Regularized procedures have also been proposed for specific choices of f and Y. For example, [44] considered the model $Y = f(\boldsymbol{X}^\top\boldsymbol{\beta}^*) + \varepsilon$ with a known continuously differentiable and monotonic f, and developed estimation and inferential procedures based on the ℓ_1 regularized quadratic loss. Non of the aforementioned works use isotonic regression to directly address prediction and simultaneous estimation of f and $\boldsymbol{\beta}^*$ for monotone SIMs. That being said, this is not the first paper to consider an application of isotonic regression to SIMs. The use of isotonic regression in SIMs dates back to [29], where the authors proposed a heuristic approach of applying isotonic regression to SIM estimation. More recently, [22, 23] formalized the Isotron algorithm, which is a combination of the perceptron and isotonic regression. The authors gave estimation guarantees which are suboptimal compared to the ones we provide. [3, 9] give predictions of monotone SIMs albeit not in the high dimensional sparse setting. A heuristic procedure for variable selection using LASSO, isotonic regression and kernel regression was given by [14].

In conclusion, even though there has been a flurry of works studying SIMs, monotone SIMs, and solving SIMs with isotonic regression, to the best of our knowledge our work is the first to derive estimation and prediction guarantees for monotone SIMs with Gaussian designs in a high dimensional setting, and more specifically to study the behavior of isotonic regression after LASSO has been applied.

1.3. Notation

Throughout the paper we adopt the convenient notation $[n] = \{1, 2, \ldots, n\}$. For a vector $\mathbf{u} = (u_1, u_2, \ldots, u_n)^\top \in \mathbb{R}^n$, let $\mathbf{u}^\uparrow$ denote the vector $(u_{(1)}, u_{(2)}, \ldots, u_{(n)})^\top$ of order statistics: $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$.

Given model (1.1), we define the shorthand notation

$$\begin{aligned}\alpha^2 = E(Y^2), \quad \eta = \mathrm{Var}(Y^2), \quad \Upsilon = \mathbb{E}(Y\boldsymbol{X}^\top\boldsymbol{\beta}^*), \quad \theta^2 = \mathrm{Var}\{(Y - \Upsilon\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2\}, \\ \gamma^2 = \mathrm{Var}(Y\boldsymbol{X}^\top\boldsymbol{\beta}^*), \quad \xi^2 = \mathbb{E}\{(Y - \Upsilon\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2\}, \qquad (1.5)\end{aligned}$$

under the assumptions that $\boldsymbol{X}^\top\boldsymbol{\beta}^* \sim \mathcal{N}(0,1)$ and $\mathbb{E}Y^4 < \infty$, so that all of the above quantities are finite and well defined. In addition we assume $\mathbb{E}Y^4$ does not scale with s, n, p asymptotically, which in turn ensures that all of the above quantities are bounded and do not scale above a constant.

For a sample of n observations $\{\boldsymbol{X}_i\}_{i\in[n]}$, of $\boldsymbol{X}_i \in \mathbb{R}^p$ vectors, we define the $n \times p$ matrix $\mathbb{X}$ which stacks the vectors $\boldsymbol{X}_i^\top$ into its rows. We will often use the sample covariance notation

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top.$$

For a symmetric and positive semi-definite matrix $\boldsymbol{\Sigma}$, we use $\boldsymbol{\Sigma}^{\frac{1}{2}}$ to denote its symmetric square root, i.e., if we have the eigendecomposition $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$

then $\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^\top$. The notations n and p are reserved for sample size and dimensionality of the signal vector $\boldsymbol{\beta}^*$. Throughout the paper we assume that the vector $\boldsymbol{\beta}^*$ is s-sparse, i.e. it has only s non-zero entries. We reserve S to denote the support of $\boldsymbol{\beta}^*$, i.e., the set of all non-zero coefficients of $\boldsymbol{\beta}^*$; hence $|S| = |\text{supp}(\boldsymbol{\beta}^*)| = s$. Similarly S^c will be used to denote $[p] \setminus S$ the set of zero coefficients of $\boldsymbol{\beta}^*$. For a vector $\mathbf{v} \in \mathbb{R}^p$ and a set $C \subset [p]$ the vector $\mathbf{v}_C$ is the C restriction of $\mathbf{v}$, i.e., it contains only the entries of $\mathbf{v}$ which belong to C. Similarly for a matrix $\mathbf{M} \in \mathbb{R}^{p_1 \times p_2}$ double indexing $\mathbf{M}_{C_1 C_2}$ takes the entries of $\mathbf{M}$ belonging to C_1 and C_2. For a vector $\mathbf{v} \in \mathbb{R}^k$ and a $q \geq 1$ we use standard notation for the ℓ_q norm $\|\mathbf{v}\|_q = \big[\sum_{i\in[k]} v_i^q\big]^{1/q}$ with the usual extension $\|\mathbf{v}\|_\infty = \max_{i\in[k]} v_i$ when $q = \infty$. For a matrix $\mathbf{M}$ we define the induced norm

$$\|\mathbf{M}\|_{p\to q} := \sup_{\|\mathbf{v}\|_p = 1} \|\mathbf{M}\mathbf{v}\|_q.$$

For brevity we let $\|\mathbf{M}\|_2 = \|\mathbf{M}\|_{2\to 2}$, and let $\|\mathbf{M}\|_{\max} = \max_{ij} |M_{ij}|$. We use $\mathbf{I}_d$ to denote a d-dimensional identity matrix, where sometimes the index d will be omitted when the dimension is clear.

We use standard asymptotic notation for sequences. Given two sequences $\{a_n\}, \{b_n\}$ we write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$; $a_n = o(b_n)$ if $a_n/b_n \to 0$, and $a_n \asymp b_n$ if there exists positive constants c and C such that $c < a_n/b_n < C$. We also use $a_n \lesssim b_n$ and $a_n \ll b_n$ as a shorthand for $a_n = O(b_n), a_n, b_n > 0$ and $a_n = o(b_n), a_n, b_n > 0$ respectively.

1.4. Organization

The paper is organized as follows. Section 2 formally states our procedure for monotone SIM prediction and gives initial motivation. Section 3 is dedicated to studying the in-sample prediction error of our procedure. Section 4 studies the prediction of our algorithm on a new observation. Finally Section 5 provides thorough numerical studies confirming the predictions of the main results of Sections 3 and 4. The majority of the proofs are relegated to the appendices.

2. Background and methodology

Suppose we observe n samples $\mathcal{D} = \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ from a monotone SIM (1.1):

$$Y = f(\boldsymbol{X}^\top \boldsymbol{\beta}^*) + \varepsilon,$$

where f is a monotone increasing and L-Lipschitz[¶] function, $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{\Sigma})$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is noise independent of $\boldsymbol{X}$. For identifiability we assume $\|\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|_2 = 1$, which implies that $\boldsymbol{X}_i^\top \boldsymbol{\beta}^* \sim \mathcal{N}(0, 1)$. We propose the following two step estimation procedure:

[¶] Recall that a function $f : \mathbb{R} \mapsto \mathbb{R}$ is L-Lipschitz when for any two values $x, y \in \mathbb{R}$ we have $|f(x) - f(y)| \leq L|x - y|$.

- **Step I.** Let $\widehat{\boldsymbol{\beta}}$ be the solution to

$$\widehat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{2.1}$$

  for an appropriately chosen λ. Let

$$\overline{\boldsymbol{\beta}} = \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \widehat{\boldsymbol{\beta}}\|_2}$$

  denote the final estimator of $\boldsymbol{\beta}^*$.
- **Step II.** Construct π, the permutation $\pi : [n] \mapsto [n]$ sorting $\{\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}\}_{i=1}^n$ in an increasing order, where ties are broken arbitrarily. In other words let π be such that $\boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}} \leq \boldsymbol{X}_{\pi_j}^\top \overline{\boldsymbol{\beta}}$ for all $i \leq j$. Fit an isotonic regression

$$\widehat{\mathbf{f}} := \operatorname*{argmin}_{\mathbf{f} \in \mathcal{S}_n^{\uparrow}} \|Y_{\pi_i} - f_i\|_2, \tag{2.2}$$

  where $\mathcal{S}_n^{\uparrow} = \{\mathbf{u}^{\uparrow} \in \mathbb{R}^n | \mathbf{u} \in \mathbb{R}^n\}$. For a given observation $\boldsymbol{X}$ set $x = \boldsymbol{X}^\top \overline{\boldsymbol{\beta}}$. The proposed final estimate of $f(\boldsymbol{X}^\top \boldsymbol{\beta}^*)$ is

$$\widehat{f}(x) = \begin{cases} \widehat{\mathbf{f}}_{\operatorname*{argmin}_{i : x \geq \boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}}} (x - \boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}})}, & x \geq \boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}} \\ \widehat{\mathbf{f}}_1, & x < \boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}} \end{cases}$$

  or in other words let $\widehat{f}(x)$ be the coefficient in the vector $\widehat{\mathbf{f}}$ corresponding to the index $\operatorname*{argmin}_{i : x \geq \boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}}} (x - \boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}})$ when $x \geq \boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}$ and $\widehat{\mathbf{f}}_1$ otherwise.

There are (at least) two ways one can go about implementing the above procedure. The first way (and the one used to state the algorithm) is to run both Step I and Step II on the entire dataset $\mathcal{D}$. The second way is to split the data $\mathcal{D}$ in two halves $\mathcal{D}_1$ and $\mathcal{D}_2$, and run Step I on $\mathcal{D}_1$, and Step II on $\mathcal{D}_2$. The compelling reason for running the full data procedure is clear; by using the full data one takes advantage of the entire dataset when estimating both $\boldsymbol{\beta}^*$ and f. A compelling reason for the second procedure is the fact that the estimate $\overline{\boldsymbol{\beta}}$ would be derived independently from the second half of the data, thus eliminating any potential "cherry-picking" when estimating $\widehat{\mathbf{f}}$.

Multiple technical challenges arise when deriving guarantees for the full data procedure. The main issue lies in the fact that the data is used once for estimating $\overline{\boldsymbol{\beta}}$, implying that the permutation π depends on the data. Consequently the distribution of the error terms $\{\varepsilon_{\pi_i}\}_{i=1}^n$ becomes complicated. The situation does not become easier even when one conditions on the random design $\boldsymbol{X}_i$, since the estimates $\{\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}\}_{i=1}^n$ depend on the error terms ε_i. Data splitting is a simple way to remedy this effect. However, while splitting the data effectively breaks the dependency of the permutation π on the error terms, it may require as much as 1.5 the number of samples to achieve the performance of the full data procedure. We will illustrate this effect on simulated data in Section 5. In theory we will argue that both procedures produce estimates satisfying the same rates, although under slightly different conditions; furthermore the data splitting procedure is significantly easier to justify compared to the full data version.

In practice, Step I may benefit from performing a careful data dependent transformation. Such transformations include for example, scaling and centering the outcome values, i.e., $Y_i \mapsto \frac{Y_i - \overline{Y}}{\operatorname{sd}(Y)}$ (here $\overline{Y} = \frac{1}{n}\sum_{i\in[n]} Y_i$, and $\operatorname{sd}(Y) = \sqrt{\frac{1}{n}\sum_{i\in[n]}(Y_i - \overline{Y})^2}$).

Before we formally justify the procedure and show some guarantees, we first provide general motivation. To see why optimizing (2.1) is useful for estimating $\boldsymbol{\beta}^*$, recall that we assume that ε is independent of $\boldsymbol{X}$ and that $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ (the latter is not crucial, as long as $\boldsymbol{X}$ has an elliptical distribution). Using the closed form solution for the least squares we have

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\beta})^2 &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}Y\\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}\widetilde{\boldsymbol{X}}f(\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*),\end{aligned}$$

where $\widetilde{\boldsymbol{X}}^\top = \boldsymbol{X}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\widetilde{\boldsymbol{\beta}}^* = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*$. In terms of this notation $\widetilde{\boldsymbol{X}} \sim \mathcal{N}(0, \mathbf{I})$. Hence by the properties of the Gaussian distribution multiplying $\mathbb{E}\widetilde{\boldsymbol{X}}f(\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*)$ by any vector perpendicular to $\widetilde{\boldsymbol{\beta}}^*$ equals 0. Therefore it follows that $\mathbb{E}\widetilde{\boldsymbol{X}}f(\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*) \propto \widetilde{\boldsymbol{\beta}}^*$ and hence

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\beta})^2 \propto \boldsymbol{\beta}^*.$$

Now, $\operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\beta})^2 \neq 0$, since supposing the contrary leads to

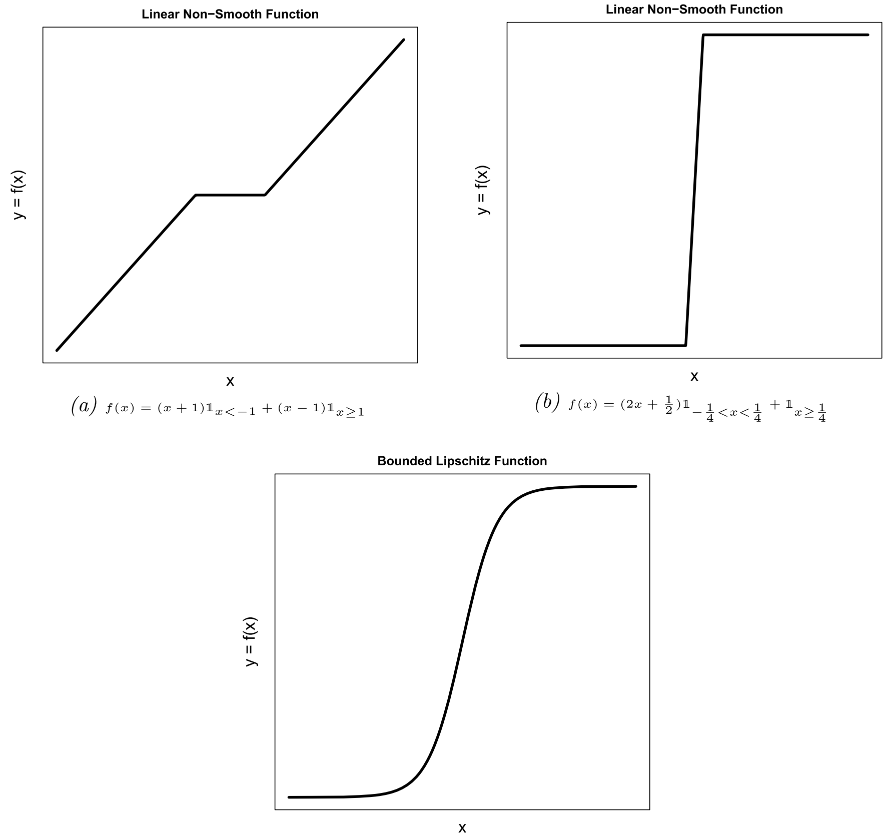
$$\mathbb{E}\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^* f(\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*) = 0,$$

which cannot hold for monotone non-constant f. Hence the population minimizer of the least squares is proportional to the true direction $\boldsymbol{\beta}^*$ which serves as a motivation to (2.1), where the ℓ_1 norm penalty is added to help with the high dimensionality of $\boldsymbol{\beta}^*$. Applying isotonic regression after (2.1) is natural since we are assuming that f is monotone increasing and (2.1) gives us a plugin estimate which we can use to reduce the dimensionality from p to 1 as required in isotonic regression.

In the next section we explore the in-sample prediction error of our procedure.

3. In-sample prediction

The main result outlined in this section shows the usefulness of the procedure motivated in Section 2, for the class of monotone increasing and Lipschitz link functions. All results presented in the main text hold when using the entire dataset for both steps I and II. In the supplement (see Appendix D) we argue



(a) $f(x) = (x+1)\mathbb{1}_{x<-1} + (x-1)\mathbb{1}_{x\geq 1}$

(b) $f(x) = (2x+\frac{1}{2})\mathbb{1}_{-\frac{1}{4}<x<\frac{1}{4}} + \mathbb{1}_{x\geq\frac{1}{4}}$

(c) $f(x) = \tanh(x)$

Fig 1. *Three examples of monotone increasing Lipschitz functions with two linear non-smooth functions.*

that if one is willing to split the data some of the assumptions can be relaxed. We begin by showcasing (see Figure 1) three simple examples of monotone increasing Lipschitz functions, which we will later on use for simulation purposes.

In order for us to guarantee that the full data procedure is successful we first ensure that for an appropriately chosen λ the estimate $\overline{\boldsymbol{\beta}}$ is sufficiently close to $\boldsymbol{\beta}^*$, and that moreover its support is a subset of the support of $\boldsymbol{\beta}^*$. The latter property of $\overline{\boldsymbol{\beta}}$ might appear unnecessary, but it allows us to carefully isolate the dependency of the error terms ε_{π_i} on $\overline{\boldsymbol{\beta}}$. To ensure that the estimate $\overline{\boldsymbol{\beta}}$, obeys these properties we will require that the entries of $\boldsymbol{X}$ are not overly correlated. Below we will use standard assumptions, which are often adopted in the high dimensional statistics literature.

To this end recall that $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$ and let $S := \operatorname{supp}(\boldsymbol{\beta}^*)$. Partition $\boldsymbol{\Sigma}$ (with a slight abuse of notation) as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{SS} & \mathbf{\Sigma}_{SS^c} \\ \mathbf{\Sigma}_{S^cS} & \mathbf{\Sigma}_{S^cS^c} \end{bmatrix},$$

where $\mathbf{\Sigma}_{SS}$ corresponds to the covariance of $\boldsymbol{X}_S$. Define the conditional covariance matrix of $\boldsymbol{X}_{S^c}|\boldsymbol{X}_S$

$$\mathbf{\Sigma}_{S^c|S} := \mathbf{\Sigma}_{S^cS^c} - \mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_{SS}^{-1}\mathbf{\Sigma}_{SS^c}.$$

We assume the following conditions

Assumption 3.1 (Weak Covariance)**.** *Suppose that for some fixed constants $\lambda_{\max}$, $\lambda_{\min}, \kappa, \Omega > 0$ we have*

$$\|\mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_{SS}^{-1}\|_{\infty\to\infty} \le (1-\kappa), \quad \lambda_{\min}\mathbf{I}_s \le \mathbf{\Sigma}_{SS} \le \lambda_{\max}\mathbf{I}_s, \quad \|\mathbf{\Sigma}_{S^c|S}\|_{\max} \le \Omega < \infty.$$

Furthermore, suppose that $\lambda_{\min}, \lambda_{\max}, \kappa, \Omega$ do not scale with (n, p, s).

Assumption 3.2 (Bounded 4th Moment)**.** *Suppose that $\mathbb{E}(Y^4) < \infty$, and $\mathbb{E}(Y^4)$ does not scale with (n, p, s).*

The first condition in Assumption 3.1 is a somewhat stringent requirement on the covariance matrix which controls how much correlation of $\boldsymbol{X}$ is allowed outside of the true support S. It is used to ensure that the support of the first step estimate is embedded within the true support of $\boldsymbol{\beta}^*$ for large enough values of λ. Furthermore, it allows us to control the dependency of the error terms ε_i on $\overline{\boldsymbol{\beta}}$. This is likely not a necessary condition, but it facilitates our analysis greatly. The remaining two conditions in Assumption 3.1 are milder and they require boundedness of the eigenvalues of $\mathbf{\Sigma}_{SS}$ and boundedness of the diagonal elements of $\mathbf{\Sigma}_{S^c|S}$. While Assumption 3.1 depends on the true support of $\boldsymbol{\beta}^*$, there are matrices which satisfy it universally for any support set S. For example the identity matrix clearly satisfies Assumption 3.1 for any S. More generally, so does any Toeplitz matrix whose i, j^{th} element is of the form $\rho^{|i-j|}$ for $0 < |\rho| < 1$. Assumption 3.2 is mild and requires that the outcome has a bounded 4th moment. The main result of this section, outlined below, shows that our estimator controls the in-sample prediction error over the observed sample.

Theorem 3.3. *Suppose Assumptions 3.1 and 3.2 hold and that the monotone increasing function f is L-Lipschitz. In addition let $\frac{s}{n} \le \frac{1}{64}$, and assume that the effective sample size $\frac{n}{s\log p}$ is large enough and the tuning parameter $\lambda = C\sqrt{\frac{\log p}{n}}$ for a sufficiently large constant $C > 0$. Then for some constants $\Omega_1, \Omega_2, \Omega_3 > 0$ the following event happens with probability at least $1 - G(n, s, p, f)$:*

$$\frac{\sum_{i=1}^n [f(\boldsymbol{X}_i^\top\boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top\overline{\boldsymbol{\beta}})]^2}{n} \le \Omega_1\sigma^2\left(\frac{\sigma + |f(\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}})|}{n\sigma}\right)^{2/3} + \Omega_2 L^2\frac{s\log p}{n} + \Omega_3\frac{s + \sigma^2\log p}{n}, \tag{3.1}$$

where $G(n, s, p, f) \to 0$ as $n \to \infty$.

Denote the rate on the right hand side of (3.1) with $\mathcal{R}_F$, where F stands for "full data" procedure.

Remark 3.4. *Condition* (3.1) *can be interpreted as a tradeoff of two distinct error rates. The first rate* $\sigma^2(\frac{\sigma+|f(\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}})-f(\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}})|}{n\sigma})^{2/3}$ *is the minimax rate of approximating a monotone function [8, 4]. The rate* $\Omega_2 L^2\frac{s\log p}{n}+\Omega_3\frac{s+\sigma^2\log p}{2n}\asymp\frac{s\log p}{n}$ *is the minimax rate (up to log factors)*[‖] *of estimating the parameter of a high dimensional linear model in* ℓ_2^2 *norm [5].*

Notably the rate in (3.1) depends on the expression $f(\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}})-f(\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}})$. In a case when f is bounded one can upper bound this difference independently of the values $\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}}$ and $\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}}$. However, in many cases such as in the linear regression model, e.g., the boundedness assumption is violated. Under the assumption that f is L-Lipschitz, we know that

$$|f(\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}})-f(\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}})|\leq L|\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}}-\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}}|. \tag{3.2}$$

Since $\overline{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}^*$, one would anticipate that the two quantities $\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}}$ behave like $\min \boldsymbol{X}_i^\top\boldsymbol{\beta}^*$ and $\max \boldsymbol{X}_i^\top\boldsymbol{\beta}^*$. However this is not immediately obvious, due to the complicated dependency of $\overline{\boldsymbol{\beta}}$ on the data. Our next result argues that under Assumptions 3.1 and 3.2 this is indeed the case assuming additionally that $\frac{s\log p}{n}=O(1)$. Before we state the result recall that the $n\times p$ matrix $\mathbb{X}$ stacks the n vectors $\boldsymbol{X}_i^\top$ into its rows.

Lemma 3.5. *Suppose the assumptions from Theorem 3.3 hold. Then for some constant* C_1 *we have*

$$\|\mathbb{X}\overline{\boldsymbol{\beta}}-\mathbb{X}\boldsymbol{\beta}^*\|_\infty\leq C_1\sqrt{\frac{s\log p}{n}} \tag{3.3}$$

with probability at least $1-H(s,n,p,f)$*, where* $H(s,n,p,f)\to 0$ *as* $n\to\infty$*.*

We now discuss (3.2) in view of Lemma 3.5. Let $Z_1,\ldots,Z_n$ be i.i.d. draws from a standard normal distribution. Since $\boldsymbol{\beta}^*$ is a fixed vector such that $\boldsymbol{X}_i^\top\boldsymbol{\beta}^*\sim\mathcal{N}(0,1)$ for all i, we have that $[\mathbb{X}\boldsymbol{\beta}^*]_{(n)}$ lies in the interval $[\mathbb{E}Z_{(n)}-t,\mathbb{E}Z_{(n)}+t]$, while $[\mathbb{X}\boldsymbol{\beta}^*]_{(1)}$ lies in the interval $[\mathbb{E}Z_{(1)}-t,\mathbb{E}Z_{(1)}+t]$ (with probability at least $1-4e^{-t^2/2}$). For a proof of this fact we refer to [Theorem 5.8 of 6]. In addition, using extreme value theory it is known that asymptotically $\mathbb{E}Z_{(n)}\approx\sqrt{2\log n}$ and $\mathbb{E}Z_{(1)}\approx-\sqrt{2\log n}$ [15]. By (3.3), it follows that with high probability (at least $1-4e^{-t^2/2}-H(s,n,p,f)$) we have

$$|[\mathbb{X}\overline{\boldsymbol{\beta}}]_{(n)}-\mathbb{E}Z_{(n)}|\leq t+C_1\sqrt{\frac{s\log p}{n}},$$

for some constant C_1. Naturally, on the same event a similar high confidence interval holds for $[\mathbb{X}\overline{\boldsymbol{\beta}}]_{(1)}$. Therefore, continuing (3.2) with probability at least

[‖] The minimax rate is $\frac{s\log p/s}{n}$ [34], which for most values of s is of the order of $\frac{s\log p}{n}$. Here to be precise we say that $\frac{s\log p}{n}$ is the minimax rate up to logarithmic factors.

$1 - 4e^{-t^2/2} - H(s,n,p,f)$ we have

$$\begin{aligned}|f(\boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}})| \leq L|\boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}} - \boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}| &= L|[\mathbb{X}\overline{\boldsymbol{\beta}}]_{(n)} - [\mathbb{X}\overline{\boldsymbol{\beta}}]_{(1)}| \\ &\leq L\left[\mathbb{E}Z_{(n)} - \mathbb{E}Z_{(1)} + 2t + 2C_1\sqrt{\frac{s\log p}{n}}\right] \\ &\approx 2L\left[t + \sqrt{2\log n} + C_1\sqrt{\frac{s\log p}{n}}\right].\end{aligned}$$

The above in conjunction with (3.2) implies that even in the unbounded f case $|f(\boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}})| \lesssim \sqrt{\frac{s\log p}{n}} + \sqrt{2\log n} \lesssim \sqrt{2\log n}$, where the last holds since we assume $\frac{s\log p}{n} = O(1)$. This discussion leads us to the following corollary of Theorem 3.3

Corollary 3.6. *Suppose the conditions of Theorem 3.3 hold. Then with high probability for some constants $\Omega_i > 0$ for $i \in [4]$ we have*

$$\begin{aligned}\frac{\sum_{i=1}^n [f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}})]^2}{n} &\leq \Omega_1\sigma^2\left(\frac{\sigma + \Omega_4 L\sqrt{\log n}}{n\sigma}\right)^{2/3} \\ &\quad + \Omega_2 L^2\frac{s\log p}{n} + \Omega_3\frac{s + \sigma^2\log p}{n}.\end{aligned}$$

4. Prediction

Inequality (3.1) ensures that the Isotonic LASSO estimate does not behave erratically on the observed data points, by guaranteeing that the average in-sample squared error is small. On the other hand, (3.1) says nothing about predicting future observations and in that sense is not completely satisfactory. What can one say about average case error when predicting a fresh new sample? In classical nonparametric theory, one would typically analyze the Mean Integrated Squared Error (MISE) [40]. In this section, we will confine to a slightly weaker, data dependent criteria which we refer to as the Conditional Mean Integrated Absolute Error (CMIAE). For a new observation $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, the CMIAE is defined conditionally on the observed data $\mathcal{D} = \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ as

$$\text{CMIAE} := \mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|\widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}^\top \boldsymbol{\beta}^*)|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]\big].$$

In the above, the conditional expectation $\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}$ is taken with respect to the new observation $\boldsymbol{X}$ conditionally on the data $\mathcal{D}$, which fixes the quantities $\widehat{f}, \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}$ and $\boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}$. Since $\boldsymbol{X}$ is independent of the dataset this notation is equivalent to simply taking the expectation with respect to $\boldsymbol{X}$: $\mathbb{E}_{\boldsymbol{X}}$. However, we will use the notation $\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}$ to underscore that this expectation is a function of the dataset $\mathcal{D}$ (hence a random variable with respect to the randomness of the data). Several remarks regarding the definition of CMIAE are in order.

First, CMIAE measures the integrated absolute error in prediction, conditionally on the new observation's projection along the estimate $\overline{\boldsymbol{\beta}}$ falling within the "data range" — $[\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]$. Importantly, the data range we use in the CMIAE is procedure dependent, as it is conditional on the first step estimate $\overline{\boldsymbol{\beta}}$. Furthermore, the condition $\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]$ excludes "extreme" new observations $\boldsymbol{X}$ from consideration in order to avoid extrapolation. Given that $\overline{\boldsymbol{\beta}}$ is a good proxy of $\boldsymbol{\beta}^*$, intuitively one should be skeptical of the estimate $\widehat{f}$ beyond the observed range $[\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]$. If one assumes that f is bounded, this condition can be traded off for an additional error rate which depends on $\|f\|_\infty$ and the probability of observing an "extreme" point $\boldsymbol{X}$ such that $\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \not\in [\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]$. We do not present this result here.

Second, if in analogy to CMIAE, we define the Conditional MISE (CMISE) as the squared error, by Jensen's inequality we obtain

$$\text{CMIAE} \le \underbrace{\big[\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[(\widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}^\top \boldsymbol{\beta}^*))^2 \big| \boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]\big]\big]^{\frac{1}{2}}}_{\text{CMISE}^{\frac{1}{2}}}. \tag{4.1}$$

The inequality above shows that any bound on CMISE implies a bound on CMIAE, and therefore CMIAE is a weaker measure of prediction. It turns out (unsurprisingly) that CMIAE is more amenable to analysis for the isotonic LASSO algorithm, and we defer to future work the analysis of CMISE.

Below we state and prove the main result on prediction. To this end recall that $\mathcal{R}_F$ denotes the right hand side of (3.1).

Theorem 4.1. *Let $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ be a new observation generated independently from the data. If the assumptions of Theorem 3.3 hold and $s \log n \ll \sqrt{n}$, then with high probability*[**] *we have:*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|\widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}^\top \boldsymbol{\beta}^*)| \big| \boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}]\big] \\ &\quad \le C\bigg(\frac{f(\boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}})}{\sqrt{n}} + \mathcal{R}_F^{\frac{1}{2}}\bigg),\end{aligned} \tag{4.2}$$

for some absolute constant C.

Before we turn to the proof of this theorem we comment on a related result, Theorem 6.1 of [3]. Cast into our framework, informally speaking, Theorem 6.1 of [3] implies that the $\text{CMISE}^{\frac{1}{2}}$ is of order $n^{-\frac{1}{3}} \log(n)^{5/3}$ (see also the remark after the Theorem) given that p is fixed, the function f is bounded, and an estimate of $\boldsymbol{\beta}^*$, $\overline{\boldsymbol{\beta}}$ is available such that $\|\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(n^{-\frac{1}{3}})$. Even if one assumes that s and p are both fixed, one cannot derive Theorem 4.1 as a Corollary of Theorem 6.1 of [3] using (4.1). The reason being that the rate [3] provide has an extraneous logarithmic factor (Theorem 4.1 gives the sharp rate of order $n^{-\frac{1}{3}}$), and the estimate $\overline{\boldsymbol{\beta}}$ in Theorem 6.1 of [3] is provided from a sample splitting procedure. Our proof strategy is fundamentally different to that of Theorem 6.1 of [3], and utilizes the properties of the Gaussian distribution.

Proof of Theorem 4.1. Without loss of generality we will assume that the vector $\overline{\boldsymbol{\beta}}$ is the same as the one generated in Step II' of the proof of Theorem 3.3. By

[**] The "high probability" here is measured in terms of the randomness of the dataset $\mathcal{D}$.

Proposition B.2 this is a high probability event, and hence we can assume it holds without loss of generality. For brevity let $\overline{x}_i := \boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}}$. Using our shorthand notation the triangle inequality yields

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|\widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}^\top \boldsymbol{\beta}^*)|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] \\
&\quad \le \mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - f(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] \qquad (4.3)\\
&\quad + \mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - \widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big]. \qquad (4.4)
\end{aligned}$$

To control the first term (4.3) by the law of total expectation we have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - f(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] &\le \frac{\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - f(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big]}{\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n])} \\
&=: I_1,
\end{aligned}$$

where $\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}$ denotes the probability of the new observation $\boldsymbol{X}$ given the dataset $\mathcal{D}$. For the second term (4.4) we first note that by the law of total expectation

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - \widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] \\
&\quad = \sum_{i=1}^{n} \mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_i, \overline{x}_{i+1})|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] \times \\
&\quad \mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}) - \widehat{f}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_i, \overline{x}_{i+1})\big] \\
&\quad =: I_2. \qquad (4.5)
\end{aligned}$$

We will now handle the terms I_1 and I_2 individually. We start with I_1 which is simpler. Using the Lipschitz property of f and (B.3) we have with high probability (the probability is measured with respect to the randomness in $\mathcal{D}$) that

$$\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]) I_1 \le L\big[\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[(\boldsymbol{X}^\top \boldsymbol{\beta}^* - \boldsymbol{X}^\top \overline{\boldsymbol{\beta}})^2\big]\big]^{\frac{1}{2}} \le C_2 L \sqrt{\frac{s \log p}{n}}.$$

Put $c = [\overline{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma} \overline{\boldsymbol{\beta}}]^{\frac{1}{2}} = \|\boldsymbol{\Sigma}^{\frac{1}{2}} \overline{\boldsymbol{\beta}}\|_2$ for brevity. Note that conditionally on the data $\mathcal{D}$, $\boldsymbol{X}^\top \overline{\boldsymbol{\beta}}|\mathcal{D} \sim \mathcal{N}(0, c^2)$, since $\boldsymbol{X}$ is independent of the data and $\overline{\boldsymbol{\beta}}$ is fixed given $\mathcal{D}$. Therefore $\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n)) = \Phi(\frac{\overline{x}_n}{c}) - \Phi(\frac{\overline{x}_1}{c})$. By (B.3) and the triangle inequality we know that $|c - 1| \le C_2 \sqrt{\frac{s \log p}{n}}$. Thus using Lemma 3.5 and the arguments after its statement, we have the following lower bound

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}(\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]) &= \Phi\big(\frac{\overline{x}_n}{c}\big) - \Phi\big(\frac{\overline{x}_1}{c}\big) \\
&\ge \Phi\Bigg(\frac{Z_{(n)} - C_1\sqrt{\frac{s \log p}{n}}}{c}\Bigg) - \Phi\Bigg(\frac{Z_{(1)} + C_1\sqrt{\frac{s \log p}{n}}}{c}\Bigg) \ge \frac{1}{2},
\end{aligned}$$

for sufficiently large n and $\frac{n}{s \log p}$, where $Z_{(1)}$ and $Z_{(n)}$ are order statistics of n standard normal samples, Φ is the cdf of a standard normal distribution. We conclude that with high probability (measured in terms of the randomness of the dataset $\mathcal{D}$)

$$I_1 \leq 2C_2 L\sqrt{\frac{s\log p}{n}}.$$

Next we tackle the term I_2. By the definition of $\widehat{f}$, when $x \in [\overline{x}_i, \overline{x}_{i+1})$ we have $\widehat{f}(x) = \widehat{f}(\overline{x}_i)$. Thus

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|f(\boldsymbol{X}^\top\overline{\boldsymbol{\beta}}) - \widehat{f}(\boldsymbol{X}^\top\overline{\boldsymbol{\beta}})|\big|\boldsymbol{X}^\top\overline{\boldsymbol{\beta}} \in [\overline{x}_i, \overline{x}_{i+1})\big] \\
&\quad\leq |f(\overline{x}_i) - \widehat{f}(\overline{x}_i)| \vee |f(\overline{x}_{i+1}) - \widehat{f}(\overline{x}_i)| \\
&\quad\leq |f(\overline{x}_i) - \widehat{f}(\overline{x}_i)| + |f(\overline{x}_{i+1}) - f(\overline{x}_i)| \\
&\quad= |f(\overline{x}_i) - \widehat{f}(\overline{x}_i)| + f(\overline{x}_{i+1}) - f(\overline{x}_i),
\end{aligned}$$

where the first and last identities hold since f is monotone increasing. Therefore by (4.5)

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top\overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] I_2 \\
&\quad\leq \sum_{i=1}^{n-1} |f(\overline{x}_i) - \widehat{f}(\overline{x}_i)| \mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top\overline{\boldsymbol{\beta}} \in [\overline{x}_i, \overline{x}_{i+1})\big] \\
&\quad+ \sum_{i=1}^{n-1} (f(\overline{x}_{i+1}) - f(\overline{x}_i)) \mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top\overline{\boldsymbol{\beta}} \in [\overline{x}_i, \overline{x}_{i+1})\big] \\
&\quad=: I_{21} + I_{22}.
\end{aligned}$$

We handle those two terms in part below, starting with I_{21}. Recall that $c = [\overline{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma}\overline{\boldsymbol{\beta}}]^{\frac{1}{2}} = \|\boldsymbol{\Sigma}^{\frac{1}{2}}\overline{\boldsymbol{\beta}}\|_2$. We have conditionally on $\mathcal{D}$ that $\boldsymbol{X}^\top\overline{\boldsymbol{\beta}}|\mathcal{D} \sim \mathcal{N}(0, c^2)$, since $\boldsymbol{X}$ is independent of the data and $\overline{\boldsymbol{\beta}}$ is fixed given $\mathcal{D}$. Define $\overline{\overline{x}}_i = \frac{\overline{x}_i}{c}$, so that $\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top\overline{\boldsymbol{\beta}} \in [\overline{x}_i, \overline{x}_{i+1})\big] = \Phi(\overline{\overline{x}}_{i+1}) - \Phi(\overline{\overline{x}}_i)$. By Cauchy-Schwartz and the triangle inequality we have:

$$\begin{aligned}
&\frac{I_{21}}{\big[n\sum_{i=1}^n(\Phi(\overline{\overline{x}}_{i+1}) - \Phi(\overline{\overline{x}}_i))^2\big]^{\frac{1}{2}}} \leq \Big[n^{-1}\sum_{i=1}^n (f(\overline{x}_i) - \widehat{f}(\overline{x}_i))^2\Big]^{\frac{1}{2}} \\
&\quad\leq \Big[n^{-1}\sum_{i=1}^n (f(\boldsymbol{X}_{\pi_i}^\top\boldsymbol{\beta}^*) - \widehat{f}(\overline{x}_i))^2\Big]^{\frac{1}{2}} + \Big[n^{-1}\sum_{i=1}^n (f(\overline{x}_i) - f(\boldsymbol{X}_{\pi_i}^\top\boldsymbol{\beta}^*))^2\Big]^{\frac{1}{2}} \\
&\quad\leq \mathcal{R}_F^{\frac{1}{2}} + L[n^{-1}\sum_{i=1}^n (\boldsymbol{X}_i^\top(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*))^2]^{\frac{1}{2}}.
\end{aligned}$$

For I_{22} we have

$$\begin{aligned}
I_{22} &\leq (f(\overline{x}_n) - f(\overline{x}_1)) \max_i [\Phi(\overline{\overline{x}}_{i+1}) - \Phi(\overline{\overline{x}}_i)] \\
&\leq \frac{f(\overline{x}_n) - f(\overline{x}_1)}{\sqrt{n}} \Big[n\sum_{i=1}^n (\Phi(\overline{\overline{x}}_{i+1}) - \Phi(\overline{\overline{x}}_i))^2\Big]^{\frac{1}{2}}.
\end{aligned}$$

Next we need the following result

Proposition 4.2. *With probability at least* $1 - 2e^{-n/2} - (1 + cn)^s e^{-\sqrt{n}}$*:*

$$\big[n \sum_{i=1}^{n} (\Phi(\overline{\overline{x}}_{i+1}) - \Phi(\overline{\overline{x}}_i))^2\big]^{\frac{1}{2}} \leq \sqrt{12} + \sqrt{1/\pi}, \tag{4.6}$$

where c *is some absolute constant.*

Since we assume $s \log(n) \ll \sqrt{n}$, the above probability converges to 1 asymptotically. Using (B.2) of Proposition B.1 on an event of high probability we have

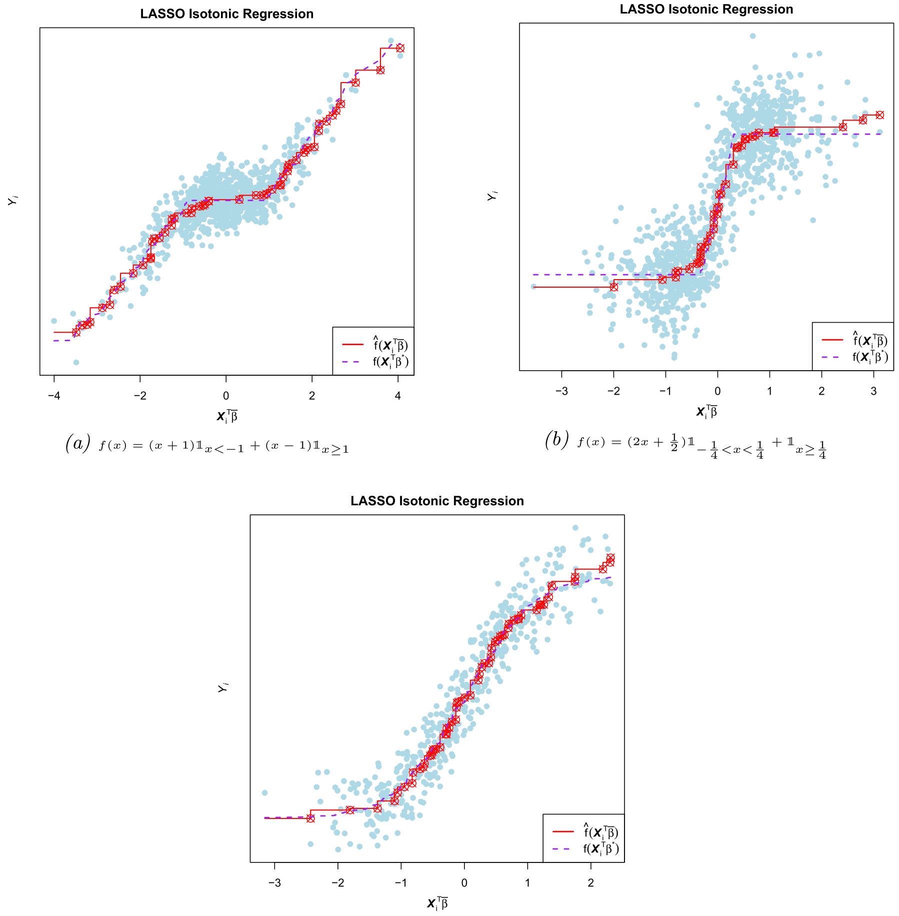
$$\begin{aligned} \mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] I_2 &\leq C\bigg(\frac{f(\overline{x}_n) - f(\overline{x}_1)}{\sqrt{n}} + \mathcal{R}_F^{\frac{1}{2}} + LC_1\sqrt{\frac{s \log p}{n}}\bigg) \\ &\leq \widetilde{C}\bigg(\frac{f(\overline{x}_n) - f(\overline{x}_1)}{\sqrt{n}} + \mathcal{R}_F^{\frac{1}{2}}\bigg). \end{aligned}$$

Combining this with our result on I_1, the fact that $\mathbb{P}_{\boldsymbol{X}|\mathcal{D}}\big[\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\overline{x}_1, \overline{x}_n]\big] \geq \frac{1}{2}$ with high probability, and adjusting the constant $\widetilde{C}$ completes the proof. □

Remark 4.3. *Aside from the conditions required by Theorem 3.3, Theorem 4.1 requires additionally that* $s \log n \ll \sqrt{n}$*. This means that the sparsity of* $\boldsymbol{\beta}^*$ *has to be significantly smaller than the square root of the sample size. This condition is likely an artifact of our proof technique. The bottleneck is Proposition 4.2 which requires* $s \log n \ll \sqrt{n}$ *in order for the term on the left hand side of* (4.6) *to be bounded. Similar result holds for the data splitting procedure, without the need to require* $s \log n \ll \sqrt{n}$*. For details we refer the reader to Appendix D.*

5. Numerical experiments

In this section we show numerical studies with the three increasing Lipschitz link functions which we showcased in Figure 1. First, in order for the reader to appreciate the results visually we attach examples using $n = 1000$ samples, $p = 2000$ dimensions, $s = 20$ non-zero equal coefficients in the signal $\boldsymbol{\beta}^*$, noise $\varepsilon \sim \frac{\mathcal{N}(0,1)}{5}$ and $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{I})$. To enforce stability and robustness in the estimation Step I, i.e., the LASSO step, we scale and center the outcome values before running LASSO. The LASSO is ran via the `glmnet` package, and we use 10-fold cross validation to select the tuning parameter by minimizing the Mean Squared Error (MSE). Isotonic regression is ran using the `isoreg` function of the `stats` package. In Figure 2 we show the estimator $\widehat{f}$ in red. The x-axis consists of estimated values $\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}$, while the y-axis are the true values Y_i. The dashed purple line connects the points $\{(\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}, f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*))\}_{i \in [n]}$. The collection of blue points is the set $\{(\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}, Y_i)\}_{i \in [n]}$. We see that in all three instances the red curve follows closely the purple dashed curve, giving visual evidence that the isotonic regression after LASSO works well to estimate the link function f on the dataset.



(a) $f(x) = (x+1)\mathbb{1}_{x<-1} + (x-1)\mathbb{1}_{x\geq 1}$

(b) $f(x) = (2x + \frac{1}{2})\mathbb{1}_{-\frac{1}{4}<x<\frac{1}{4}} + \mathbb{1}_{x\geq\frac{1}{4}}$

(c) $f(x) = \tanh(x)$

Fig 2. *Three typical examples of the full data LASSO isotonic procedure on the monotone increasing Lipschitz functions from Figure 1. The points shown on this figure are* $\{(\boldsymbol{X}_i^\top\overline{\boldsymbol{\beta}}, Y_i)\}_{i\in[n]}$*, i.e., their y-axis is the true Y_i value and their x-axis is the estimated value* $\boldsymbol{X}_i^\top\overline{\boldsymbol{\beta}}$ *(not the actual value* $\boldsymbol{X}_i^\top\boldsymbol{\beta}^*$*). The red curve is the isotonic regression fit; it contains the points* $\{(\boldsymbol{X}_i^\top\overline{\boldsymbol{\beta}}, \widehat{f}(\boldsymbol{X}_i^\top\overline{\boldsymbol{\beta}}))\}_{i\in[n]}$*; the purple dashed curve connects the points* $\{(\boldsymbol{X}_i^\top\overline{\boldsymbol{\beta}}, f(\boldsymbol{X}_i^\top\boldsymbol{\beta}^*))\}_{i\in[n]}$.

Next we provide numerical evidence which directly corroborates with the theoretical findings of Sections 3 and 4. Our setup is the following. For each of the three link functions of Figure 1, we set the sample size $n = 500$, and vary freely the dimension p and the sparsity s in the sets $\{100, 200, 500, 1000\}$ and $\{5, 10, 15, 20, 25, 30, 40\}$ respectively. This results in a range of possible values for the adjusted sample size $\frac{n}{s\log p}$. The error term, as in the previous example, is set to $\varepsilon \sim \frac{\mathcal{N}(0,1)}{5}$. We consider three possibilities for the design $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\rho)$

where $\rho = 0, 0.2$ and 0.5. The covariance matrix is given by $\mathbf{\Sigma}_{\rho,ij} = \rho^{|i-j|}$ (where we understand $0^0 = 1$). In all occasions the signal vector $\boldsymbol{\beta}^*$ is initially selected with equal positive entries, and is properly normalized so that $\|\mathbf{\Sigma}_\rho^{\frac{1}{2}}\boldsymbol{\beta}^*\|_2 = 1$. As before the LASSO step is ran using scaled and centered outcome values; the tuning parameter λ is selected via 10-fold cross validation of the MSE.

To evaluate the accuracy of the algorithm, we calculated both the square root in-sample prediction error (the square root of the LHS of (3.1)) over 500 repetitions and the CMIAE over 500 repetitions and 1000 fresh samples in each repetition. Since Theorems 3.3 and 4.1 both bound the square root in-sample prediction and the CMIAE with $\sqrt{\frac{s \log p}{n}}$, we compared the observed values to the value of $\sqrt{\frac{s \log p}{n}}$, see Figure 3. All trends appear relatively linear hence confirming the predictions of Theorems 3.3 and 4.1. We also notice that there is a significant difference between the in-sample errors and CMIAE. For example, it turns out that the smooth function (tanh) has smaller in-sample error compared to the linear non-smooth function 1, but has consistently larger CMIAE compared to this function. Furthermore, the slopes of all lines are steeper in the CMIAE simulation compared to the in-sample errors simulation; in addition, the slope of the linear non-smooth function 2 is much steeper for the CMIAE, implying that this function is easier to predict on the dataset at hand, and one can predict well even with relatively small adjusted sample size, but this is not the case for CMIAE.

In addition to presenting simulations for the full data procedure, we also present results on the sample splitting version. Figure 4 demonstrates how sample splitting performs with $n = 500$ observations in comparison to Figure 3. We can see a difference in the "in-sample" errors in comparison to the results in Figure 3. The in-sample error error is larger in comparison to the full procedure by about 1.2 to 1.3 times. The CMIAE is higher compared to the CMIAE using the full data procedure by roughly 1.3 to 1.5 times. This is to be expected as the full data uses double the observations for the two steps. However, in defense of sample splitting, as we argued in Appendix D some assumptions can be lifted when running the two steps independently.

Additional simulation results can be found in Section E.

6. Discussion

This paper focused on a two step procedure applying LASSO and isotonic regression for monotone increasing SIMs. We showed guarantees for both the in-sample and out-of-sample prediction errors.

One interesting technical question is whether we can relax the condition $s \log n \ll \sqrt{n}$ required in Theorem 4.1. The fact that sample splitting does not require this condition suggests that perhaps it is not necessary for the full data procedure either.

Another interesting direction is to explore whether the condition of Gaussian covariates can be relaxed. It is known, see e.g., [26, 25] that if the predictors have

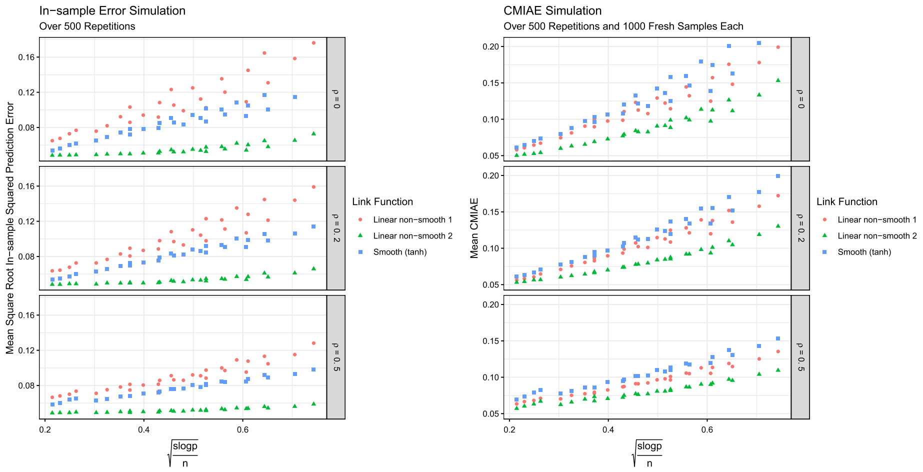


FIG 3. *The left panel of this figure shows* $\sqrt{\frac{s\log p}{n}}$ *versus the square root of in-sample prediction error (the square root of the LHS of* (3.1)*), over 500 repetitions under different link functions and covariance settings. We see that in the three cases there appears to be a relatively linear trend relating the two quantities which corroborates the RHS of* (3.1)*. Similarly, in the right panel we plot the* $\sqrt{\frac{s\log p}{n}}$ *versus CMIAE. We observe relatively linear trends relatating CMIAE to* $\sqrt{\frac{s\log p}{n}}$ *which confirms the prediction of* (4.2).

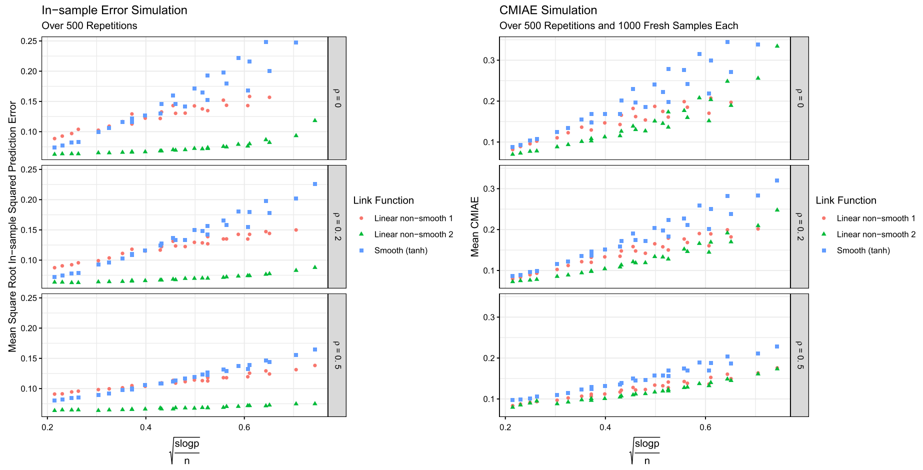


FIG 4. *This figure shows the same simulation results as in Figure* 3*, with the main difference being that we use sample splitting. This means that for* $n = 500$ *observations we use* 250 *observations for the LASSO and the remaining* 250 *for the isotonic regression. We can see that the errors here are about* 1.2 *to* 1.5 *times larger in comparison to the results in Figure* 3.

elliptically symmetric distributions, linear regression can recover the predictor up to a proportionality constant. Such an extension lies beyond the scope of the present manuscript. In our current proof of the full data version of the procedure we are relying on some results which require the Gaussianity of the covariates. We do believe however, that the procedure should work for elliptical distributed covariates, and defer this important question for future investigations.

A technical question regarding CMIAE was raised by one of the referees. It was asked whether one can remove the condition $\boldsymbol{X}^\top \overline{\boldsymbol{\beta}} \in [\min \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}},\ \max \boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}}]$ from the expectation. Currently we do not know how to remove this condition, unless we impose boundedness on the function f. This represents a challenge that could be addressed in our future work.

It is also of interest to see whether one can extend our results on CMIAE to the stronger criterion – CMISE – defined in Section 4. Our preliminary simulations (not reported) show that CMISE behaves comparably to CMIAE suggesting that similar result to Theorem 4.1 might hold for CMISE. Another interesting question is whether consistency can be established without imposing Assumption 3.1 on the covariance.

We also conjecture that the rates we derived are minimax optimal. This conjecture is based off on the fact that when f is the linear, the rate $\frac{s \log p}{n}$ is known to be minimax for the ℓ_2^2 error, while even if we know $\boldsymbol{\beta}^*$ the other rate is minimax when estimating an increasing f. We defer investigating the validity of this conjecture to future work.

Appendix A: Auxiliary results

First we state a very useful inequality from random matrix theory regarding the singular values of a Gaussian matrix.

Lemma A.1 (Corollary 5.35 [41])**.** *Let* $\mathbf{A}_{n\times s}$ *matrix whose entries are i.i.d. standard normal random variables. Then for every* $t \geq 0$*, with probability at least* $1 - 2\exp(-t^2/2)$ *one has:*

$$\sqrt{n} - \sqrt{s} - t \leq s_{\min}(\mathbf{A}) \leq s_{\max}(\mathbf{A}) \leq \sqrt{n} + \sqrt{s} + t,$$

where $s_{\min}(\mathbf{A})$ *and* $s_{\max}(\mathbf{A})$ *are the smallest and largest singular values of* $\mathbf{A}$ *correspondingly.*

Next, we remind the reader of a powerful result of sub-Gaussian concentration of non-Lipschitz functions proved by [1]. Before that we give definitions of $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ norms which will be later used in our analysis. For a real random variable X, define

$$\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}, \quad \|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}.$$

Recall that a random variable is called *sub-Gaussian* if $\|X\|_{\psi_2} < \infty$ and *sub-exponential* if $\|X\|_{\psi_1} < \infty$.

First we introduce the notation of [1]. For an integer ℓ, let P_ℓ denote the set of partitions of $[\ell]$ into non-empty and non-intersecting disjoint sets. For a partition $\mathcal{J} = \{J_1, \dots, J_k\}$, and an ℓ-indexed matrix $\mathbf{A} = (a_{\mathbf{i}})_{\mathbf{i}\in[n]^\ell}$, define the norm:

$$\|\mathbf{A}\|_{\mathcal{J}} = \sup\Big\{\sum_{\mathbf{i}\in[n]^\ell} a_{\mathbf{i}} \prod_{l=1}^{k} x^{(l)}_{\mathbf{i}_{J_l}} : \|x^{(l)}_{\mathbf{i}_{J_l}}\|_2 \le 1, 1 \le l \le k\Big\},$$

where the indexing should be understood as $\mathbf{i}_I := (i_k)_{k\in I}$. Given the convention that $|\mathcal{J}|$ is the cardinality of the set $\mathcal{J}$ we restate a version of Theorem 1.4 of [1].

Theorem A.2 (Theorem 1.4 [1])**.** *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random vector with independent components, such that for all $i \le n$, $\|X_i\|_{\psi_2} \le \Gamma$. Then for every polynomial $f : \mathbb{R}^n \mapsto \mathbb{R}$ of degree L we have:*

$$\mathbb{P}(|f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})| \ge t) \le 2\exp\big(-\frac{1}{C_L}\eta_f(t)\big),$$

where

$$\eta_f(t) = \min_{1\le\ell\le L}\min_{\mathcal{J}\in P_\ell}\big(\frac{t}{\Gamma^\ell\|\mathbb{E}\mathbf{D}^\ell f(\boldsymbol{X})\|_{\mathcal{J}}}\big)^{2/|\mathcal{J}|}.$$

In the above, $\mathbf{D}^\ell$ is the ℓ^{th} derivative of f.

Appendix B: In-sample prediction proofs (full data)

Proof of Theorem 3.3. Our first step is to show that $\widehat{\boldsymbol{\beta}}$ can be identified with the solution to the following program with high probability:

$$\widetilde{\boldsymbol{\beta}}_S = \operatorname*{argmin}_{\boldsymbol{\beta}_S\in\mathbb{R}^s} \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_{iS}^\top\boldsymbol{\beta}_S)^2 + \lambda\|\boldsymbol{\beta}_S\|_1, \tag{B.1}$$

where S is the support of $\boldsymbol{\beta}^*$ and $\boldsymbol{X}_{iS}$ is $\boldsymbol{X}_i$ restricted to the set S. In addition will we argue that the vector $\breve{\boldsymbol{\beta}}_S = \frac{\widetilde{\boldsymbol{\beta}}_S}{\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2}$ is consistent for $\boldsymbol{\beta}^*_S$. We have the following result, which describes the relationship between $\overline{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}_S$.

Proposition B.1. *Suppose all conditions of Theorem 3.3 hold. Then with probability at least $1 - F(n, s, p, f)$ for some constants $C_1, C_2 > 0$ we have*

$$\operatorname{supp}(\overline{\boldsymbol{\beta}}) \subseteq S, \quad \overline{\boldsymbol{\beta}}_S = \breve{\boldsymbol{\beta}}_S, \quad \big\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big\|_2 \le C_1\sqrt{\frac{s\log p}{n}} \tag{B.2}$$

$$\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big\|_2 \le C_2\sqrt{\frac{s\log p}{n}} \tag{B.3}$$

where $F(n, s, p, f) \to 0$ as $n \to \infty$.

The first two claims involving the support of $\overline{\boldsymbol{\beta}}$ have already been established in [30], while similar claims to the third and fourth one (although for the square-root LASSO) were given by [32, 38]. Since we were not able to find a proof of (B.2) and (B.3) in the LASSO literature (in particular under Assumptions 3.2 and 3.1), we provide a standalone proof of Proposition B.1 in Appendix B. This proof will also be useful later when we discuss the predictive properties of our procedure. Let $\mathcal{E}$ denote the event on which (B.2) and (B.3) hold.

To make our analysis simpler we will now isolate the dependency of $\overline{\boldsymbol{\beta}}$ on ε and will argue (in a roundabout manner) that it will not cause issues in the final estimator that we proposed.

Step II'. Construct the permutation $\pi : [n] \mapsto [n]$ which sorts $\{\boldsymbol{X}_{iS}^\top \breve{\boldsymbol{\beta}}_S\}_{i=1}^n$ in increasing order[††]. In other words let π be such that $\boldsymbol{X}_{\pi_i S}^\top \breve{\boldsymbol{\beta}}_S \le \boldsymbol{X}_{\pi_j S}^\top \breve{\boldsymbol{\beta}}_S$ for all $\pi_i \le \pi_j$. Fit isotonic regression as in (2.2). For a given $\boldsymbol{X}$ set $x = \boldsymbol{X}_S^\top \breve{\boldsymbol{\beta}}_S$. The proposed final estimate of $f(\boldsymbol{X}^\top \boldsymbol{\beta}^*)$ is

$$\widehat{f}(x) = \begin{cases} \widehat{\mathbf{f}}_{\underset{i: x \ge \boldsymbol{X}_{\pi_i S}^\top \breve{\boldsymbol{\beta}}_S}{\operatorname{argmin}} (x - \boldsymbol{X}_{\pi_i S}^\top \breve{\boldsymbol{\beta}}_S)}, & x \ge \boldsymbol{X}_{\pi_1 S}^\top \breve{\boldsymbol{\beta}}_S \\ \widehat{\mathbf{f}}_1, & x < \boldsymbol{X}_{\pi_1 S} \breve{\boldsymbol{\beta}}_S \end{cases}$$

Since Proposition B.1 argues that $\mathcal{E}$ is a high probability event, analyzing Step II' ensures that similar guarantees hold for Step II with slightly smaller probability. From now on we focus on analyzing Step II'. We note that:

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_S &= \underset{\boldsymbol{\beta}_S \in \mathbb{R}^s}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^n (Y_i - \boldsymbol{X}_{iS}^\top \boldsymbol{\beta}_S)^2 + \lambda \|\boldsymbol{\beta}_S\|_1 \\ &= \underset{\boldsymbol{\beta}_S \in \mathbb{R}^s}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^n (f(\boldsymbol{X}_{iS}^\top \boldsymbol{\beta}_S^*) - \boldsymbol{X}_{iS}^\top \boldsymbol{\beta}_S)^2 - \frac{1}{n} \sum_{i=1}^n \varepsilon_i \boldsymbol{X}_{iS}^\top \boldsymbol{\beta}_S + \lambda \|\boldsymbol{\beta}_S\|_1 \right\}. \end{aligned} \tag{B.4}$$

The above representation makes it apparent that in fact:

$$\widetilde{\boldsymbol{\beta}}_S := \widetilde{\boldsymbol{\beta}}_S\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_{iS}, \{\boldsymbol{X}_{iS}\}_{i=1}^n\right), \text{ and thus } \breve{\boldsymbol{\beta}}_S := \breve{\boldsymbol{\beta}}_S\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_{iS}, \{\boldsymbol{X}_{iS}\}_{i=1}^n\right).$$

For the following argument let us first condition on the matrix $\mathbb{X}$.

Denote with $\mathbb{X}_S$ the $n \times s$ matrix whose rows are the row vectors $\{\boldsymbol{X}_{1S}^\top, \ldots, \boldsymbol{X}_{nS}^\top\}$. Define the projections

$$\mathbf{P}_{\mathbb{X}_S} = \mathbb{X}_S(\mathbb{X}_S^\top \mathbb{X}_S)^{-1}\mathbb{X}_S^\top, \quad \mathbf{P}_{\mathbb{X}_S^\perp} = \mathbf{I} - \mathbf{P}_{\mathbb{X}_S}.$$

Since the error vector $\boldsymbol{\varepsilon}$ has a $\mathcal{N}(0, \sigma^2 \mathbf{I})$ distribution by assumption, we have that conditionally on $\mathbb{X}$ the random vector $\mathbf{P}_{\mathbb{X}_S^\perp}\boldsymbol{\varepsilon}$ is independent of $\breve{\boldsymbol{\beta}}_S$ (note

[††] Here we use π with a slight abuse of notation to denote both permutations in Step II and Step II'; the two coincide on the high probability event $\mathcal{E}$ by Proposition B.1.

that the covariance $\mathbb{E}\mathbf{P}_{\mathbb{X}_S^\perp}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbb{X}_S = 0$). Let $\boldsymbol{\varepsilon}'$ be an independent copy of $\boldsymbol{\varepsilon}$ such that $\boldsymbol{\varepsilon}' \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We can rewrite our model as

$$Y_{\pi_i} = f(\boldsymbol{X}_{\pi_i S}^\top \boldsymbol{\beta}_S^*) + \varepsilon_{\pi_i} = f(\boldsymbol{X}_{\pi_i S}^\top \boldsymbol{\beta}_S^*) + [\mathbf{P}_{\mathbb{X}_S}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}')]_{\pi_i} + \xi_i,$$

where $\xi_i = [\mathbf{P}_{\mathbb{X}_S^\perp}\boldsymbol{\varepsilon} + \mathbf{P}_{\mathbb{X}_S}\boldsymbol{\varepsilon}']_{\pi_i}$. Importantly, notice the random variables $\xi_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d, since by definition $\mathbf{P}_{\mathbb{X}_S^\perp}\boldsymbol{\varepsilon} + \mathbf{P}_{\mathbb{X}_S}\boldsymbol{\varepsilon}' \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ is independent of $\breve{\boldsymbol{\beta}}_S$ and therefore of π_i. The motivation for defining ξ_i is that the aforementioned property is not true for the original noise variables ε_i, i.e., the distribution of ε_{π_i} is complicated. One disadvantage of ξ_i however is that they depend on the term $[\mathbf{P}_{\mathbb{X}_S}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}')]_{\pi_i}$.

We now introduce some shorthand notation. Let $u_i := f(\boldsymbol{X}_{\pi_i S}^\top \breve{\boldsymbol{\beta}}_S)$ and let $\gamma_i := \nu_i + \zeta_i$ where $\nu_i = f(\boldsymbol{X}_{\pi_i S}^\top \boldsymbol{\beta}_S^*)$ and $\zeta_i = [\mathbf{P}_{\mathbb{X}_S}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}')]_{\pi_i}$. The function $f(x)$ is increasing, and moreover by definition $\{\boldsymbol{X}_{\pi_i S}^\top \breve{\boldsymbol{\beta}}_S\}_{i=1}^n$ is an increasing sequence; we therefore have $\mathbf{u} \in \mathcal{S}_n^\uparrow$.

Even conditionally on the design $\mathbb{X}_S$, the terms γ_i for $i \in [n]$ remain random, and depend on ξ_i. Therefore known results such as Corollary 2.2 of [4] are not directly applicable in this situation. Following (1.21) of [4], using the cosine theorem and the fact that $\widehat{\mathbf{f}}$ is the projection of $\boldsymbol{Y} = \boldsymbol{\gamma} + \boldsymbol{\xi}$ on the cone $\mathcal{S}_n^\uparrow$, for any $\mathbf{v} \in \mathcal{S}_n^\uparrow$:

$$\|\widehat{\mathbf{f}} - \boldsymbol{\gamma} - \boldsymbol{\xi}\|_2^2 \leq \|\mathbf{v} - \boldsymbol{\gamma} - \boldsymbol{\xi}\|_2^2 - \|\mathbf{v} - \widehat{\mathbf{f}}\|_2^2.$$

In particular the above inequality holds for $\mathbf{v} = \mathbf{u}$, and after expanding the norms we obtain

$$\|\widehat{\mathbf{f}} - \boldsymbol{\gamma}\|_2^2 - \|\mathbf{u} - \boldsymbol{\gamma}\|_2^2 \leq 2\boldsymbol{\xi}^\top(\widehat{\mathbf{f}} - \mathbf{u}) - \|\widehat{\mathbf{f}} - \mathbf{u}\|_2^2.$$

By Cauchy-Schwartz we can lower bound the left hand side of the preceding display as

$$\|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2^2 - \|\mathbf{u} - \boldsymbol{\nu}\|_2^2 - 2\|\boldsymbol{\zeta}\|_2(\|\mathbf{u} - \boldsymbol{\nu}\|_2 + \|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2) \leq \|\widehat{\mathbf{f}} - \boldsymbol{\gamma}\|_2^2 - \|\mathbf{u} - \boldsymbol{\gamma}\|_2^2. \tag{B.5}$$

To this end we remind the reader of Theorem 2.3 of [4] which shows that if there exists $t_*(\mathbf{u})$ is such that

$$2\mathbb{E} \sup_{\mathbf{v} \in S_n^\uparrow : \|\mathbf{v} - \mathbf{u}\|_2 \leq t_*(\mathbf{u})} \boldsymbol{\xi}^\top(\mathbf{v} - \mathbf{u}) \leq t_*(\mathbf{u})^2,$$

then with probability at least $1 - e^{-t}$

$$2\boldsymbol{\xi}^\top(\widehat{\mathbf{f}} - \mathbf{u}) - \|\widehat{\mathbf{f}} - \mathbf{u}\|_2^2 \leq (2t_*^2(\mathbf{u}) + 4\sigma^2 t).$$

For such a $t^*(\mathbf{u})$ inequality (B.5) shows that with probability at least $1 - p^{-1}$

$$\begin{aligned}\|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2^2 &\leq \|\mathbf{u} - \boldsymbol{\nu}\|_2^2 + 2\|\boldsymbol{\zeta}\|_2(\|\mathbf{u} - \boldsymbol{\nu}\|_2 + \|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2) + 2t_*^2(\mathbf{u}) + 4\sigma^2 \log p \\ &\leq 2\|\mathbf{u} - \boldsymbol{\nu}\|_2^2 + 3\|\boldsymbol{\zeta}\|_2^2 + \frac{1}{2}\|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2^2 + 2t_*^2(\mathbf{u}) + 4\sigma^2 \log p,\end{aligned}$$

where we used that $2\|\boldsymbol{\zeta}\|_2\|\mathbf{u}-\boldsymbol{\nu}\|_2 \leq \|\boldsymbol{\zeta}\|_2^2 + \|\mathbf{u}-\boldsymbol{\nu}\|_2^2$ and $2\|\boldsymbol{\zeta}\|_2\|\widehat{\mathbf{f}}-\boldsymbol{\nu}\|_2 \leq 2\|\boldsymbol{\zeta}\|_2^2 + \frac{1}{2}\|\widehat{\mathbf{f}}-\boldsymbol{\nu}\|_2^2$. It follows that

$$\|\widehat{\mathbf{f}}-\boldsymbol{\nu}\|_2^2 \leq 4\|\mathbf{u}-\boldsymbol{\nu}\|_2^2 + 6\|\boldsymbol{\zeta}\|_2^2 + 4t_*^2(\mathbf{u}) + 8\sigma^2\log p.$$

[8] showed that such $t^*(\mathbf{u})$ can be taken as $t^*(\mathbf{u}) = \sqrt{c}\sigma(1+\frac{(u_n-u_1)}{\sigma})^{1/3}n^{1/6}$ for some absolute constant c. Now we will control the terms $\|\mathbf{u}-\boldsymbol{\nu}\|_2^2$ and $\|\boldsymbol{\zeta}\|_2^2$ which appear in (B.5). Since f is L-Lipschitz we have that

$$\|\mathbf{u}-\boldsymbol{\nu}\|_2^2 = \sum_{i=1}^{n}[f(\boldsymbol{X}_{iS}^\top\boldsymbol{\beta}_S^*) - f(\boldsymbol{X}_{iS}^\top\breve{\boldsymbol{\beta}}_S)]^2 \leq nL^2\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}(\boldsymbol{\beta}_S^* - \breve{\boldsymbol{\beta}}_S)\|_2^2.$$

Furthermore since $\mathbf{P}_{\mathbb{X}_S}$ is a projection matrix, we have

$$\|\boldsymbol{\zeta}\|_2^2 = (\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}')^\top\mathbf{P}_{\mathbb{X}_S}(\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}') \sim 2\chi^2(s).$$

Thus $\mathbb{E}\frac{1}{n}(\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}')^\top\mathbf{P}_{\mathbb{X}_S}(\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}') = \frac{2s}{n}$. Furthermore $\|\mathbf{P}_{\mathbb{X}_S}\|_F = \sqrt{s}$, and $\|\mathbf{P}_{\mathbb{X}_S}\|_2 = 1$. Using the Hanson-Wright inequality [35] for some absolute constant c we have for any $\mathbb{X}$ that

$$\mathbb{P}\big(\big|n^{-1}\|\boldsymbol{\zeta}\|_2^2 - \frac{2s}{n}\big| > t\big) \leq 2\exp\big(-c\frac{(nt)^2}{s} \wedge nt\big).$$

Setting $t = \frac{s}{n}$ gives that with probability at least $1-2\exp(-cs)$ we have $n^{-1}\|\boldsymbol{\zeta}\|_2^2 \leq \frac{3s}{n}$. To summarize conditionally on $\mathbb{X}$ we have established that with an overwhelming probability

$$\begin{aligned} n^{-1}\|\widehat{\mathbf{f}}-\boldsymbol{\nu}\|_2^2 \leq 4\sigma^2\big[c(1+\frac{(u_n-u_1)}{\sigma})^{2/3}n^{-2/3}\big] \\ + 4L^2\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}(\boldsymbol{\beta}_S^* - \breve{\boldsymbol{\beta}}_S)\|_2^2 + \frac{18s+8\sigma^2\log p}{n}. \end{aligned}$$

Note that the above bound has been established conditionally on the design matrix $\mathbb{X}$ and holds with high probability (independent of the design $\mathbb{X}$) for any design. Therefore it holds also unconditionally with high probability. Denote the event on which the above bound holds with $\mathcal{E}'$. It follows that on the intersection event $\mathcal{E}\cap\mathcal{E}'$ (recall that $\mathcal{E}$ is the event where (B.2) holds) we can further bound:

$$n^{-1}\|\widehat{\mathbf{f}}-\boldsymbol{\nu}\|_2^2 \leq 4\sigma^2\big[c(\frac{\sigma+(u_n-u_1)}{n\sigma})^{2/3}\big] + 8L^2C_1\frac{s\log p}{n} + \frac{18s+8\sigma^2\log p}{n}.$$

This completes the proof. □

Proof of Proposition B.1. Recall that $\mathbb{X}$ denotes the matrix stacking the vectors $\{\boldsymbol{X}_i^\top\}_{i\in[n]}$ into its rows, and that $\boldsymbol{Y}$ is the vector stacking the values $\{Y_i\}_{i\in[n]}$. Denote by $\mathbb{X}_S$ the restriction of $\mathbb{X}$ on the set S, i.e., $\mathbb{X}_S$ is the matrix with rows $\{\boldsymbol{X}_{iS}^\top\}_{i\in[n]}$.

Set $\lambda = C\sqrt{\frac{\log p}{n}}$. Theorem 2.3.4 i. of [30] states that under the conditions of Proposition B.1 if

$$\frac{n}{s\log p} \geq \frac{4\|\boldsymbol{\Sigma}_{S^c|S}\|_{\max}\left(\frac{4}{\lambda_{\min}} + \frac{2\xi^2}{\lambda^2 s}\right)}{\kappa^2} = \frac{4\|\boldsymbol{\Sigma}_{S^c|S}\|_{\max}\left(\frac{4}{\lambda_{\min}} + \frac{2\xi^2\frac{n}{s\log p}}{C}\right)}{\kappa^2},$$

then with high probability $\operatorname{supp}(\overline{\boldsymbol{\beta}}) \subseteq S$ and $\overline{\boldsymbol{\beta}}_S = \breve{\boldsymbol{\beta}}_S$ (the latter is not stated in the theorem but is implied by the proof). Note that under the assumptions of the Proposition this condition is satisfied when $\frac{n}{s\log p}$ is sufficiently large and $\lambda = C\sqrt{\frac{\log p}{n}}$ for a sufficiently large constant C. This completes the proof of our first claim. We will now show that the vector $\widetilde{\boldsymbol{\beta}}_S$ is close to $\boldsymbol{\beta}^*_S$ in Euclidean distance. We start by using the inequality:

$$\frac{1}{2n}\|\boldsymbol{Y} - \mathbb{X}_S\widetilde{\boldsymbol{\beta}}_S\|_2^2 + \lambda\|\widetilde{\boldsymbol{\beta}}_S\|_1 \leq \frac{1}{2n}\|\boldsymbol{Y} - \Upsilon\mathbb{X}_S\boldsymbol{\beta}^*_S\|_2^2 + \lambda\|\Upsilon\boldsymbol{\beta}^*_S\|_1,$$

where recall the definition of Υ from (1.5). Expanding the norms leads to

$$\begin{aligned}\frac{1}{2n}\|\mathbb{X}_S(\Upsilon\boldsymbol{\beta}^*_S - \widetilde{\boldsymbol{\beta}}_S)\|_2^2 + \lambda\|\widetilde{\boldsymbol{\beta}}_S\|_1 &\leq \frac{1}{n}\mathbf{w}^\top\mathbb{X}_S(\widetilde{\boldsymbol{\beta}}_S - \Upsilon\boldsymbol{\beta}^*_S) + \lambda\|\Upsilon\boldsymbol{\beta}^*_S\|_1\\ &\leq \frac{1}{n}\|\mathbf{w}^\top\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}}\|_\infty\|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}(\Upsilon\boldsymbol{\beta}^*_S - \widetilde{\boldsymbol{\beta}}_S)\|_1 + \lambda\|\Upsilon\boldsymbol{\beta}^*_S\|_1, \end{aligned} \tag{B.6}$$

where $\mathbf{w} = \boldsymbol{Y} - \Upsilon\mathbb{X}_S\boldsymbol{\beta}^*_S$. The vector $\mathbf{w}^\top\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}}$ is mean 0. We will now control $n^{-1}\|\mathbf{w}^\top\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}}\|_\infty$. First suppose that $\boldsymbol{\Sigma}_{SS} = \mathbf{I}$ (hence by the assumption $\|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}\boldsymbol{\beta}^*_S\|_2 = 1$ it follows $\|\boldsymbol{\beta}^*_S\|_2 = 1$). We have

$$n^{-1}\|\mathbf{w}^\top\mathbb{X}_S\|_\infty \leq n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}^{*\perp}_S}\mathbb{X}_S^\top\mathbf{w}\|_\infty + n^{-1}\|\boldsymbol{\beta}^*_S\boldsymbol{\beta}^{*\top}_S\mathbb{X}_S^\top\mathbf{w}\|_\infty, \tag{B.7}$$

where $\mathbf{P}_{\boldsymbol{\beta}^{*\perp}_S} = \mathbf{I}_s - \boldsymbol{\beta}^*_S\boldsymbol{\beta}^{*\top}_S$. Note that $\mathbf{P}_{\boldsymbol{\beta}^{*\perp}_S}\mathbb{X}_S^\top$ and $\mathbf{w}$ are independent. Thus it is simple to check that conditionally on $\mathbf{w}$ the vector

$$n^{-1}\mathbf{P}_{\boldsymbol{\beta}^{*\perp}_S}\mathbb{X}_S^\top\mathbf{w} \sim \mathcal{N}(0, \mathbf{P}_{\boldsymbol{\beta}^{*\perp}_S}n^{-2}\|\mathbf{w}\|_2^2).$$

We now argue that the term $n^{-1}\|\mathbf{w}\|_2^2 \leq 2\xi^2$ with probability at least $1 - \frac{\theta^2}{n\xi^2}$ (recall the definitions of θ^2, ξ^2 in (1.5)). Note that $\mathbf{w} = \boldsymbol{Y} - \Upsilon\mathbb{X}_S\boldsymbol{\beta}^*_S$ is not necessarily a zero mean vector. However, by Chebyshev's inequality we have:

$$\mathbb{P}\left(\left|\frac{\|\mathbf{w}\|_2^2}{n} - \xi^2\right| \geq t\right) \leq \frac{\theta^2}{nt^2}.$$

Then setting $t = \xi^2$ brings the above probability to 0 at a rate $\frac{\theta^2}{n\xi^2}$. Let $\mathcal{W} = \{\mathbf{w} : \frac{\|\mathbf{w}\|_2^2}{n} \leq 2\xi^2\}$. We just showed that $\mathbb{P}(\mathcal{W}) \geq 1 - \frac{\theta^2}{n\xi^2}$.

The diagonal entries of the covariance matrix $n^{-2}\|\mathbf{w}\|_2^2\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}$ are less than $n^{-2}\|\mathbf{w}\|_2^2$. Hence by a standard Gaussian tail bound and a union bound, we have that

$$\mathbb{P}(n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}\mathbb{X}_S^\top\mathbf{w}\|_\infty \geq t|\mathbf{w}) \leq 2se^{-2\bar{c}t^2n^2/\|\mathbf{w}\|_2^2},$$

for some universal constant $\bar{c}$. By the law of total probability

$$\begin{aligned}\mathbb{P}(n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}\mathbb{X}_S^\top\mathbf{w}\|_\infty \geq t) &= \int_{\mathcal{W}}\mathbb{P}(n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}\mathbb{X}_S^\top\mathbf{w}\|_\infty \geq t|\mathbf{w})p(\mathbf{w})d\mathbf{w}\\ &\quad+\int_{\mathcal{W}^c}\mathbb{P}(n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}\mathbb{X}_S^\top\mathbf{w}\|_\infty \geq t|\mathbf{w})p(\mathbf{w})d\mathbf{w}\\ &\leq \max_{\mathbf{w}\in\mathcal{W}}\mathbb{P}(n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}\mathbb{X}_S^\top\mathbf{w}\|_\infty \geq t|\mathbf{w}) + \mathbb{P}(\mathcal{W}^c)\\ &\leq 2se^{-\bar{c}nt^2/\xi^2} + \frac{\theta^2}{n\xi^2},\end{aligned}$$

where $p(\mathbf{w})$ denotes the density of $\mathbf{w}$.

Therefore setting $t \geq \sqrt{\frac{2\xi^2\log p}{\bar{c}n}}$ bounds the first term in the above probability by $\frac{2s}{p^2} + \frac{\theta^2}{n\xi^2} \leq 2p^{-1} + \frac{\theta^2}{n\xi^2}$. We now move to the second term of (B.7). Since $\|\boldsymbol{\beta}_S^*\|_\infty \leq \|\boldsymbol{\beta}_S^*\|_2 \leq 1$

$$n^{-1}\|\boldsymbol{\beta}_S^*\boldsymbol{\beta}_S^{*\top}\mathbb{X}_S^\top\mathbf{w}\|_\infty \leq n^{-1}|\boldsymbol{\beta}_S^{*\top}\mathbb{X}_S^\top\mathbf{w}|.$$

Next we have the elementary inequality

$$\begin{aligned}\mathbb{P}(n^{-1}|\boldsymbol{\beta}_S^{*\top}\mathbb{X}_S^\top\boldsymbol{Y} - \Upsilon\|\mathbb{X}_S\boldsymbol{\beta}_S^*\|_2^2| \geq t) &\leq \mathbb{P}(|n^{-1}\boldsymbol{\beta}_S^{*\top}\mathbb{X}_S^\top\boldsymbol{Y} - \Upsilon| \geq t/2)\\ &\quad+\mathbb{P}(|n^{-1}\|\mathbb{X}_S\boldsymbol{\beta}_S^*\|_2^2 - 1| \geq t/(2\Upsilon))^{\ddagger\ddagger},\end{aligned}$$

By Chebyshev's inequality

$$\mathbb{P}(|n^{-1}\boldsymbol{\beta}_S^{*\top}\mathbb{X}_S^\top\boldsymbol{Y} - \Upsilon| \geq t/2) \leq \frac{4\gamma^2}{nt^2}, \tag{B.8}$$

Setting $t = 2\gamma\sqrt{\frac{\log p}{n}}$ bounds the above probability by $(\log p)^{-1}$. By Lemma 1 of [24]

$$\mathbb{P}(|n^{-1}\|\mathbb{X}_S\boldsymbol{\beta}_S^*\|_2^2 - 1| \geq t/(2\Upsilon)) \leq 2\exp\bigg(-n\frac{t}{8\Upsilon}\wedge\frac{t^2}{64\Upsilon^2}\bigg),$$

Setting $t = 8\Upsilon\sqrt{\frac{\log p}{n}}$ bounds the above probability by $2p^{-1}$. We conclude that with probability at least $1 - 4p^{-1} - (\log p)^{-1} - \frac{\theta^2}{n\xi^2}$

$$n^{-1}\|\mathbf{w}^\top\mathbb{X}_S\|_\infty \leq \overline{C}\sqrt{\frac{\log p}{n}}, \tag{B.9}$$

[‡‡] Recall that $\Upsilon > 0$.

where $\overline{C}(\Upsilon, \gamma, \xi) = 8|\Upsilon| + 2\gamma + c_0\xi$ and $c_0 = \sqrt{2/\overline{c}}$ is a universal constant.

For the general $\mathbf{\Sigma}_{SS}$ case, we can rewrite the term $n^{-1}\|\mathbf{w}^\top \mathbb{X}_S \mathbf{\Sigma}_{SS}^{-\frac{1}{2}}\|_\infty$ as

$$n^{-1}\|\mathbf{w}^\top \mathbb{X}_S \mathbf{\Sigma}_{SS}^{-\frac{1}{2}}\|_\infty = n^{-1}\|(\boldsymbol{Y} - \Upsilon\widetilde{\mathbb{X}}_S\widetilde{\boldsymbol{\beta}}^*_S)^\top \widetilde{\mathbb{X}}_S\|_\infty$$

where $\widetilde{\mathbb{X}}_S = \mathbb{X}_S \mathbf{\Sigma}_{SS}^{-\frac{1}{2}} \sim \mathcal{N}(0, \mathbf{I})$ and $\widetilde{\boldsymbol{\beta}}^*_S = \mathbf{\Sigma}_{SS}^{\frac{1}{2}}\boldsymbol{\beta}^*_S$ so that $\|\widetilde{\boldsymbol{\beta}}^*_S\|_2 = 1$. Using the previous argument it follows that $n^{-1}\|(\boldsymbol{Y} - \Upsilon\widetilde{\mathbb{X}}_S\widetilde{\boldsymbol{\beta}}^*_S)^\top \widetilde{\mathbb{X}}_S\|_\infty \le \overline{C}\sqrt{\frac{\log p}{n}}$ with high probability. Hence we conclude

$$n^{-1}\|\mathbf{w}^\top \mathbb{X}_S \mathbf{\Sigma}_{SS}^{-\frac{1}{2}}\|_\infty \le \overline{C}\sqrt{\frac{\log p}{n}},$$

with high probability. Going back to (B.6) we have established that with high probability

$$\begin{aligned}\frac{1}{2n}\|\mathbb{X}_S(\Upsilon\boldsymbol{\beta}^*_S - \widetilde{\boldsymbol{\beta}}_S)\|_2^2 &\le \overline{C}\sqrt{\frac{\log p}{n}}\|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}(\Upsilon\boldsymbol{\beta}^*_S - \widetilde{\boldsymbol{\beta}}_S)\|_1 + \lambda(\|\Upsilon\boldsymbol{\beta}^*_S\|_1 - \|\widetilde{\boldsymbol{\beta}}_S\|_1) \\ &\le \left(\overline{C}\sqrt{\frac{s\log p}{n}} + \sqrt{s}\lambda\|\mathbf{\Sigma}_{SS}^{-\frac{1}{2}}\|_2\right)\|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}(\Upsilon\boldsymbol{\beta}^*_S - \widetilde{\boldsymbol{\beta}}_S)\|_2,\end{aligned} \tag{B.10}$$

where we used the inequality $\|\mathbf{v}\|_1 \le \sqrt{s}\|\mathbf{v}\|_2$ for $\mathbf{v} \in \mathbb{R}^s$. Lemma A.1 guarantees that

$$\frac{\lambda_{\min}(\mathbf{\Sigma}_{SS}^{-\frac{1}{2}}\mathbb{X}_S^\top\mathbb{X}_S\mathbf{\Sigma}_{SS}^{-\frac{1}{2}})}{n} \ge \frac{(\sqrt{n} - 2\sqrt{s})^2}{n}, \tag{B.11}$$

with probability at least $1 - 2e^{-s/2}$. Hence, when the above two events happen, i.e., events (B.10) and (B.11) (which happens with probability at least $1 - 4p^{-1} - (\log p)^{-1} - \frac{\theta^2}{n\xi^2} - 2e^{-s/2}$) we have

$$\|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}(\Upsilon\boldsymbol{\beta}^*_S - \widetilde{\boldsymbol{\beta}}_S)\|_2 \le r, \tag{B.12}$$

where $r := \frac{2\left(\overline{C}\sqrt{\frac{s\log p}{n}} + \sqrt{s}\lambda\lambda_{\min}^{-\frac{1}{2}}\right)}{\left(1 - 2\sqrt{\frac{s}{n}}\right)^2} = \overline{\overline{C}}\sqrt{\frac{s\log p}{n}}$. This completes the proof of (B.3). By the triangle inequality and (B.12)

$$\Upsilon - r \le \|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2 \le \Upsilon + r. \tag{B.13}$$

Now we will control the following term:

$$|\|\widehat{\mathbf{\Sigma}}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2 - \|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2|$$

By the Courant-Fischer minmax theorem for singular values we have

$$|\|\widehat{\mathbf{\Sigma}}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2 - \|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2| = |n^{-\frac{1}{2}}\|\mathbb{X}_S^\top\mathbf{\Sigma}_{SS}^{-\frac{1}{2}}\mathbf{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2 - \|\mathbf{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2|$$

$$
\begin{aligned}
&\leq \big\{|s_{\max}(n^{-\frac{1}{2}}\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}}) - 1| \vee |s_{\min}(n^{-\frac{1}{2}}\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}}) - 1|\big\}\|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2 \\
&\leq 2\sqrt{\frac{s}{n}}(\Upsilon + r),
\end{aligned}
$$

where the last inequality holds with probability at least $1 - 2e^{-s/2}$ by Lemma A.1. Recall that $\breve{\boldsymbol{\beta}}_S = \frac{\widetilde{\boldsymbol{\beta}}_S}{\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2}$. Next when $\Upsilon - r - 2\sqrt{\frac{s}{n}}(\Upsilon + r) > 0$ (which holds for sufficiently large $\frac{n}{s\log p}$), we have

$$
\begin{aligned}
&\|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}(\breve{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)\|_2 \\
&\qquad \leq \frac{\|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}_S - \Upsilon\boldsymbol{\beta}_S^*)\|_2 + |\Upsilon - \|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2| + |\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2 - \|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2|}{\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_S\|_2} \\
&\qquad \leq \frac{2r + 2\sqrt{\frac{s}{n}}(\Upsilon + r)}{\Upsilon - r - 2\sqrt{\frac{s}{n}}(\Upsilon + r)} = \frac{\overline{r}}{1 - \overline{r}},
\end{aligned}
\tag{B.14}
$$

where $\overline{r} = 2(\frac{r/\Upsilon}{1+r/\Upsilon} + \sqrt{\frac{s}{n}})$. It follows that (B.14) is smaller than $2\overline{r}$ when $1 - \overline{r} \leq 1/2$. Furthermore, by Lemma A.1

$$
\frac{\lambda_{\max}(\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}}\mathbb{X}_S^\top\mathbb{X}_S\boldsymbol{\Sigma}_{SS}^{-\frac{1}{2}})}{n} \leq \Big(1 + 2\sqrt{\frac{s}{n}}\Big)^2,
$$

with probability at least $1 - 2e^{-s/2}$. It follows that

$$
\|\widehat{\boldsymbol{\Sigma}}_{SS}^{\frac{1}{2}}(\breve{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)\|_2^2 = n^{-1}\|\mathbb{X}_S(\breve{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)\|_2^2 \leq \Big(1 + 2\sqrt{\frac{s}{n}}\Big)^2\|\boldsymbol{\Sigma}_{SS}^{\frac{1}{2}}(\breve{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)\|_2^2,
$$

which completes the proof of (B.2) in view of (B.14) and the assumptions of the Proposition. □

Proof of Lemma 3.5. We have the following chain of inequalities

$$
\|\mathbb{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_\infty \leq \|\mathbb{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 \leq C_1\sqrt{\frac{s\log p}{n}}
$$

where the last inequality holds by (B.2). □

Appendix C: Prediction proofs (full data)

Proof of Proposition 4.2. Since we are assuming that $\operatorname{supp}(\overline{\boldsymbol{\beta}}) \subseteq S$ holds, we can restrict our analysis only to the vector $\overline{\boldsymbol{\beta}}_S$, and $\boldsymbol{X}_{iS}$. In order not to burden the notation in this proof we will still refer to $\overline{\boldsymbol{\beta}}$ as $\overline{\boldsymbol{\beta}}_S$ and $\boldsymbol{X}_i$ as $\boldsymbol{X}_{iS}$.

Before we begin the proof we note that $\overline{\overline{x}}_i = \boldsymbol{X}_{\pi_i}^\top\overline{\boldsymbol{\beta}}/c = \boldsymbol{X}_{\pi_i}^\top\overline{\overline{\boldsymbol{\beta}}}$, where $\overline{\overline{\boldsymbol{\beta}}} = \overline{\boldsymbol{\beta}}/c$ satisfies $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\overline{\overline{\boldsymbol{\beta}}}\|_2 = 1$.

Let $\mathcal{N}_\epsilon$ be an ϵ-net on the s-dimensional unit sphere $\mathcal{S}^{s-1} = \{\mathbf{v} : \|\mathbf{v}\|_2 = 1\}$, where ϵ will be determined. By a standard volume argument [see, e.g., Lemma 5.2 in 41] we know that

$$|\mathcal{N}_\epsilon| \leq \left(1 + \frac{2}{\epsilon}\right)^s.$$

Convert the sphere covering to a covering on the ellipsoid $\mathcal{E}^{s-1} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{S}^{s-1} = \{\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{v} : \mathbf{v} \in \mathcal{S}^{s-1}\}$, by taking the set $\mathcal{N}_\epsilon^e = \mathbf{\Sigma}^{-\frac{1}{2}}\mathcal{N}_\epsilon = \{\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{v} : \mathbf{v} \in \mathcal{N}_\epsilon\}$. Since $\mathbf{\Sigma}^{\frac{1}{2}}\mathcal{E}^{s-1} = \mathcal{S}^{s-1}$, clearly for any vector $\mathbf{v} \in \mathcal{E}^{s-1}$ there exists $\mathbf{w} \in \mathcal{N}_\epsilon^e$ satisfying $\|\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{v} - \mathbf{w})\|_2 \leq \epsilon$. Note that by our argument above it follows that the vector $\overline{\overline{\boldsymbol{\beta}}} \in \mathcal{E}^{s-1}$.

For any two vectors $\boldsymbol{\gamma} \in \mathcal{E}^{s-1}, \boldsymbol{\beta} \in \mathcal{N}_\epsilon^e$ let $\mathbf{u}_{\boldsymbol{\gamma}} = [\mathbb{X}\boldsymbol{\gamma}]^{\uparrow}$ and $\mathbf{v}_{\boldsymbol{\beta}} = [\mathbb{X}\boldsymbol{\beta}]^{\uparrow}$. Below we will suppress the dependency of $\mathbf{u}_{\boldsymbol{\gamma}}$ and $\mathbf{v}_{\boldsymbol{\beta}}$ on $\boldsymbol{\gamma}, \boldsymbol{\beta}$ respectively. Using Lemma C.1 we know that:

$$\begin{aligned}\sup_{\boldsymbol{\gamma} \in \mathcal{E}^{s-1}} \left[n \sum_{i=1}^n [\Phi(u_i) - \Phi(u_{i-1})]^2\right]^{\frac{1}{2}} \leq & \sup_{\boldsymbol{\beta} \in \mathcal{N}_\epsilon^e} \left[n \sum_{i=1}^n [\Phi(v_i) - \Phi(v_{i-1})]^2\right]^{\frac{1}{2}} \\ & + \sqrt{2n\pi^{-1}}\|\mathbb{X}\mathbf{\Sigma}^{-\frac{1}{2}}\|_2\epsilon. \qquad \text{(C.1)}\end{aligned}$$

For any fixed vector $\boldsymbol{\gamma} \in \mathcal{E}^{s-1}$ we have that $\boldsymbol{X}_i^\top \boldsymbol{\gamma} \sim \mathcal{N}(0, 1)$ and therefore $\Phi((\boldsymbol{X}^\top\boldsymbol{\gamma})_i) \sim U(0, 1)$ are i.i.d. for $i \in [n]$. Hence we are in a position to apply Lemma C.2. Using Lemma C.2 together with the union bound we can ensure

$$\mathbb{P}\big(\sup_{\boldsymbol{\beta} \in \mathcal{N}_\epsilon}\big[n \sum_i (\Phi(v_i) - \Phi(v_{(i-1)})^2\big]^{\frac{1}{2}} \geq \sqrt{12}\big) \leq \left(1 + \frac{2}{\epsilon}\right)^s e^{-c_0\sqrt{n}}.$$

Additionally, by Lemma A.1 we have:

$$\|\mathbb{X}\mathbf{\Sigma}^{-\frac{1}{2}}\|_2 \leq (2\sqrt{n} + \sqrt{s}),$$

with probability at least $1 - 2e^{-n/2}$. Select $\epsilon = \frac{c}{n}$ for some constant c. We obtain that with probability at least $1 - 2e^{-n/2} - \left(1 + \frac{2n}{c}\right)^s e^{-c_0\sqrt{n}}$ we have that the RHS of (C.1) is bounded by $\sqrt{12} + c\sqrt{2}(2 + \sqrt{\frac{s}{n}})/\sqrt{\pi}$. Selecting c appropriately we can bound this quantity by $\sqrt{12} + \frac{1}{\sqrt{\pi}}$. Furthermore since we are assuming that $s \log n \ll \sqrt{n}$ the above probability will converge to 1. This completes the proof. □

Lemma C.1. *Suppose that $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{\Sigma})$ are i.i.d. for $i \in [n]$. Suppose that we have two vectors $\mathbf{w}, \mathbf{z}$, so that 1) $\|\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{w}\|_2 = 1, \|\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{z}\|_2 = 1$ and 2) $\|\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{z} - \mathbf{w})\|_2 \leq \epsilon$. Denote $\mathbf{u} = [\mathbb{X}\mathbf{z}]^{\uparrow}$ and $\mathbf{v} = [\mathbb{X}\mathbf{w}]^{\uparrow}$. Then*

$$\left[\sum_{i=1}^n [\Phi(u_i) - \Phi(u_{i-1})]^2\right]^{\frac{1}{2}} \leq \left[\sum_{i=1}^n [\Phi(v_i) - \Phi(v_{i-1})]^2\right]^{\frac{1}{2}} + \sqrt{2\pi^{-1}}\|\mathbb{X}\mathbf{\Sigma}^{-\frac{1}{2}}\|_2\epsilon.$$

Proof of Lemma C.1. By the triangle inequality

$$\Big[\sum_{i=1}^n \big[\Phi(u_i)-\Phi(u_{i-1})\big]^2\Big]^{\frac{1}{2}} \le \Big[\sum_{i=1}^n \big[\Phi(v_i)-\Phi(v_{i-1})\big]^2\Big]^{\frac{1}{2}}$$
$$+\Big[\sum_{i=1}^n \big[\Phi(u_i)-\Phi(u_{i-1})-\Phi(v_i)+\Phi(v_{i-1})\big]^2\Big]^{\frac{1}{2}}.$$

Using the elementary inequality $(a-b)^2 \le 2(a^2+b^2)$ and the fact that $\Phi(x)$ is a $\frac{1}{\sqrt{2\pi}}$-Lipschitz function, we have

$$\begin{aligned}\sum_{i=1}^n \big[\Phi(u_i)-\Phi(u_{i-1})-\Phi(v_i)+\Phi(v_{i-1})\big]^2 &\le \pi^{-1}\|\mathbf{u}-\mathbf{v}\|_2^2+\pi^{-1}\|\mathbf{u}-\mathbf{v}\|_2^2\\ &\le 2\pi^{-1}\|\mathbb{X}(\mathbf{z}-\mathbf{w})\|_2^2\\ &= 2\pi^{-1}\|\mathbb{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2^2\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\mathbf{z}-\mathbf{w})\|_2^2.\end{aligned}$$

The next to last inequality follows since for any two vectors $\mathbf{a}\in\mathbb{R}^n$ and $\boldsymbol{b}\in\mathbb{R}^n$ we have

$$\|\mathbf{a}^{\uparrow}-\boldsymbol{b}^{\uparrow}\|_2^2 \le \|\mathbf{a}-\boldsymbol{b}\|_2^2.$$

The latter be easily seen upon observing that if we have the ordering $a_i \le a_j$ but $b_j \le b_i$ then $(a_j-a_i)(b_j-b_i)\le 0$ or equivalently $(a_i-b_j)^2+(a_j-b_i)^2 \le (a_i-b_i)^2+(a_j-b_j)^2$. This is what we wanted to show. □

Lemma C.2. *Let $\{U_i\}_{i\in[n]}$ be i.i.d. samples from a uniform $U[0,1]$ distribution, and $\{U_{(i)}\}_{i\in[n]}$ be their order statistics. Then*

$$\mathbb{P}(n\sum_{i=0}^n (U_{(i+1)}-U_{(i)})^2 \ge 12) \le \exp(-c_0\sqrt{n}),$$

where $U_{(0)}=0$ and $U_{(n+1)}=1$, and c_0 is an absolute constant.

Proof of Lemma C.2. Let $W_i = U_{(i+1)}-U_{(i)}$. Recall that we have the representation [see Theorem 6.6. in 11, e.g.]

$$(W_0,W_1,\dots,W_{n-1}) = \left(\frac{X_0}{\sum_{i=0}^n X_i},\frac{X_1}{\sum_{i=0}^n X_i},\dots,\frac{X_{n-1}}{\sum_{i=0}^n X_i}\right),$$

where $X_i \sim \operatorname{Exp}(1)$ are i.i.d. This is equivalent to $(W_0,W_1,\dots,W_{n-1}) \sim \mathcal{D}((1,1,\dots,1))$, i.e., $(W_0,W_1,\dots,W_{n-1},W_n)$ are uniform on the n-dimensional simplex. It follows that

$$n\sum_{i=0}^n W_i^2 = \frac{n}{n+1}\frac{\sum_{i=0}^n X_i^2}{n+1}\frac{(n+1)^2}{(\sum_{i=0}^n X_i)^2} \le \frac{\sum_{i=0}^n X_i^2}{n+1}\frac{(n+1)^2}{(\sum_{i=0}^n X_i)^2}.$$

Since X_i are exponential, by definition they are sub-exponential, i.e., we have $\|X_i\|_{\psi_1}\le c$ for some absolute constant c.

Next we construct the random variables $Z_i = X_i^{\frac{1}{2}}$. By definition $\mathbb{E}Z_i^2 = 1$. We will now argue that Z_i are sub-Gaussian random variables. By Jensen's inequality

$$\mathbb{E}|Z_i|^p \le \sqrt{\mathbb{E}|X_i|^p} \le (p\|X_i\|_{\psi_1})^{p/2} \le (\sqrt{p}(\|X_i\|_{\psi_1})^{\frac{1}{2}})^p,$$

Hence $\|Z_i\|_{\psi_2} \le \|X_i\|_{\psi_1}^{1/2}$, and therefore Z is sub-Gaussian as claimed.

For the remaining part recall the notation preceding Theorem A.2. For $f(x) = x^4$ and $F(\boldsymbol{x}) = \sum_{i=1}^n f(x_i)$ we have $\mathbf{D}^\ell F(\boldsymbol{x}) = \operatorname{diag}_\ell(f^{(\ell)}(x_1), \ldots, f^{(\ell)}(x_n))$ for $\ell \in [4]$. Using the definition of ψ_2 norm we can easily estimate $\mathbb{E}[|Z|^\ell] \le (\sqrt{\ell})^\ell \|Z\|_{\psi_2}^\ell$. To this end we observe the following:

$$\|\operatorname{diag}_\ell\{x_1, \ldots, x_n\}\|_{\mathcal{J}} = \mathbb{1}(|\mathcal{J}| = 1)\|\boldsymbol{x}\|_2 + \mathbb{1}(|\mathcal{J}| \ge 2)\|\boldsymbol{x}\|_{\max}.$$

Hence:

$$\|\mathbb{E}\mathbf{D}^\ell F(\boldsymbol{Z})\|_{\mathcal{J}} \le [\mathbb{1}(|\mathcal{J}| = 1)\sqrt{n} + \mathbb{1}(|\mathcal{J}| \ge 2)]\frac{4!}{(4-\ell)!}(\sqrt{4-\ell})^{4-\ell}\|Z\|_{\psi_2}^{4-\ell},$$

for $\ell \in [4]$, where with a slight abuse of notation we understand $(\sqrt{4-\ell})^{(4-\ell)} = 1$ when $\ell = 4$. Using Theorem A.2 we obtain:

$$\begin{aligned}\mathbb{P}(|F(\boldsymbol{Z}) - \mathbb{E}F(\boldsymbol{Z})| \ge t) &\le 2\exp\Bigg(-\frac{1}{C_4}\min_{1\le\ell\le L}\min_{\mathcal{J}\in P_\ell}\Bigg(\frac{t}{\|Z\|_{\psi_2}^\ell \|\mathbb{E}\mathbf{D}^\ell f(\boldsymbol{X})\|_{\mathcal{J}}}\Bigg)^{2/|\mathcal{J}|}\Bigg)\\ &\le 2\exp\big(-\widetilde{C}_4\big(\frac{t^2}{n} \wedge \sqrt{t}\big)\big)\end{aligned}$$

where $\mathcal{P}_\ell$ is the set of partitions of $[\ell]$, the absolute constant C_4 depends solely on the dimension 4, and $\widetilde{C}_4$ depends on C_4 and the ψ_2-norm: $\|Z\|_{\psi_2} \le \|X_i\|_{\psi_1}^{1/2}$. Recalling that $\mathbb{E}X_i^2 = 2$, it follows that

$$\mathbb{P}\Bigg(\frac{\sum_{i=0}^n X_i^2}{n+1} - 2 \ge t\Bigg) \le 2\exp(-\widetilde{C}_4(nt^2 \wedge \sqrt{nt})).$$

Setting $t = 1$ gives a lower bound on the probability of at least $\exp(-\widetilde{C}_4\sqrt{n})$. Moreover since X_i are i.i.d. sub-exponential

$$\mathbb{P}\Bigg(-t \ge \frac{\sum_{i=0}^n X_i}{n+1} - 1\Bigg) \le 2\exp(-\widetilde{C}_2(nt^2 \wedge nt)),$$

for some absolute constant $\widetilde{C}_2$. Selecting $t = \frac{1}{2}$ completes the proof. □

Appendix D: Sample split proofs

Proposition D.1. *Under the conditions of Theorem 3.3 for the sample splitting version we have for some constant Ω_1*

$$\frac{\sum_{i=1}^n [f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}})]^2}{2n} \le \sigma^2\big[c\big(\frac{\sigma + |f(\boldsymbol{X}_{\pi_n}^\top \overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}_{\pi_1}^\top \overline{\boldsymbol{\beta}})|}{n\sigma}\big)^{2/3}\big]$$

$$+ \Omega_1 \frac{s \log p}{n} + \frac{2\sigma^2 \log p}{n}. \tag{D.1}$$

Denote with $\mathcal{R}_S$ the RHS of (D.1).

Proof of Proposition D.1. We now introduce some shorthand notation. Let $u_i := f(\boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}})$ and let $\nu_i := f(\boldsymbol{X}_{\pi_i}^\top \boldsymbol{\beta}^*)$. Since by definition $\{\boldsymbol{X}_{\pi_i}^\top \overline{\boldsymbol{\beta}}\}_{i=1}^n$ is an increasing sequence we have $\mathbf{u} \in \mathcal{S}_n^\uparrow$.

Conditionally on the design $\mathbb{X}$, the terms ν_i satisfy the conditions of Corollary 2.2 of [4] and hence we can directly apply this result. For convenience of the reader we spell out the full details of the application. Following (1.21) of [4], using the cosine theorem and the fact that $\widehat{\mathbf{f}}$ is the projection of $\boldsymbol{Y} = \boldsymbol{\nu} + \boldsymbol{\varepsilon}$ on the cone $\mathcal{S}_n^\uparrow$, for any $\mathbf{v} \in \mathcal{S}_n^\uparrow$:

$$\|\widehat{\mathbf{f}} - \boldsymbol{\nu} - \boldsymbol{\varepsilon}\|_2^2 \le \|\mathbf{v} - \boldsymbol{\nu} - \boldsymbol{\varepsilon}\|_2^2 - \|\mathbf{v} - \widehat{\mathbf{f}}\|_2^2.$$

In particular the above inequality holds for $\mathbf{v} = \mathbf{u}$, and after expanding the norms we obtain

$$\|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2^2 - \|\mathbf{u} - \boldsymbol{\nu}\|_2^2 \le 2\boldsymbol{\varepsilon}^\top(\widehat{\mathbf{f}} - \mathbf{u}) - \|\widehat{\mathbf{f}} - \mathbf{u}\|_2^2$$

Now we will show the first bound, but before that we remind the reader of Theorem 2.3 of [4] which shows that if $t_*(\mathbf{u})$ is such that

$$2\mathbb{E} \sup_{\mathbf{v} \in S_n^\uparrow : \|\mathbf{v} - \mathbf{u}\|_2 \le t_*(\mathbf{u})} \boldsymbol{\varepsilon}^\top(\mathbf{v} - \mathbf{u}) \le t_*(\mathbf{u})^2,$$

then with probability at least $1 - e^{-t}$

$$2\boldsymbol{\varepsilon}^\top(\widehat{\mathbf{f}} - \mathbf{u}) - \|\widehat{\mathbf{f}} - \mathbf{u}\|_2^2 \le 2t_*^2(\mathbf{u}) + 4\sigma^2 t.$$

For such a $t^*(\mathbf{u})$ with probability at least $1 - p^{-1}$

$$\|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2^2 \le \|\mathbf{u} - \boldsymbol{\nu}\|_2^2 + 2t_*^2(\mathbf{u}) + 4\sigma^2 \log p.$$

[8] showed that such $t^*(\mathbf{u})$ can be taken as $t^*(\mathbf{u}) = \sqrt{c}\sigma(1 + \frac{(u_n - u_1)}{\sigma})^{1/3} n^{1/6}$ for some absolute constant c.

Now we will control the term $\|\mathbf{u} - \boldsymbol{\nu}\|_2^2$. Since f is L-Lipschitz we have that

$$\|\mathbf{u} - \boldsymbol{\nu}\|_2^2 = \sum_{i=1}^n [f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - f(\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}})]^2 \le L^2 \|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\boldsymbol{\beta}^* - \overline{\boldsymbol{\beta}})\|_2^2.$$

Note that the above bound has been established conditionally on the design matrix $\mathbb{X}$ and holds with high probability (independent of the design $\mathbb{X}$) for any design. Therefore it holds also unconditionally with high probability. Denote the event on which the above bound holds with $\mathcal{E}'$. It follows that on the intersection event $\mathcal{E} \cap \mathcal{E}'$ (recall that $\mathcal{E}'$ is the event where (B.2) holds) we can further bound:

$$n^{-1}\|\widehat{\mathbf{f}} - \boldsymbol{\nu}\|_2^2 \le 2\sigma^2 t\big[c(\frac{\sigma + (u_n - u_1)}{n\sigma})^{2/3}\big]$$

$$+ C_1 L^2 \frac{s \log p}{n} + \frac{4\sigma^2 \log p}{n}.$$

This completes the proof. □

Remark D.2. *$[\mathbb{X}\overline{\boldsymbol{\beta}}]_{(1)}$ and $[\mathbb{X}\overline{\boldsymbol{\beta}}]_{(n)}$ are automatically of the order $\pm\sqrt{2\log(n/2)}$ since $\mathbb{X}$ and $\overline{\boldsymbol{\beta}}$ are independent.*

Proposition D.3. *Let $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ be a new observation generated independently from the data. Under the assumptions of Proposition D.1 hold, conditionally on the data and with high probability for some absolute constant C we have:*

$$\mathbb{E}_{\boldsymbol{X}|\mathcal{D}}\big[|\widehat{f}(\boldsymbol{X}^\top\overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}^\top\boldsymbol{\beta}^*)|\big|\boldsymbol{X}^\top\overline{\boldsymbol{\beta}} \in [\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}}, \boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}}]\big]$$
$$\leq C\bigg(\frac{f(\boldsymbol{X}_{\pi_n}^\top\overline{\boldsymbol{\beta}}) - f(\boldsymbol{X}_{\pi_1}^\top\overline{\boldsymbol{\beta}})}{\sqrt{n}} + \mathcal{R}_S^{\frac{1}{2}}\bigg).$$

Proof of Proposition D.3. The proof follows along the lines of the proof of Theorem 4.1. Due to the data split we can avoid using Proposition 4.2, and we can directly apply Lemma C.2 to control

$$\big[n\sum_{i=1}^{n}(\Phi(\overline{\overline{x}}_{i+1}) - \Phi(\overline{\overline{x}}_i))^2\big]^{\frac{1}{2}}.$$

This is true since $\Phi(\overline{\overline{x}}_i)$ are the order statistics $U_{(i)}$ conditionally on $\overline{\boldsymbol{\beta}}$. □

Appendix E: Additional simulation results

Here we provide additional numerical evidence. In particular we replicate the results from Section 5 but this time we use $n = 1000$ instead of $n = 500$ as before. The results are very similar so we do not comment further and refer the reader to commentary in Section 5.

Acknowledgements

The author is grateful to the Editor, Associate Editor and two anonymous referees for their numerous comments and suggestions which helped to significantly improve this manuscript.

References

[1] Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3-4):531–586, 2015. MR3383337

[2] Pierre Alquier and Gérard Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(Jan):243–280, 2013. MR3033331

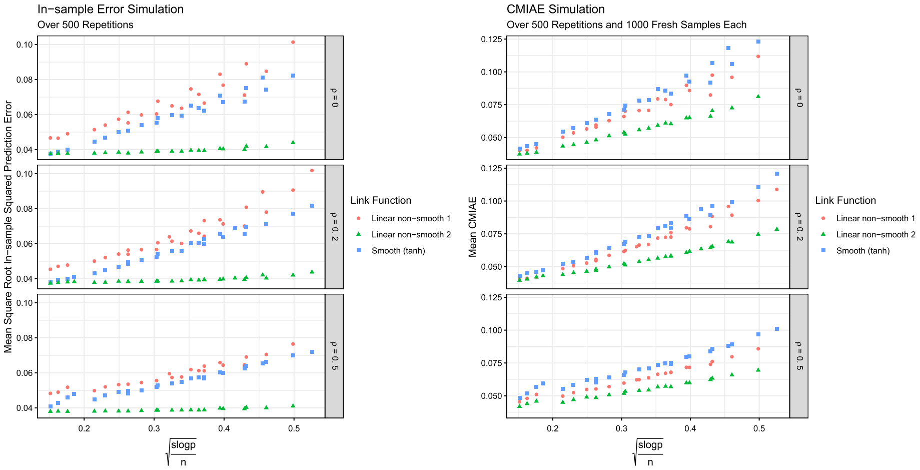


FIG 5. *This figure shows the same simulation results as in Figure 3, the difference being that $n = 1000$ through all simulation settings. We observe similar linear trends in all settings confirming our theoretical predictions.*

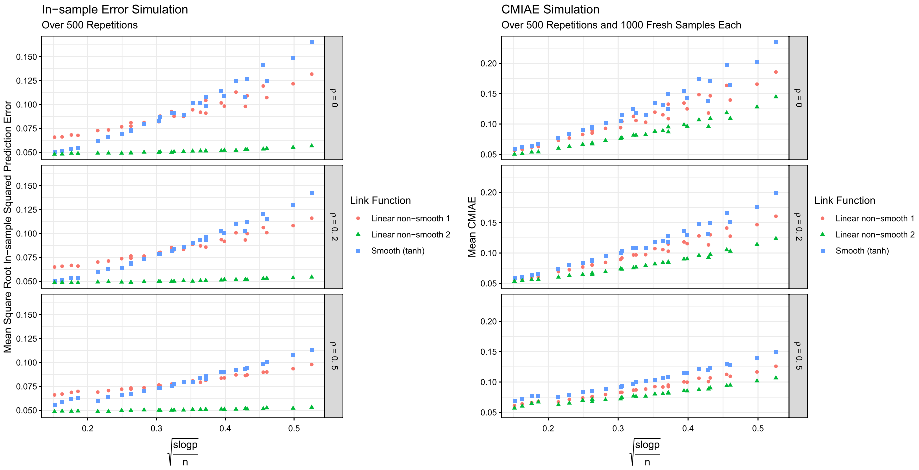


FIG 6. *This figure shows the same simulation results as in Figure 4, the difference being that $n = 1000$ through all simulation settings. We observe similar linear trends in all settings confirming our theoretical predictions.*

[3] Fadoua Balabdaoui, Cécile Durot, and Hanna Jankowski. Least squares estimation in the monotone single index model. *arXiv preprint arXiv:1610.06026*, 2016.

[4] Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*, 2015. MR3782383

[5] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009. MR2533469

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. MR3185193

[7] Sabyasachi Chatterjee, Adityanand Guntuboyina, Bodhisattva Sen, et al. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015. MR3357878

[8] Sourav Chatterjee et al. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014. MR3269982

[9] Yining Chen and Richard J Samworth. Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):729–754, 2016. MR3534348

[10] R Dennis Cook and Liqiang Ni. Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470), 2005. MR2160547

[11] Anirban DasGupta. Finite sample theory of order statistics and extremes. In *Probability for Statistics and Machine Learning*, pages 221–248. Springer, 2011. MR2807365

[12] Cécile Durot. Sharp asymptotics for isotonic regression. *Probability theory and related fields*, 122(2):222–240, 2002. MR1894068

[13] Cécile Durot et al. On the-error of monotonicity constrained estimators. *The Annals of Statistics*, 35(3):1080–1104, 2007. MR2341699

[14] Jared C Foster, Jeremy MG Taylor, and Bin Nan. Variable selection in monotone single-index models via the adaptive lasso. *Statistics in medicine*, 32(22):3944–3954, 2013. MR3102450

[15] Janos Galambos. Extreme value theory for applications. In *Extreme Value Theory and Applications*, pages 1–14. Springer, 1994.

[16] Larry Goldstein, Stanislav Minsker, and Xiaohan Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *arXiv preprint arXiv:1609.01025*, 2016. MR3832320

[17] Fang Han, Hongkai Ji, Zhicheng Ji, Honglang Wang, et al. A provable smoothing approach for high dimensional generalized regression with applications in genomics. *Electronic Journal of Statistics*, 11(2):4347–4403, 2017. MR3724223

[18] Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pages 505–531, 1992. MR1162997

[19] Joel L Horowitz. Optimal rates of convergence of parameter estimators in

the binary response model with weak distributional assumptions. *Econometric Theory*, 9(1):1–18, 1993. MR1221420

[20] Joel L Horowitz. *Semiparametric and nonparametric methods in econometrics*. Springer, 2009. MR2535631

[21] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

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

[23] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[24] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. MR1805785

[25] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. MR1137117

[26] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, pages 1009–1052, 1989. MR1015136

[27] Charles F Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228, 1975. MR0436905

[28] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall/CRC, 1989. MR3223057

[29] Prasad A Naik and Chih-Ling Tsai. Isotonic single-index model for high-dimensional database marketing. *Computational statistics &amp; data analysis*, 47(4):775–790, 2004. MR2101551

[30] Matey Neykov, Jun S Liu, and Tianxi Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *Journal of Machine Learning Research*, 17(87):1–37, 2016. MR3517110

[31] Heng Peng and Tao Huang. Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141(4):1362–1379, 2011. MR2747907

[32] Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016. MR3472264

[33] Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015. MR3349492

[34] Philippe Rigollet, Alexandre Tsybakov, et al. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011. MR2816337

[35] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013. MR3125258

[36] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003. MR1998720

[37] Robert P Sherman. Maximum score methods. In *Microeconometrics*, pages

122–128. Springer, 2010.

[38] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

[39] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. MR1379242

[40] Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. MR2724359

[41] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. MR2963170

[42] Yingcun Xia and WK Li. On single-index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285, 1999. MR1731489

[43] Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pages 3851–3860, 2017.

[44] Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina C. Eldar, and Tong Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *arXiv; 1511:04514*, 2015.

[45] Cun-Hui Zhang et al. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002. MR1902898