

# CANONICAL THRESHOLDING FOR NON-SPARSE HIGH DIMENSIONAL LINEAR REGRESSION

BY IGOR SILIN AND JIANQING FAN

Princeton University

We consider a high-dimensional linear regression problem. Unlike many papers on the topic, we do not require sparsity of the regression coefficients; instead, our main structural assumption is a decay of eigenvalues of the covariance matrix of the data. We propose a new family of estimators, called the canonical thresholding estimators, which pick largest regression coefficients in the canonical form. The estimators admit an explicit form and can be linked to LASSO and Principal Component Regression (PCR). A theoretical analysis for both fixed design and random design settings is provided. Obtained bounds on the mean squared error and the prediction error of a specific estimator from the family allow to clearly state sufficient conditions on the decay of eigenvalues to ensure convergence. In addition, we promote the use of the relative errors, strongly linked with the out-of-sample  $R^2$ . The study of these relative errors leads to a new concept of joint effective dimension, which incorporates the covariance of the data and the regression coefficients simultaneously, and describes the complexity of a linear regression problem. Some minimax lower bounds are established to showcase the optimality of our procedure. Numerical simulations confirm good performance of the proposed estimators compared to the previously developed methods.

## 1. Introduction and Setup. Consider the standard linear regression model

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon,$$

where  $\mathbf{x} \in \mathbb{R}^d$  is a vector of covariates,  $\boldsymbol{\beta} \in \mathbb{R}^d$  is a vector of coefficients,  $\varepsilon \in \mathbb{R}$  is a noise term, and  $y \in \mathbb{R}$  is a response. Suppose we observe  $n$  pairs  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  from this model with the assumption that the underlying noise terms  $\{\varepsilon_i\}_{i=1}^n$  are i.i.d. random variables with mean zero. In matrix notations, introducing

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n,$$

we rewrite our model as

$$(1.1) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Define the covariance matrix of the data  $\hat{\Sigma} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top = n^{-1} \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$ . Our goal is to estimate the unknown  $\boldsymbol{\beta}$  and analyze the quality of estimation in two different settings:

- *Fixed design.* That means, the vectors of covariates  $\{\mathbf{x}_i\}_{i=1}^n$  are deterministic (without loss of generality we assume  $\sum_{i=1}^n \mathbf{x}_i = 0$ ). A standard way to measure the error of an

---

*MSC 2010 subject classifications:* Primary 62J05; secondary 62H12, 62H25.

*Keywords and phrases:* High-dimensional linear regression, covariance eigenvalues decay, thresholding, relative errors, principal component regression.

estimator  $\tilde{\beta}$  in this case is the *mean squared error* (MSE):

$$\text{MSE}(\tilde{\beta}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \tilde{\beta} - \mathbf{x}_i^\top \beta)^2 = \frac{1}{n} \|\mathbf{X}\tilde{\beta} - \mathbf{X}\beta\|_2^2 = (\tilde{\beta} - \beta)^\top \hat{\Sigma}(\tilde{\beta} - \beta).$$

This differs from the prediction error for the fixed design by an amount of  $\mathbb{E}[\varepsilon^2]$  (independent of the model) and reflects the model error in the prediction for this case.

- *Random design.* In this scenario the vectors of covariates  $\{\mathbf{x}_i\}_{i=1}^n$  come independently from some unknown distribution with mean zero (for simplicity) and the covariance matrix  $\Sigma \stackrel{\text{def}}{=} \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{d \times d}$ . We are interested in the performance of an estimator  $\tilde{\beta}$  measured by the *expected prediction error* (PE):

$$\text{PE}(\tilde{\beta}) \stackrel{\text{def}}{=} \mathbb{E} \left[ (\mathbf{x}^\top \tilde{\beta} - \mathbf{x}^\top \beta)^2 \right] = (\tilde{\beta} - \beta)^\top \Sigma (\tilde{\beta} - \beta).$$

This quantity differs also from the prediction error for random design by  $\mathbb{E}[\varepsilon^2]$  and equals the excess risk

$$\text{PE}(\tilde{\beta}) = \mathbb{E} \left[ (y - \mathbf{x}^\top \tilde{\beta})^2 \right] - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta)^2 \right].$$

In the sequel we refer to these quantities simply as the (*absolute*) MSE and PE. The reason we give two names is to differentiate their statistical behavior in high dimensions and to avoid confusions at various discussions. We will also motivate and analyze the *relative errors*  $\text{MSE}(\tilde{\beta})/\text{MSE}(0)$  and  $\text{PE}(\tilde{\beta})/\text{PE}(0)$ . Surprisingly, the relative errors in this form, appearing naturally and being well-motivated, have not gained much attention in the literature (some related, but still quite different relative measures of performance in different contexts were considered in [Dobriban and Liu \(2019\)](#); [Dobriban and Sheng \(2020, 2021+\)](#)). As we will see, the importance of the relative errors arises as a high-dimensional effect.

Being a fundamental statistical problem, the *high-dimensional* linear regression has been approached in various ways. Probably the simplest method is Principal Component Regression (PCR). The idea is to reduce the dimension first via Principal Component Analysis (PCA) ([Pearson, 1901](#)), and then use several leading principal components as covariates to construct the least squares estimator. This approach heavily relies on a very strong assumption that the response depends on just a few leading principal components of the data. Various examples were provided where PCR performs poorly, see [Jolliffe \(1982\)](#). Another related idea to use supervised principal components was proposed by [Bair et al. \(2006\)](#). See also Chapters 10 and 11 of [Fan et al. \(2020\)](#) for further discussions and applications.

Over the past two decades, the main approach to tackle high-dimensionality of the problem has been the sparsity assumption on  $\beta$ , which is reasonable for many real-world applications. This has given rise to such model selection procedures as LASSO ([Tibshirani, 1996](#)), SCAD ([Fan and Li, 2001](#)), Least Angle Regression ([Efron et al., 2004](#)), Dantzig selector ([Candes and Tao, 2007](#)), SLOPE ([Bogdan et al., 2015](#)). The list of papers devoted to these methods is too long to be presented here, so we just mention some of them: [Greenshtein and Ritov \(2004\)](#); [Paul et al. \(2008\)](#); [Bickel, Ritov and Tsybakov \(2009\)](#); [Dalalyan, Hebiri and Lederer \(2017\)](#); [Bellec, Lecué and Tsybakov \(2018\)](#). Typically, a theoretical analysis of such procedures requires assumptions on the design like restricted isometry property (RIP), restricted eigenvalue (RE) condition, incoherence. These assumptions are needed to make sure that the correlations among subsets of features are small. See [van de Geer and Bühlmann \(2009\)](#) for an overview of conditions used in the theoretical analysis of sparse linear regression. We also refer to Chapters 3–5 of [Fan et al. \(2020\)](#) for an overview of existing methods and theoretical results for high-dimensional linear regression under sparsity.

The methods from the previous two paragraphs were developed (partially) due to a belief that the unconstrained least squares estimator is hopeless in high dimensions. Recent papers

by Bartlett et al. (2020) and later Chinot and Lerasle (2020) have shown that the minimum  $\ell_2$ -norm least squares estimator  $\tilde{\beta}^{LS} \stackrel{\text{def}}{=} (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{Y}$  (where  $(\mathbf{X}^\top \mathbf{X})^+$  is the generalized inverse of the matrix  $\mathbf{X}^\top \mathbf{X}$ ) can generalize well (i.e. have small absolute PE) even interpolating the training data – they call this phenomenon “benign overfitting”. To deliver convergence of  $\text{PE}(\tilde{\beta}^{LS})$  to zero they require quite specific conditions on  $\Sigma$ : the decay of its eigenvalues should be fast, but not too fast. These requirements are quantified by two notions of effective rank of  $\Sigma$ . A closely related paper by Hastie et al. (2019) also studies  $\tilde{\beta}^{LS}$ , but in the regime  $p/n \rightarrow \gamma$ . They are interested in the dependence of  $\text{PE}(\tilde{\beta}^{LS})$  on  $\gamma$ , and focus on the case  $\lambda_{\min}(\Sigma) \geq c > 0$ , e.g. considering isotropic and equicorrelation covariances. One more work on the topic is Belkin, Hsu and Xu (2019), where the authors try to mathematically explain double descent phenomenon in several different models.

Going beyond the linear regression, one basic idea to approach general (nonlinear) regression problem  $y = f(\mathbf{x}) + \varepsilon$  is to decompose the regression function  $f(\mathbf{x}) \approx \sum_{j=1}^D \beta_j \psi_j(\mathbf{x}) = \psi(\mathbf{x})^\top \beta$  over a Fourier basis, wavelet basis, or basis of eigenfunctions in reproducing kernel Hilbert space (RKHS), denoted here by  $\psi_1(\cdot), \dots, \psi_D(\cdot)$ . This reduces the nonlinear regression problem to a linear one (potentially very high-dimensional or even infinite-dimensional), and allows to apply the existing methods. Though we do not pursue the analysis of nonlinear regression in our work, this setting provides an excellent motivation for the main structural assumptions we make in our results. One of them is *fast decay of the eigenvalues* of  $\Sigma$  (or  $\hat{\Sigma}$ ). For instance, we require that the *effective rank*

$$r^{\text{eff}}[\Sigma] \stackrel{\text{def}}{=} \frac{\text{Tr}[\Sigma]}{\|\Sigma\|} \quad \left( \text{or } r^{\text{eff}}[\hat{\Sigma}] \stackrel{\text{def}}{=} \frac{\text{Tr}[\hat{\Sigma}]}{\|\hat{\Sigma}\|} \right)$$

can be well-controlled. The spectral decay has been observed in real-world datasets (e.g. MNIST, see Figure 5 in Liang and Rakhlin (2020); financial data in Zumbach (2009), Figure 5; economics data in Fan, Ke and Wang (2020), Figure 5), which makes our assumption reasonable. Importance of the eigenvalue decay (not only of the covariance, but of general kernel matrices) is highlighted in Liang and Rakhlin (2020), where such kind of conditions on the spectral decay is called “favorable data geometry”. Moreover, Ma and Belkin (2017); Belkin (2018) analyze the super-polynomial decay of eigenvalues of smooth kernel matrices. Going even further in deep learning literature, neural tangent kernels also exhibit the spectral decay, as shown by Bietti and Mairal (2019), among others. However, the fast eigenvalue decay is not the only motivation behind our work; another structural assumption that can make our results meaningful is a fast decay of regression coefficients in eigenbasis. This is a very well-understood condition as well: it is well-known that Fourier coefficients decay at a polynomial rate, where the degree depends on the smoothness of the underlying regression function. In addition, the decay of coefficients in RKHS was studied by Belkin (2018).

With these structural assumptions, the idea behind our family of estimators  $\hat{\beta}$  is quite natural: in some eigendirections (e.g. the ones that correspond to small eigenvalues of  $\Sigma$ ) we do not gain much by estimating the associated coefficient, so it makes sense to estimate only those components that allow to significantly reduce the error; specifically, we use thresholding to cut the components associated with the insignificant directions off. When applied to the nonlinear regression with wavelet basis, one estimator from the proposed family coincides with the soft thresholding approach studied in the series of papers by Donoho and Johnstone (1994); Donoho (1995); Donoho and Johnstone (1995); Donoho et al. (1995); Donoho and Johnstone (1998), among others. We highlight that we will not require sparsity of  $\beta$  or any restrictive conditions on the design.

Let us summarize some motivations behind our work:

- Our methods can be viewed as an attempt to fix PCR by relaxing its restrictive assumptions. Instead of working with the several leading principal components, our estimators automatically screen for the most important principal components, not necessarily the leading ones.
- Remarkably, the procedures that we propose are a modification of LASSO, so one can view this work as an attempt to extend LASSO to non-sparse high-dimensional linear regression.
- Though the papers by [Bartlett et al. \(2020\)](#) and [Chinot and Lerasle \(2020\)](#) do not advocate the use of interpolating estimators rather justify why the overfitting may not be harmful (very relevant question in modern deep learning research), we aim to show that there is no necessity to give up the in-sample denoising quality to get good bounds on the prediction error. In fact, our numerical results show that our method is better the least squares estimator in various situations.

Main contributions of this paper are:

- We propose a new method for high-dimensional linear regression, called *Natural Canonical Thresholding* (NCT), in Section 2. The connection of this approach to LASSO and PCR is discussed in Section 2.1 and Section 2.2. In Section 2.3 we extend the suggested procedure and present a richer family of estimators, called *Generalized Canonical Thresholding* (GCT). Our estimators  $\hat{\beta}$  are given via an explicit expression and do not require any optimization. Though each estimator from the family has one hyperparameter, it can be tuned in an efficient way via cross-validation as shown in Section 6. An optimality result for the cross-validation is also presented.
- We provide theoretical guarantees for the NCT estimator in the fixed design and random design settings in Section 3. The presented tight bounds have two-fold meaning:
  - For the absolute errors  $\text{MSE}(\hat{\beta})$  and  $\text{PE}(\hat{\beta})$ , studied in Section 3.1 and Section 3.2, we state explicit sufficient conditions of the form “the eigenvalues of  $\hat{\Sigma}$  or  $\Sigma$  decay fast enough” to ensure convergence in high dimensions. No conditions on  $\beta$  are imposed in this case.
  - For the relative errors  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  and  $\text{PE}(\hat{\beta})/\text{PE}(0)$ , motivated in Section 3.3.1, our bounds factorize into the newly defined notion of the *joint effective dimension*, the signal-to-noise ratio, and a vanishing factor. To get good rates for the relative errors in high dimensions it is not enough to assume fast decay of eigenvalues of  $\hat{\Sigma}$  or  $\Sigma$  alone, and we need to impose conditions of  $\Sigma$  and  $\beta$  together (Section 3.3.2), which is reflected by the joint effective dimension that we analyze (Section 3.3.3).

We introduce parameter classes for linear regression problems with bounded joint effective dimension, and demonstrate a minimax optimality of the NCT estimator over this classes in the fixed design setting (Section 4).

Theoretical analysis of the GCT estimator is not that insightful, however we still present and discuss a tight bound on the absolute error  $\text{MSE}(\hat{\beta})$  (Section 5).

- Numerical experiments, conducted in Section S.1 of the Supplementary Material [Silin and Fan \(2021\)](#), confirm good performance of NCT and especially GCT in comparison with other existing methods.

All the proofs are collected in the Supplementary Material [Silin and Fan \(2021\)](#). We conclude this section with defining some notations used throughout the work.

For a positive integer  $k$ , we write  $[k]$  as shorthand for the set  $\{1, 2, \dots, k\}$ . We use  $\mathbb{O}_{k \times l}$  for  $k \times l$  matrix of zeros and  $\mathbb{I}_k$  for the identity matrix of size  $k \times k$ . For a vector  $\mathbf{a} = [a_1, \dots, a_k]^\top \in \mathbb{R}^k$  and  $q > 0$ , the standard  $\ell_q$ -(pseudo)norm in  $\mathbb{R}^k$  is  $\|\mathbf{a}\|_q \stackrel{\text{def}}{=} \left( \sum_{j=1}^k |a_j|^q \right)^{1/q}$ . We use the following convention for the  $\ell_0$ -pseudonorm:  $\|\mathbf{a}\|_0^0 =$

$\|\mathbf{a}\|_0 \stackrel{\text{def}}{=} \sum_{j=1}^k \mathbb{1}\{a_j \neq 0\}$ . Also, the  $\ell_\infty$ -norm is  $\|\mathbf{a}\|_\infty = \max_{j \in [k]} |a_j|$ . For a matrix  $\mathbf{A}$ , let  $\|\mathbf{A}\|$  be the spectral norm (the largest singular value),  $\text{rank}[\mathbf{A}]$  be the rank, and (if  $\mathbf{A}$  is square)  $\text{Tr}[\mathbf{A}]$  be the trace.

For sequences  $a_n$  and  $b_n$  the relation  $a_n \lesssim b_n$  means that there exists a positive absolute constant  $C$  such that  $a_n \leq Cb_n$  for all  $n$ , while  $a_n \asymp b_n$  means that  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ . Oftentimes, similar notations will be used to denote inequalities/equalities up to a multiplicative constant not across the sample size  $n$ , but across all indices  $j \in [d]$  or  $j \in [\min(d, n)]$ . The exact meaning in each case will be clear from the context. By  $c, C$  we denote absolute constants which may differ from place to place.

Throughout the work,  $\beta$  stands for the true vector of regression coefficients in the model (1.1),  $\hat{\beta}$  stands for our NCT or GCT estimators proposed in the next section, and a generic estimator is denoted as  $\tilde{\beta}$ . If we want to refer to an abstract vector, for example, as a variable in an optimization problem, we will be using  $\beta'$ .

**2. Estimators.** Let  $r \stackrel{\text{def}}{=} \text{rank}[\hat{\Sigma}]$ . Typically,  $r = \min(d, n)$ . Consider the SVD of the data matrix  $\mathbf{X}$  (scaled by  $n^{-1/2}$ ):

$$\frac{1}{\sqrt{n}} \mathbf{X} = \hat{\mathbf{V}} \hat{\mathbf{\Lambda}} \hat{\mathbf{U}}^\top,$$

where  $\hat{\mathbf{\Lambda}} = \text{diag}(\hat{\lambda}_1^{1/2}, \dots, \hat{\lambda}_r^{1/2}) \in \mathbb{R}^{r \times r}$  is a diagonal matrix consisting of the non-zero singular values of  $n^{-1/2} \mathbf{X}$  in non-increasing order, the columns of  $\hat{\mathbf{V}} \in \mathbb{R}^{n \times r}$  are the left singular vectors of  $\mathbf{X}$ , and the columns of  $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_r] \in \mathbb{R}^{d \times r}$  are the right singular vectors of  $\mathbf{X}$ . Alternatively, it is also convenient to think of the eigendecomposition of  $\hat{\Sigma}$ :

$$\hat{\Sigma} = \hat{\mathbf{U}} \hat{\mathbf{\Lambda}}^2 \hat{\mathbf{U}}^\top,$$

where now the diagonal entries  $\hat{\lambda}_1, \dots, \hat{\lambda}_r$  of  $\hat{\mathbf{\Lambda}}^2$  are interpreted as the non-zero eigenvalues of  $\hat{\Sigma}$  in non-increasing order, and the columns  $\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_r$  of  $\hat{\mathbf{U}}$  are the corresponding eigenvectors of  $\hat{\Sigma}$ . Similarly, in what follows we will actively use the eigendecomposition of  $\Sigma$ :

$$\Sigma = \mathbf{U} \mathbf{\Lambda}^2 \mathbf{U}^\top,$$

where  $\mathbf{\Lambda}^2 = \text{diag}(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^{d \times d}$  is a diagonal matrix consisting of the eigenvalues of  $\Sigma$  in non-increasing order, and  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_d] \in \mathbb{R}^{d \times d}$  consists of the corresponding eigenvectors.

We introduce the following definition, which will be extensively used throughout the work.

**DEFINITION 2.1.** Rewrite the linear regression model  $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$  as

$$\mathbf{Y} = (\mathbf{X} \hat{\mathbf{U}} \hat{\mathbf{\Lambda}}^{-1}) (\hat{\mathbf{\Lambda}} \hat{\mathbf{U}}^\top \beta) + \varepsilon = \mathbf{Z} \theta + \varepsilon.$$

We call this representation the canonical form of the linear regression model. Here  $\mathbf{Z} \stackrel{\text{def}}{=} \mathbf{X} \hat{\mathbf{U}} \hat{\mathbf{\Lambda}}^{-1} \in \mathbb{R}^{n \times r}$  is the standardized design matrix and

$$\theta \stackrel{\text{def}}{=} \hat{\mathbf{\Lambda}} \hat{\mathbf{U}}^\top \beta \in \mathbb{R}^r$$

is the new vector of coefficients, called the canonical regression coefficients vector, or simply canonical coefficients.

Note that the standardized design coincides with the left singular vectors  $\mathbf{Z} = n^{1/2}\widehat{\mathbf{V}}$  and satisfies the orthonormality constraints:  $n^{-1}\mathbf{Z}^\top\mathbf{Z} = \mathbb{I}_r$ . Hence, the least-squares estimator for the canonical parameter is  $\widetilde{\boldsymbol{\theta}}^{LS} = n^{-1}\mathbf{Z}^\top\mathbf{Y} = n^{-1}\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\mathbf{U}}^\top\mathbf{X}^\top\mathbf{Y}$ . As in Fan (1996), we further regularize the estimated canonical coefficients vector by either thresholding or truncation (setting higher indices to zero), depending whether  $\boldsymbol{\theta}$  is approximately sparse or concentrates on the leading principal components. Transforming the canonical parameter back to the original domain leads to the canonical thresholding estimator or principal component regression estimator, as to be further elaborated below. Our work pushes forward the interactions between the canonical parameters and the design matrix.

More specifically, in the canonical domain our estimator looks like

$$\widehat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \text{SOFT}_\tau \left[ \widetilde{\boldsymbol{\theta}}^{LS} \right],$$

which in the original domain brings us to the *Natural Canonical Thresholding* (NCT) estimator of  $\boldsymbol{\beta}$ , defined as

$$(2.1) \quad \widehat{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}^{-1} \text{SOFT}_\tau \left[ \widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\mathbf{U}}^\top \frac{\mathbf{X}^\top\mathbf{Y}}{n} \right],$$

where  $\text{SOFT}_\tau[z] \stackrel{\text{def}}{=} z \cdot \max(1 - \tau/|z|, 0)$  is the *soft thresholding* function applied component-wise and  $\tau \geq 0$  is a hyperparameter to be chosen. Note that in an overparameterized setting  $d > n$  there are infinitely many  $d$ -dimensional vectors leading to the same canonical coefficients. Specifically, for any vector  $\mathbf{h} \in \mathbb{R}^d$  the estimator  $\widehat{\boldsymbol{\beta}} + (\mathbb{I}_d - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top)\mathbf{h}$  leads to the same vector of canonical coefficients:  $\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^\top[\widehat{\boldsymbol{\beta}} + (\mathbb{I}_d - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top)\mathbf{h}] = \widehat{\boldsymbol{\theta}}$ . While having the same in-sample fit, these estimators may produce different predictions for new points  $\mathbf{x}$ . Among all of these estimators,  $\widehat{\boldsymbol{\beta}}$  from (2.1), namely the estimator with  $\mathbf{h} = 0$ , has the smallest  $\ell_2$ -norm, arguably being the most reasonable choice.

Let us explain some intuition behind the NCT estimator. Neglecting the noise term, we plug  $\mathbf{Y} \approx \mathbf{Z}\boldsymbol{\theta}$  in, use the eigendecomposition of  $\widehat{\boldsymbol{\Sigma}}$  and get

$$\widehat{\boldsymbol{\theta}} \approx \text{SOFT}_\tau[\boldsymbol{\theta}].$$

Due to the structure of our error (e.g. in the fixed design case)

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \|\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^\top\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^\top\boldsymbol{\beta}\|_2^2 = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2$$

and since we assume the eigenvalue decay, it is likely that some components  $\theta_j$  do not play role, and the estimation of them with  $\widehat{\theta}_j$  is not that important. Hence, it is reasonable to cut such insignificant components off, and this is exactly what the thresholding does. This reduces the variance of the estimator, while not increasing the bias by too much.

We also mention that when  $\tau = 0$ , our estimator reduce to the minimum  $\ell_2$ -norm least squares solution

$$\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}^{LS} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{LS} = \widehat{\boldsymbol{\Sigma}}^+ \frac{\mathbf{X}^\top\mathbf{Y}}{n} = (\mathbf{X}^\top\mathbf{X})^+ \mathbf{X}^\top\mathbf{Y}$$

(unbiased or slightly biased, but with large variance), while  $\tau = +\infty$  corresponds to the trivial solution  $\widehat{\boldsymbol{\theta}} = 0$  and  $\widehat{\boldsymbol{\beta}} = 0$  (very biased, but with zero variance).

**2.1. Relation to LASSO.** Recall that the standard LASSO estimator is a solution of the following optimization problem:

$$\widetilde{\boldsymbol{\beta}}^{LASSO} \in \arg \min_{\boldsymbol{\beta}' \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}'\|_2^2 + \tau \|\boldsymbol{\beta}'\|_1 \right\}.$$



In practice one usually standardizes the columns of  $\mathbf{X}$  so that they are on the same scale and the coefficients corresponding to different covariates are penalized equally. Now imagine that we standardize our  $\mathbf{X}$  in the canonical manner as in Definition 2.1. If we run LASSO for the vector of coefficients  $\boldsymbol{\theta}$ , then the solution is expressed via the soft thresholding:

$$\tilde{\boldsymbol{\theta}}^{LASSO} = \arg \min_{\boldsymbol{\theta}' \in \mathbb{R}^r} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}'\|_2^2 + \tau \|\boldsymbol{\theta}'\|_1 \right\} = \text{SOFT}_\tau \left[ \frac{\mathbf{Z}^\top \mathbf{Y}}{n} \right],$$

which is exactly our estimator  $\hat{\boldsymbol{\theta}}$  in the canonical domain. Going back to  $\hat{\boldsymbol{\beta}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{\theta}}^{LASSO}$  we recover the NCT estimator (2.1). The solution is the soft thresholding on the canonical regression coefficients. This is why we call the method canonical thresholding.

We also note that the NCT estimator can be represented as the min- $\ell_2$ -norm solution of the optimization problem

$$\hat{\boldsymbol{\beta}} \in \arg \min_{\boldsymbol{\beta}' \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}'\|_2^2 + \tau \|\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top \boldsymbol{\beta}'\|_1 \right\}.$$

Our estimator is nothing more than LASSO penalized on the canonical regression coefficients.

**2.2. Relation to PCR.** Principal Component Regression (PCR) approaches the high dimensionality of the problem by taking only  $m$  ( $m < r = \min(d, n)$ ) leading principal components of the original data. The new design matrix becomes

$$\mathbf{Z}_m = \mathbf{X}\hat{\mathbf{U}}_{\leq m}\hat{\boldsymbol{\Lambda}}_{\leq m}^{-1} \in \mathbb{R}^{n \times m},$$

where  $\hat{\mathbf{U}}_{\leq m} \in \mathbb{R}^{d \times m}$  consists of the first  $m$  columns of  $\hat{\mathbf{U}}$  and  $\hat{\boldsymbol{\Lambda}}_{\leq m} \in \mathbb{R}^{m \times m}$  is  $m \times m$  leading principal submatrix of  $\hat{\boldsymbol{\Lambda}}$ . The new regression problem

$$\mathbf{Y} = \mathbf{Z}_m \boldsymbol{\theta}_m + \boldsymbol{\epsilon}$$

is solved via the least squares, yielding the solution

$$\tilde{\boldsymbol{\theta}}_m^{LS} = \frac{\mathbf{Z}_m^\top \mathbf{Y}}{n} \in \mathbb{R}^m,$$

and thus

$$\tilde{\boldsymbol{\beta}}^{PCR} \stackrel{\text{def}}{=} \hat{\mathbf{U}}_{\leq m} \hat{\boldsymbol{\Lambda}}_{\leq m}^{-1} \tilde{\boldsymbol{\theta}}_m^{LS} \in \mathbb{R}^d.$$

Note that  $\tilde{\boldsymbol{\theta}}_m^{LS}$  is essentially the first  $m$  components of  $\tilde{\boldsymbol{\theta}}^{LS}$ , and we can express

$$\tilde{\boldsymbol{\beta}}^{PCR} = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}^{-1} \text{ZERO}_m \left[ \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{U}}^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n} \right],$$

where  $\text{ZERO}_m[\mathbf{z}] = [z_1, \dots, z_m, 0, \dots, 0]^\top$  is the operator zeroing out all the components of a vector  $\mathbf{z} \in \mathbb{R}^r$  except the first  $m$ . (Again, in an overparameterized case this estimator has the smallest  $\ell_2$ -norm among all estimators leading to the canonical coefficients  $\tilde{\boldsymbol{\theta}}_m^{LS}$ .) The similarity of the PCR estimator to the NCT estimator (2.1) is now clear. While PCR blindly selects the coefficients corresponding to the first  $m$  principal components, our procedure screens for the most important principal directions, which may be different from the leading ones, and leaves only those with significant contribution exceeding  $\tau$ . However, if there is a strong prior indicating the canonical coefficients spike at the principal directions, then of course PCR should also be a suitable procedure, and NCT will simply mimic its behavior, with some small costs. The above contrasts between truncation and thresholding appear in Fan (1996) under a simpler model.

2.3. *Extension to family of canonical thresholding estimators.* PCR focuses only on the estimators on the principal component directions, and NCT does not have any preferences. We now propose a family of canonical thresholding estimators to bridge these two extremes, progressively putting more preferences on the principal directions. In addition, we also generalize the thresholding function.

First, the soft thresholding can be replaced by *generalized thresholding rules* (see e.g. Definition 9.3 in Fan et al. (2020)), introduced for completeness in the following definition.

DEFINITION 2.2. *The function  $T_\tau : \mathbb{R} \rightarrow \mathbb{R}$  is called a generalized thresholding function, if*

- (i)  $|T_\tau[z]| \leq c|z'|$  for all  $z, z'$  satisfying  $|z - z'| \leq \tau/2$  and some constant  $c$ ;
- (ii)  $|T_\tau[z] - z| \leq \tau$  for all  $z \in \mathbb{R}$ .

The parameter  $\tau$  is called the thresholding level.

Second, if there is some prior that the spike canonical coefficients are more likely to be in the lower principal components rather than in the higher ones, we can introduce additional multiplicative weights, equal to the eigenvalues raised to a non-negative power  $\varphi/2$ , under thresholding to reflect this preference. This is equivalent to applying a larger thresholding on higher principal components, with  $\varphi$  controlling the degree. Implementing this strategy, we propose the following more general family of estimators, parameterized by  $\varphi \geq 0$ :

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \hat{\boldsymbol{\Lambda}}^{-\varphi} T_\tau \left[ \hat{\boldsymbol{\Lambda}}^\varphi \tilde{\boldsymbol{\theta}}^{LS} \right]$$

in the canonical domain, or

$$(2.2) \quad \hat{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \hat{\mathbf{U}} \hat{\boldsymbol{\Lambda}}^{-1-\varphi} T_\tau \left[ \hat{\boldsymbol{\Lambda}}^{-1+\varphi} \hat{\mathbf{U}}^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n} \right]$$

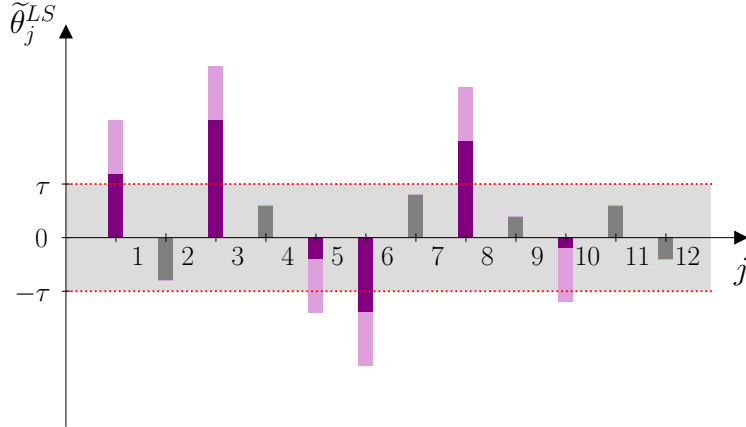
in the original domain. Here  $T_\tau[\mathbf{z}]$  is a generalized thresholding function from Definition 2.2 applied component-wise and  $\tau \geq 0$  is a hyperparameter to be chosen. The estimators from this family are called the *Generalized Canonical Thresholding* (GCT) estimators.

When  $\varphi = 0$  and the soft thresholding function is used, GCT reduces to the NCT estimator (2.1). The intuition behind GCT is somewhat similar to NCT: the estimators automatically screen the most significant principal components. However, the choice of  $\varphi$  allows to put different importance to eigenvalue  $\hat{\lambda}_j$  and projection  $\hat{\mathbf{u}}_j^\top \boldsymbol{\beta}$  when deciding whether to threshold  $j$ -th principal component or not. While in NCT this importance is calibrated in accordance to the scaling appearing in the decomposition of the absolute MSE (this justifies the word ‘‘Natural’’ in the name), GCT with  $\varphi > 0$  gives more weight to the leading canonical coefficients, making the method closer to PCR, and selects other components when absolutely necessary.

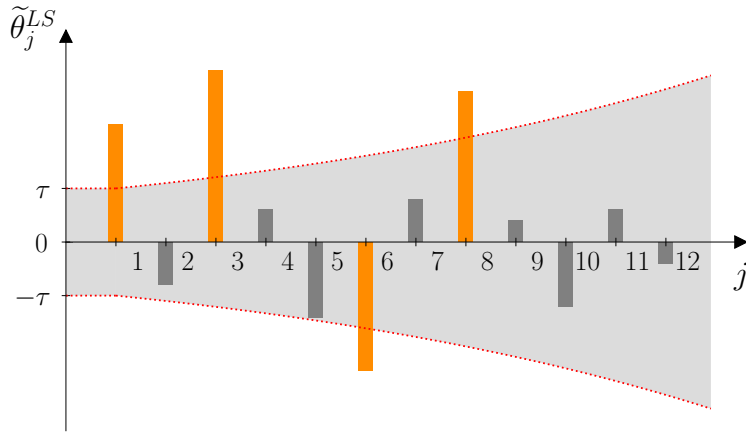
There is one common situation where PCR is preferable: pervasive latent factors that drives both the covariates and the response (Fan, Wang and Yao, 2017). In this case, the principal components are used to learn latent factors and these learned factors are used as the covariates for regressing the response  $y$ . This leads to PCR. We introduce GCT to better accommodate this situation. Figure 1 visually illustrates the conceptual difference between NCT, GCT, and PCR approaches. We highlight once again that the selection of the components in NCT and GCT is data-driven, unlike in PCR where the selected components are fixed before the data are observed (modulo cross-validation, which helps to select the number of leading components, but not the components itself).

It turns out that the theoretical results for GCT are not that nice and insightful as for NCT. However, in practice we observed that GCT may behave much better in some scenarios, as to be shown in the corresponding section. In principle, one can even tune  $\varphi$  in addition to tuning  $\tau$  via cross-validation, which might further enhance the practical utility of the procedure.

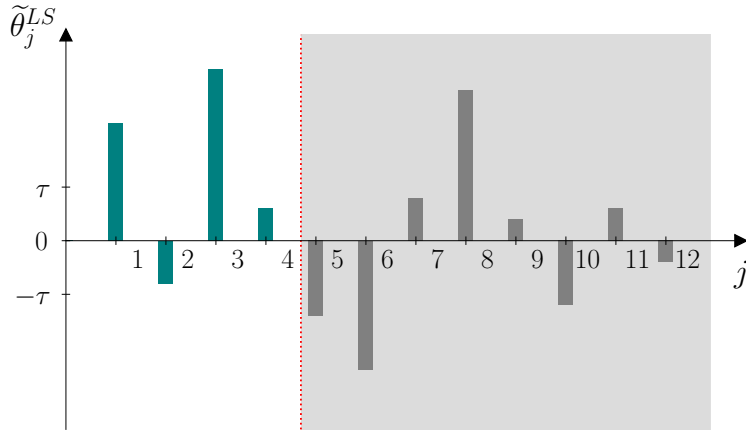




(a) Natural Canonical Thresholding: light purple indicates the portion of the active coefficients subtracted during soft thresholding, and purple depicts the remaining surviving coefficients;



(b) Generalized Canonical Thresholding with the hard thresholding rule,  $\varphi = 1$ , and assumed eigenvalue decay  $\hat{\lambda}_j = 1.21^{(-j+1)}$ ;



(c) Principal Component Regression with  $m = 4$ ;

Fig 1: Comparison of NCT, GCT, and PCR estimators on an artificial example with  $r = 12$ . On the horizontal axes – an index of the component  $j$ , on the vertical axes – the coefficient of the canonical least squares  $\tilde{\theta}_j^{LS}$ . The red dotted lines depict the thresholding boundaries. The coefficients falling into the shaded area are thresholded/truncated to zero and depicted in gray. The coefficients surviving the thresholding/truncation are depicted in purple, orange, and teal, respectively.

**3. Theoretical properties of the NCT estimator.** The first condition needed for our theoretical analysis is the following assumption on the noise, which will be used in both fixed design and random design settings.

**ASSUMPTION 3.1 (Sub-Weibull noise).** *The noise vector  $\varepsilon$  is independent of  $\mathbf{X}$  and is jointly sub-Weibull random vector with parameter  $0 < \alpha \leq 2$  (see [Kuchibhotla and Chakraborty \(2018\)](#)). That is, there exists  $\sigma < \infty$  such that*

$$\sup_{\|\mathbf{w}\|_2=1} \|\mathbf{w}^\top \varepsilon\|_{\psi_\alpha} \leq \sigma,$$

where  $\|\cdot\|_{\psi_\alpha}$  is the Orlicz norm for  $\psi_\alpha = e^{x^\alpha} - 1$ . The following tail bound takes place:

$$\mathbb{P} \left[ |\mathbf{w}^\top \varepsilon| \geq t \right] \leq 2 \exp \left( - (t/\sigma)^\alpha \right)$$

for all  $\mathbf{w}$ ,  $\|\mathbf{w}\|_2 = 1$  and  $t > 0$ .

This allows to go slightly beyond sub-Gaussian and sub-Exponential tails. For i.i.d. sub-Gaussian noise,  $\sigma^2$  coincides (up to a multiplicative constant) with the variance of a single  $\varepsilon_i$ . So,  $\sigma$  can be interpreted as the magnitude of the noise.

Define the *signal-to-noise ratio* of the linear regression problem in the fixed design setting as

$$\text{SNR} \stackrel{\text{def}}{=} \left( \frac{n^{-1} \sum_{i=1}^n (\mathbf{x}_i^\top \boldsymbol{\beta})^2}{\sigma^2} \right)^{1/2} = \left( \frac{\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}}{\sigma^2} \right)^{1/2} = \frac{\|\boldsymbol{\theta}\|_2}{\sigma},$$

while for the random design we use

$$\text{SNR} \stackrel{\text{def}}{=} \left( \frac{\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\beta})^2]}{\sigma^2} \right)^{1/2} = \left( \frac{\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}}{\sigma^2} \right)^{1/2} = \frac{\|\mathbf{A} \mathbf{U}^\top \boldsymbol{\beta}\|_2}{\sigma}.$$

For our problem to be meaningful, we assume for the rest of the work that  $\text{SNR} > 0$  in both settings.

Recall the canonical regression coefficients  $\boldsymbol{\theta} = \widehat{\mathbf{A}} \widehat{\mathbf{U}}^\top \boldsymbol{\beta}$  for the fixed design. Its normalized version  $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$  has the first  $k$  components  $\boldsymbol{\theta}_{\leq k}/\|\boldsymbol{\theta}\|_2$  where  $\boldsymbol{\theta}_{\leq k} = \widehat{\mathbf{A}}_{\leq k} \widehat{\mathbf{U}}_{\leq k}^\top \boldsymbol{\beta}$ . Here as usual  $\widehat{\mathbf{A}}_{\leq k} \in \mathbb{R}^{k \times k}$  is the leading principal submatrix of  $\widehat{\mathbf{A}}$  (containing the square roots of the first  $k$  eigenvalues of  $\widehat{\boldsymbol{\Sigma}}$  on the diagonal) and  $\widehat{\mathbf{U}}_{\leq k} \in \mathbb{R}^{d \times k}$  is the matrix consisting of the first  $k$  columns of  $\widehat{\mathbf{U}}$  (which are the  $k$  leading eigenvectors of  $\widehat{\boldsymbol{\Sigma}}$ ). Measuring the first  $k$  normalized components  $\boldsymbol{\theta}_{\leq k}/\|\boldsymbol{\theta}\|_2$  in  $\ell_q$ -(pseudo)norm gives

$$\text{D}_{q,k}^{\text{eff}}(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\beta}) \stackrel{\text{def}}{=} \frac{\|\boldsymbol{\theta}_{\leq k}\|_q^q}{\|\boldsymbol{\theta}\|_2^q}.$$

We call this quantity the *joint effective dimension of order  $q$  up to index  $k$*  of  $\widehat{\boldsymbol{\Sigma}}$  and  $\boldsymbol{\beta}$ . Note that when  $q = 2$ , it measures the proportion of  $\boldsymbol{\theta}$  explained by  $\boldsymbol{\theta}_{\leq k}$ ; when  $q = 0$ , it counts the sparsity among  $\boldsymbol{\theta}_{\leq k}$ . Similar quantity can be defined for the random design setting:

$$\text{D}_{q,k}^{\text{eff}}(\boldsymbol{\Sigma}, \boldsymbol{\beta}) \stackrel{\text{def}}{=} \frac{\|\mathbf{A}_{\leq k} \mathbf{U}_{\leq k}^\top \boldsymbol{\beta}\|_q^q}{\|\mathbf{A} \mathbf{U}^\top \boldsymbol{\beta}\|_2^q}.$$

It turns out that this joint effective dimension will play crucial role in our bounds for the NCT estimator (2.1).

For shortness, we introduce the following quantity that will be appearing regularly throughout the section:

$$(3.1) \quad \rho \stackrel{\text{def}}{=} \frac{2}{\sqrt{n}} (\log(2r/\delta))^{1/\alpha},$$

where  $\delta$  is from the statements “with probability  $1 - \delta$ ...”. The thresholding level  $\tau$  for both NCT and GCT will be expressed in terms of  $\rho$ .

**3.1. Fixed design.** We first provide a simple guarantee on the mean squared error  $\text{MSE}(\hat{\beta})$  of the NCT estimator (2.1).

**THEOREM 3.1.** *Suppose Assumption 3.1 is satisfied. Take  $\tau = \sigma\rho$  with  $\rho$  given by (3.1). Then, with probability  $1 - \delta$ , the NCT estimator  $\hat{\beta}$  from (2.1) with thresholding at level  $\tau$  satisfies*

$$(3.2) \quad \text{MSE}(\hat{\beta}) \asymp \sum_{j=1}^r \min(\sigma\rho, |\theta_j|)^2$$

$$(3.3) \quad \leq \inf_{q \in [0,2]} \{ \|\boldsymbol{\theta}\|_q^q (\sigma\rho)^{2-q} \}$$

$$(3.4) \quad = \text{MSE}(0) \inf_{q \in [0,2]} \left\{ D_{q,r}^{\text{eff}}(\hat{\Sigma}, \beta) \left[ \text{SNR}^{-2} \frac{(\log(2r/\delta))^{2/\alpha}}{n} \right]^{1-q/2} \right\}.$$

The proof of this result almost repeats the classical proof for hard and soft thresholding in case of orthonormal design.

**REMARK 3.1 (Choice of  $\tau$ ).** *The choice of  $\tau$  in the above theorem depends on the noise magnitude  $\sigma$ , the probability  $\delta$ , and the quantity  $\alpha$ , but this is not a significant problem. Later in Section 6 we will show how to tune  $\tau$  using an efficient cross-validation procedure.*

In Theorem 3.1 we present several bounds on  $\text{MSE}(\hat{\beta})$ . Note that the bound (3.2) is tight. One can argue that the matching lower bound is due to the excessive bias introduced by soft thresholding. However, even if we replace soft thresholding with hard thresholding  $\text{HARD}_\tau[z] \stackrel{\text{def}}{=} z \cdot \mathbb{1}\{|z| \geq \tau\}$ , the upper bound in probability stays the same as in (3.2), and at the same time it is not difficult to get a nearly matching lower bound (for Gaussian noise  $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$ ) in expectation:

$$\mathbb{E} [\text{MSE}(\hat{\beta})] \gtrsim \sum_{j=1}^r \min \left( \frac{\sigma}{\sqrt{n}}, |\theta_j| \right)^2.$$

In any case, the bound (3.2) is not really interpretable. The bounds (3.3) and (3.4) have much deeper intuition as we will see later. Nevertheless, a reasonable question is how tight the inequality leading from (3.2) to (3.3) is. Though in general the opposite inequality (up to a multiplicative constant) does not seem to hold, the following proposition states that in several important cases the inequality is actually tight (or almost tight).

**PROPOSITION 3.2.** *(i) Sparsity. Assume  $\|\boldsymbol{\theta}\|_0 = s \ll r$  with  $|\theta_j| \gtrsim \sigma\rho$  for all  $j$  such that  $\theta_j \neq 0$ . Then (taking  $q = 0$ )*

$$\sum_{j=1}^r \min(\sigma\rho, |\theta_j|)^2 \gtrsim \inf_{q \in [0,2]} \{ \|\boldsymbol{\theta}\|_q^q (\sigma\rho)^{2-q} \}.$$

(ii) *Approximate sparsity.* Assume there exists a small  $\nu > 0$  such that  $\|\boldsymbol{\theta}\|_\nu \lesssim \sigma$ , and there are at least  $s$  significant canonical coefficients:  $|\{j \in [r] : |\theta_j| \gtrsim \sigma\rho\}| \geq s$ . Then (taking  $q = \nu$ )

$$\sum_{j=1}^r \min(\sigma\rho, |\theta_j|)^2 \gtrsim \inf_{q \in [0,2]} \{ \|\boldsymbol{\theta}\|_q^q (\sigma\rho)^{2-q} \} \times s\rho^\nu.$$

(iii) *Polynomial decay.* Let  $|\theta_{(1)}| \geq |\theta_{(2)}| \geq \dots \geq |\theta_{(r)}|$  be the absolute values of the components of  $\boldsymbol{\theta}$  arranged in descending order. Assume polynomial decay of the ordered canonical coefficients: for some  $a > 0$  holds

$$\frac{|\theta_{(j)}|}{|\theta_{(1)}|} \asymp j^{-a} \quad \text{for all } j \in [r].$$

Then (taking  $q = \min(1/a, 2)$ )

$$\sum_{j=1}^r \min(\sigma\rho, |\theta_j|)^2 \gtrsim \inf_{q \in [0,2]} \{ \|\boldsymbol{\theta}\|_q^q (\sigma\rho)^{2-q} \} \times \begin{cases} 1/\log(r), & a \geq 1/2, \\ 1, & a < 1/2. \end{cases}$$

Continuing the discussion on the bounds presented in Theorem 3.1, we claim that though (3.3) and (3.4) coincide, the way we state them reflects two different messages. The first one, if we take  $q = 1$  and apply the inequality  $\|\boldsymbol{\theta}\|_1 \leq \|\widehat{\boldsymbol{\Sigma}}\|^{1/2} \|\boldsymbol{\beta}\|_2 r^{\text{eff}}[\widehat{\boldsymbol{\Sigma}}]^{1/2}$  (which follows from the Cauchy-Schwarz inequality), then from the bound (3.3) we get

$$(3.5) \quad \text{MSE}(\widehat{\boldsymbol{\beta}}) \lesssim \sigma \|\widehat{\boldsymbol{\Sigma}}\|^{1/2} \|\boldsymbol{\beta}\|_2 \sqrt{\frac{r^{\text{eff}}[\widehat{\boldsymbol{\Sigma}}] (\log(2r/\delta))^{2/\alpha}}{n}}$$

with high probability. This means that if one is interested in the absolute error  $\text{MSE}(\widehat{\boldsymbol{\beta}})$ , then essentially  $r^{\text{eff}}[\widehat{\boldsymbol{\Sigma}}]/n = o(1)$  is enough to guarantee  $\text{MSE}(\widehat{\boldsymbol{\beta}}) = \sigma \|\widehat{\boldsymbol{\Sigma}}\|^{1/2} \|\boldsymbol{\beta}\|_2 \cdot o(1)$ , omitting logarithmic terms. No additional assumptions on  $\boldsymbol{\beta}$  are required, and there is no necessity to worry about the joint effective dimension and the signal-to-noise ratio from the bound (3.4) in this situation. Note that the bound (3.5) is tight up to a logarithmic factor when  $\widehat{\lambda}_j = \widehat{\lambda}_1/j$  and  $|\widehat{\mathbf{u}}_j^\top \boldsymbol{\beta}| \asymp \|\boldsymbol{\beta}\|_2 / (\sqrt{j \log(r)})$ .

Nevertheless, the bound (3.4) is useful to better understand the structure of the error. Taking into account that the main motivation in our work is the decay of eigenvalues of  $\widehat{\boldsymbol{\Sigma}}$  or  $\boldsymbol{\Sigma}$ , it may easily be the case that even the trivial estimator  $\widetilde{\boldsymbol{\beta}} = 0$  has a very small error  $\text{MSE}(0) = \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}$ . Hence, it makes sense to care more about the relative error  $\text{MSE}(\widetilde{\boldsymbol{\beta}})/\text{MSE}(0)$ . In this case, the joint effective dimension and the signal-to-noise ratio control the upper bound on the relative error. We will get back to the analysis of the relative error and the joint effective dimension after we state an analogous result for the random design case.

**3.2. Random design.** In addition to the noise assumption, to study the performance of the NCT estimator in the random design setting we need to impose a couple more conditions on the distribution of the covariates.

**ASSUMPTION 3.2 (Sub-Gaussian covariates).** *The scaled generic random vector of covariates  $\boldsymbol{\Sigma}^{-1/2} \mathbf{x}$  is sub-Gaussian.*

ASSUMPTION 3.3 (Convex decay of eigenvalues). *There exists a convex decreasing function  $\lambda(\cdot)$  such that the eigenvalues of  $\Sigma$  satisfy  $\lambda_j = \lambda(j)$  for  $j \in [d]$ .*

The previous assumption is technical and we impose it in our main result just for concreteness. Later in Remark 3.9 we mention how our result can be modified if this assumption does not hold.

One more assumption is needed just to make the rates more friendly-looking. If it is not satisfied, our result below will be meaningless, so there is no loss of generality in this condition.

ASSUMPTION 3.4 (Technical conditions). *The effective rank satisfies  $r^{\text{eff}}[\Sigma] \leq n$ . Also, whenever we say “with probability  $1 - \delta$ ...”, we suppose that the quantity*

$$\epsilon \stackrel{\text{def}}{=} \sqrt{\frac{\log(d/\delta)}{n}}$$

*satisfies  $\epsilon \leq c$  for properly chosen implicit absolute constant  $c > 0$  (this constant comes from the proof).*

In addition to the assumptions above, in the sequel we take the convention  $\lambda_k = 0$  for all  $k > d$ . Now we are ready to present the following result.

THEOREM 3.3. *Suppose Assumptions (3.1) – (3.4) are fulfilled. Recall  $\rho$  from (3.1),  $\epsilon \stackrel{\text{def}}{=} \sqrt{\log(d/\delta)/n}$ , and define  $k^* \stackrel{\text{def}}{=} (\epsilon \log(1/\epsilon))^{-2/3}$  (essentially,  $k^* \asymp n^{1/3}$  up to a logarithmic term). Then:*

(i) *With probability  $1 - \delta$ , the NCT estimator  $\hat{\beta}$  from (2.1) with thresholding at level*

$$\tau = \sigma\rho$$

*satisfies*

$$\begin{aligned} \text{PE}(\hat{\beta}) &\lesssim \inf_{q \in [0, 2]} \left\{ \|\Lambda_{\leq k} \mathbf{U}_{\leq k}^\top \beta\|_q^q (\sigma\rho)^{2-q} \right\} + \\ &\quad + \|\Sigma\| \|\beta\|_2^2 \left( \frac{\lambda_k}{\lambda_1} + \sqrt{\frac{r^{\text{eff}}[\Sigma] + \log(1/\delta)}{n}} + \epsilon \sum_{j=1}^k \frac{\lambda_j(1 + \epsilon j^2)}{\lambda_1} \right) \end{aligned}$$

*for all  $1 \leq k \leq k^*$ .*

(ii) *With probability  $1 - \delta$ , the NCT estimator  $\hat{\beta}$  from (2.1) with thresholding at level*

$$\bar{\tau} = \sigma\rho + C \|\Sigma\|^{1/2} \|\beta\|_2 \epsilon^{1/2} \max_{j \in [k]} \left( \frac{\lambda_j(1 + \epsilon j^2)}{\lambda_1} \right)^{1/2}$$

*for some  $C$  satisfies*

$$\text{PE}(\hat{\beta}) \lesssim \inf_{q \in [0, 2]} \left\{ \|\Lambda_{\leq k} \mathbf{U}_{\leq k}^\top \beta\|_q^q \bar{\tau}^{2-q} \right\} + \|\Sigma\| \|\beta\|_2^2 \left( \frac{\lambda_k}{\lambda_1} + \sqrt{\frac{r^{\text{eff}}[\Sigma] + \log(1/\delta)}{n}} \right)$$

*for all  $1 \leq k \leq k^*$ .*

Several comments are in order.

REMARK 3.2 (Why two bounds?). We present two separate bounds in (i) and (ii) for distinct thresholds  $\tau$  and  $\bar{\tau}$ , because they behave differently for various eigenvalue regimes. The bound from (i) outperforms the bound from (ii) in a wide variety of settings (e.g. in polynomial and superpolynomial decay scenario), however there are cases when the bound (ii) can be better (e.g. specific cases in factor model regime).

REMARK 3.3 (Meaning of terms). We call the term  $\inf_{q \in [0,2]} \{\dots\}$  in the bounds of Theorem 3.3 (i) and (ii) the “main term”, since it is almost the same as what we had in Theorem 3.1 for the fixed design. The other terms in these bounds are referred to as “additional terms” as they appear only in the random design case. Allowing  $1 \leq k \leq k^*$  provides a tradeoff, as some of the terms increase with growing  $k$ , while others decrease. In what follows we are typically interested in  $k = k^*$  just for concreteness. The meaning of different parts of the “additional terms” is the following. The parts including  $r^{\text{eff}}[\Sigma]$  are the payment for the covariance matrix estimation. The part  $\lambda_{k^*}/\lambda_1$  appears due to the difficulty of control of the empirical eigenvectors beyond  $k^*$ -th. The parts with  $\sum_{j=1}^{k^*} \lambda_j (1 + \epsilon j^2)/\lambda_1$  and  $\max_{j \in [k^*]} \lambda_j (1 + \epsilon j^2)/\lambda_1$  are the payment for the control of the sample eigenvalues and eigenvectors up to index  $k^*$ .

REMARK 3.4 (Moderate noise: simplifications and sufficient conditions for convergence). Consider the moderate noise situation  $\sigma^2 \lesssim \|\Sigma\| \|\beta\|_2^2$ . In this case the “additional terms” become dominating: simply taking  $k = k^*$  and  $q = 1$ , applying  $\|\Lambda_{\leq k} \mathbf{U}_{\leq k}^\top \beta\|_1 \leq \|\Sigma\|^{1/2} \|\beta\|_2 r^{\text{eff}}[\Sigma]^{1/2}$  and plugging in  $\tau$  and  $\bar{\tau}$  makes the “main term” negligible. Omitting logarithmic terms, the bound (i) reduces to

$$(3.6) \quad \text{PE}(\hat{\beta}) \lesssim \|\Sigma\| \|\beta\|_2^2 \left( \frac{\lambda_{k^*}}{\lambda_1} + \sqrt{\frac{r^{\text{eff}}[\Sigma]}{n}} + \epsilon \sum_{j=1}^{k^*} \frac{\lambda_j (1 + \epsilon j^2)}{\lambda_1} \right),$$

while the bound (ii) reduces to

$$\text{PE}(\hat{\beta}) \lesssim \|\Sigma\| \|\beta\|_2^2 \left( \frac{\lambda_{k^*}}{\lambda_1} + \left( \frac{r^{\text{eff}}[\Sigma]^2}{n} \right)^{1/4} \max_{j \in [k^*]} \left( \frac{\lambda_j (1 + j^2/\sqrt{n})}{\lambda_1} \right)^{1/2} \right)$$

with high probability.

From here we can easily deduce sufficient conditions to ensure the convergence of the absolute error  $\text{PE}(\hat{\beta}) = \|\Sigma\| \|\beta\|_2^2 \cdot o(1)$  as  $n \rightarrow \infty$  without any conditions on  $\beta$ . In particular,  $\lambda_n = o(1)$ ,  $r^{\text{eff}}[\Sigma] = o(n)$ , and  $\sum_{j=1}^{k^*} \frac{\lambda_j (1 + j^2/\sqrt{n})}{\lambda_1} = o(n^{1/2})$  is enough (again, up to logarithmic factors). Essentially, these sufficient conditions require the decay of eigenvalues to be fast enough (but in a more sophisticated fashion than for MSE).

REMARK 3.5 (Moderate noise: further simplifications in specific examples). Continuing the moderate noise situation, for the sake of exposition, let us consider a couple of specific examples of eigenvalue regimes and illustrate how the bound from Theorem 3.3 (i) simplifies. As above, we omit logarithmic terms.

- *Polynomial decay.* If  $\lambda_j \lesssim j^{-a}$  with  $a \geq 1$  or  $d \lesssim n^{(3-2a)/(3-3a)}$  with  $a < 1$ , it is easy to verify that the bracket factor of (3.6) is dominated by  $\lambda_{k^*}/\lambda_1 + n^{-1/2}$  (again ignoring logarithmic terms), and with high probability

$$\text{PE}(\hat{\beta}) \lesssim \frac{\|\Sigma\| \|\beta\|_2^2}{n^{\min(a/3, 1/2)}}.$$



In particular, when  $a = 1$  (the boundary that has a good control of  $r^{\text{eff}}[\Sigma]$  in high dimensions), we have with high probability

$$\text{PE}(\hat{\beta}) \lesssim \frac{\|\Sigma\| \|\beta\|_2^2}{n^{1/3}}.$$

When  $a \geq 3/2$ , we have with high probability

$$\text{PE}(\hat{\beta}) \lesssim \frac{\|\Sigma\| \|\beta\|_2^2}{n^{1/2}}.$$

- *Factor model regime.* If  $\lambda_1 \asymp \dots \asymp \lambda_m \asymp d$ ,  $\lambda_{m+1} \asymp \dots \asymp \lambda_d \asymp 1$  for some  $m \lesssim k^*$ , then taking  $k = m + 1$  yields with high probability

$$\text{PE}(\hat{\beta}) \lesssim \|\Sigma\| \|\beta\|_2^2 \left( \frac{1}{d} + \frac{m}{\sqrt{n}} + \frac{m^3}{n} \right).$$

REMARK 3.6 (Large noise). In the large noise case, when  $\sigma^2 \gg \|\Sigma\| \|\beta\|_2^2$ , the “main term” dominates. Similarly to the fixed design case, we can factorize the “main term” into the error of the trivial estimator  $\text{PE}(0) = \beta^\top \Sigma \beta$ , the joint effective dimension and the signal-to-noise ratio: the “main term” from (i) becomes

$$\text{PE}(0) \inf_{q \in [0,2]} \left\{ D_{q,k^*}^{\text{eff}}(\Sigma, \beta) \left[ \text{SNR}^{-2} \frac{(\log(2d/\delta))^{2/\alpha}}{n} \right]^{1-q/2} \right\},$$

and the “main term” from (ii) can be rewritten as

$$\begin{aligned} \text{PE}(0) \inf_{q \in [0,2]} \left\{ D_{q,k^*}^{\text{eff}}(\Sigma, \beta) \left[ \text{SNR}^{-2} \frac{(\log(2d/\delta))^{2/\alpha}}{n} + \right. \right. \\ \left. \left. + \frac{\|\Sigma\| \|\beta\|_2^2}{\beta^\top \Sigma \beta} \epsilon \max_{j \in [k^*]} \left( \frac{\lambda_j (1 + \epsilon j^2)}{\lambda_1} \right) \right]^{1-q/2} \right\}. \end{aligned}$$

In this regime, the relative error  $\text{PE}(\hat{\beta})/\text{PE}(0)$  is essentially controlled by the joint effective dimension and the signal-to-noise ratio. More detailed analysis of the joint effective dimension  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$  is conducted in the next section.

REMARK 3.7 (Comparison with the least squares). It is straightforward to notice that the faster decay of eigenvalues, the better bound we obtain. This contrasts the min-norm least squares estimator considered in [Bartlett et al. \(2020\)](#); [Chinot and Lerasle \(2020\)](#), where the decay is required to be not too fast. It also reveals the benefits of the thresholding even in such a situation.

REMARK 3.8 (Dependence on the dimension  $d$ ). One can see that while the bound of Theorem 3.1 contains  $\log(r)$  term with  $r = \min(d, n)$  and gives meaningful result even in ultra-high or infinite dimension, the results of Theorem 3.3 contain the factor  $\log(d)$  directly. This logarithmic factor appears through the definition of  $\epsilon$ , which is an upper bound in Lemma S.2.3 of the Supplementary Material [Silin and Fan \(2021\)](#): with probability  $1 - \delta$

$$\max_{l, l' \in [d]} \left| (\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - \mathbb{I}_d)_{l, l'} \right| \leq \epsilon.$$

This condition is crucial for application of the relative perturbation bounds from [Jirak and Wahl \(2018\)](#), which are the foundation of our proof technique for Theorem 3.3. If we could choose  $\epsilon$  in a completely dimension-free way to satisfy the above, it would yield totally dimension-free bounds in Theorem 3.3, but currently applying the union bound inevitably brings  $\log(d)$  factor.

REMARK 3.9 (Relaxing Assumption 3.3). *Assumption 3.3 can be avoided. If one defines  $k^*$  as*

$$k^* = \min \left( (\epsilon \log(1/\epsilon))^{-2/3}, \max \left\{ j \in [d] \left| \sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_l}{|\lambda_j - \lambda_l|} + \frac{\lambda_j}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} \leq \frac{1}{3\epsilon} \right\} \right), \right)$$

*then the same conclusion as in Theorem 3.3 is true with a slightly different rate.*

To compare the natural canonical thresholding with the canonical truncation of higher index components, i.e. PCR, we state the next proposition.

PROPOSITION 3.4. *Assume the conditions of Theorem 3.3 hold and let  $k^*$  be defined in the same way. Then, with probability  $1 - \delta$ , the PCR estimator  $\tilde{\beta}^{PCR}$  with the number of leading principal components set to  $m \lesssim k^*$  satisfies*

$$\text{PE}(\tilde{\beta}^{PCR}) \lesssim \|\Sigma\| \|\beta\|_2^2 \left( \frac{\lambda_m}{\lambda_1} + \sqrt{\frac{r^{\text{eff}}[\Sigma] + \log(1/\delta)}{n}} \right) + \frac{\sigma^2 m}{n} (\log(2m/\delta))^{1/\alpha}.$$

We omit the proof of this result, since it essentially uses the same techniques and follows the same strategy as the proof of Theorem 3.3. Note that in the moderate noise scenario the rate essentially coincides with what we obtained for the NCT estimator estimator in Remark 3.4. The adaptivity of our estimator  $\hat{\beta}$  comes into play in large noise case: the “main term” in the bounds on  $\text{PE}(\hat{\beta})$  is better than  $\sigma^2 m/n$  in situations when  $\mathbf{U}_{\leq k}^\top \beta$  is approximately sparse.

3.3. *Relative errors and joint effective dimension.* So far we were able to establish some sufficient conditions for convergence of absolute errors  $\text{MSE}(\hat{\beta})$  and  $\text{PE}(\hat{\beta})$  of the NCT estimator without any assumptions on  $\beta$  by simply taking  $q = 1$  (bound (3.5) and Remark 3.4). The analysis of the relative errors for fixed design and (in large noise case) random design requires more careful study of  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$ . Let us motivate why relative errors  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  and  $\text{PE}(\hat{\beta})/\text{PE}(0)$  might be of interest in the first place.

3.3.1. *Motivation for relative errors.* One reason behind studying the relative errors was already mentioned previously. Note that if there is no relation between  $\hat{\Sigma}$  and  $\beta$ , meaning that  $\hat{\mathbf{U}}^\top \beta$  is a “random” vector, then we can expect  $\|\hat{\mathbf{U}}^\top \beta\|_\infty \asymp \|\beta\|_2 \sqrt{\log(d)/d}$ . In this case, the trivial estimator  $\tilde{\beta} = 0$  achieves error

$$\text{MSE}(0) = \beta^\top \hat{\Sigma} \beta \leq \text{Tr}[\hat{\Sigma}] \|\hat{\mathbf{U}}^\top \beta\|_\infty^2 \asymp \|\hat{\Sigma}\| \|\beta\|_2^2 \frac{r^{\text{eff}}[\hat{\Sigma}] \log(d)}{d}.$$

As long as the eigenvalues of  $\hat{\Sigma}$  decay fast, even the trivial estimator gives error close to zero in high dimensions. Here we should highlight that this effect does not appear in low dimensions (and even in high-dimensional but isotropic situations), where the absolute and relative errors are just a multiplicative constant apart. (Same reasoning works for the PE.) Hence, it is not satisfactory for us to show that the absolute error of our estimator goes to zero with growing sample size and dimension. We would like to get more meaningful conclusions from our results, which would confirm that the proposed estimator does better than the trivial estimator. This naturally leads to the relative errors.

Another motivation comes from the way statisticians evaluate and compare estimators in practical applications. In particular, a widely used performance measure is the *coefficient*

of determination, or simply  $R^2$ . For instance, the in-sample version for an estimator  $\tilde{\beta}$  is defined as

$$R_{in}^2(\tilde{\beta}) \stackrel{\text{def}}{=} 1 - \frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \tilde{\beta})^2}{\sum_{i=1}^n y_i^2}.$$

The larger this quantity is, the better method we have; its largest possible value is 1, and the value of 0 indicates that the estimator does no better than the trivial estimator. Maximization of  $R_{in}^2(\tilde{\beta})$  would try to fit the observed data perfectly, and in this sense it is not equivalent to minimizing  $\text{MSE}(\tilde{\beta})/\text{MSE}(0)$ . Nevertheless, a crucial observation is that  $R_{in}^2(\tilde{\beta})$  is a relative quantity, which takes into account the performance of the trivial estimator. This supports our choice of the relative MSE as an error measure.

Similar intuition applies to the out-of-sample  $R_{out}^2$  and the relative prediction error  $\text{PE}(\tilde{\beta})/\text{PE}(0)$ , and intuitively it seems that they are linked even stronger. Note that in applications it is often the case that even small but positive  $R^2$  (e.g. 0.05) can be considered a success. Therefore, the hope to have  $\text{MSE}(\tilde{\beta})/\text{MSE}(0)$  or  $\text{PE}(\tilde{\beta})/\text{PE}(0)$  converging to 0 might be too optimistic in some situations. Having these relative errors smaller than 1 already means that the procedure is able to extract some useful signal from the data.

**3.3.2. Why joint conditions on design and regression coefficients?.** Prior to describing the properties of  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$ , let us show why imposing conditions on the design alone, or imposing conditions on  $\beta$  alone can be not enough to establish convergence of the relative errors. It is easier to do for the fixed design case, so let us focus on this setting for now.

For a given design matrix  $\mathbf{X}$  and an estimator  $\tilde{\beta}$  we can construct another estimator  $\tilde{\theta} = \hat{\Lambda} \hat{\mathbf{U}}^\top \tilde{\beta}$ . Therefore,

$$\frac{(\tilde{\beta} - \beta)^\top \hat{\Sigma} (\tilde{\beta} - \beta)}{\beta^\top \hat{\Sigma} \beta} = \frac{\|\tilde{\theta} - \theta\|_2^2}{\|\theta\|_2^2}.$$

The relative MSE in canonical domain has nothing to do with  $\hat{\Sigma}$ . This demonstrates that getting a good rate is hopeless in high dimension assuming only fast decay of eigenvalues of  $\hat{\Sigma}$ .

On the other hand, we might impose strong conditions on  $\beta$ , such as sparsity, in which case one could expect even  $\text{MSE}(\tilde{\beta}) \asymp \sigma^2 s/n$  (up to a logarithmic factor) for some appropriate estimator  $\tilde{\beta}$ , where  $s$  measures the degree of sparsity. However, as we mentioned previously, if  $\hat{\Sigma}$  and  $\beta$  are not related and the eigenvalues of  $\hat{\Sigma}$  decay fast, we might have  $\text{MSE}(0) \asymp \|\hat{\Sigma}\| \|\beta\|_2^2/d$  (again up to logarithmic factors) for the trivial estimator. This implies that there is no much hope in getting vanishing relative error  $\text{MSE}(\tilde{\beta})/\text{MSE}(0)$  in high dimensions. That is why it seems natural that our bound on the relative error depends on the joint effective dimension, that takes into account not only decay of eigenvalues or only assumptions on  $\beta$ , but the joint structure of  $\hat{\Sigma}$  and  $\beta$ .

**3.3.3. Joint effective dimension.** Now, once we supported the appearance of the joint effective dimension, let us mention its basic properties. (For concreteness we choose the random design setting and consider  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$ , though the following ideas apply to  $D_{q,r}^{\text{eff}}(\hat{\Sigma}, \beta)$  appearing in the fixed design case.)

- $D_{q,k}^{\text{eff}}(\Sigma, \beta) \leq k$ .
- $D_{q,k}^{\text{eff}}(\Sigma, \beta)$  is decreasing in  $q$  and increasing in  $k$ .
- $D_{2,k}^{\text{eff}}(\Sigma, \beta) \leq 1$ ,  $D_{2,d}^{\text{eff}}(\Sigma, \beta) = 1$ .

- $D_{0,k}^{\text{eff}}(\Sigma, \beta) = \|\Lambda_{\leq k} \mathbf{U}_{\leq k}^\top \beta\|_0$  is essentially the sparsity of  $\mathbf{U}_{\leq k}^\top \beta$ .

For now let us assume  $\text{SNR} \geq c > 0$  and focus on  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$  only. Recall that the “main term” in the relative bounds looks like

$$\inf_{q \in [0,2]} \left\{ D_{q,k}^{\text{eff}}(\Sigma, \beta) \zeta_n^{1-q/2} \right\}$$

where  $\zeta_n$  is some vanishing rate like  $1/n$ . Hence, the properties above reveal a tradeoff in the main term:

- When  $q$  is large, i.e. closer to 2, it is easier to control  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$ ; however, the vanishing rate  $\zeta_n$  is raised to a small power, making the convergence slow.
- When  $q$  is small, i.e. closer to 0, it is more difficult to control  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$ ; in contrast,  $\zeta_n$  is raised to a large power potentially enabling fast convergence rate.

So, the bound allows to find largest  $q$  for which  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$  can be bounded in dimension-free and sample size-free manner (or at least the dependence on  $d$  and  $n$  should not be that severe) to facilitate faster convergence rate.

Some scenarios where  $D_{q,k}^{\text{eff}}(\Sigma, \beta)$  can be bounded more explicitly (for some  $q < 2$ ) are discussed below:

- Sparsity of  $\mathbf{U}_{\leq k}^\top \beta$ . Denote  $s \stackrel{\text{def}}{=} \|\mathbf{U}_{\leq k}^\top \beta\|_0$  to be the sparsity level. Then, as already mentioned previously,  $D_{0,k}^{\text{eff}}(\Sigma, \beta) = s$ .
- Approximate sparsity of  $\Lambda_{\leq k} \mathbf{U}_{\leq k}^\top \beta$ . Suppose there exists a small set  $\mathcal{J} \subseteq [k]$  (of size  $|\mathcal{J}| = s$ ) of significant components, so that the rest of the components satisfy

$$\lambda_j^{1/2} |\mathbf{u}_j^\top \beta| \lesssim \frac{1}{d} \|\Lambda \mathbf{U}^\top \beta\|_2 \quad \text{for all } j \in [k] \setminus \mathcal{J}.$$

Then  $D_{1,k}^{\text{eff}}(\Sigma, \beta) \lesssim s$ .

- Polynomial decay. Let  $\lambda_j/\lambda_1 \asymp j^{-a}$ ,  $|\mathbf{u}_j^\top \beta|/|\mathbf{u}_1^\top \beta| \asymp j^{-b}$ , where  $a \geq 0$ . We have several cases:
  - If  $a + 2b \geq 1$  and  $q \geq 2/(a + 2b)$ , then  $D_{q,k}^{\text{eff}}(\Sigma, \beta) \asymp 1$ .
  - If  $a + 2b \geq 1$  and  $q < 2/(a + 2b)$ , then  $D_{q,k}^{\text{eff}}(\Sigma, \beta) \asymp k^{1-(a+2b)q/2}$ .
  - If  $a + 2b < 1$ , then  $D_{q,k}^{\text{eff}}(\Sigma, \beta) \asymp k^{1-(a+2b)q/2} / d^{(1-a-2b)q/2}$ .

Here we omitted logarithmic factors. To better understand the dependence of  $D_{q,d}^{\text{eff}}(\Sigma, \beta)$  on  $d$  in specific case  $k = d$ , in Figure 2 we depict this dependence in  $(a + 2b) - q$  axes. In the green region  $D_{q,d}^{\text{eff}}(\Sigma, \beta)$  does not grow with  $d$ , while in the yellow region  $D_{q,d}^{\text{eff}}(\Sigma, \beta)$  grows with  $d$  polynomially, and the contours of constant power are illustrated with different colors.

**3.3.4. Bounds on relative errors in polynomial decay scenario.** Since the “main term” conveniently decomposes into several factors, among which the most interesting one – the joint effective dimension – was discussed above, the analysis of the relative errors of the NCT estimator in the fixed design and (for large noise case) random design is pretty much complete. It is intriguing though, what happens to the bounds on PE in moderate noise case: recall that in this scenario the “main term” is absorbed by the “additional term”, which does not have a structure allowing a direct analysis of the bounds on the relative error  $\text{PE}(\hat{\beta})/\text{PE}(0)$ . It is not clear whether they can be stated in a way that will provide better understanding of the relative error. Instead, we can take a look at the particular case of polynomial decay of eigenvalues and regression coefficients:  $\lambda_j/\lambda_1 \asymp j^{-a}$ ,  $|\mathbf{u}_j^\top \beta|/|\mathbf{u}_1^\top \beta| \asymp j^{-b}$ . After tedious

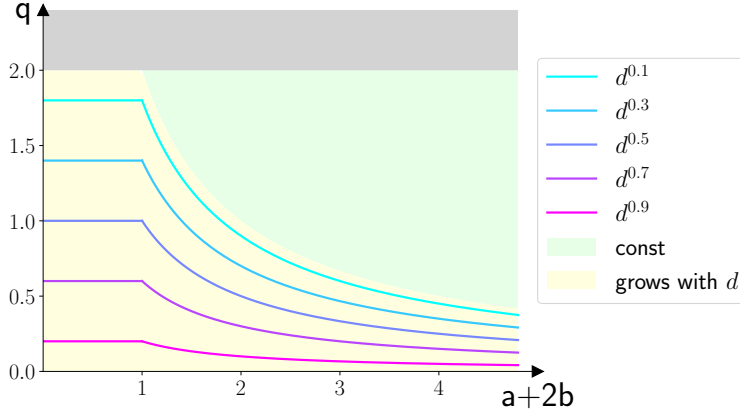


Fig 2: Dependence of  $D_{q,d}^{\text{eff}}(\Sigma, \beta)$  on  $d$ .

calculations, one may express the bounds from Theorem 3.1 and Theorem 3.3 (i), (ii) in terms of  $d, n, a, b$  only. It turns out, that in this scenario the bound from Theorem 3.3 (ii) is always worse than the bound from Theorem 3.3 (i), so we exclude it from consideration. In Figure 3 we plot the contours of constant convergence rate on  $a - b$  plane for the bounds from Theorem 3.1 and Theorem 3.3 (i). Different colors of the contours correspond to different rates. The background color describes the assumptions on  $d$  and  $n$  that we make in different regions: in light green zones  $d$  can be much larger than  $n$  (though this is not necessary), in light yellow zones  $d$  is allowed to be at most of the same order as  $n$ , i.e.  $d \lesssim n$ , and in grey zone the rates do not go to zero unless  $d$  is significantly smaller than  $n$ . We again disregard the logarithmic terms.

**4. (Near) minimax optimality of the NCT estimator.** This section focuses on the fixed design setting. In light of the above discussion, we introduce the following parameter classes for the fixed design linear regression problem (with Gaussian noise, for simplicity): for any design matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$

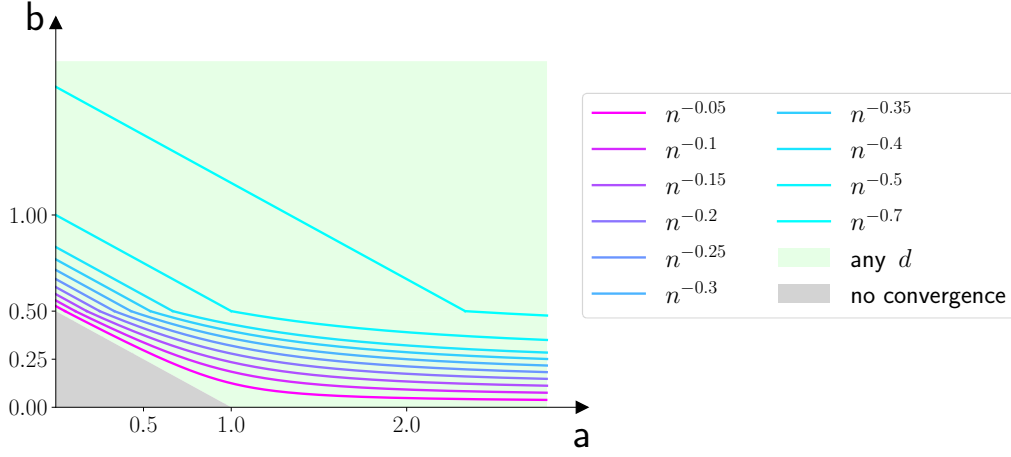
$$\mathcal{P}^{\mathbf{X}}(q, D, S) \stackrel{\text{def}}{=} \left\{ (\beta, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+ : D_{q,r}^{\text{eff}}(\widehat{\Sigma}, \beta) \leq D, \text{SNR} \geq S \right\}.$$

Clearly, we have the following relations:

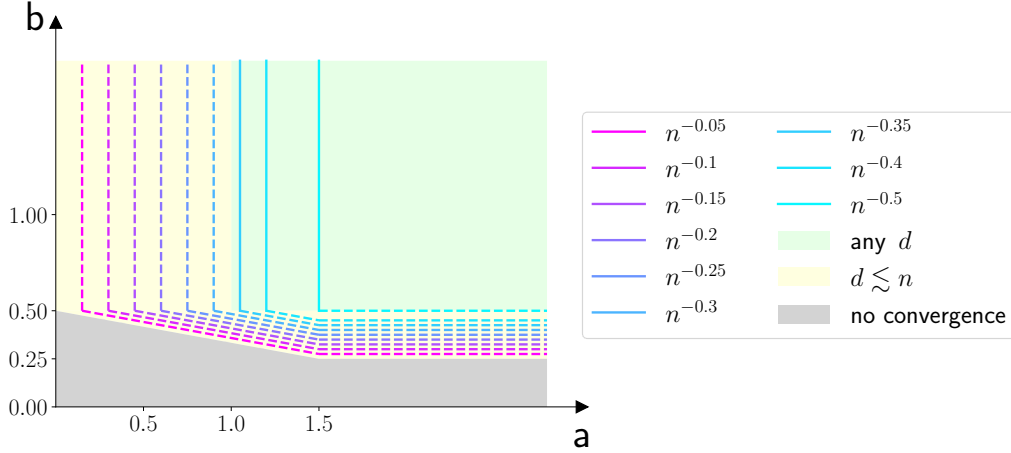
$$\begin{aligned} \mathcal{P}^{\mathbf{X}}(q, D, S) &\subset \mathcal{P}^{\mathbf{X}}(q', D, S) \quad \text{for } q < q', \\ \mathcal{P}^{\mathbf{X}}(q, D, S) &\subset \mathcal{P}^{\mathbf{X}}(q, D', S) \quad \text{for } D < D', \\ \mathcal{P}^{\mathbf{X}}(q, D, S) &\subset \mathcal{P}^{\mathbf{X}}(q, D, S') \quad \text{for } S > S'. \end{aligned}$$

In a sense, this orders instances of problems by difficulty. Note that for fixed  $d, n$  the family of classes is parameterized by three quantities:  $q$ ,  $D$  and  $S$ . While  $S$  is independent of the other two (since it is the only quantity related to the magnitude of noise),  $q$  and  $D$  bring some ambiguity. More specifically, if a specific instance of a problem belongs to  $\mathcal{P}^{\mathbf{X}}(q, D, S)$ , it also belongs to  $\mathcal{P}^{\mathbf{X}}(q - \Delta q, D \exp(\Delta \log D), S)$  for some perturbations  $\Delta q$  and  $\Delta \log D$ . To reduce this indeterminacy, we restrict  $D$  to be at most of logarithmic order when  $q > 0$ , i.e.  $D \lesssim \log(r)$ . Hence, just two quantities essentially control a “complexity” of an instance of the linear regression problem: smallest  $q$  for which  $D_{q,r}^{\text{eff}}(\widehat{\Sigma}, \beta) \asymp \log(r)$  and the signal-to-noise ratio.

It is reasonable to ask whether our NCT estimator is minimax optimal over these classes with respect to the relative MSE. The following theorem answers this question.



(a) Bound on the relative mean squared error  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  from Theorem 3.1;



(b) Bound on the relative prediction error  $\text{PE}(\hat{\beta})/\text{PE}(0)$  from Theorem 3.3 (i);

Fig 3: Rates for the relative errors of the NCT estimator in polynomial decay scenario.

**THEOREM 4.1.** (i) Let  $q \in (0; 2)$  and  $D \lesssim \log(r)$  be large enough. Then for any design matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$

$$\inf_{\tilde{\beta}} \sup_{\mathcal{P}^{\mathbf{X}}(q, D, S)} \mathbb{E} \left[ \frac{\text{MSE}(\tilde{\beta})}{\text{MSE}(0)} \right] \gtrsim \frac{1}{(S^2 n)^{1-q/2}}.$$

(ii) Let  $q = 0$  (in this case  $D$  plays role of the sparsity of the principal components). Then for any design matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$

$$\inf_{\tilde{\beta}} \sup_{\mathcal{P}^{\mathbf{X}}(q, D, S)} \mathbb{E} \left[ \frac{\text{MSE}(\tilde{\beta})}{\text{MSE}(0)} \right] \gtrsim \frac{D \log(er/D)}{S^2 n}.$$

This result almost directly follows from classical minimax lower bounds for Gaussian sequence model, with some slight adjustments. The bounds align well with [Raskutti, Wain-](#)



wright and Yu (2011), which studies minimax lower bounds over  $\ell_q$ -balls. We emphasize that the lower bounds presented here hold for any design matrix  $\mathbf{X}$ , unlike a lot of minimax optimality results that just find one “difficult” design to show the lower bound or impose restrictive assumptions on  $\mathbf{X}$ . This supports our intuition that for the relative bounds design matrix does not play role as much as the interactions of the covariance and the regression coefficients.

One could notice that in (i) the rate does not completely match the upper bound of Theorem 3.1, which also contains a factor  $D$  and a logarithmic factor. However, as we highlighted earlier, without much loss of generality we restrict  $D$  to be of at most of order  $\log(r)$ . Therefore, the discrepancy between the upper bound for NCT and the associated minimax lower bounds is just in the logarithmic factors.

**5. Theoretical properties of the GCT estimator.** Now we move to a brief study of the theoretical guarantees for the GCT estimator (2.2). The results are not as insightful as the ones for the NCT estimator, and we focus on the fixed design setting only. However, once we state the MSE bound, the theory for the random design can be developed in the same way as for the NCT estimator.

**THEOREM 5.1.** *Suppose Assumption 3.1 is satisfied. Let  $\varphi \geq 0$ . Take  $\tau = \hat{\lambda}_1^{\varphi/2} \sigma \rho$  with  $\rho = \frac{2}{\sqrt{n}} (\log(2r/\delta))^{1/\alpha}$ . Then, with probability  $1 - \delta$ , the GCT estimator  $\hat{\beta}$  from (2.2) with parameter  $\varphi$  and thresholding at level  $\tau$  satisfies*

$$\text{MSE}(\hat{\beta}) \lesssim \sum_{j=1}^r \min \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, |\theta_j| \right)^2.$$

Moreover, when  $T_\tau[\cdot] = \text{SOFT}_\tau[\cdot]$ , the matching lower bound takes place, i.e. the above upper bound is tight:

$$\text{MSE}(\hat{\beta}) \gtrsim \sum_{j=1}^r \min \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, |\theta_j| \right)^2$$

with probability  $1 - \delta$ .

The obtained bound may be difficult to comprehend, but we state it in the most general form to make sure it is tight and applicable in wide range of scenarios. When  $\theta_j = 0$  for  $j \geq m$ , it has no estimation errors beyond the first  $m$  principal components, adapting very well to focusing only on the low dimensions estimation like PCR. If, in addition,  $\hat{\lambda}_1/\hat{\lambda}_m$  is bounded, we have

$$\text{MSE}(\hat{\beta}) \lesssim \sum_{j=1}^m \min(\sigma \rho, |\theta_j|)^2,$$

which is not much larger than the MSE of PCR. It can even be much smaller than PCR when  $|\theta_j|$  are small for many indices  $j$ . In general, GCT outperforms NCT when  $|\theta_j|$  decays fast enough.

The next corollary allows to make sure that in general the bound is essentially dimension-free: as for Theorem 3.1, the rate can be expressed in terms of the effective rank  $r^{\text{eff}}[\hat{\Sigma}]$ .

**COROLLARY 5.2.** *Under assumptions of Theorem 5.1, with probability  $1 - \delta$ , the GCT estimator  $\hat{\beta}$  from (2.2) with parameter  $\varphi \geq 0$  satisfies*

$$\text{MSE}(\hat{\beta}) \lesssim \left( \|\hat{\Sigma}\| \|\beta\|^2 + \sigma \|\hat{\Sigma}\|^{1/2} \|\beta\| r^{\text{eff}}[\hat{\Sigma}]^{1/2} \right) \frac{(\log(2r/\delta))^{2/(2+\varphi)\alpha}}{n^{1/(2+\varphi)}}.$$

Note that when  $\varphi = 0$  we essentially recover the rate obtained after Theorem 3.1. The rate deteriorates when  $\varphi$  is far from 0, and this is explainable: the GCT procedure significantly deviates from the natural one, leading to a worse bound in the worst case, i.e. when the only assumption is the control of effective rank, and spikiness of canonical coefficients is not justified.

REMARK 5.1 (Simplifications in specific cases). *In several specific cases the rate from Theorem 5.1 can be made much more explicit. We omit logarithmic terms.*

- *Polynomial decay.* If  $\widehat{\lambda}_j/\widehat{\lambda}_1 \asymp j^{-a}$  and  $|\mathbf{u}_j^\top \boldsymbol{\beta}|/|\mathbf{u}_1^\top \boldsymbol{\beta}| \asymp j^{-b}$  for  $a \geq 0, a + 2b \geq 1$ , then with high probability

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) \lesssim \left( \frac{\sigma^2}{n} \right)^{\frac{a+2b-1}{a(\varphi+1)+2b}}.$$

- *Sparsity.* If there exists a set  $\mathcal{J}$  of size  $|\mathcal{J}| = s$  such that  $\mathbf{u}_j^\top \boldsymbol{\beta} = 0$  for  $j \notin \mathcal{J}$  and  $\widehat{\lambda}_{\max \mathcal{J}} \gtrsim \widehat{\lambda}_1$ , then with high probability

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) \lesssim \frac{s\sigma^2}{n}.$$

- *Approximate sparsity.* If there exists a set  $\mathcal{J}$  of size  $|\mathcal{J}| = s$  such that  $\widehat{\lambda}_j^{1/2} |\mathbf{u}_j^\top \boldsymbol{\beta}| \lesssim \|\boldsymbol{\theta}\|_2/d$  for  $j \notin \mathcal{J}$  and  $\widehat{\lambda}_{\max \mathcal{J}} \gtrsim \widehat{\lambda}_1$ , then with high probability

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) \lesssim \frac{s\sigma^2}{n} + \frac{\|\boldsymbol{\theta}\|_2^2}{d}.$$

- *Factor Model regime.* If  $\lambda_1 \asymp \dots \asymp \lambda_m \asymp d$ ,  $\lambda_{m+1} \asymp \dots \asymp \lambda_d \asymp 1$  for some  $m$ , then with high probability

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) \lesssim \frac{m\sigma^2}{n} + \frac{\|\boldsymbol{\beta}\|_2^2}{d}.$$

Therefore, the rate from Theorem 5.1 can adapt well to these specific structures despite the deteriorating rate of Corollary 5.2, which is only an upper bound.

## 6. Miscellaneous aspects.

6.1. *Computational complexity for single  $\tau$ .* To start with, we focus on the case when a good value of  $\tau$  is somehow known, and analyze the computational complexity of the GCT estimators. In particular, this includes the NCT estimator. The computation of SVD of  $\mathbf{X}$  (specifically,  $\widehat{\mathbf{\Lambda}}$  and  $\widehat{\mathbf{U}}$ ) takes  $O(dn \min(d, n))$  time. Once we have SVD, computing the matrix  $\mathbf{A} \stackrel{\text{def}}{=} \widehat{\mathbf{U}} \widehat{\mathbf{\Lambda}}^{-1-\varphi}$  and the vector  $\mathbf{b} \stackrel{\text{def}}{=} \widehat{\mathbf{\Lambda}}^{-1+\varphi} \widehat{\mathbf{U}}^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n}$  needed before the generalized thresholding takes  $O(dn)$  time. Obtaining  $\widehat{\boldsymbol{\beta}}$  for already computed  $\mathbf{A}$  and  $\mathbf{b}$  takes  $O(d \min(d, n))$  time. Therefore, the total computational time of our procedure is  $O(dn \min(d, n))$ . The computation is as fast as the SVD of the design matrix.

Note that computational complexity of the LASSO is  $O(n \min(d, n)^2)$ , when we compute its solution path via a modification of Least Angle Regression, see Efron et al. (2004).

6.2. *Efficient tuning of thresholding level  $\tau$ .* Our approach requires to tune the hyperparameter  $\tau$ . Whatever  $\tau$  is, we anyway have to compute  $\mathbf{A}$  and  $\mathbf{b}$ . This already takes  $O(dn \min(d, n))$  time. Applying the generalized thresholding and combining the result into the vector  $\widehat{\boldsymbol{\beta}}$  takes  $O(d \min(d, n))$  time. This means that we can try  $n$  different values of

$\tau$  “for free” — the computational complexity will be still of the same order as computing  $\hat{\beta}$  for a single value of  $\tau$ . But we can go even further, if we focus on the GCT estimators with the soft or hard thresholding.

Notice that varying  $\tau$  continuously from 0 to  $+\infty$ , we still can get only  $(\min(d, n) + 1)$  different solutions  $\hat{\beta}$  for given  $\mathbf{X}, \mathbf{Y}$  because we threshold a vector of size  $\min(d, n)$ . Therefore, we can compute the whole solution path for  $\tau$  from 0 to  $+\infty$ . However, we are not that interested in the solution path, since we do not expect to get coefficients entering the picture one by one as in LASSO for sparse regression. Instead, this can be useful for  $L$ -fold cross validation, where we will have at most  $(L \min(d, n) + 2)$  “interesting” values of  $\tau$  giving different solutions  $\hat{\beta}$ . In total, this implies that we need

$O(dn \min(d, n)) + (L \min(d, n) + 2) \cdot O(d \min(d, n)) = O(dn \min(d, n) + Ld \min(d, n)^2)$  operations to find the best  $\tau$  (providing smallest cross-validation error). In practice, we typically use  $L$  of constant order, e.g.  $L = 5$  or  $L = 10$ , which leads to the total computational complexity of  $O(dn \min(d, n))$  for our optimally tuned estimator – same as for a single value of  $\tau$ . Leave-one-out cross validation takes slightly more computations, namely,  $O(dn \min(d, n)^2)$ .

Similar “free tuning” property holds for LASSO (we again refer to [Efron et al. \(2004\)](#)). However, for example the ridge regression does not possess this nice properties: different values of regularization parameter will lead to different estimators, and one has to “guess” a discrete set of values to be tried.

**6.3. Optimality of cross-validation.** More formally, let  $\{\mathcal{B}_l\}_{l=1}^L$  be the split of the data point indices  $[n]$  into  $L$  approximately equally-sized disjoint blocks, i.e.

$$\mathcal{B}_l \cap \mathcal{B}_{l'} = \emptyset \quad \text{for all } l \neq l', l, l' \in [L] \quad \text{and} \quad \bigcup_{l=1}^L \mathcal{B}_l = [n]$$

satisfying  $\lfloor n/L \rfloor \leq |\mathcal{B}_l| \leq \lfloor n/L \rfloor + 1$  for all  $l \in [L]$ . The  $L$ -fold cross-validation leads to the following choice of the hyperparameter  $\tau$ :

$$\begin{aligned} (6.1) \quad \tau^{cv} &\stackrel{\text{def}}{=} \arg \min_{\tau \geq 0} \frac{1}{L} \sum_{l=1}^L \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} \left( y_i - \mathbf{x}_i^\top \hat{\beta}_\tau^{(l)} \right)^2 \\ &= \arg \min_{\tau \in \mathcal{T}} \frac{1}{L} \sum_{l=1}^L \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} \left( y_i - \mathbf{x}_i^\top \hat{\beta}_\tau^{(l)} \right)^2, \end{aligned}$$

where  $\hat{\beta}_\tau^{(l)}$  is the NCT estimator with thresholding at level  $\tau$  computed on the part of the sample  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ . Note that here  $\mathcal{T}$  is a set of  $|\mathcal{T}| \leq Lr + 2$  known “interesting” values giving all possible variety of estimators (i.e.  $\{\hat{\beta}_\tau^{(l)}\}_{l \in [L], \tau \geq 0} = \{\hat{\beta}_\tau^{(l)}\}_{l \in [L], \tau \in \mathcal{T}}$ ), as discussed in Section 6.2. An oracle counterpart of  $\tau^{cv}$ , defining optimal value of the hyperparameter w.r.t. the expected cross-validation error, is given by

$$\begin{aligned} (6.2) \quad \tau^{oracle} &\stackrel{\text{def}}{=} \arg \min_{\tau \geq 0} \frac{1}{L} \sum_{l=1}^L \mathbb{E} \left[ \left( y - \mathbf{x}^\top \hat{\beta}_\tau^{(l)} \right)^2 \right] \\ &= \arg \min_{\tau \in \mathcal{T}} \frac{1}{L} \sum_{l=1}^L \text{PE}(\hat{\beta}_\tau^{(l)}). \end{aligned}$$

We have the following result stating that the choice of  $\tau$  based on cross-validation performs as well as the oracle choice (in terms of the expected cross-validation error).

**THEOREM 6.1.** *Suppose Assumptions (3.1) – (3.4) are fulfilled. For technical simplicity, consider the truncated NCT estimator:*

$$\hat{\beta}_\tau \stackrel{\text{def}}{=} \hat{\mathbf{U}}_{\leq k^*} \hat{\mathbf{\Lambda}}_{\leq k^*}^{-1} \text{SOFT}_\tau \left[ \hat{\mathbf{\Lambda}}_{\leq k^*}^{-1} \hat{\mathbf{U}}_{\leq k^*}^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n} \right]$$

with  $k^*$  from Theorem 3.3. Then, with probability  $1 - \delta$ , the above estimator with thresholding at level  $\tau^{cv}$  chosen by  $L$ -fold cross-validation satisfies

$$\begin{aligned} \frac{1}{L} \sum_{l=1}^L \text{PE}(\hat{\beta}_{\tau^{cv}}^{(l)}) &\lesssim \frac{1}{L} \sum_{l=1}^L \text{PE}(\hat{\beta}_{\tau^{oracle}}^{(l)}) \\ &+ \left( \|\Sigma\| \|\beta\|_2^2 + \sigma^2 (\log(2r/\delta))^{2/\alpha} \right) \left( \sqrt{\frac{L \log(Lr/\delta)}{n}} + \frac{L (\log(Lr/\delta))^{2/\alpha}}{n} \right). \end{aligned}$$

The result resembles, for instance, Györfi et al. (2002); Dudoit and van der Laan (2005) (see Theorem 7.1 in the former and Theorem 1 in the latter) and many other works, and the logic behind it is quite standard. However, since our framework is not as general as in Györfi et al. (2002) or Dudoit and van der Laan (2005), we state the high probability bound rather than in expectation, and we avoid almost sure boundedness condition on the response and the possible predictions of our estimator (unlike the aforementioned literature). We also emphasize once again that  $\tau^{cv}$  and  $\tau^{oracle}$  from (6.1) and (6.2) are minimizers across all  $\tau \geq 0$ , and the structure of our estimator allows to compute  $\tau^{cv}$  defined in such a way in a reasonable time, which is quite unusual feature.

From Theorem 3.3 and Theorem 6.1 it follows that  $L^{-1} \sum_{l=1}^L \text{PE}(\hat{\beta}_{\tau^{cv}}^{(l)}) \lesssim \diamond_\delta$  with probability  $1 - \delta$ , where  $\diamond_\delta$  is the error bound from Theorem 3.3 (best of (i) and (ii)). This bound holds since  $\tau^{oracle}$  is no worse than the choices of  $\tau$  from Theorem 3.3 in terms of the expected cross-validation error, and because the extra term in Theorem 6.1 does not exceed the error bounds from Theorem 3.3 (treating  $L$  as constant).

It worth mentioning that the result of Theorem 6.1 is not really what one aims for. Ideally, we would like to obtain a high probability bound of the form  $\text{PE}(\hat{\beta}_{\tau^{cv}}) \lesssim \text{PE}(\hat{\beta}_{\tau^{opt}}) + \Delta_\delta$  with small  $\Delta_\delta$ , where  $\tau^{opt} \stackrel{\text{def}}{=} \arg \min_{\tau \geq 0} \text{PE}(\hat{\beta}_\tau)$  is the optimal value of the hyperparameter, which may differ from  $\tau^{oracle}$ . This would consequently imply  $\text{PE}(\hat{\beta}_{\tau^{cv}}) \lesssim \diamond_\delta + \Delta_\delta$  with high probability. Instead of analyzing the expected prediction error of the estimator trained on the whole sample, Theorem 6.1 is concerned with the expected cross-validation error. However, as we already mentioned, even classical works on cross-validation, such as Györfi et al. (2002) and Dudoit and van der Laan (2005), also state results of this flavor, and many papers focusing specifically on cross-validation actually work with the expected cross-validation error, which anyway is believed to be a good proxy for the expected prediction error. With this in mind, we hope our result gives a convincing confirmation that the cross-validation procedure applied to our estimator is reasonable, even though in terms of the expected cross-validation error. Since the main focus of this paper is not on the cross-validation, we do not go beyond this.

To conclude the discussion on the cross-validation, we mention that in principle one may want to tune  $\varphi$  in addition to  $\tau$  using cross-validations as well. In that case, pairs of the hyperparameters  $(\tau, \varphi)$  should be chosen from some prespecified candidate set  $\mathcal{T} \times \Phi \subset \mathbb{R}_+ \times \mathbb{R}_+$  (e.g. the Cartesian product of two grids) of a finite size  $|\mathcal{T}| \times |\Phi| < \infty$ . Optimal pairs  $(\tau^{cv}, \varphi^{cv})$  and  $(\tau^{oracle}, \varphi^{oracle})$  should be defined as the solutions to similar optimization problems as (6.1) and (6.2), respectively, but this time the estimator  $\hat{\beta}_{\tau, \varphi}$  depends also on  $\varphi$ , and the optimization is over the finite candidate set. Then a result similar to Theorem 6.1 holds, but with  $\log(Lr)$  in the bound replaced by  $\log(|\mathcal{T}| \times |\Phi|)$ .

**7. Discussion.** We provide a new prospective on the non-sparse high-dimensional linear regression problem. The proposed family of GCT estimators serves as a bridge between two classical paradigms: sparse regression and principal components regression. W.r.t. the absolute errors, the fast decay of eigenvalues of the covariance is enough to ensure convergence even in high dimensions without any assumptions on the regression coefficients. Moreover, we argue that the relative errors are more appropriate in the high-dimensional regression with the eigenvalue decay, and that the complexity of a linear regression problem is characterized by the signal-to noise ratio (instead of the magnitude of noise) and the interaction between the covariance and regression coefficients, expressed by the joint effective dimension (instead of assumptions on the regression coefficients). It is not really important what the design matrix is, and we do not need to impose restrictive assumptions on it, if we choose relative errors as a measure of performance and standardize the data properly. The NCT estimator is minimax optimal for any design over suitable parameter classes in this paradigm. Hopefully, our insights shed some light on the nature of the non-sparse high-dimensional linear regression.

We leave several important directions for further investigation. First of all, despite our joint effective dimension is quite well-motivated, it does not mean that there are no other structural assumptions related to the eigenvalue decay. New discoveries in the structure of the high-dimensional linear regression can potentially lead to other procedures, whose minimax optimality should be analyzed over appropriate parameter classes.

Also, even though the analysis of our structural assumptions and procedures for fixed design seems quite complete, there are unanswered questions in the random design setting. It is not clear whether the bounds of Theorem 3.3 can be improved and whether the associated relative errors can be represented in a convenient way. An uncertainty brought by the covariance matrix requires developing and applying new advanced statistical tools for the analysis of minimax optimality in random design linear regression in high dimensions.

Furthermore, from the numerical experiments, postponed to the Supplementary Material [Silin and Fan \(2021\)](#), we observe that even regularized estimators (such as NCT) behave unexpectedly around the interpolation threshold  $d = n$ . Our theoretical results do not predict the bumps that errors as functions of the dimension exhibit in this region. This definitely worth studying in the future.

In addition, the uncertainty quantification for the estimated parameters and function values is of significant interest, as well as possible extensions of our ideas to nonparametric regression in reproducing kernel Hilbert space (RKHS). Beyond the linear model, can penalized quasi-likelihood on canonical parameters share similar properties to those in the regression problem?

**Acknowledgements.** We thank gratefully the Editor, the Associate Editor and the Referees for constructive comments and valuable suggestions which led to significant improvements on the paper. The research was supported by ONR grant N00014-19-1-2120, NSF grant DMS-1662139, and NIH grant 2R01-GM072611-14.

## SUPPLEMENTARY MATERIAL

**Supplement to “Canonical thresholding for non-sparse high-dimensional linear regression”.** Simulation studies and proofs of the results presented in this paper can be found in the Supplementary Material ([Silin and Fan, 2021](#)).

()

## REFERENCES

- BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.*, **101**, 473, 119–137.
- BARTLETT, P., LONG, P., LUGOSI, G. and TSIGLER, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*.
- BELKIN, M. (2018). Approximation beats concentration? An approximation view on inference with smooth radial kernels. *Proc. Mach. Learn. Res.*, **75**, 1–18.
- BELKIN, M., HSU, D. and XU, J. (2019). Two models of double descent for weak features. *ArXiv:1903.07571*.
- BELLEÇ, P., LECUÉ, G. and TSYBAKOV, A. (2018). SLOPE meets Lasso: improved oracle bounds and optimality. *Ann. Statist.*, **46**, 6B, 3603–3642.
- BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 4, 1705–1732.
- BIETTI, A. and MAIRAL, J. (2019). On the inductive bias of Neural Tangent Kernels. *Advances in Neural Information Processing Systems*, 12893–12904.
- BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. (2015). SLOPE – adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, **9**, 3, 1103–1140.
- CANDÈS, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Ann. Statist.*, **35**, 6, 2313–2351.
- CHINOT, G. and LERASLE, M. (2020). Benign overfitting in the large deviation regime. *ArXiv:2003.05838*.
- DALALYAN, A., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli*, **23**, 1, 552–581.
- DOBRIAN, E. and LIU, S. (2019). Asymptotics for sketching in least squares regression *Advances in Neural Information Processing Systems*.
- DOBRIAN, E. and SHENG, Y. (2020). WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions. *J. Mach. Learn. Res.*, **21**(66), 1–52.
- DOBRIAN, E. and SHENG, Y. (2021+). Distributed linear regression by averaging. *Ann. Statist.*, to appear.
- DONOHU, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
- DONOHU, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, **41**, 3, 613–627.
- DONOHU, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.
- DONOHU, D. L., JOHNSTONE, I. M., KERKYCHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **57**, 301–369.
- DONOHU, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **3**, 879–921.
- DUDOIT, S. and VAN DER LAAN, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statist. Methodol.*, **2**, 2, 131–154.
- EFRON, B., HASTIE, T., JOHNSTONE, I. M. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.*, **32**, 2, 407–499.
- FAN, J. (1996). Test of significance based on wavelet thresholding and Neyman’s truncation *J. Amer. Statist. Assoc.*, **91**, 434, 674–688.
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 456, 1348–1360.
- FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). Statistical foundations of data science. *CRC Press*.
- FAN, J., KE, Y. and WANG, K. (2020). Factor-adjusted regularized model selection. *J. Econometrics*, **216**, 71–85.
- FAN, J., WANG, W. and YAO, J. (2017). Sufficient forecasting using factor models. *J. Econometrics*, **201**, 292–306.
- GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**, 6, 971–988.
- GYÖFRI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). A distribution-free theory of nonparametric regression. *Springer, New York*.
- HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *ArXiv:1903.08560*.
- JIRAK, M. and WAHL, M. (2018). Relative perturbation bounds with applications to empirical covariance operators. *ArXiv:1802.02869*.
- JOLLIFFE, I. (1982). A note on the use of principal components in regression. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **31**, 3, 300–303.
- KUCHIBHOTLA, A. K. and CHAKRABORTY, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *ArXiv:1804.02605*.



- LIANG, T. and RAKHLIN, A. (2020). Just interpolate: kernel "ridgeless" regression can generalize. *Ann. Statist.*, **48**, 3, 1329–1347.
- MA, S. and BELKIN, M. (2017). Diving into the shallows: a computational perspective on large-scale shallow learning. *Advances in Neural Information Processing Systems*, 3781–3790.
- PAUL, D., BAIR, E., HASTIE, T. and TIBSHIRANI, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.*, **36**, 4, 1595–1618.
- PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science.*, **2**, 559–572.
- RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls. *IEEE Trans. Inform. Theory*, **57**, 10, 6976–6994.
- SILIN, I. and FAN, J. (2021). Supplement to "Canonical thresholding for non-sparse high dimensional linear regression".
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **58**, 1, 267–288.
- VAN DE GEER, S. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, **3**, 1360–1392.
- ZUMBACH, G. (2009). The empirical properties of large covariance matrices. *RiskMetrics Journal*, **9**, 1.

# SUPPLEMENT TO “CANONICAL THRESHOLDING FOR NON-SPARSE HIGH DIMENSIONAL LINEAR REGRESSION”

BY IGOR SILIN AND JIANQING FAN

*Princeton University*

## CONTENTS

S.1 Simulation studies . . . . .	1
S.2 Main proofs . . . . .	4
S.2.1 Proof of Theorem 3.1 . . . . .	5
S.2.2 Proof of Theorem 3.3 . . . . .	6
S.2.3 Proof of Theorem 4.1 . . . . .	11
S.2.4 Proof of Theorem 5.1 . . . . .	13
S.2.5 Proof of Corollary 5.2 . . . . .	14
S.2.6 Proof of Theorem 6.1 . . . . .	14
S.3 Additional proofs . . . . .	17
S.3.1 Proof of Proposition 3.2 . . . . .	17
S.3.2 Proof of Lemma S.2.1 . . . . .	18
S.3.3 Proof of Lemma S.2.3 . . . . .	18
S.3.4 Proof of Lemma S.2.4 . . . . .	19
S.3.5 Proof of Lemma S.2.5 . . . . .	19
S.3.6 Proof of Lemma S.2.6 . . . . .	20
References . . . . .	20

**S.1. Simulation studies.** We compare the following methods:

- “NCT”: Natural Canonical Thresholding estimator (2.1) with efficient hyperparameter tuning by 10-fold CV.
- “GCT”: Generalized Canonical Thresholding estimator (2.2) with  $\varphi = 1$ , the soft thresholding, and with efficient hyperparameter tuning by 10-fold CV.
- “OLS”: Ordinary Least Squares. When  $d > n$ , the min norm solution is considered.
- “PCR”: Principal Component Regression. The number of PCs is chosen by 10-fold CV.
- “Ridge”: Ridge regression with 10-fold CV (default implementation from R-package glmnet).
- “LASSO”: LASSO with 10-fold CV (default implementation from R-package glmnet).

We fix  $n = 200$ ,  $\text{SNR} = 10$  or  $\text{SNR} = 1$ , and focus on how the relative errors  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  and  $\text{PE}(\hat{\beta})/\text{PE}(0)$  of these methods behave when the dimension  $d$  grows. The covariates  $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{N}(0, \Sigma)$ , where  $\Sigma$  depends on the eigenvalue scenario, and the noise vector  $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$  (where  $\sigma^2$  is chosen to ensure  $\text{SNR} = 10$  or  $\text{SNR} = 1$  for given  $\Sigma$  and  $\beta$ ). Without loss of generality we take  $\Sigma$  a diagonal matrix, or equivalently  $\mathbf{U} = \mathbb{I}_d$ . The results are presented in Figure S1–S4. Figure S1 and Figure S2 correspond to  $\text{SNR} = 10$ , Figure S3 and Figure S4 correspond to  $\text{SNR} = 1$ . Figure S1 and Figure S3 cover the scenarios of polynomial decay of the eigenvalues  $\lambda_j = j^{-a}$  and the coefficients  $\mathbf{u}_j^\top \beta = j^{-b}$  with

---

*MSC 2010 subject classifications:* Primary 62J05; secondary 62H12, 62H25.

*Keywords and phrases:* High-dimensional linear regression, covariance eigenvalues decay, thresholding, relative errors, principal component regression.

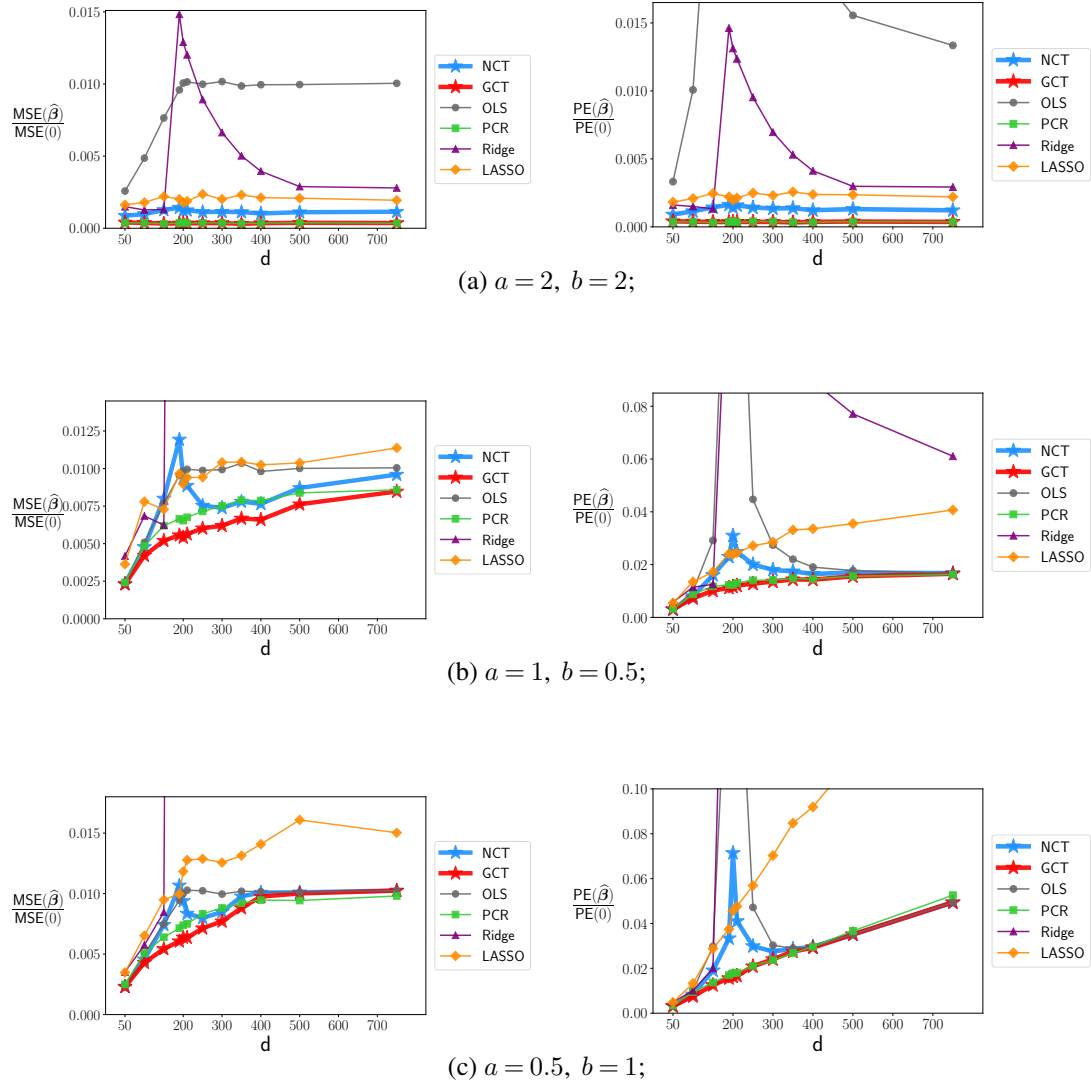


Fig S1: The relative errors  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  (left) and  $\text{PE}(\hat{\beta})/\text{PE}(0)$  (right) for different estimators with  $n = 200$ ,  $\text{SNR} = 10$ . Polynomial decay of eigenvalues and coefficients in eigenbasis:  $\lambda_j = j^{-a}$ ,  $\mathbf{u}_j^\top \beta = j^{-b}$ .

(a)  $a = 2, b = 2$ ;

(b)  $a = 1, b = 0.5$ ;

(c)  $a = 0.5, b = 1$ ;

while Figure S2 and Figure S4 also consider polynomial decay of the eigenvalues  $\lambda_j = j^{-a}$  but  $\mathbf{U}^\top \beta$  is different:

(a)  $a = 1$ ;  $\mathbf{u}_j^\top \beta = 1$  for  $j = 1, \dots, 10$  and 0 otherwise;

(b)  $a = 0.1$ ;  $\mathbf{u}_j^\top \beta = 1$  for 10 randomly chosen  $j \in \{d-25, \dots, d\}$ , and the rest components are i.i.d.  $\mathcal{N}(0, d^{-1})$ ;

(c)  $a = 0.5$ ;  $\mathbf{U}^\top \beta \sim \mathcal{N}(0, \mathbb{I}_d)$ .

In each scenario, for each method and dimension we run the corresponding experiment 100 times and plot the median errors.

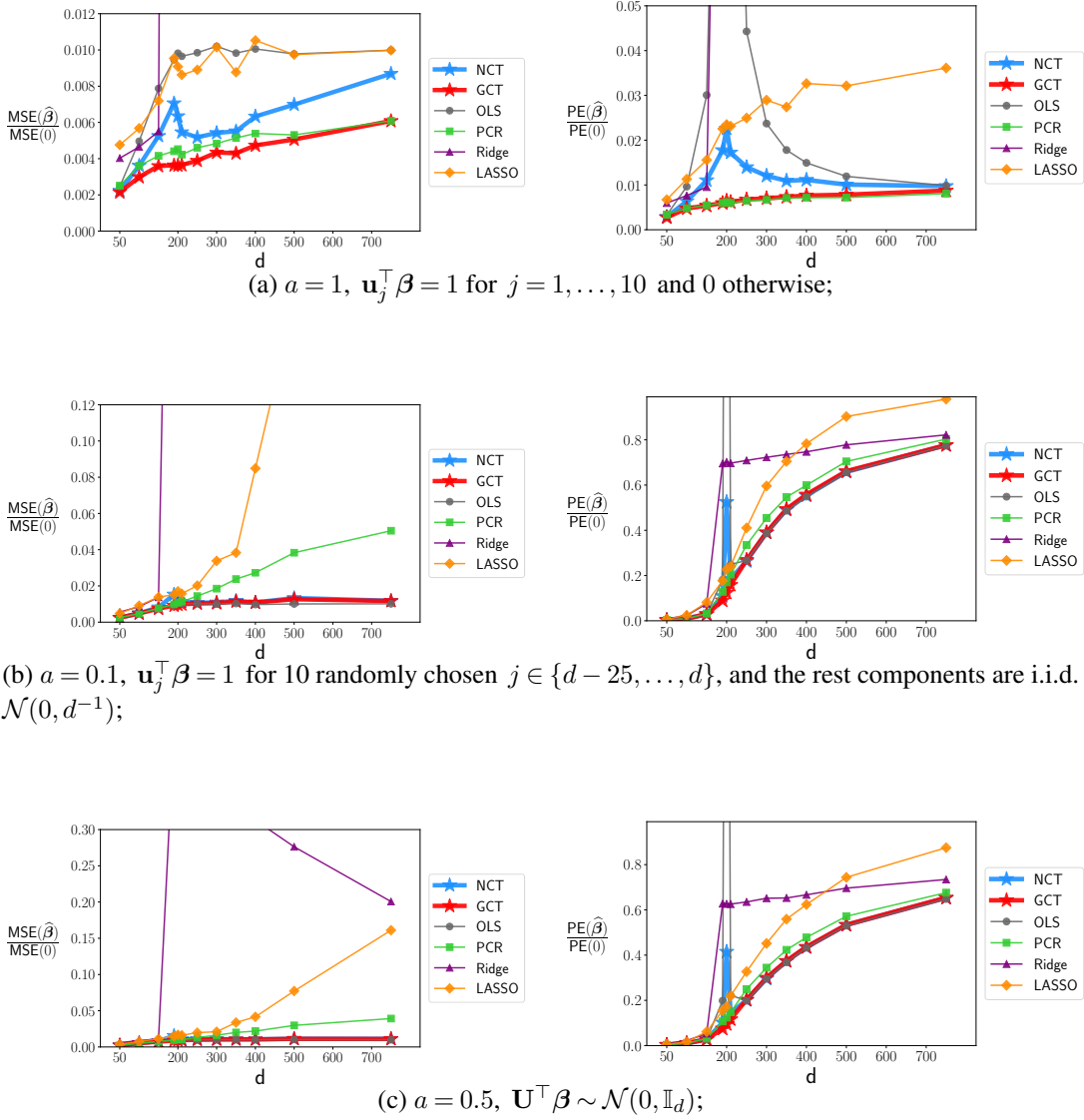


Fig S2: The relative errors  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  (left) and  $\text{PE}(\hat{\beta})/\text{PE}(0)$  (right) for different estimators with  $n = 200$ ,  $\text{SNR} = 10$ . Polynomial decay of eigenvalues  $\lambda_j = j^{-a}$  and different regimes of coefficients in eigenbasis  $\mathbf{u}_j^\top \beta$ .

We notice that the NCT estimator (among some others) in some settings suffer around  $d = n$ . This is so called “interpolation threshold” – when the dimension exceeds the number of data points, a model has enough features to interpolate training points. The behavior around this point and the associated “double descent” phenomenon has been an active area of research for the last couple of years. We do not focus on this in our work.

Otherwise, from the plots it is clear that in the presented settings the proposed procedure performs quite good compared to the other methods. In particular, the persistent performance of GCT suggests the benefit of varying thresholding to better adapt to various scenarios with different priors. However, it is worth mentioning that other methods also perform quite unexpectedly well in a variety of settings, though previous theoretical results for them do not

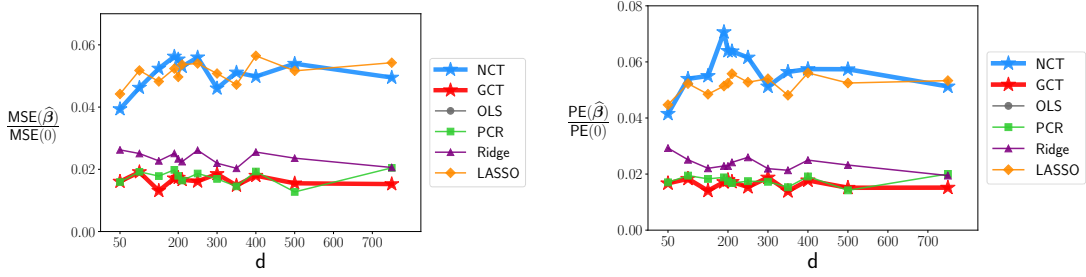
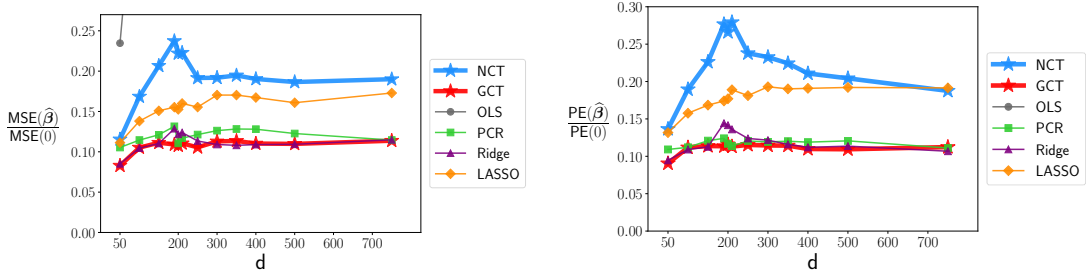
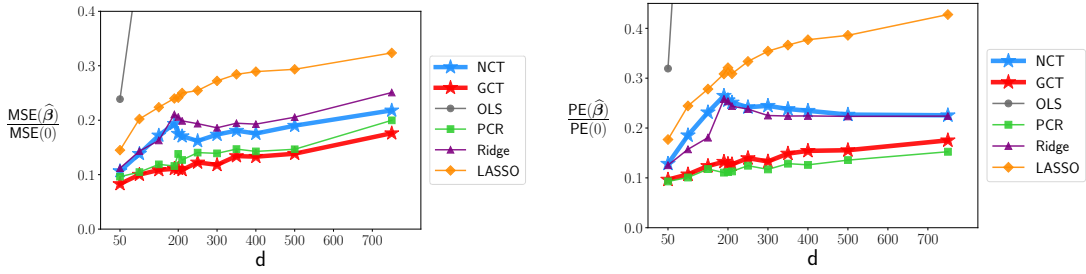
(a)  $a = 2, b = 2$ ;(b)  $a = 1, b = 0.5$ ;(c)  $a = 0.5, b = 1$ ;

Fig S3: The relative errors  $\text{MSE}(\hat{\beta})/\text{MSE}(0)$  (left) and  $\text{PE}(\hat{\beta})/\text{PE}(0)$  (right) for different estimators with  $n = 200$ ,  $\text{SNR} = 1$ . Polynomial decay of eigenvalues and coefficients in eigenbasis:  $\lambda_j = j^{-a}$ ,  $\mathbf{u}_j^\top \beta = j^{-b}$ .

predict such performance. This may engender an interest in more thorough study of classical linear regression methods in high-dimensional setting under different structural assumptions.

**S.2. Main proofs.** We start with the following lemma that allows to bound the properly scaled noise vector in  $\ell_\infty$ -norm. The lemma simultaneously deals with both fixed and random design settings.

LEMMA S.2.1. Suppose Assumption 3.1 is fulfilled. Let  $\rho$  be as in (3.1). Define the event

$$\Omega_1 \stackrel{\text{def}}{=} \left\{ \|\xi\|_\infty \leq \frac{\sigma\rho}{2} \right\} \quad \text{with} \quad \xi \stackrel{\text{def}}{=} \frac{\mathbf{Z}^\top \varepsilon}{n} = \frac{\hat{\Lambda}^{-1} \hat{\mathbf{U}}^\top \mathbf{X}^\top \varepsilon}{n}.$$

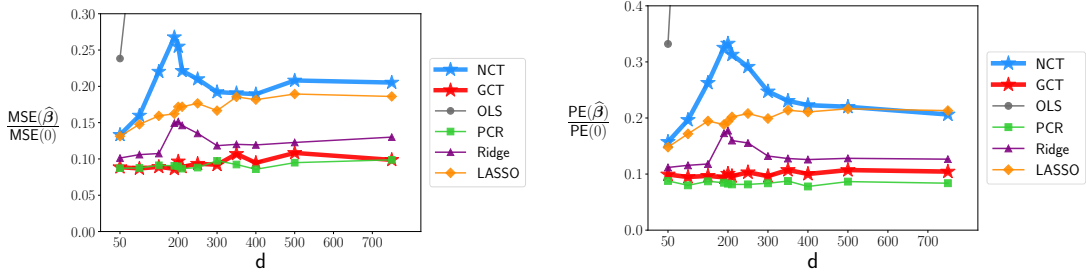
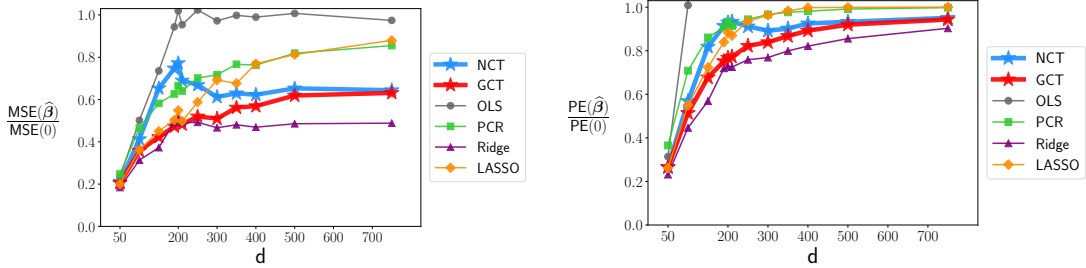
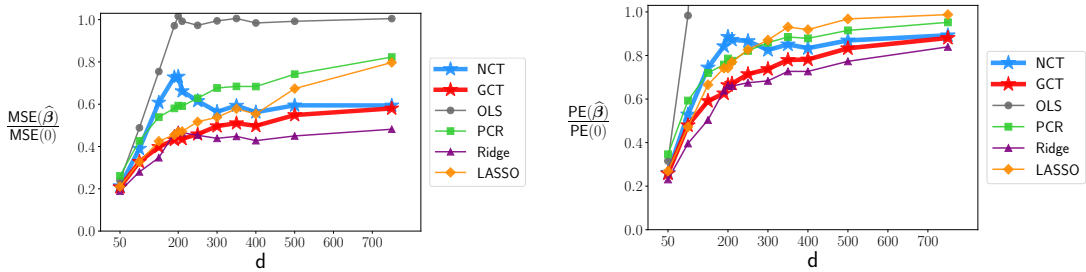
(a)  $a = 1$ ,  $\mathbf{u}_j^\top \boldsymbol{\beta} = 1$  for  $j = 1, \dots, 10$  and 0 otherwise;(b)  $a = 0.1$ ,  $\mathbf{u}_j^\top \boldsymbol{\beta} = 1$  for 10 randomly chosen  $j \in \{d - 25, \dots, d\}$ , and the rest components are i.i.d.  $\mathcal{N}(0, d^{-1})$ ;(c)  $a = 0.5$ ,  $\mathbf{U}^\top \boldsymbol{\beta} \sim \mathcal{N}(0, \mathbb{I}_d)$ ;

Fig S4: The relative errors  $\text{MSE}(\hat{\boldsymbol{\beta}})/\text{MSE}(0)$  (left) and  $\text{PE}(\hat{\boldsymbol{\beta}})/\text{PE}(0)$  (right) for different estimators with  $n = 200$ ,  $\text{SNR} = 1$ . Polynomial decay of eigenvalues  $\lambda_j = j^{-a}$  and different regimes of coefficients in eigenbasis  $\mathbf{u}_j^\top \boldsymbol{\beta}$ .

Then

$$\mathbb{P}[\Omega_1] \geq 1 - \delta.$$

S.2.1. *Proof of Theorem 3.1.* Using  $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}^2\hat{\mathbf{U}}^\top$  we write for  $\hat{\boldsymbol{\beta}}$  from (2.1)

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \|\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top \boldsymbol{\beta}\|_2^2.$$



Now we plug our estimator  $\hat{\beta}$  and  $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$  in to get

$$\begin{aligned} \text{MSE}(\hat{\beta}) &= \left\| \text{SOFT}_{\tau} \left[ \hat{\Lambda}^{-1} \hat{\mathbf{U}}^{\top} \frac{\mathbf{X}^{\top} \mathbf{Y}}{n} \right] - \hat{\Lambda} \hat{\mathbf{U}}^{\top} \beta \right\|_2^2 \\ &= \left\| \text{SOFT}_{\tau} \left[ \hat{\Lambda} \hat{\mathbf{U}}^{\top} \beta + \frac{\hat{\Lambda}^{-1} \hat{\mathbf{U}}^{\top} \mathbf{X}^{\top} \varepsilon}{n} \right] - \hat{\Lambda} \hat{\mathbf{U}}^{\top} \beta \right\|_2^2 = \left\| \text{SOFT}_{\tau} [\theta + \xi] - \theta \right\|_2^2, \end{aligned}$$

where we recall the canonical coefficients  $\theta = \hat{\Lambda} \hat{\mathbf{U}}^{\top} \beta$  from Definition 2.1 and  $\xi$  from Lemma S.2.1. From now on, the proof basically repeats the classical derivation for the soft and hard thresholding. Let us analyze its  $j$ -th component on  $\Omega_1$  from Lemma S.2.1 of probability at least  $1 - \delta$ .

- If  $|\theta_j + \xi_j| > \tau$ , then  $|\theta_j| \geq \tau - |\xi_j| \geq \tau/2$  and

$$|\text{SOFT}_{\tau} [\theta_j + \xi_j] - \theta_j| = |\theta_j + \xi_j \pm \tau - \theta_j| = |\xi_j \pm \tau| \leq |\xi_j| + \tau \leq \frac{3\tau}{2} \leq 3 \min(\tau, |\theta_j|),$$

where  $\pm$  means that we take either  $+$  or  $-$  depending on the sign of  $(\theta_j + \xi_j)$ , but this doesn't play any role. For the lower bound,

$$|\text{SOFT}_{\tau} [\theta_j + \xi_j] - \theta_j| = |\xi_j \pm \tau| \geq \tau - |\xi_j| \geq \tau/2 \geq \frac{1}{2} \min(\tau, |\theta_j|).$$

- If  $|\theta_j + \xi_j| \leq \tau$ , then  $|\theta_j| \leq \tau + |\xi_j| \leq 3\tau/2$  and

$$|\text{SOFT}_{\tau} [\theta_j + \xi_j] - \theta_j| = |0 - \theta_j| = |\theta_j| \leq 3 \min(\tau, |\theta_j|).$$

For the lower bound,

$$|\text{SOFT}_{\tau} [\theta_j + \xi_j] - \theta_j| = |\theta_j| \geq \frac{1}{2} \min(\tau, |\theta_j|).$$

Hence, on  $\Omega_1$

$$\frac{1}{4} \sum_{j=1}^r \min(\tau, |\theta_j|)^2 \leq \|\text{SOFT}_{\tau} [\theta + \xi] - \theta\|_2^2 \leq 9 \sum_{j=1}^r \min(\tau, |\theta_j|)^2.$$

Continuing the upper bound, note that for any  $0 \leq q \leq 2$  we have  $\min(\tau, |\theta_j|) \leq \tau^{1-q/2} |\theta_j|^{q/2}$  (we use convention  $0^0 = 0$ ). Thus,

$$\sum_{j=1}^r \min(\tau, |\theta_j|)^2 \leq \tau^{2-q} \sum_{j=1}^r |\theta_j|^q = \tau^{2-q} \|\theta\|_q^q,$$

using convention  $\|\cdot\|_0^0 = \|\cdot\|_0$ . Taking infimum over  $q \in [0, 2]$ , extracting  $\beta^{\top} \hat{\Sigma} \beta = \text{MSE}(0)$  and recalling the definitions of SNR and  $\text{D}_{q,d}^{\text{eff}}(\hat{\Sigma}, \beta)$ , we conclude the proof.

**S.2.2. Proof of Theorem 3.3.** To begin with, we state the following well-known result on the concentration of the sample covariance around the true covariance in terms of the effective rank. See Koltchinskii and Lounici (2017), Theorem 9; also, Vershynin (2018), Theorem 9.2.4 and Exercise 9.2.5.

**LEMMA S.2.2.** *Suppose Assumption 3.2 is fulfilled. Then, with probability  $1 - \delta$*

$$\|\hat{\Sigma} - \Sigma\| \leq C \|\Sigma\| \left( \sqrt{\frac{r^{\text{eff}}[\Sigma] + \log(1/\delta)}{n}} + \frac{r^{\text{eff}}[\Sigma] + \log(1/\delta)}{n} \right).$$

Using Assumption 3.4 we can leave only the first term in the bound above. Let  $\Omega_2$  be the event on which this bound holds.

Our main tools to prove the main result is the beautiful work by [Jirak and Wahl \(2018\)](#) that develops tight *relative perturbation bounds for eigenvalues and eigenvectors* of covariance matrix. Let us describe the framework of that paper. By Assumption 3.3 we consider the case of simple eigenvalues of  $\Sigma$ . The following quantities play important role: the *relative rank*

$$\mathbf{r}_j(\Sigma) \stackrel{\text{def}}{=} \sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_l}{|\lambda_j - \lambda_l|} + \frac{\lambda_j}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} \quad \text{for } j \in [d]$$

(here  $\lambda_0 = +\infty$  and  $\lambda_{d+1} = 0$  for convenience) and the entries of  $\Sigma^{-1/2}(\widehat{\Sigma} - \Sigma)\Sigma^{-1/2}$

$$\bar{\eta}_{ll'} \stackrel{\text{def}}{=} \frac{\mathbf{u}_l^\top (\widehat{\Sigma} - \Sigma) \mathbf{u}_{l'}}{\sqrt{\lambda_l \lambda_{l'}}} \quad \text{for } l, l' \in [d].$$

Relative perturbation bounds for  $j$ -th eigenvalue and eigenvector hold under the condition that there exist  $x$  such that

$$|\bar{\eta}_{ll'}| \leq x \quad \text{for all } l, l' \in [d], \quad \text{and} \quad \mathbf{r}_j(\Sigma) \leq \frac{1}{3x}.$$

The following lemma helps to control the first condition.

LEMMA S.2.3. *Suppose Assumption 3.2 and Assumption 3.4 hold. Then, with probability  $1 - \delta$*

$$\max_{l, l' \in [d]} |\bar{\eta}_{ll'}| \leq \epsilon,$$

where

$$\epsilon = \epsilon_{n,d,\delta} \stackrel{\text{def}}{=} C \sqrt{\frac{\log(d/\delta)}{n}}.$$

for some  $C$ .

Define  $\Omega_3$  to be the event where the inequality from the previous lemma holds.

So, the relative perturbation bounds hold true on  $\Omega_3$  for indices  $j$  for which  $\mathbf{r}_j(\Sigma) \leq 1/(3\epsilon)$ . We would like to have this property for as many indices as possible. Under Assumption 3.3 we have (see [Jirak and Wahl \(2018\)](#), inequalities (3.30); [Jirak \(2016\)](#), Lemma 7.13; [Cardot, Mas and Sarda \(2007\)](#), Lemma 6.1)

$$\mathbf{r}_j(\Sigma) \leq 1 + 2 \sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_l}{|\lambda_j - \lambda_l|} \leq 1 + 2Cj \log(j).$$

Note that with

$$k^* \stackrel{\text{def}}{=} (\epsilon \log(1/\epsilon))^{-2/3}$$

we indeed have  $\mathbf{r}_j(\Sigma) \leq 1/(3\epsilon)$  for all  $j \in [k^*]$  due to Assumption 3.4. Hence, the following relative perturbation bounds from [Jirak and Wahl \(2018\)](#) hold true.

LEMMA S.2.4. *For all  $j \in [k^*]$  on  $\Omega_3$  holds*

$$|\widehat{\lambda}_j - \lambda_j| \leq C\epsilon\lambda_j \quad \text{and} \quad \|\widehat{\mathbf{u}}_j - \mathbf{u}_j\|_2 \leq C\epsilon \sqrt{\sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_j \lambda_l}{(\lambda_j - \lambda_l)^2}}.$$

Furthermore, for all  $j \in [k^*]$  and  $l \in [d], l \neq j$  on  $\Omega_3$  holds

$$|\hat{\mathbf{u}}_j^\top \mathbf{u}_l| \leq C\epsilon \frac{\sqrt{\lambda_j \lambda_l}}{|\lambda_j - \lambda_l|}.$$

Note that the bounds from the previous lemma apply even for larger indices  $j$ , which can be up  $(\epsilon \log(1/\epsilon))^{-1}$  (of order  $n^{1/2}$ ), while we restrict  $k^*$  to be of order  $n^{1/3}$ . Later in the proof it will be clear how this specific  $k^*$  arises.

Now we are ready to proceed to the main part of the proof.

**PROOF OF THEOREM 3.3.** We prove the theorem for  $k = k^*$ , and it will be clear that the same proof works with any  $k < k^*$ . The proof for part (i) and part (ii) coincides up to the last step. Denote  $\tau'$  to be the general thresholding level, which is  $\tau$  for part (i) and  $\bar{\tau}$  for part(ii). Using the definition of  $\hat{\beta}$  given in (2.1), the eigendecompositions  $\Sigma = \mathbf{U}\Lambda^2\mathbf{U}^\top$ ,  $\hat{\Sigma} = \hat{\mathbf{U}}\hat{\Lambda}^2\hat{\mathbf{U}}^\top$  and the model  $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ , write the prediction error as

$$\begin{aligned} \text{PE}(\hat{\beta}) &= (\hat{\beta} - \beta)^\top \Sigma (\hat{\beta} - \beta) \\ &= \|\Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\Lambda}^{-1} \text{SOFT}_{\tau'} \left[ \hat{\Lambda}^{-1} \hat{\mathbf{U}}^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n} \right] - \Lambda \mathbf{U}^\top \beta\|_2^2 \\ &= \|\Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\Lambda}^{-1} \text{SOFT}_{\tau'} [\hat{\Lambda} \hat{\mathbf{U}}^\top \beta + \boldsymbol{\xi}] - \Lambda \mathbf{U}^\top \beta\|_2^2 \end{aligned}$$

with  $\boldsymbol{\xi}$  from Lemma S.2.1. Let us add and subtract  $\Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \beta$  inside the norm and apply  $\|\mathbf{a} + \mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ :

$$\begin{aligned} \text{PE}(\hat{\beta}) &\leq 2\|\Lambda \mathbf{U}^\top \beta - \Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \beta\|_2^2 + \\ &\quad + 2\|\Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\Lambda}^{-1} \text{SOFT}_{\tau'} [\hat{\Lambda} \hat{\mathbf{U}}^\top \beta + \boldsymbol{\xi}] - \Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \beta\|_2^2 =: I_1 + I_2. \end{aligned}$$

We first deal with  $I_1$ :

$$\begin{aligned} \frac{I_1}{2} &= \|\Lambda \mathbf{U}^\top \beta - \Lambda \mathbf{U}^\top \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \beta\|_2^2 = \beta^\top (\mathbb{I}_d - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top) \Sigma (\mathbb{I}_d - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top) \beta \\ &= \beta^\top (\mathbb{I}_d - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top) (\Sigma - \hat{\Sigma}) (\mathbb{I}_d - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top) \beta \leq \|\hat{\Sigma} - \Sigma\| \|(\mathbb{I}_d - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top) \beta\|_2^2 \\ &\leq \|\hat{\Sigma} - \Sigma\| \|\beta\|_2^2 \leq C \|\Sigma\| \|\beta\|_2^2 \sqrt{\frac{\text{r}^{\text{eff}}[\Sigma] + \log(1/\delta)}{n}}, \end{aligned}$$

where the last inequality holds on  $\Omega_2$  due to Lemma S.2.2 and Assumption 3.4.

Next, we focus on  $I_2$ . We will decompose it into two parts: one will correspond to the first  $k^*$  eigenvectors and eigenvalues, while the other will correspond to the rest  $(r - k^*)$ . Let us split

$$\hat{\mathbf{U}} = [\hat{\mathbf{U}}_{\leq k^*} \hat{\mathbf{U}}_{> k^*}] \quad \text{and} \quad \hat{\Lambda} = \begin{bmatrix} \hat{\Lambda}_{\leq k^*} & \mathbb{O}_{k^* \times (r-k^*)} \\ \mathbb{O}_{(r-k^*) \times k^*} & \hat{\Lambda}_{> k^*} \end{bmatrix},$$

where  $\hat{\Lambda}_{\leq k^*} \in \mathbb{R}^{k^* \times k^*}$ ,  $\hat{\mathbf{U}}_{\leq k^*} \in \mathbb{R}^{d \times k^*}$  correspond to the first  $k^*$  eigenvalues and eigenvectors, while  $\hat{\Lambda}_{> k^*} \in \mathbb{R}^{(r-k^*) \times (r-k^*)}$ ,  $\hat{\mathbf{U}}_{> k^*} \in \mathbb{R}^{d \times (r-k^*)}$  correspond to the rest. Also let  $\boldsymbol{\xi} = [\boldsymbol{\xi}_{\leq k^*}^\top \boldsymbol{\xi}_{> k^*}^\top]^\top$  with  $\boldsymbol{\xi}_{\leq k^*} \in \mathbb{R}^{k^*}$  and  $\boldsymbol{\xi}_{> k^*} \in \mathbb{R}^{r-k^*}$ . Then

$$\begin{aligned} &\hat{\mathbf{U}} \hat{\Lambda}^{-1} \text{SOFT}_{\tau'} [\hat{\Lambda} \hat{\mathbf{U}}^\top \beta + \boldsymbol{\xi}] - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \beta = \\ &= \hat{\mathbf{U}}_{\leq k^*} \hat{\Lambda}_{\leq k^*}^{-1} \text{SOFT}_{\tau'} [\hat{\Lambda}_{\leq k^*} \hat{\mathbf{U}}_{\leq k^*}^\top \beta + \boldsymbol{\xi}_{\leq k^*}] - \hat{\mathbf{U}}_{\leq k^*} \hat{\mathbf{U}}_{\leq k^*}^\top \beta \\ &\quad + \hat{\mathbf{U}}_{> k^*} \hat{\Lambda}_{> k^*}^{-1} \text{SOFT}_{\tau'} [\hat{\Lambda}_{> k^*} \hat{\mathbf{U}}_{> k^*}^\top \beta + \boldsymbol{\xi}_{> k^*}] - \hat{\mathbf{U}}_{> k^*} \hat{\mathbf{U}}_{> k^*}^\top \beta. \end{aligned}$$

Again applying  $\|\mathbf{a} + \mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$  we obtain

$$\begin{aligned} \frac{I_2}{2} &= \|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}} \widehat{\mathbf{\Lambda}}^{-1} \text{SOFT}_{\tau'} [\widehat{\mathbf{\Lambda}} \widehat{\mathbf{U}}^\top \boldsymbol{\beta} + \boldsymbol{\xi}] - \mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}} \widehat{\mathbf{U}}^\top \boldsymbol{\beta}\|_2^2 \\ &\leq 2\|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1} \text{SOFT}_{\tau'} [\widehat{\mathbf{\Lambda}}_{\leq k^*} \widehat{\mathbf{U}}_{\leq k^*}^\top \boldsymbol{\beta} + \boldsymbol{\xi}_{\leq k^*}] - \mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{U}}_{\leq k^*}^\top \boldsymbol{\beta}\|_2^2 \\ &\quad + 2\|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{> k^*} \widehat{\mathbf{\Lambda}}_{> k^*}^{-1} \text{SOFT}_{\tau'} [\widehat{\mathbf{\Lambda}}_{> k^*} \widehat{\mathbf{U}}_{> k^*}^\top \boldsymbol{\beta} + \boldsymbol{\xi}_{> k^*}] - \mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{> k^*} \widehat{\mathbf{U}}_{> k^*}^\top \boldsymbol{\beta}\|_2^2 \\ &=: I_3 + I_4. \end{aligned}$$

So, to upper bound  $I_2$  we will upper bound  $I_3$  and  $I_4$  separately.

Consider  $I_4$ . Denote

$$\boldsymbol{\gamma} \stackrel{\text{def}}{=} \widehat{\mathbf{\Lambda}}_{> k^*}^{-1} \text{SOFT}_{\tau'} [\widehat{\mathbf{\Lambda}}_{> k^*} \widehat{\mathbf{U}}_{> k^*}^\top \boldsymbol{\beta} + \boldsymbol{\xi}_{> k^*}] - \widehat{\mathbf{U}}_{> k^*}^\top \boldsymbol{\beta} \in \mathbb{R}^{r-k^*}.$$

Let us analyze  $j$ -th component  $\gamma_j$ , for  $j \in [r - k^*]$ , using the definition of  $\text{SOFT}_{\tau'}[\cdot]$ . We have two cases:

- If  $|\widehat{\lambda}_{j+k^*}^{1/2} \widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta} + \xi_{j+k^*}| > \tau'$ , then  $\gamma_j = \widehat{\lambda}_{j+k^*}^{-1/2} (\xi_{j+k^*} \pm \tau')$  (the actual sign will play no role). Since by Lemma S.2.1 on  $\Omega_1$

$$\widehat{\lambda}_{j+k^*}^{1/2} |\widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta}| \geq \tau' - |\xi_{j+k^*}| \geq \tau'/2,$$

we have

$$|\gamma_j| \leq \widehat{\lambda}_{j+k^*}^{-1/2} (|\xi_{j+k^*}| + \tau') \leq \widehat{\lambda}_{j+k^*}^{-1/2} \cdot \frac{3\tau'}{2} \leq 3|\widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta}|.$$

- If  $|\widehat{\lambda}_{j+k^*}^{1/2} \widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta} + \xi_{j+k^*}| \leq \tau'$ , then  $\gamma_j = -\widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta}$ , and we directly get

$$|\gamma_j| = |\widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta}|.$$

In any case,  $|\gamma_j| \leq 3|\widehat{\mathbf{u}}_{j+k^*}^\top \boldsymbol{\beta}|$  for all  $j \in [r - k^*]$ , and therefore  $\|\boldsymbol{\gamma}\|_2^2 \leq 9\|\boldsymbol{\beta}\|_2^2$  on  $\Omega_1$ . Hence,

$$\begin{aligned} \frac{I_4}{2} &= \|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{> k^*} \boldsymbol{\gamma}\|_2^2 = \boldsymbol{\gamma}^\top \widehat{\mathbf{U}}_{> k^*}^\top \boldsymbol{\Sigma} \widehat{\mathbf{U}}_{> k^*} \boldsymbol{\gamma} = \boldsymbol{\gamma}^\top \widehat{\mathbf{U}}_{> k^*}^\top (\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}) \widehat{\mathbf{U}}_{> k^*} \boldsymbol{\gamma} + \boldsymbol{\gamma}^\top \widehat{\mathbf{U}}_{> k^*}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{U}}_{> k^*} \boldsymbol{\gamma} \\ &\leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \|\widehat{\mathbf{U}}_{> k^*} \boldsymbol{\gamma}\|_2^2 + \boldsymbol{\gamma}^\top \widehat{\boldsymbol{\Lambda}}_{> k^*}^2 \boldsymbol{\gamma} \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \|\boldsymbol{\gamma}\|_2^2 + \widehat{\lambda}_{k^*+1} \|\boldsymbol{\gamma}\|_2^2. \end{aligned}$$

We bound the first term on the right-hand side on  $\Omega_2$  by Lemma S.2.2, and for the second term on  $\Omega_3$  holds  $\widehat{\lambda}_{k^*+1} \leq \widehat{\lambda}_{k^*} \leq (1 + C\epsilon)\lambda_{k^*} \leq C'\lambda_{k^*}$  due to Lemma S.2.4 and Assumption 3.4. Taking into account  $\|\boldsymbol{\gamma}\|_2 \leq C\|\boldsymbol{\beta}\|_2$  on  $\Omega_1$ , we get on  $\Omega_1 \cap \Omega_2 \cap \Omega_3$

$$I_4 \leq C\|\boldsymbol{\Sigma}\| \|\boldsymbol{\beta}\|_2^2 \left( \sqrt{\frac{\mathbf{r}^{\text{eff}}[\boldsymbol{\Sigma}] + \log(1/\delta)}{n}} + \frac{\lambda_{k^*}}{\lambda_1} \right).$$

Finally, it is left to bound  $I_3$ . Denote

$$\widehat{\boldsymbol{\omega}} \stackrel{\text{def}}{=} \text{SOFT}_{\tau'} [\widehat{\mathbf{\Lambda}}_{\leq k^*} \widehat{\mathbf{U}}_{\leq k^*}^\top \boldsymbol{\beta} + \boldsymbol{\xi}_{\leq k^*}] - \widehat{\mathbf{\Lambda}}_{\leq k^*} \widehat{\mathbf{U}}_{\leq k^*}^\top \boldsymbol{\beta} \in \mathbb{R}^{k^*}.$$

Then,

$$\frac{I_3}{2} = \|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1} \widehat{\boldsymbol{\omega}}\|_2^2 \leq \|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1}\|^2 \|\widehat{\boldsymbol{\omega}}\|_2^2.$$

An upper bound on  $\|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1}\|$  is provided in the next lemma.

LEMMA S.2.5. Suppose Assumption 3.2 holds. Then on  $\Omega_3$  holds

$$\|\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1}\| \leq C.$$

REMARK S.2.1. The previous lemma is the only place where we use  $k^* = (\epsilon \log(1/\epsilon))^{-2/3}$ . The rest of the proof would go through if  $k^*$  was defined as  $(\epsilon \log(1/\epsilon))^{-1}$ .

REMARK S.2.2. Interestingly, a closely related to  $\mathbf{A}\mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1}$  matrix appears also in Bartlett et al. (2020). The main difficulty of their proof is to find regimes of eigenvalues such that for  $\mathbf{C}$  defined as

$$\mathbf{C} \stackrel{\text{def}}{=} (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X} \Sigma \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}$$

holds  $\text{Tr}[\mathbf{C}] = o(1)$  as  $n \rightarrow \infty$ . Using SVD  $n^{-1/2} \mathbf{X} = \widehat{\mathbf{V}} \widehat{\mathbf{\Lambda}} \widehat{\mathbf{U}}^\top$  one can show

$$\text{Tr}[\mathbf{C}] = \frac{1}{n} \text{Tr}[\widehat{\mathbf{\Lambda}}^{-1} \widehat{\mathbf{U}}^\top \mathbf{U} \mathbf{\Lambda}^2 \mathbf{U}^\top \widehat{\mathbf{U}} \widehat{\mathbf{\Lambda}}^{-1}],$$

while in Lemma S.2.5 we essentially upper bound the operator norm of somewhat simpler (in a sense that we truncate the sample eigenvalues and eigenvector beyond  $k^*$ -th) matrix  $\widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1} \widehat{\mathbf{U}}_{\leq k^*}^\top \mathbf{U} \mathbf{\Lambda}^2 \mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\mathbf{\Lambda}}_{\leq k^*}^{-1}$ . The latter task turns out to be much easier and does not require specific regimes of eigenvalues, unlike the former one.

To deal with  $\|\widehat{\omega}\|_2^2$ , we first state the following lemma which has two parts, one of which will help to conclude the proof of claim (i), and the other one will be useful for claim (ii).

LEMMA S.2.6. On  $\Omega_3$  holds

(i)

$$\|\widehat{\mathbf{\Lambda}}_{\leq k^*} \widehat{\mathbf{U}}_{\leq k^*}^\top \beta - \mathbf{\Lambda}_{\leq k^*} \mathbf{U}_{\leq k^*}^\top \beta\|_2^2 \leq C \|\Sigma\| \|\beta\|_2^2 \epsilon \left( r^{\text{eff}}[\Sigma] + \epsilon \sum_{j=1}^{k^*} \frac{\lambda_j j^2}{\lambda_1} \right).$$

(ii)

$$\|\widehat{\mathbf{\Lambda}}_{\leq k^*} \widehat{\mathbf{U}}_{\leq k^*}^\top \beta - \mathbf{\Lambda}_{\leq k^*} \mathbf{U}_{\leq k^*}^\top \beta\|_\infty \leq \frac{\tau^\infty}{2} \stackrel{\text{def}}{=} C \|\Sigma\|^{1/2} \|\beta\|_2 \epsilon^{1/2} \max_{j \in [k^*]} \left( \frac{\lambda_j (1 + \epsilon j^2)}{\lambda_1} \right)^{1/2}.$$

So, to deal with part (ii) of Theorem 3.3, we notice that our thresholding level  $\tau' = \bar{\tau} = \tau + \tau^\infty$ , where  $\tau$  from Lemma S.2.1 is responsible for the noise and  $\tau^\infty$  from Lemma S.2.6 is responsible for the estimation of the eigenvalues and eigenvectors. Now we analyze the components  $\widehat{\omega}_j$  of  $\widehat{\omega}$  for  $j \in [k^*]$ .

- If  $|\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \beta + \xi_j| > \bar{\tau}$ , then  $\widehat{\omega}_j = \xi_j \pm \bar{\tau}$  (the sign again doesn't matter). Moreover, due to Lemma S.2.1 and Lemma S.2.6 (ii) on  $\Omega_1 \cap \Omega_3$

$$|\lambda_j^{1/2} \mathbf{u}_j^\top \beta| \geq \bar{\tau} - |\xi_j| - |\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \beta - \lambda_j^{1/2} \mathbf{u}_j^\top \beta| \geq \bar{\tau} - \frac{\tau}{2} - \frac{\tau^\infty}{2} = \frac{\bar{\tau}}{2}.$$

Hence,

$$|\widehat{\omega}_j| \leq 3 \min(\bar{\tau}, |\lambda_j^{1/2} \mathbf{u}_j^\top \beta|).$$

- If  $|\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \boldsymbol{\beta} + \xi_j| \leq \bar{\tau}$ , then  $\widehat{\omega}_j = -\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \boldsymbol{\beta}$ . Furthermore, again by Lemma S.2.1 and Lemma S.2.6 (ii) on  $\Omega_1 \cap \Omega_3$

$$|\lambda_j^{1/2} \mathbf{u}_j^\top \boldsymbol{\beta}| \leq \bar{\tau} + |\xi_j| + |\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \boldsymbol{\beta} - \lambda_j^{1/2} \mathbf{u}_j^\top \boldsymbol{\beta}| \leq \bar{\tau} + \frac{\tau}{2} + \frac{\tau^\infty}{2} = \frac{3\bar{\tau}}{2}.$$

Thus,

$$|\widehat{\omega}_j| \leq 3 \min(\bar{\tau}, |\lambda_j^{1/2} \mathbf{u}_j^\top \boldsymbol{\beta}|).$$

In both cases,  $|\widehat{\omega}_j| \leq 3 \min(\bar{\tau}, |\lambda_j^{1/2} \mathbf{u}_j^\top \boldsymbol{\beta}|)$ , and based on the same derivation as in the proof of Theorem 3.1, we obtain on  $\Omega_1 \cap \Omega_3$

$$\|\widehat{\boldsymbol{\omega}}\|_2^2 \leq 9 \inf_{q \in [0,2]} \left\{ \bar{\tau}^{2-q} \|\mathbf{A}_{\leq k^*} \mathbf{U}_{\leq k^*}^\top \boldsymbol{\beta}\|_q^q \right\}.$$

For part (i) we act slightly differently. Now  $\tau' = \tau$ . We decompose

$$\widehat{\boldsymbol{\omega}} = \boldsymbol{\omega} + \Delta \boldsymbol{\omega},$$

where

$$\boldsymbol{\omega} \stackrel{\text{def}}{=} \text{SOFT}_\tau \left[ \mathbf{A}_{\leq k^*} \mathbf{U}_{\leq k^*}^\top \boldsymbol{\beta} + \boldsymbol{\xi}_{\leq k^*} \right] - \mathbf{A}_{\leq k^*} \mathbf{U}_{\leq k^*}^\top \boldsymbol{\beta} \in \mathbb{R}^{k^*}.$$

By Lemma S.2.6 (i) it is easy to bound on  $\Omega_3$

$$\|\Delta \boldsymbol{\omega}\|_2^2 = \|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_2^2 \leq C \|\boldsymbol{\Sigma}\| \|\boldsymbol{\beta}\|_2^2 \epsilon \left( r^{\text{eff}}[\boldsymbol{\Sigma}] + \epsilon \sum_{j=1}^{k^*} \frac{\lambda_j j^2}{\lambda_1} \right).$$

The norm  $\|\boldsymbol{\omega}\|_2^2$  with thresholding at level  $\tau$  can be bounded as in the proof of Theorem 3.1: on  $\Omega_1$

$$\|\boldsymbol{\omega}\|_2^2 \leq 9 \inf_{q \in [0,2]} \left\{ \tau^{2-q} \|\mathbf{A}_{\leq k^*} \mathbf{U}_{\leq k^*}^\top \boldsymbol{\beta}\|_q^q \right\},$$

which, together with the bound on  $\|\Delta \boldsymbol{\omega}\|_2^2$ , gives bound on  $\|\widehat{\boldsymbol{\omega}}\|_2^2$  on  $\Omega_1 \cap \Omega_3$ .

Putting all the bounds for  $I_1$  and  $I_2$  (in particular, for  $I_3$  and  $I_4$ ) together on the intersection of high probability events  $\Omega_1 \cap \Omega_2 \cap \Omega_3$ , adjusting  $\delta \rightarrow \delta/3$  so that the intersection has probability at least  $1 - \delta$ , we conclude the proof.  $\square$

**S.2.3. Proof of Theorem 4.1.** We first reduce the general linear regression model to the Gaussian sequence model, and then apply some classical results from the literature. Our original linear regression problem (restricted to the Gaussian noise case)

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_r)$$

can be rewritten in the canonical form

$$\mathbf{Y} = \sqrt{n} \widehat{\mathbf{V}} \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_r)$$

with  $\boldsymbol{\theta} = \widehat{\mathbf{A}} \widehat{\mathbf{U}}^\top \boldsymbol{\beta} \in \mathbb{R}^r$ , and then as the Gaussian sequence model

$$\frac{\widehat{\mathbf{V}}^\top \mathbf{Y}}{\sqrt{n}} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_r).$$

Recall that the joint effective dimension and the signal-to-noise ratio can be expressed in terms of the canonical parameter as  $D_{q,r}^{\text{eff}}(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\beta}) = \|\boldsymbol{\theta}\|_q^q / \|\boldsymbol{\theta}\|_2^q$  and  $\text{SNR} = \|\boldsymbol{\theta}\|_2 / \sigma$ . Hence, the parameter space  $\mathcal{P}^{\mathbf{X}}(q, D, S)$  for the initial model translates into the parameter space

$$\mathcal{Q}_{r,n}(q, D, S) \stackrel{\text{def}}{=} \{(\boldsymbol{\theta}, \sigma) \in \mathbb{R}^r \times \mathbb{R}_+ : \|\boldsymbol{\theta}\|_q^q / \|\boldsymbol{\theta}\|_2^q \leq D, \|\boldsymbol{\theta}\|_2 / \sigma \geq S\}$$

for the Gaussian sequence model. Also, any estimator  $\tilde{\beta} = \tilde{\beta}(\mathbf{X}, \mathbf{Y})$  in the original problem corresponds to an estimator  $\tilde{\theta} = \hat{\Lambda} \hat{\mathbf{U}}^\top \tilde{\beta}(\mathbf{X}, \mathbf{Y})$ , and since  $\mathbf{X}$  is a fixed known design, we can write  $\tilde{\theta} = \tilde{\theta}(\hat{\mathbf{V}}^\top \mathbf{Y} / \sqrt{n})$ , so that it is indeed an estimator in the Gaussian sequence model (the reverse is also true). Therefore,

$$\inf_{\tilde{\beta}} \sup_{\mathcal{P}^{\mathbf{X}}(q, D, S)} \mathbb{E} \left[ \frac{\text{MSE}(\tilde{\beta})}{\text{MSE}(0)} \right] = \inf_{\tilde{\theta}} \sup_{\mathcal{Q}_{r,n}(q, D, S)} \mathbb{E} \left[ \frac{\|\tilde{\theta} - \theta\|_2^2}{\|\theta\|_2^2} \right],$$

and to establish the desired minimax lower bounds for the general problem it is enough to study the minimax lower bound for the Gaussian sequence model in the right-hand side, which we will do next.

(i) We take the following subset of  $\mathcal{Q}_{r,n}(q, D, S)$  to prove the minimax lower bound:

$$\mathcal{Q}_{r,n}^{\text{poly}}(q, S) \stackrel{\text{def}}{=} \left\{ (\theta, \sigma) \in \mathbb{R}^r \times \mathbb{R}_+ : |\theta_j| = j^{-1/q}, \sigma = \|\theta\|_2 / S \right\}.$$

It is easy to check that indeed  $\mathcal{Q}_{r,n}^{\text{poly}}(q, S) \subseteq \mathcal{Q}_{r,n}(q, D, S)$  for large enough constant  $D$ . Also, all  $\theta$  from this new set of parameters have the same  $\ell_2$ -norm of constant order, which we denote  $h$  for concreteness, even though its value will not play a role. We can write

$$\inf_{\tilde{\theta}} \sup_{\mathcal{Q}_{r,n}(q, D, S)} \mathbb{E} \left[ \frac{\|\tilde{\theta} - \theta\|_2^2}{\|\theta\|_2^2} \right] \geq \inf_{\tilde{\theta}} \sup_{\mathcal{Q}_{r,n}^{\text{poly}}(q, S)} \mathbb{E} \left[ \frac{\|\tilde{\theta} - \theta\|_2^2}{\|\theta\|_2^2} \right] = \frac{1}{h^2} \inf_{\tilde{\theta}} \sup_{\mathcal{Q}_{r,n}^{\text{poly}}(q, S)} \mathbb{E} \left[ \|\tilde{\theta} - \theta\|_2^2 \right].$$

The minimax risk on the right-hand side is easy to deal with using, for instance, [Johnstone \(2019\)](#). In particular, by Proposition 4.16 and (4.47) of [Johnstone \(2019\)](#) the minimax risk decomposes into the sum of univariate minimax risks, which are given in (4.40) of [Johnstone \(2019\)](#). Thus,

$$\inf_{\tilde{\theta}} \sup_{\mathcal{Q}_{r,n}^{\text{poly}}(q, S)} \mathbb{E} \left[ \|\tilde{\theta} - \theta\|_2^2 \right] \gtrsim \sum_{j=1}^r \min \left( \frac{\sigma}{\sqrt{n}}, j^{-1/q} \right)^2.$$

One subtlety is that the results of [Johnstone \(2019\)](#) that we used are derived for hyperrectangles, i.e. in the definition of  $\mathcal{Q}_{r,n}^{\text{poly}}(q, S)$  we should have  $|\theta_j| \leq j^{-1/q}$  instead of  $|\theta_j| = j^{-1/q}$ . (It is important for us to use equality here, because otherwise  $\|\theta\|_2$  cannot be bounded from below, and it is not clear how to get the minimax lower bound for the relative error.) However, the analysis of their proof shows that the lower bound holds also for  $\mathcal{Q}_{r,n}^{\text{poly}}(q, S)$  defined in our way, since the underlying least favorable prior used to obtain the lower bound for the univariate minimax risk puts mass  $1/2$  at the extremes of the interval, forcing its support to be contained in our parametric set.

The right-hand side of the above display can be computed similarly to the proof of Proposition 3.2 (iii), and we have

$$\sum_{j=1}^r \min \left( \frac{\sigma}{\sqrt{n}}, j^{-1/q} \right)^2 \gtrsim \left( \frac{\sigma}{\sqrt{n}} \right)^{2-q}.$$

Putting this all together, we obtain

$$\inf_{\tilde{\theta}} \sup_{\mathcal{Q}_{r,n}(q, D, S)} \mathbb{E} \left[ \frac{\|\tilde{\theta} - \theta\|_2^2}{\|\theta\|_2^2} \right] \gtrsim \frac{1}{h^2} \left( \frac{\sigma}{\sqrt{n}} \right)^{2-q} = \frac{1}{h^q} \frac{1}{(S^2 n)^{1-q/2}} \asymp \frac{1}{(S^2 n)^{1-q/2}}.$$



(ii) The classical minimax lower bound for sparse linear regression are derived via a reduction to multiple hypothesis testing. In particular, one constructs a specific finite set of hypotheses  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)}\}$  (each of which has the desired sparsity) and using techniques from [Tsybakov \(2009\)](#) shows (see, for instance, [Rigollet \(2019\)](#), Corollary 4.15 together with equivalence of Definition 4.1 and Definition 4.2 by (4.5))

$$\inf_{\tilde{\boldsymbol{\theta}}} \sup_{\substack{\boldsymbol{\theta} \in \mathbb{R}^r \\ \|\boldsymbol{\theta}\|_0 \leq D}} \mathbb{E} \left[ \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right] \gtrsim \inf_{\tilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{(l)}\}_{l=1}^M} \mathbb{E} \left[ \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right] \gtrsim \sigma^2 \frac{D \log(er/D)}{n}.$$

To get a minimax lower bound for the relative error, we need to make one slight modification. The construction of  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)}\}$  is based on the sparse Varshamov-Gilbert lemma (e.g. Lemma 4.14 in [Rigollet \(2019\)](#)): it produces binary vectors  $\boldsymbol{\omega}^{(l)} \in \{0, 1\}^r$ ,  $l \in [M]$ , satisfying some properties, and then one sets  $\boldsymbol{\theta}^{(l)} = C(r, n, \sigma, \dots) \boldsymbol{\omega}^{(l)}$  for all  $l \in [m]$ , where  $C(r, n, \sigma, \dots)$  is carefully chosen and may depend on  $r, n, \sigma$ , etc. With a simple linear transform  $\boldsymbol{\omega} \rightarrow 2\boldsymbol{\omega} - \mathbf{e}$  (here  $\mathbf{e} = [1, \dots, 1]^\top \in \mathbb{R}^r$ ) we modify the binary vectors produced by the Varshamov-Gilbert lemma so that they belong to  $\{-1, 1\}^r$  after this modification. This forces the transformed vectors  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)}\}$  to have the same  $\ell_2$ -norm without changing the essence of the argument. Now we can deal with the relative errors. Define for concreteness  $h = \|\boldsymbol{\theta}^{(1)}\|_2 = \dots = \|\boldsymbol{\theta}^{(M)}\|_2$ , and write

$$\begin{aligned} \inf_{\tilde{\boldsymbol{\theta}}} \sup_{\mathcal{Q}_{r,n}(0,D,S)} \mathbb{E} \left[ \frac{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2}{\|\boldsymbol{\theta}\|_2^2} \right] &\gtrsim \inf_{\tilde{\boldsymbol{\theta}}} \sup_{\substack{\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{(l)}\}_{l=1}^M \\ \sigma = \|\boldsymbol{\theta}\|_2/S}} \mathbb{E} \left[ \frac{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2}{\|\boldsymbol{\theta}\|_2^2} \right] \\ &\gtrsim \frac{1}{h^2} \inf_{\tilde{\boldsymbol{\theta}}} \sup_{\substack{\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{(l)}\}_{l=1}^M \\ \sigma = h/S}} \mathbb{E} \left[ \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right] \gtrsim \frac{1}{h^2} \frac{h^2}{S^2} \frac{D \log(er/D)}{n} = \frac{D \log(er/D)}{S^2 n}, \end{aligned}$$

as desired.

**S.2.4. Proof of Theorem 5.1.** Similarly to the proof of Theorem 3.1, we write for  $\hat{\beta}$  from (2.2)

$$\text{MSE}(\hat{\beta}) = \left\| \hat{\Lambda}^{-\varphi} \mathbf{T}_\tau \left[ \hat{\Lambda}^\varphi(\boldsymbol{\theta} + \boldsymbol{\xi}) \right] - \boldsymbol{\theta} \right\|_2^2 = \sum_{j=1}^r \left| \hat{\lambda}_j^{-\varphi/2} \mathbf{T}_\tau[\hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j)] - \theta_j \right|^2.$$

For each individual term we apply the following two bounds. On one hand,

$$\begin{aligned} \left| \hat{\lambda}_j^{-\varphi/2} \mathbf{T}_\tau[\hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j)] - \theta_j \right| &= \left| \hat{\lambda}_j^{-\varphi/2} \left( \mathbf{T}_\tau[\hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j)] - \hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j) \right) + \xi_j \right| \\ &\leq \hat{\lambda}_j^{-\varphi/2} \tau + |\xi_j| \leq \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} + \frac{1}{2} \right) \sigma \rho \leq \frac{3}{2} \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, \end{aligned}$$

where the first inequality uses property (ii) from Definition 2.2 and the triangle inequality, and the second inequality holds on  $\Omega_1$  by Lemma S.2.1. On the other hand,

$$\begin{aligned} \left| \hat{\lambda}_j^{-\varphi/2} \mathbf{T}_\tau[\hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j)] - \theta_j \right| &\leq \hat{\lambda}_j^{-\varphi/2} \left| \mathbf{T}_\tau[\hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j)] \right| + |\theta_j| \\ &\leq \hat{\lambda}_j^{-\varphi/2} \cdot c |\hat{\lambda}_j^{\varphi/2} \theta_j| + |\theta_j| = (c+1) |\theta_j|, \end{aligned}$$

where the second inequality is due to property (i) from Definition 2.2 applied to  $z = \hat{\lambda}_j^{\varphi/2}(\theta_j + \xi_j)$  and  $z' = \hat{\lambda}_j^{\varphi/2} \theta_j$  satisfying  $|z - z'| = \hat{\lambda}_j^{\varphi/2} |\xi_j| \leq \hat{\lambda}_1^{\varphi/2} \sigma \rho / 2 = \tau / 2$  on  $\Omega_1$

by Lemma S.2.1. Thus, on  $\Omega_1$

$$\text{MSE}(\hat{\beta}) = \left\| \hat{\Lambda}^{-\varphi} \mathsf{T}_\tau \left[ \hat{\Lambda}^\varphi (\boldsymbol{\theta} + \boldsymbol{\xi}) \right] - \boldsymbol{\theta} \right\|_2^2 \lesssim \sum_{j=1}^r \min \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, |\theta_j| \right)^2.$$

Now we choose  $\mathsf{T}_\tau[\cdot] = \text{SOFT}_\tau[\cdot]$  to prove the desired lower bound. We again consider two cases:

- If  $\left| \hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j) \right| > \tau$ , then on  $\Omega_1$  using Lemma S.2.1

$$\begin{aligned} & \left| \hat{\lambda}_j^{-\varphi/2} \text{SOFT}_\tau[\hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j)] - \theta_j \right| \\ &= \left| \hat{\lambda}_j^{-\varphi/2} \left( \text{SOFT}_\tau[\hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j)] - \hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j) \right) + \xi_j \right| \\ &\geq \hat{\lambda}_j^{-\varphi/2} \tau - |\xi_j| \geq \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} - \frac{1}{2} \right) \sigma \rho \geq \frac{1}{2} \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho. \end{aligned}$$

- If  $\left| \hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j) \right| \leq \tau$ , then

$$\left| \hat{\lambda}_j^{-\varphi/2} \text{SOFT}_\tau[\hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j)] - \theta_j \right| = |\hat{\lambda}_j^{-\varphi/2} \cdot 0 - \theta_j| = |\theta_j|$$

In any case,

$$\left| \hat{\lambda}_j^{-\varphi/2} \text{SOFT}_\tau[\hat{\lambda}_j^{\varphi/2} (\theta_j + \xi_j)] - \theta_j \right| \gtrsim \min \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, |\theta_j| \right),$$

implying the matching lower bound.

**S.2.5. Proof of Corollary 5.2.** We need to upper bound the right-hand side of the inequality obtained in Theorem 5.1. Let us define the following auxiliary set:

$$\mathcal{P} \stackrel{\text{def}}{=} \left\{ j \in [r] \mid \hat{\lambda}_j \geq \hat{\lambda}_1 \rho^{2/(2+\varphi)} \right\}.$$

We first bound

$$\sum_{j \notin \mathcal{P}} \min \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, |\theta_j| \right)^2 \leq \sum_{j \notin \mathcal{P}} |\theta_j|^2 = \sum_{j \notin \mathcal{P}} \hat{\lambda}_j (\hat{\mathbf{u}}_j^\top \boldsymbol{\beta})^2 \leq \hat{\lambda}_1 \|\boldsymbol{\beta}\|_2^2 \rho^{2/(2+\varphi)},$$

where we used only the definition of  $\mathcal{P}$  (more specifically its complement). Then, we bound

$$\sum_{j \in \mathcal{P}} \min \left( \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho, |\theta_j| \right)^2 \leq \sum_{j \in \mathcal{P}} \frac{\hat{\lambda}_1^{\varphi/2}}{\hat{\lambda}_j^{\varphi/2}} \sigma \rho |\theta_j| \leq \sigma \rho^{-\frac{2}{2+\varphi} \cdot \frac{\varphi}{2} + 1} \sum_{j \in \mathcal{P}} |\theta_j| \leq \sigma \rho^{2/(2+\varphi)} \|\boldsymbol{\theta}\|_1,$$

where we again used the definition of  $\mathcal{P}$ . Now we apply  $\|\boldsymbol{\theta}\|_1 \leq \|\hat{\boldsymbol{\Sigma}}\|^{1/2} \|\boldsymbol{\beta}\|_2 \text{r}^{\text{eff}}[\hat{\boldsymbol{\Sigma}}]^{1/2}$ , and adding the above two inequalities yields the desired statement.

**S.2.6. Proof of Theorem 6.1.** Using standard “empirical risk minimization” reasoning, we write

$$\frac{1}{L} \sum_{l=1}^L \mathbb{E} \left[ \left( y - \mathbf{x}^\top \hat{\boldsymbol{\beta}}_{\tau^{cv}}^{(l)} \right)^2 \right] \leq \frac{1}{L} \sum_{l=1}^L \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} \left( y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{\tau^{cv}}^{(l)} \right)^2$$

$$\begin{aligned}
& + \frac{1}{L} \sum_{l=1}^L \sup_{\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}} \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right| \\
& \leq \frac{1}{L} \sum_{l=1}^L \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \hat{\beta}_{\tau^{oracle}}^{(l)})^2 \\
& \quad + \frac{1}{L} \sum_{l=1}^L \sup_{\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}} \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right| \\
& \leq \frac{1}{L} \sum_{l=1}^L \mathbb{E} \left[ (y - \mathbf{x}^\top \hat{\beta}_{\tau^{oracle}}^{(l)})^2 \right] \\
& \quad + \frac{2}{L} \sum_{l=1}^L \sup_{\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}} \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right|.
\end{aligned}$$

Here in the second inequality we used the definition of  $\tau^{cv}$ , namely the fact that it minimizes the cross-validation error. The expectations in the above expressions are over  $(\mathbf{x}, y)$  only, so using  $\mathbb{E} \left[ (y - \mathbf{x}^\top \hat{\beta}_\tau^{(l)})^2 \right] = \text{PE}(\hat{\beta}_\tau^{(l)}) + \sigma^2$  the only thing left is to bound the supremums in the right-hand side. We will do this for each  $l \in [L]$  similarly, so from now on we fix  $l \in [L]$ . Note that due to the structure of our estimator  $\hat{\beta}_\tau^{(l)}$ , the set  $\{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}$  contains at most  $(r+1)$  distinct estimators, each derived from the training sample of the  $l$ -th fold  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ , and thus independent of the validation set of the  $l$ -th fold  $\{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{B}_l}$ . Working conditionally on  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ , we can bound

$$\left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right|$$

with high probability for each single  $\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}$  (which are treated as deterministic vectors), and the rest of the proof will easily follow. Let us focus on an arbitrary  $\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}$ .

To apply some concentration results, we first show that  $(y_i - \mathbf{x}_i^\top \beta')$  are sub-Weibull random variables with parameter  $\alpha/2$ . Indeed,  $\|\varepsilon_i\|_{\psi_\alpha} \leq \sigma$  by Assumption 3.1 and

$$\begin{aligned}
\|\mathbf{x}_i^\top (\beta - \beta')\|_{\psi_\alpha} & \leq \|\mathbf{x}_i^\top (\beta - \beta')\|_{\psi_2} = \|(\Sigma^{-1/2} \mathbf{x}_i)^\top \Sigma^{1/2} (\beta - \beta')\|_{\psi_2} \\
& = \left\| (\Sigma^{-1/2} \mathbf{x}_i)^\top \frac{\Sigma^{1/2} (\beta - \beta')}{\|\Sigma^{1/2} (\beta - \beta')\|_2} \right\|_{\psi_2} \cdot \sqrt{(\beta - \beta')^\top \Sigma (\beta - \beta')} \\
& \lesssim \sqrt{(\beta - \beta')^\top \Sigma (\beta - \beta')},
\end{aligned}$$

where in the first inequality we used the monotonicity of the Orlicz norm w.r.t. parameter  $\alpha$ , and in the last inequality we used Assumption 3.2 together with the fact that the inner product of a sub-Gaussian vector with a unit vector is a sub-Gaussian random variable. By properties of the Orlicz norm we further have

$$\begin{aligned}
\|y_i - \mathbf{x}_i^\top \beta'\|_{\psi_\alpha} & = \|\mathbf{x}_i^\top (\beta - \beta') + \varepsilon_i\|_{\psi_\alpha} \\
& \lesssim \|\mathbf{x}_i^\top (\beta - \beta')\|_{\psi_\alpha} + \|\varepsilon_i\|_{\psi_\alpha} \lesssim \sqrt{(\beta - \beta')^\top \Sigma (\beta - \beta')} + \sigma.
\end{aligned}$$

and due to Proposition D.2 of [Kuchibhotla and Chakraborty \(2018\)](#)

$$\begin{aligned} \left\| (y_i - \mathbf{x}_i^\top \beta')^2 \right\|_{\psi_{\alpha/2}} &\leq \|y_i - \mathbf{x}_i^\top \beta'\|_{\psi_\alpha}^2 \\ &\lesssim (\beta - \beta')^\top \Sigma (\beta - \beta') + \sigma^2 \lesssim \beta^\top \Sigma \beta + \beta'^\top \Sigma \beta' + \sigma^2. \end{aligned}$$

After subtracting the expectation, the same bound holds for the centered random variable, but with a different hidden constant. Now we readily apply Theorem 3.1 of [Kuchibhotla and Chakraborty \(2018\)](#) with  $X_i \stackrel{\text{def}}{=} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E}[(y - \mathbf{x}^\top \beta')^2]$ ,  $a_i = 1/|\mathcal{B}_l|$  for all  $i \in [|\mathcal{B}_l|]$  and parameter  $\alpha/2$ :

$$\begin{aligned} \text{(S.1)} \quad \mathbb{P} \left[ \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E}[(y - \mathbf{x}^\top \beta')^2] \right| \geq \right. \\ \left. \gtrsim (\beta^\top \Sigma \beta + \beta'^\top \Sigma \beta' + \sigma^2) \left( \sqrt{\frac{t}{|\mathcal{B}_l|}} + \frac{t^{2/\alpha}}{|\mathcal{B}_l|} \right) \mid \{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l} \right] \leq 2e^{-t} \end{aligned}$$

for all  $t \geq 0$ . Now we need to carefully integrate out  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ , and before this, we need to bound  $\beta'^\top \Sigma \beta'$  with high probability over  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ .

In this paragraph, to keep the notation light, we drop the superscript  $(l)$ , but keep in mind that we work with the sample  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ , and for the purposes of this paragraph the quantities  $\hat{\Sigma}, \hat{\Lambda}, \hat{\mathbf{U}}, \theta, \xi$  correspond to the training sample of  $l$ -th fold  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$  only. With any  $\tau \geq 0$ , we have the following:

$$\begin{aligned} \beta'^\top \Sigma \beta' &= \hat{\beta}_\tau^{(l)\top} \Sigma \hat{\beta}_\tau^{(l)} \\ &= \text{SOFT}_\tau[\theta_{\leq k^*} + \xi_{\leq k^*}]^\top \hat{\Lambda}_{\leq k^*}^{-1} \hat{\mathbf{U}}_{\leq k^*}^\top \mathbf{U} \Lambda^2 \mathbf{U}^\top \hat{\mathbf{U}}_{\leq k^*} \hat{\Lambda}_{\leq k^*}^{-1} \text{SOFT}_\tau[\theta_{\leq k^*} + \xi_{\leq k^*}] \\ &\leq \left\| \Lambda \mathbf{U}^\top \hat{\mathbf{U}}_{\leq k^*} \hat{\Lambda}_{\leq k^*}^{-1} \right\|^2 \left\| \text{SOFT}_\tau[\theta_{\leq k^*} + \xi_{\leq k^*}] \right\|_2^2. \end{aligned}$$

By Lemma S.2.5,

$$\left\| \Lambda \mathbf{U}^\top \hat{\mathbf{U}}_{\leq k^*} \hat{\Lambda}_{\leq k^*}^{-1} \right\| \lesssim 1$$

on some set  $\Omega_3^{(l)}$ , defined similarly to  $\Omega_3$  (which is introduced after Lemma S.2.3) but for  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$  rather than for the whole sample. Also,

$$\left\| \text{SOFT}_\tau[\theta_{\leq k^*} + \xi_{\leq k^*}] \right\|_2^2 \leq \left\| \theta_{\leq k^*} + \xi_{\leq k^*} \right\|_2^2 \lesssim \left\| \theta_{\leq k^*} \right\|_2^2 + \left\| \xi_{\leq k^*} \right\|_2^2$$

with

$$\left\| \theta_{\leq k^*} \right\|_2^2 = \beta^\top \hat{\Sigma} \beta \leq \beta^\top \Sigma \beta + \|\Sigma\| \|\beta\|_2^2 \sqrt{\frac{r^{\text{eff}}[\Sigma] + \log(1/\delta)}{n - |\mathcal{B}_l|}} \lesssim \|\Sigma\| \|\beta\|_2^2$$

by Lemma S.2.2 on  $\Omega_2^{(l)}$  (defined similarly to  $\Omega_2$  after Lemma S.2.2, but for  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ ) and Assumption 3.4, and

$$\left\| \xi_{\leq k^*} \right\|_2^2 \leq r \left\| \xi_{\leq k^*} \right\|_\infty^2 \lesssim \sigma^2 (\log(2r/\delta))^{2/\alpha}$$

by Lemma S.2.1 on  $\Omega_1^{(l)}$  (defined similarly to  $\Omega_1$  after Lemma S.2.1, but for  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ ). Putting this all together,

$$\beta'^\top \Sigma \beta' \lesssim \|\Sigma\| \|\beta\|_2^2 + \sigma^2 (\log(2r/\delta))^{2/\alpha}$$

on  $\Omega_1^{(l)} \cap \Omega_2^{(l)} \cap \Omega_3^{(l)}$ . Note that  $\mathbb{P}[\Omega_1^{(l)} \cap \Omega_2^{(l)} \cap \Omega_3^{(l)}] \geq 1 - 3\delta$ , with the probability taken over the randomness of  $\{(\mathbf{x}_i, y_i)\}_{i \in [n] \setminus \mathcal{B}_l}$ .

Taking the bound in the previous display into account, we integrate (S.1). We split the integral into two parts: the integration over  $\Omega_1^{(l)} \cap \Omega_2^{(l)} \cap \Omega_3^{(l)}$  allows to replace  $\beta'^\top \Sigma \beta'$  inside the conditional probability with its deterministic bound on this set and gives  $2e^{-t}$  in the right-hand side, and the integration over  $(\Omega_1^{(l)} \cap \Omega_2^{(l)} \cap \Omega_3^{(l)})^c$ , where the bound on  $\beta'^\top \Sigma \beta'$  may be violated, adds at most  $3\delta$  to the probability of the bad event. Therefore,

$$\begin{aligned} \mathbb{P} \left[ \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right| \gtrsim \right. \\ \left. \gtrsim \left( \|\Sigma\| \|\beta\|_2^2 + \sigma^2 (\log(2r/\delta))^{2/\alpha} \right) \left( \sqrt{\frac{t}{|\mathcal{B}_l|}} + \frac{t^{2/\alpha}}{|\mathcal{B}_l|} \right) \right] \leq 2e^{-t} + 3\delta. \end{aligned}$$

Now we apply the union bound (recall that the cardinality of  $\{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}$  does not exceed  $r+1$ )

$$\begin{aligned} \mathbb{P} \left[ \sup_{\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}} \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right| \gtrsim \right. \\ \left. \gtrsim \left( \|\Sigma\| \|\beta\|_2^2 + \sigma^2 (\log(2r/\delta))^{2/\alpha} \right) \left( \sqrt{\frac{t}{|\mathcal{B}_l|}} + \frac{t^{2/\alpha}}{|\mathcal{B}_l|} \right) \right] \leq 2(r+1)e^{-t} + 3\delta. \end{aligned}$$

Finally, by yet another application of the union bound we obtain

$$\begin{aligned} \mathbb{P} \left[ \frac{2}{L} \sum_{l=1}^L \sup_{\beta' \in \{\hat{\beta}_\tau^{(l)}\}_{\tau \geq 0}} \left| \frac{1}{|\mathcal{B}_l|} \sum_{i \in \mathcal{B}_l} (y_i - \mathbf{x}_i^\top \beta')^2 - \mathbb{E} \left[ (y - \mathbf{x}^\top \beta')^2 \right] \right| \gtrsim \right. \\ \left. \gtrsim \left( \|\Sigma\| \|\beta\|_2^2 + \sigma^2 (\log(2r/\delta))^{2/\alpha} \right) \left( \sqrt{\frac{t}{|\mathcal{B}_l|}} + \frac{t^{2/\alpha}}{|\mathcal{B}_l|} \right) \right] \leq 2L(r+1)e^{-t} + 3L\delta. \end{aligned}$$

We conclude the proof by picking  $t = \log(2Lr/\delta)$  and adjusting the constants throughout the proof to make sure that the desired result holds with probability  $1 - \delta$ .

### S.3. Additional proofs.

**S.3.1. Proof of Proposition 3.2.** (i) and (ii) follow trivially by taking  $q = 0$  and  $q = \nu$ , respectively. Now we focus on (iii). Define  $j^*$  such that  $|\theta_{(j^*)}| \asymp \sigma\rho$ , i.e. take  $j^* = \lfloor (\sigma\rho/|\theta_{(1)}|)^{-1/a} \rfloor$ .

If  $j^* \geq r$ , then taking  $q = 0$  we have

$$\sum_{j=1}^r \min(\sigma\rho, |\theta_{(j)}|)^2 \asymp r(\sigma\rho)^2 \geq \inf_{q \in [0, 2]} \{ \|\theta\|_q^q (\sigma\rho)^{2-q} \}.$$

For the rest of the proof assume  $j^* < r$ . Then

$$\sum_{j=1}^r \min(\sigma\rho, |\theta_{(j)}|)^2 \asymp \sum_{j=1}^{j^*} (\sigma\rho)^2 + \sum_{j=j^*+1}^r |\theta_{(j)}|^2 \asymp j^*(\sigma\rho)^2 + |\theta_{(1)}| \sum_{j=j^*+1}^r j^{-2a}$$

$$\begin{aligned}
&\asymp \begin{cases} j^*(\sigma\rho)^2 + |\theta_{(1)}| (j^{*-2a+1} - r^{-2a+1}), & a > 1/2, \\ j^*(\sigma\rho)^2 + |\theta_{(1)}| \log(r/j^*), & a = 1/2, \\ j^*(\sigma\rho)^2 + |\theta_{(1)}| (r^{-2a+1} - j^{*-2a+1}), & a < 1/2; \end{cases} \\
&\asymp \begin{cases} |\theta_{(1)}|^{1/a} (\sigma\rho)^{2-1/a}, & a > 1/2, \\ |\theta_{(1)}| \log(er/j^*), & a = 1/2, \\ |\theta_{(1)}| r^{-2a+1}, & a < 1/2; \end{cases}
\end{aligned}$$

At the same time, taking  $q = 1/a \in (0, 2]$  for  $a \geq 1/2$  and  $q = 2$  for  $a < 1/2$ , we get

$$\begin{aligned}
\inf_{q \in [0, 2]} \{ \|\boldsymbol{\theta}\|_q^q (\sigma\rho)^{2-q} \} &\lesssim \begin{cases} (\sigma\rho)^{2-1/a} |\theta_{(1)}|^{1/a} \sum_{j=1}^r (j^{-a})^{1/a}, & a \geq 1/2, \\ |\theta_{(1)}| \sum_{j=1}^r j^{-2a}, & a < 1/2; \end{cases} \\
&\asymp \begin{cases} |\theta_{(1)}|^{1/a} (\sigma\rho)^{2-1/a} \log(r), & a \geq 1/2, \\ |\theta_{(1)}| r^{-2a+1}, & a < 1/2; \end{cases}
\end{aligned}$$

Comparing this to the expressions for  $\sum_{j=1}^r \min(\sigma\rho, |\theta_{(j)}|)^2$  above, we conclude the proof.

**S.3.2. Proof of Lemma S.2.1.** It is straightforward to verify that  $\boldsymbol{\xi}$  is also a sub-Weibull random vector conditionally on  $\mathbf{X}$ :

$$\sup_{\|\mathbf{w}\|_2=1} \left\| \mathbf{w}^\top \frac{\widehat{\boldsymbol{\Lambda}}^{-1} \widehat{\mathbf{U}}^\top \mathbf{X}^\top \boldsymbol{\varepsilon}}{n} \right\|_{\psi_\alpha} \leq \frac{1}{\sqrt{n}} \sup_{\|\mathbf{w}\|_2=1} \|\mathbf{w}^\top \boldsymbol{\varepsilon}\|_{\psi_\alpha} \leq \frac{\sigma}{\sqrt{n}},$$

where the first inequality holds given  $\mathbf{X}$  since  $\|\mathbf{X} \widehat{\mathbf{U}} \widehat{\boldsymbol{\Lambda}}^{-1} \mathbf{w} / \sqrt{n}\|_2 = 1$ , and the last inequality is due to Assumption 3.1. Then, taking  $\mathbf{e}_1, \dots, \mathbf{e}_r$  (the standard basis in  $\mathbb{R}^r$ ) we have

$$\mathbb{P} \left[ |\mathbf{e}_j^\top \boldsymbol{\xi}| \geq t \mid \mathbf{X} \right] \leq 2 \exp \left( - (t\sqrt{n}/\sigma)^\alpha \right) \quad \text{for } j \in [r].$$

Applying the union bound and plugging in  $t = \sigma\rho/2$ , we get the desired.

**S.3.3. Proof of Lemma S.2.3.** The proof is pretty standard and can be found in numerous papers. Fix  $l, l' \in [d]$ . We have

$$\begin{aligned}
\bar{\eta}_{ll'} &= \frac{\mathbf{u}_l^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \mathbf{u}_{l'}}{\sqrt{\lambda_l \lambda_{l'}}} = \mathbf{u}_l^\top (\boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1/2} - \mathbb{I}_d) \mathbf{u}_{l'} \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ (\mathbf{u}_l^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i) \cdot (\mathbf{u}_{l'}^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i) - \mathbb{E} \left[ (\mathbf{u}_l^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i) \cdot (\mathbf{u}_{l'}^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i) \right] \right\}.
\end{aligned}$$

Since by Assumption 3.2  $\boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i$  is sub-Gaussian for  $i \in [n]$ , then by Lemma 2.7.7 of [Vershynin \(2018\)](#)  $(\mathbf{u}_l^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i) \cdot (\mathbf{u}_{l'}^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i)$  is sub-Exponential and by Exercise 2.7.10 of [Vershynin \(2018\)](#) its centered version is also sub-Exponential for  $i \in [n]$ . Bernstein's inequality (e.g. Corollary 2.8.3 of [Vershynin \(2018\)](#)) applied to this centered random variables implies

$$\mathbb{P} [|\bar{\eta}_{ll'}| \geq t] \leq 2 \exp \left( -cn \min \left( \frac{t^2}{C^2}, \frac{t}{C} \right) \right).$$

By union bound,

$$\mathbb{P} \left[ \max_{l, l' \in [d]} |\bar{\eta}_{ll'}| \geq t \right] \leq 2d^2 \exp \left( -cn \min \left( \frac{t^2}{C^2}, \frac{t}{C} \right) \right).$$

Taking  $t = C\sqrt{\log(2d^2/\delta)/n}$  for some other properly chosen  $C$  and using Assumption 3.4 to make sure  $t^2/c^2 \leq t/c$ , we conclude the proof.

**S.3.4. Proof of Lemma S.2.4.** The first part follows from Corollary 2 of Jirak and Wahl (2018), and the second part follows from Lemma 4 of Jirak and Wahl (2018). Conditions (2.1) and  $|\eta_{ll'}| \leq x$  for all  $l, l' \in [d]$  are satisfied since we work on  $\Omega_3$  from Lemma S.2.3 and consider  $j \in [k^*]$  with properly defined  $k^*$ .

**S.3.5. Proof of Lemma S.2.5.** We first apply inequalities

$$\begin{aligned} \|\Lambda \mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \widehat{\Lambda}_{\leq k^*}^{-1}\| &= \|\Lambda \mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \Lambda_{\leq k^*}^{-1} \Lambda_{\leq k^*} \widehat{\Lambda}_{\leq k^*}^{-1}\| \leq \|\Lambda \mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \Lambda_{\leq k^*}^{-1}\| \|\Lambda_{\leq k^*} \widehat{\Lambda}_{\leq k^*}^{-1}\| \\ &\leq \left( \left\| \Lambda \mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \Lambda_{\leq k^*}^{-1} - \begin{bmatrix} \mathbb{I}_{k^*} \\ \mathbb{O}_{(d-k^*) \times k^*} \end{bmatrix} \right\| + 1 \right) \|\Lambda_{\leq k^*} \widehat{\Lambda}_{\leq k^*}^{-1}\|. \end{aligned}$$

The spectral norm of  $\Lambda_{\leq k^*} \widehat{\Lambda}_{\leq k^*}^{-1}$  is easy to control:

$$\|\Lambda_{\leq k^*} \widehat{\Lambda}_{\leq k^*}^{-1}\| = \left( \max_{j \in [k^*]} \frac{\lambda_j}{\widehat{\lambda}_j} \right)^{1/2} \leq \frac{1}{(1 - C\epsilon)^{1/2}} \leq 4,$$

where the first inequality is due to Lemma S.2.4 on  $\Omega_3$  and in the last inequality we assumed that  $\epsilon$  is small enough by Assumption 3.4.

Next, let us focus on  $\Lambda \mathbf{U}^\top \widehat{\mathbf{U}}_{\leq k^*} \Lambda_{\leq k^*}^{-1} - [\mathbb{I}_{k^*} \mathbb{O}_{k^* \times (d-k^*)}]^\top$ , which we denote by  $\mathbf{H}$  for shortness. Denote its columns as  $\mathbf{h}_1, \dots, \mathbf{h}_{k^*}$ . We can bound  $\ell_1$ -norm of each column, using Lemma S.2.4, as

$$\begin{aligned} \|\mathbf{h}_j\|_1 &= \sum_{\substack{l=1 \\ l \neq j}}^d \left| \frac{\lambda_l^{1/2} \mathbf{u}_l^\top \widehat{\mathbf{u}}_j}{\lambda_j^{1/2}} \right| + |\mathbf{u}_j^\top \widehat{\mathbf{u}}_j - 1| = \sum_{\substack{l=1 \\ l \neq j}}^d \left| \frac{\lambda_l^{1/2} \mathbf{u}_l^\top \widehat{\mathbf{u}}_j}{\lambda_j^{1/2}} \right| + \frac{1}{2} \|\widehat{\mathbf{u}}_j - \mathbf{u}_j^\top\|_2^2 \\ &\leq \epsilon \sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_l^{1/2} \lambda_j^{1/2} \lambda_l^{1/2}}{\lambda_j^{1/2} |\lambda_j - \lambda_l|} + C\epsilon^2 \sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_j \lambda_l}{(\lambda_j - \lambda_l)^2} \end{aligned}$$

on  $\Omega_3$ . Applying Jirak and Wahl (2018), inequalities (3.30), or Jirak (2016), Lemma 7.13, together with Assumption 3.3, we get

$$\|\mathbf{h}_j\|_1 \leq \epsilon j \log(j) + C\epsilon^2 j^2.$$

Since  $j \leq k^*$  and  $\epsilon k^* \leq C$ , we have

$$\|\mathbf{h}_j\|_1 \leq C\epsilon j \log(k^*) \quad \text{for all } j \in [k^*].$$

Finally, we have for the Frobenius norm  $\|\mathbf{H}\|_F$  on  $\Omega_3$

$$\|\mathbf{H}\|_F^2 = \sum_{j=1}^{k^*} \|\mathbf{h}_j\|_2^2 \leq C \sum_{j=1}^{k^*} \epsilon^2 j^2 \log(k^*)^2 = C\epsilon^2 k^{*3} \log(k^*)^2 \leq C',$$

where we used the definition of  $k^*$ . The inequality between the spectral and the Frobenius norms completes the proof.



S.3.6. *Proof of Lemma S.2.6.* Fix arbitrary  $j \in [k^*]$ . We have the following chain of inequalities:

$$\begin{aligned}
|\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \boldsymbol{\beta} - \lambda_j^{1/2} \mathbf{u}_j^\top \boldsymbol{\beta}| &\leq \|\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j - \lambda_j^{1/2} \mathbf{u}_j\|_2 \|\boldsymbol{\beta}\|_2 \\
&\leq \|\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j - \lambda_j^{1/2} \widehat{\mathbf{u}}_j + \lambda_j^{1/2} \widehat{\mathbf{u}}_j - \lambda_j^{1/2} \mathbf{u}_j\|_2 \|\boldsymbol{\beta}\|_2 \\
&\leq \left( \left| \widehat{\lambda}_j^{1/2} - \lambda_j^{1/2} \right| \|\widehat{\mathbf{u}}_j\|_2 + \lambda_j^{1/2} \|\widehat{\mathbf{u}}_j - \mathbf{u}_j\|_2 \right) \|\boldsymbol{\beta}\|_2 \\
&\leq \left( |\widehat{\lambda}_j - \lambda_j|^{1/2} + \lambda_j^{1/2} \|\widehat{\mathbf{u}}_j - \mathbf{u}_j\|_2 \right) \|\boldsymbol{\beta}\|_2.
\end{aligned}$$

Applying Lemma S.2.4 we have on  $\Omega_3$

$$\begin{aligned}
|\widehat{\lambda}_j^{1/2} \widehat{\mathbf{u}}_j^\top \boldsymbol{\beta} - \lambda_j^{1/2} \mathbf{u}_j^\top \boldsymbol{\beta}| &\leq C \|\boldsymbol{\beta}\|_2 \left( \lambda_j^{1/2} \epsilon^{1/2} + \lambda_j^{1/2} \epsilon \sqrt{\sum_{\substack{l=1 \\ l \neq j}}^d \frac{\lambda_j \lambda_l}{(\lambda_j - \lambda_l)^2}} \right) \\
&\leq C \|\boldsymbol{\beta}\|_2 \left( \lambda_j^{1/2} \epsilon^{1/2} + \lambda_j^{1/2} \epsilon j \right),
\end{aligned}$$

where in the second inequality we used [Jirak and Wahl \(2018\)](#), inequalities (3.30), or [Jirak \(2016\)](#), Lemma 7.13, together with Assumption 3.3. Taking maximum over  $j \in [k^*]$  we obtain the claim (i), and raising to the square and summing over  $j \in [k^*]$  we get the claim (ii).

## REFERENCES

- BARTLETT, P., LONG, P., LUGOSI, G. and TSIGLER, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*.
- CARDOT, H., MAS, A. and SARDA, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields*, **138**, 325–361.
- DUDOIT, S. and VAN DER LAAN, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statist. Methodol.*, **2**, 2, 131–154.
- JIRAK, M. (2016). Optimal eigen expansions and uniform bounds. *Probab. Theory Related Fields*, **166**, 753–799.
- JIRAK, M. and WAHL, M. (2018). Relative perturbation bounds with applications to empirical covariance operators. *ArXiv:1802.02869*.
- JOHNSTONE, I. M. (2019). Gaussian estimation: Sequence and wavelet models. *Draft*. [statweb.stanford.edu/~imj/GE\\_09\\_16\\_19.pdf](http://statweb.stanford.edu/~imj/GE_09_16_19.pdf).
- KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, **23**, 1, 110–133.
- KUCHIBHOTLA, A. K. and CHAKRABORTTY, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *ArXiv:1804.02605*.
- RIGOLLET, P. (2019). *Lecture notes on High-dimensional statistics*.
- TSYBAKOV, A. B. (2009). Introduction to nonparametric estimation. *Springer Series in Statistics*, Springer, New York.
- VERSHYNIN, R. (2018). High-dimensional probability. An introduction with applications in data science. *Cambridge Series in Statistical and Probabilistic Mathematics*.