Electronic Journal of Statistics
ISSN: 1935-7524

Understanding Best Subset Selection: A Tale of Two C(omplex)ities

Saptarshi Roy[1], **Ambuj Tewari**[1] **and Ziwei Zhu**[2]

[1]*Department of Statistics, University of Michigan, Ann Arbor, USA*

[2]*Quantitative Research, Radix Trading, Chicago, USA*

Abstract: We consider the problem of best subset selection (BSS) under the well known high-dimensional sparse linear regression model. Recently, Guo et al. (2020) [10] showed that the model selection performance of BSS depends on a certain *identifiability margin*, a measure that captures the model discriminative power of BSS under a general correlation structure that is robust to the design dependence, unlike its computational surrogates such as LASSO, SCAD, MCP, etc. Expanding on this, we further broaden the theoretical understanding of BSS in this paper and show that the complexities of the *residualized signals*, the portion of the signals orthogonal to the true active features, and *spurious projections*, describing the projection operators associated with the irrelevant features, also play fundamental roles in characterizing the margin condition for model consistency of BSS. In particular, we establish both necessary and sufficient margin conditions depending only on the identifiability margin and the two complexity measures. We also partially extend our sufficiency result to the case of high-dimensional sparse generalized linear models (GLMs).

MSC2020 subject classifications: Primary 62J05 ; 65C20 ; 62G32 .
Keywords and phrases: High-dimensional statistics, Model consistency, Variable selection.

Contents

1. Introduction

Variable selection in high-dimensional sparse regression has been one of the central topics in statistical research over the past few decades. Consider n observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ following the linear model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{1}$$

where $\{\mathbf{x}_i\}_{i\in[n]}$ are *fixed* p-dimensional feature vectors, $\{\varepsilon_i\}_{i\in[n]}$ are i.i.d. *mean-zero* noise, and the signal vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown but is assumed to have a sparse support. In matrix notation, the observations can be represented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. We consider the standard *high-dimensional sparse* setup where $n < p$, and possibly $n \ll p$, and the vector $\boldsymbol{\beta}$ is sparse in the sense that $\|\boldsymbol{\beta}\|_0 := \sum_{j=1}^p \mathbb{1}(\beta_j \neq 0) = s$, which is much smaller than p. In this paper, we focus on the variable selection problem, i.e., identifying the active set $\mathcal{S} := \{j : \beta_j \neq 0\}$. We primarily use the 0-1 loss, i.e., $\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S})$, to assess the quality of the selected model $\widehat{\mathcal{S}}$.

One of the well-studied methods for variable selection in high-dimensional sparse regression is to penalize the empirical risk by model complexity, thereby encouraging sparse solutions. Specifically, consider

$$\widehat{\boldsymbol{\beta}}^{\text{pen}} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\beta}) + \text{pen}_\lambda(\boldsymbol{\beta}),$$

were $\mathcal{L}(\boldsymbol{\beta})$ is a loss function and $\text{pen}_\lambda(\boldsymbol{\beta})$ is the penalization term that controls the model complexity. Classical methods such as AIC [2, 3], BIC [18], Mallow's C_p [16] use model complexity as penalty term, i.e., ℓ_0-norm of the regression coefficient, to penalize the negative log-likelihood. Although these methods enjoy nice sampling properties [4, 26], such ℓ_0 regularized methods are known to suffer from huge computational bottleneck [9]. This motivated a whole generation of statisticians to develop alternative penalization methods such as LASSO [21], SCAD [8], MC+ [24], and many others that have both strong statistical guarantees and computational expediency.

However, after recent computational advancements in solving BSS [5, 6, 29], there has been growing acknowledgment that BSS enjoys significant statistical superiority over its computational surrogates and has inevitably motivated statisticians to investigate the properties of BSS. For example, through extensive simulations, [11] shows that BSS performs better than LASSO in high signal-to-noise ratio regime in terms of the prediction risk. [12] showed that a wide family of iterative hard thresholding (IHT) algorithms can approximately solve the BSS problem, in the sense that they can achieve similar goodness of fit with the best subset with slight violation of the sparsity constraint. [15] studied the optimal thresholding operator for such iterative thresholding algorithms, which manages to exploit fewer variables than IHT to achieve the same goodness fit as BSS. Recently, [19] proposed an algorithmic framework based on quantile-thresholding that iteratively optimizes ℓ_2-penalized BSS objective function and can achieve model consistency under certain regularity conditions on the design. On the theoretical side, [10] showed that the model selection behavior of BSS does

not explicitly depend on the restricted eigenvalue condition for the design [7, 22], a condition which appears unavoidable (assuming a standard computational complexity conjecture) for any polynomial-time method [27]. Specifically, they show that BSS is robust to design collinearity. Under a particular asymptotic regime and independent design, [17] further established information-theoretic optimality of BSS in terms of precise constants for the signal strength parameter under weak and heterogeneous signal regimes.

In this paper, we also study the variable selection property of BSS and identify novel quantities that are fundamental to understanding the model consistency of BSS. Specifically, we take the geometric alignment of the feature vectors $\{\mathbf{X}_j\}_{j\in[p]}$ into consideration to produce a more refined analysis of BSS, and show that on top of a certain identifiability margin [10], the following two geometric quantities also control the model selection performance of BSS: (a) Geometric complexity of the space of *residualized signals*, and (b) Geometric complexity of *spurious projections*. We show the explicit dependence of these two complexity measures in our main results and demonstrate the interplay between the margin condition and the underlying geometric structure of the features through some illustrative examples. In the process, we also point out the existence of a design that is more favorable to BSS than the orthogonal design, which is commonly believed to be the easiest case for model selection. To the best of our knowledge, this is the first work that identifies the underlying geometric complexity of the feature space as a governing force behind the performance of BSS.

The rest of the paper is organized as follows. In Section 2 we discuss the preliminaries of BSS. Section 3 is devoted to the discussion of the key quantities, i.e., identifiability margin and the two complexities. In particular, Section 3.1-3.3 carefully introduce the notion of identifiability margin and the two novel complexity measures. In Section 3.4, we build intuition for understanding the effect of these two complexities with varying correlation. In Section 4, we present both sufficient (Section 4.1) and necessary (Section 4.3) conditions for model consistency of BSS. We also partially extend our result to GLMs and present a similar sufficiency result for model consistency in Section S2 of the supplementary material.

Notation. Let $\mathbb{R}$ denote the set of real numbers. Denote by $\mathbb{R}^p$ the p-dimensional Euclidean space and by $\mathbb{R}^{p\times q}$ the space of real matrices of order $p\times q$. For a positive integer K, denote by $[K]$ the set $\{1,2,\ldots,K\}$.

Regarding vectors and matrices, for a vector $v\in\mathbb{R}^p$, we denote by $\|v\|_2$ the ℓ_2-norm of v. $\mathbb{I}_p$ denotes the p-dimensional identity matrix, and $\mathbf{1}_p\in\mathbb{R}^p$ denotes the p-dimensional vector with all entries equal to 1. For $\mathbf{A}\in\mathbb{R}^{p\times p}$, we denote by $\mathbf{A}_j$ and $\mathbf{a}_j$ the jth column and the transposed jth row of $\mathbf{A}$ respectively. We use $\mathrm{col}(\mathbf{A})$ and $\mathrm{col}(\mathbf{A})^\perp$ to denote the columnspace of $\mathbf{A}$ and its orthogonal complement respectively.

Let (M,d) be a metric space where M is a set endowed with the metric d. For a subset $T\subseteq M$, we denote by $\mathcal{N}(T,d,\varepsilon)$ the ε-covering number of T. Similarly, we denote by $\mathcal{M}(T,d,\varepsilon)$ the ε-packing number of T.

Throughout the paper, let $O(\cdot)$ (respectively $\Omega(\cdot)$) denote the standard big-O (respectively big-Omega) notation, i.e., we say $a_n=O(b_n)$ if there exists a universal constant $C>0$, such that $a_n\le Cb_n$ (respectively $a_n\ge Cb_n$) for all $n\in\mathbb{N}$. Sometimes for notational convenience, we write $a_n\lesssim b_n$ in place of $a_n=O(b_n)$ and $a_n\gtrsim b_n$ in place of $a_n=\Omega(b_n)$. We write $a_n\asymp b_n$ if $a_n=O(b_n)$ and $a_n=\Omega(b_n)$. Finally, we write $a_n\sim b_n$ if $\lim_{n\to\infty}a_n/b_n=1$, and $a_n=o(b_n)$ if $\lim_{n\to\infty}a_n/b_n=0$.

Table 1
Notations for the geometric sets and corresponding complexity measures.

Geometric Sets	Residualized signals (scaled)	Spurious projections
Notation	$\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}$ for all $\mathcal{D} \neq \mathcal{S}$ (Eq. (7))	$\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{D}\cap\mathcal{S}}$ for all $\mathcal{D} \neq \mathcal{S}$ (Eq. (2) and (9))
Collection sets	$\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$ such that $\mathcal{I} = \mathcal{S} \cap \mathcal{D}$ (Eq. (7))	$\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$ such that $\mathcal{I} = \mathcal{S} \cap \mathcal{D}$ (Eq. (7))
Upper scaled complexities	$\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$ (Eq. (8))	$\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ (Eq. (10))
Lower scaled complexity	$\mathscr{E}^{*}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$ (Eq. (16))	$\mathscr{E}^{*}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ (Eq. (18))
Diameter	$\mathsf{D}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$	$\mathsf{D}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$
Minimal separation	$\mathsf{d}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$	$\mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$

Brief summary of main contributions. Before diving into the mathematical details, we first lay out a brief summary of the main results of the paper. As mentioned before, the main contribution of the paper is to identify the role of two quantities related to the complexities of the residualized signals and the spurious projections in the model recovery performance of BSS. To elaborate more on this, we revisit a specific identifiability margin $\tau_*(s)$ (introduced in [10] and Equation (4)) that essentially captures the joint effect of the signal strength and the collinearity among the true and spurious features arising from mutual correlation between them. If the minimum signal strength $\beta_{\min} := \min\{|\beta_j| : j \in \mathcal{S}\}$ is large and the correlation between true and spurios features are small, then $\tau_*(s)$ is large, which makes it easier for BSS to identify the true support $\mathcal{S}$. Next, we introduce the two complexity measures of the set of residualized signals $\mathcal{T}_{\mathcal{I}}^{(s)}$ and the set of spurious projections $\mathcal{G}_{\mathcal{I}}^{(s)}$ (see Table 1):

1. *Complexity of residualized signals*: In Section 3.2, we introduce the complexity measure for the class of residualized signals $\boldsymbol{\gamma}_{\mathcal{D}}$ (see Equation (6)) originating from the part of the true signal $\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}$ that can not be *linearly explained* by a model $\mathcal{D}$ with $\mathcal{D} \cap \mathcal{S} = \mathcal{I}$. To be prices, we consider a scaled log-entropy integral of the space of resulting unit vectors $\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}$ which we denote by $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(s)}}$ (see Equation (8)).
2. *Complexity of spurious projections*: Section 3.3 introduces the complexity measure for the the spurious projection operators $\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}}$ (see Equation (2) and (9)) which are the orthogonal projection matrices onto the subspace $\mathrm{col}(\mathbf{X}_{\mathcal{D}}) \cap \mathrm{col}(\mathbf{X}_{\mathcal{S}})^{\perp}$ with $\mathcal{D} \cap \mathcal{S} = \mathcal{I}$. Similar to the previous case, the proposed complexity measure is a scaled version of the log-entropy (under operator norm) integral of the space of spurious projection (see Equation (10)) which is denoted by $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$.

Both of this complexities can be linked to the diameter of the set of residualized signals $\mathcal{T}_{\mathcal{I}}^{(s)}$ and the spurious projections $\mathcal{G}_{\mathcal{I}}^{(s)}$ respectively under proper choices of metrics. It is important to note that depending on the structural properties of the design matrix $\mathbf{X}$, the intrinsic complexity of the sets might be much smaller compared to the vanilla complexity measure that is often associated with just the cardinality of the sets. We elaborate on this through the following example.

Example 1. *Let us consider the case when the dimension $p = 3$, i.e,*

$$\mathbf{X} = [\mathbf{e}_1, (\mathbf{e}_1 + \delta\mathbf{e}_2)/(\sqrt{1+\delta^2}), (\mathbf{e}_1 + \delta\mathbf{e}_3)/(\sqrt{1+\delta^2})] \in \mathbb{R}^{n\times 3}$$

where $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are the first three canonical basis of $\mathbb{R}^n$, and $\delta \approx 0$. Also assume that the true support is $\mathcal{S} = \{1\}$. For any other candidate model $\mathcal{D}$ of unit size, we have $\mathcal{D} \cap \mathcal{S} = \emptyset$. Therefore, we have the set of spurious projections to be $\mathcal{G}_\emptyset^{(1)} := \{\mathbf{P}_{\{2\}}, \mathbf{P}_{\{3\}}\}$. Then, we have $\left\|\mathbf{P}_{\{2\}} - \mathbf{P}_{\{3\}}\right\|_{\text{op}} = \delta/\sqrt{1+\delta^2} < \delta$, and hence $\log \mathcal{N}(\mathcal{G}_\emptyset^{(1)}, \|\cdot\|_{\text{op}}, \delta') = 0$ for all $\delta' \geq \delta$. Therefore, overall complexity of the spurious projections, which is integral of the log-entropy, is much smaller compared to the log-cardinality of the set which is $\log 2$ in this case. A more detailed study can be found in Section 4.2.

Therefore, considerations of these quantities will result into sharper results on the margin conditions required for $\tau_*(s)$ (see Theorem 1) in order to have model consistency of BSS. Informally, one of our our main results Theorem 1 shows that

$$\frac{\tau_*(s)}{\sigma^2} \gtrsim \max\left\{\max_{I\subset\mathcal{S}} \mathscr{E}^2_{\mathcal{T}_I^{(s)}}, \max_{I\subset\mathcal{S}} \mathscr{E}^2_{\mathcal{G}_I^{(s)}}\right\}\frac{\log p}{n}$$

is sufficient for model consistency of BSS. The above condition reveals an interesting interplay between the two complexity measures that shows that *only the set of dominating complexity measure* characterizes the model recovery performance. Moreover, the above condition allows us to artificially construct an example with a specific correlation structure in the design matrix which is more favorable for BSS compared to the independent Gaussian design (see Section 4.2) case which is popularly believed to be the easiest setting for model selection. Finally, in Theorem 2, we also present a somewhat similar necessary condition that also involves the complexity measures (fourth row of Table 1) of the residualized signals and spurious projections. In the supplementary material, we also present an extension of Theorem 1 under the GLM case, i.e., we also provide similar complexity measures tailored to the GLM models under some regularity assumptions on the design and the link function.

2. Best subset selection

We briefly review the preliminaries of BSS, one of the most classical variable selection approaches. For a given sparsity level $\widehat{s}$, BSS solves for

$$\widehat{\boldsymbol{\beta}}_{\text{best}}(\widehat{s}) := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \|\boldsymbol{\beta}\|_0 \leq \widehat{s}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 .$$

For model selection purposes, we can choose the best fitting model to be $\widehat{\mathcal{S}}_{\text{best}}(\widehat{s}) := \{j : [\hat{\boldsymbol{\beta}}_{\text{best}}(\widehat{s})]_j \neq 0\}$. For a subset $\mathcal{D} \subseteq [p]$, define the matrix $\mathbf{X}_\mathcal{D} := (\mathbf{X}_j; j \in \mathcal{D})$. In addition, we denote by $\mathbf{P}_\mathcal{D}$ the orthogonal projection operator onto the column space of $\mathbf{X}_\mathcal{D}$, i.e.,

$$\mathbf{P}_\mathcal{D} := \mathbf{X}_\mathcal{D}(\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^\top. \tag{2}$$

Next, we define the corresponding residual sum of squares (RSS) for model $\mathcal{D}$ as

$$R_\mathcal{D} := \mathbf{y}^\top(\mathbb{I}_n - \mathbf{P}_\mathcal{D})\mathbf{y}.$$

With this notation, the $\widehat{\mathcal{S}}_{\text{best}}(\widehat{s})$ can be alternatively written as

$$\widehat{\mathcal{S}}_{\text{best}}(\widehat{s}) := \arg\min_{\mathcal{D}\subseteq[p]:|\mathcal{D}|\leq\widehat{s}} R_{\mathcal{D}}. \tag{3}$$

Given any candidate model $\mathcal{D} \subset [p]$, we can rewrite the model (1) as

$$\mathbf{y} = \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} + \boldsymbol{\varepsilon} = \mathbf{P}_{\mathcal{D}}\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} + (\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} + \boldsymbol{\varepsilon}.$$

The term $(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}$ is the residual part of the signal that can not be linearly explained by $\mathbf{X}_{\mathcal{D}}$. We refer to this part as the *residualized signals*. We can thus measure the discrimination between the true model $\mathcal{S}$ and a different candidate model $\mathcal{D}$ through the quantity $n^{-1}\|(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}\|_2^2$.

Let $\widehat{\boldsymbol{\Sigma}} := n^{-1}\mathbf{X}^\top\mathbf{X}$ be the sample covariance matrix and for any two sets $\mathcal{D}_1, \mathcal{D}_2 \subset [p]$, $\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}_1,\mathcal{D}_2}$ denotes the submatrix of $\boldsymbol{\Sigma}$ with row indices in $\mathcal{D}_1$ and column indices in $\mathcal{D}_2$. Next, we define the collection $\mathscr{A}_{\widehat{s}} := \{\mathcal{D} \subset [p] : \mathcal{D} \neq \mathcal{S}, |\mathcal{D}| = \widehat{s}\}$, and for $\mathcal{D} \in \mathscr{A}_{\widehat{s}}$ write

$$\Gamma(\mathcal{D}) = \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\setminus\mathcal{D},\mathcal{S}\setminus\mathcal{D}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\setminus\mathcal{D},\mathcal{D}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{S}\setminus\mathcal{D}}.$$

In the Gaussian design case, the above quantity can be identified as the empirical version of the conditional variance-covariance matrix $\text{cov}(\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}} \mid \mathbf{X}_{\mathcal{D}})$. Therefore, $\Gamma(\mathcal{D})$ roughly captures the degree correlation between the features in $\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}$ and $\mathbf{X}_{\mathcal{D}}$. To understand the above quantity more clearly, note that

$$\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} = n^{-1}\|(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}\|_2^2 = n^{-1}\left\|(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\right\|_2^2.$$

This shows that $\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}$ captures the goodness-of-fit for model $\mathcal{D}$. Intuitively, if $\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}$ is very close to zero, then there exists $\mathbf{b} \in \mathbb{R}^{|\mathcal{D}|}$ such that $\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} \approx \mathbf{X}_{\mathcal{D}}\mathbf{b}$. Hence, $\mathcal{S}$ and $\mathcal{D}$ have similar linear explanatory power, and the true model $\mathcal{S}$ becomes practically indistinguishable from $\mathcal{D}$. In fact, the following lemma shows that $\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}$ needs to be at least bounded away from 0 for all $\mathcal{D} \in \mathscr{A}_{\widehat{s}}$ to make $\mathcal{S}$ identifiable.

Lemma 1. *For any given $\widehat{s} > 0$, if there exists a $\mathcal{D} \in \mathscr{A}_{\widehat{s}}$ such that $\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} = 0$, then there exists $\mathbf{b} \in \mathbb{R}^{\widehat{s}}$ such that $\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} = \mathbf{X}_{\mathcal{D}}\mathbf{b}$. Hence, both $\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}$ and $\mathbf{X}_{\mathcal{D}}\mathbf{b}$ generates the same probability distribution for $\mathbf{y}$, and $\mathcal{S}$ becomes non-identifiable.*

Now we are ready to introduce the identifiability margin that characterizes the *model discriminative power* of BSS and the two complexity measures.

3. Identifiability margin and two complexities

3.1. Identifiability margin

The discussion in Section 2 motivates us to define the following *identifiability margin*:

$$\tau_*(\widehat{s}) := \min_{\mathcal{D}\in\mathscr{A}_{\widehat{s}}} \frac{\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}}{|\mathcal{S}\setminus\mathcal{D}|}. \tag{4}$$

If we define $\mathscr{A}_{\widehat{s},k} := \{\mathcal{D} \in \mathscr{A}_{\widehat{s}} : |\mathcal{S}\setminus\mathcal{D}| = k\}$, then the above can be rewritten as

$$\tau_*(\widehat{s}) = \min_{k\in[\widehat{s}]}\min_{\mathcal{D}\in\mathscr{A}_{\widehat{s},k}} \frac{\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}}{k}.$$

As mentioned earlier, the quantity $\tau_*(\widehat{s})$ captures the model discriminative power of BSS. To add more perspective, note that if the features are highly correlated among themselves then it is expected that $\tau_*(\widehat{s})$ is very close to 0. Hence, any candidate model $\mathcal{D}$ is practically indistinguishable from the actual model $\mathcal{S}$ which in turn makes the problem of exact model recovery harder. On the contrary, if the features are uncorrelated then $\tau_*(\widehat{s})$ becomes bounded away from 0 making the true model $\mathcal{S}$ easily recoverable. For example, [10] showed that under the condition

$$\tau_*(s) \gtrsim \sigma^2 \frac{\log p}{n}, \tag{5}$$

BSS is able to achieve model consistency. In general, Condition (5) is less restrictive than the well known β-min condition which demands

$$\beta_{\min} = \min_{j \in \mathcal{S}} |\beta_j| \gtrsim \sigma \left(\frac{\log p}{n} \right)^{1/2}.$$

To see this, let $\hat{\lambda}_m := \min_{\mathcal{D} \in \mathscr{A}_s} \lambda_{\min}(\Gamma(\mathcal{D}))$ where $\lambda_{\min}(\Gamma(\mathcal{D}))$ denotes the minimum eigenvalue of $\Gamma(\mathcal{D})$, and note that $\tau_*(s) \geq \hat{\lambda}_m \beta_{\min}^2$. Thus a sufficient condition for (5) to hold is $\beta_{\min} \gtrsim \sigma \{\log p/(n\hat{\lambda}_m)\}^{1/2}$. In comparison, [26] showed that the ℓ_0-regularized least square estimator is able to achieve model consistency when $\beta_{\min} \gtrsim \sigma \{\log p/(n\kappa_-)\}^{1/2}$, where $\kappa_- := \min_{\mathcal{D}:|\mathcal{D}| \leq s, \mathcal{D} \subset [p]} \lambda_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{D}})$. The latter condition is very sensitive to the feature correlation as κ_- can vary drastically depending on the degree of correlation between the features. In contrast, $\hat{\lambda}_m$ is robust against design dependence; rather, it reflects how spurious variables can approximate the true model, which implies much less restriction than that induced by κ_-. For more details on the identifiability margin, we point the readers to Section 2.1 of [10]. From now on, unless otherwise mentioned, we will assume that the margin quantity $\tau_*(\widehat{s}) > 0$ to avoid the non-identifiability issue as pointed out in Lemma 1.

Next, we will shift focus on the underlying geometric structures of two spaces that govern the difficulty of the BSS problem (3). We identify the complexities of two types of sets that control the hardness of BSS: (i) the set of residualized signals, and (ii) the set of spurious projections. We discuss these two sets, and the associated complexities in detail below and Table 1 compiles important notations and quantities related to the aforementioned sets.

3.2. *Complexity of residualized signals*

We start with the definition of the residualized signal. For a candidate model $\mathcal{D} \in \mathscr{A}_{\widehat{s}}$, define

$$\boldsymbol{\gamma}_{\mathcal{D}} := n^{-1/2}(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S} \setminus \mathcal{D}} \boldsymbol{\beta}_{\mathcal{S} \setminus \mathcal{D}}, \tag{6}$$

and the corresponding unit vector $\widehat{\boldsymbol{\gamma}}_{\mathcal{D}} := \boldsymbol{\gamma}_{\mathcal{D}} / \|\boldsymbol{\gamma}_{\mathcal{D}}\|_2$. Note that $\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}$ is well-defined as $\|\boldsymbol{\gamma}_{\mathcal{D}}\|_2^2 \geq \tau_*(\widehat{s}) > 0$. As mentioned before, $\boldsymbol{\gamma}_{\mathcal{D}}$ represents the part of the signal that can not be linearly explained by the features in model $\mathcal{D}$. Note that the margin condition (5) essentially tells that the vectors $\boldsymbol{\gamma}_{\mathcal{D}}$ are well bounded away from the origin. However, this property does not quite capture the degree of their radial spread in $\mathbb{R}^n$. It may happen that despite being well bounded away from the origin, the vectors are clustered along one common unit direction. For example, in Figure 1, the distance between $\boldsymbol{\gamma}_{\mathcal{D}_1}$ and $\boldsymbol{\gamma}_{\mathcal{D}_2}$ are large as the vectors are at a larger distance from the origin, although the angular separation between them

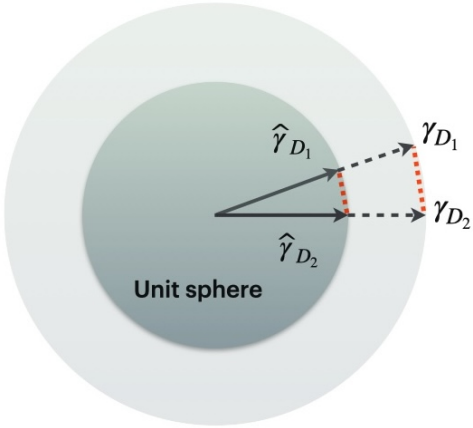


Fig 1: Distance between $\widehat{\gamma}_{\mathcal{D}_1}$ and $\widehat{\gamma}_{\mathcal{D}_2}$ correctly captures the angular separation.

is small. To capture this notion of separation within the vectors $\{\gamma_{\mathcal{D}}\}_{\mathcal{D}\in\mathscr{A}_{\widehat{s}}}$, we also need to capture the spatial alignment of their corresponding unit vectors $\{\widehat{\gamma}_{\mathcal{D}}\}_{\mathcal{D}\in\mathscr{A}_{\widehat{s}}}$. This motivates us to consider the geometric complexities of this set of unit vectors. Specifically, for a set $\mathcal{I} \subset \mathcal{S}$, we define

$$\mathcal{T}_{\mathcal{I}}^{(\widehat{s})} := \{\widehat{\gamma}_{\mathcal{D}} : \mathcal{D} \in \mathscr{A}_{\widehat{s}}, \mathcal{S} \cap \mathcal{D} = \mathcal{I}\} \subseteq \mathbb{R}^n, \tag{7}$$

which is the set of all the normalized forms of the residualized signals corresponding to the models $\mathcal{D} \in \mathscr{A}_{\widehat{s}}$ with $\mathcal{I}$ as the common part with true model $\mathcal{S}$. To capture the complexity of these spaces, we look at the scaled entropy integral

$$\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} := \frac{\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}, \|\cdot\|_2, \varepsilon)}\, \mathrm{d}\varepsilon}{\sqrt{\log |\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}|}}. \tag{8}$$

The numerator in the above display is commonly known as entropy integral which captures the topological complexity of $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$. In literature, this quantity has a connection to the well-known *Talagrand's complexity* [20], which often comes up in controlling the expectation of the supremum of Gaussian processes [13, 14, 1]. In this paper, we look at the above scaled version of the entropy integral which allows us to compare the quantity with the diameter and minimum pairwise distance between the elements of the set $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$. To elaborate on this point, define the diameter and minimum pairwise distance of $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$ as follows:

$$\mathsf{D}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} := \max_{\mathbf{u},\mathbf{v}\in\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} \|\mathbf{u}-\mathbf{v}\|_2, \quad \text{and} \quad \mathsf{d}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} := \min_{\mathbf{u},\mathbf{v}\in\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} \|\mathbf{u}-\mathbf{v}\|_2.$$

Now notice the following two simple facts:

$$\log \mathcal{N}(\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}, \|\cdot\|_2, \mathsf{D}_{\mathcal{T}_{\mathcal{I}}}) = 0, \quad \log \mathcal{N}(\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}, \|\cdot\|_2, \mathsf{d}_{\mathcal{T}_{\mathcal{I}}}) = \log |\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}|.$$

Noting that $\log \mathcal{N}(\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}, \|\cdot\|_2, \delta)$ is a decreasing function over δ, we finally get

$$\mathsf{d}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} \le \mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}} \le \mathsf{D}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}.$$

This shows that the quantity $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$ roughly captures the average separation of the elements within $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$. In our subsequent discussion, we will show that $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$ heavily influences the margin condition for exact recovery. This is indeed an important observation, as the complexity of the set of residualized signals depends heavily on the association between the features. For example, they can differ vastly for highly correlated designs compared to almost uncorrelated designs. Hence, the effect of $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$ on the exact model recovery also varies significantly across different classes of distributions and leads to sharper margin conditions for exact model recovery.

3.3. Complexity of spurious projections

In this section, we will introduce the space of projection operators that also control the level of difficulty of the true model recovery. Similar to the previous section, for a fixed set $\mathcal{I} \subset \mathcal{S}$, we consider the set

$$\mathcal{G}_{\mathcal{I}}^{(\widehat{s})} := \{\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}} : \mathcal{D} \in \mathscr{A}_{\widehat{s}}, \mathcal{S} \cap \mathcal{D} = \mathcal{I}\} \subseteq \mathbb{R}^{n\times n}. \tag{9}$$

It is a well-known fact that every projection operator of the form $\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}} \in \mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$ has a one-to-one correspondence with the subspace $\mathrm{col}(\mathbf{X}_{\mathcal{D}}) \cap \mathrm{col}(\mathbf{X}_{\mathcal{I}})^{\perp}$. Thus, $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$ can be thought of as the collection of all linear subspaces of the form $\mathrm{col}(\mathbf{X}_{\mathcal{D}}) \cap \mathrm{col}(\mathbf{X}_{\mathcal{I}})^{\perp}$, which is essentially the set of spurious features that can not be linearly explained by the set of features in model $\mathcal{I} \subset \mathcal{S}$. To capture the proper measure of complexity of the set $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$, it is crucial to induce the space of projection operators with a proper metric. It turns out that Grassmannian distance is the correct distance to consider in this context. Specifically, for two linear subspaces U, V we look at their *maximum sin-theta distance*:

$$\mathrm{d}(U, V) := \|\mathbf{\Pi}_U - \mathbf{\Pi}_V\|_{\mathrm{op}},$$

where $\mathbf{\Pi}_U, \mathbf{\Pi}_V$ are the orthogonal projection operators of U, V respectively. It turns out that $\mathrm{d}(U, V)$ evaluates the trigonometric sine function at the maximum principal angle between the subspaces U and V (see Figure 2). We point the readers to [23] for a more detailed discussion on this topic.

Under this distance, we define the scaled entropy integral as

$$\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} := \frac{\int_0^{\infty} \sqrt{\log \mathcal{N}(\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}, \|\cdot\|_{\mathrm{op}}, \varepsilon)}\, \mathrm{d}\varepsilon}{\sqrt{\log |\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}|}}. \tag{10}$$

Note that, unlike $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$, the complexity measure $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ has no dependence on $\boldsymbol{\beta}$ or the residualized signal. Thus, $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ roughly captures the geometric complexity of only the spurious features. In fact, via a similar argument as in Section 3.2, it can be shown that $\mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} \le \mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} \le \mathsf{D}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$, where

$$\mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} := \min_{\mathbf{U},\mathbf{V}\in\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} \|\mathbf{U} - \mathbf{V}\|_{\mathrm{op}}, \quad \mathsf{D}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} := \max_{\mathbf{U},\mathbf{V}\in\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} \|\mathbf{U} - \mathbf{V}\|_{\mathrm{op}}.$$

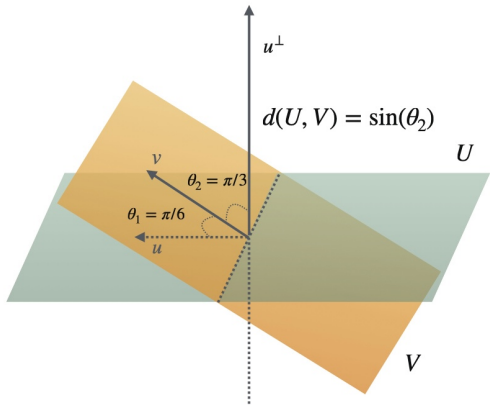


Fig 2: The figure shows the two principal angles between two subspaces U and V. $\{u, u^\perp\}$ are the two orthonormal basis of U, and v is an orthonormal basis of V. θ_2 is the maximum principal angle between U and V.

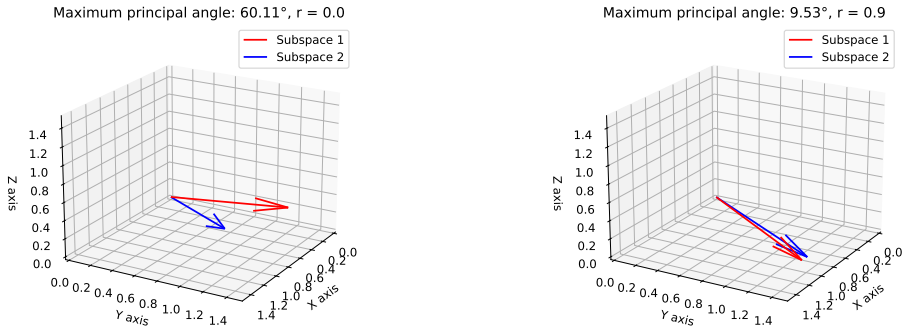
Thus, $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$ only captures the separability in the set of subspaces generated by the spurious features. The main motivation behind considering such quantity is to capture the influence of the effective size of the set $\{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}\}_{\mathcal{I}\subset\mathcal{S}}$ in the analysis of BSS. A naive union bound only uses $|\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}| = \binom{p-\widehat{s}}{\widehat{s}-|\mathcal{I}|}$ as a measure of complexity of the set $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$. This is rather loose, as the effective complexity of the set is much smaller if $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ is small. Thus, taking $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ into account unravels a broader picture of the effect incurred by the underlying geometry of the feature space.

3.4. Correlation and complexities

From the discussion on the two complexities, it is quite evident that both of the complexity measures heavily rely on the alignment of the feature vectors $\{\mathbf{X}_j : j \in [p]\}$, which directly depends on the correlation structure among the features in the model. Below, we discuss how these two types of complexities may vary with correlation among the features.

Correlation and spurious projection operators: We first focus on the set $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$, as it is relatively easy to understand its behavior across different correlation structures. Recall that for a fixed choice of $\mathcal{I}$, the set $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$ is the collection of all the projection operators of the form $\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}}$ for all $\mathcal{D} \in \mathscr{A}_{\widehat{s}}$, which can be thought of as the collection of different subspaces generated by the linear combination of the spurious features. If the spurious features are highly correlated then these subspaces may be essentially indistinguishable from each other, i.e., the mutual distance between the projection operators $\{\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}}\}_{\mathcal{D}\in\mathscr{A}_{\widehat{s}}}$ is significantly smaller compared to the case when they are weakly correlated. As an example, let us consider the equi-correlated Gaussian design, i.e., the row vectors $\{x_i\}_{i\in[n]}$ of $\mathbf{X}$ in (1) follows i.i.d. mean-zero Gaussian distribution with covariance matrix

$$\mathbf{\Sigma} = (1-r)\mathbb{I}_p + r\mathbf{1}_p\mathbf{1}_p^\top.$$



(a) (b)

Fig 3: (a) shows the angle between spurious features in $\mathcal{G}_{\emptyset}^{(1)}$ for $r = 0$, (b) shows the angle between spurious features in $\mathcal{G}_{\emptyset}^{(1)}$ for $r = 0.9$.

For the sake of simplicity, we also assume that the true model is a singleton set. In particular, we consider $\mathcal{S} = \{1\}$ and set $\widehat{s} = 1$. Also, note that in this case $\mathscr{A}_{\widehat{s}} = \{j \in [p] : j \neq 1\}$ and $\mathcal{I} = \emptyset$. Under this setup, we have $\mathcal{G}_{\emptyset}^{(1)} = \{\mathbf{X}_j \mathbf{X}_j^{\top} / \|\mathbf{X}_j\|_2^2 : j \notin \mathcal{S}\}$ and $n^{-1} \|\mathbf{X}_j - \mathbf{X}_k\|_2^2 \approx 2(1-r)$, for all $j, k \neq 1$. If r is very close to 1 in the above display, then it follows that the vectors $\{\mathbf{X}_j/\sqrt{n}\}_{j\neq 1}$ are extremely clustered towards each other, and as a result, the spurious projection operators are also very close to each other in operator norm. Due to this, the complexity measure $\mathscr{E}_{\mathcal{G}_{\emptyset}^{(1)}}$ becomes extremely small and the subspaces become almost indistinguishable. In contrast, when the features are approximately uncorrelated, i.e., $r \approx 0$, the scaled features $\{\mathbf{X}_j/\sqrt{n}\}_{j\neq 1}$ are roughly orthogonal. In that case the

$$n^{-1} \|\mathbf{X}_j - \mathbf{X}_k\|_2^2 \approx 2, \quad \text{for all } j, k \neq 1.$$

As an example, for $p = 6$ and $s = 1$, Figure 3 illustrates a similar phenomenon in 3-dimension. Figure 3(a) clearly shows that for the case $r = 0$ the angle is larger compared to the $r = 0.9$ case in Figure 3(b). This suggests that the linear spans generated by each of the set of features $\{n^{-1/2}\mathbf{X}_j\}_{j\neq 1}$ are well separated and $\mathscr{E}_{\mathcal{G}_{\emptyset}^{(1)}}$ is well bounded away from zero. Thus, it follows that the features are well spread out in $\mathbb{R}^n$. This phenomenon indicates that a higher correlation may aid the model recovery chance for BSS by reducing the search space over the features. As we will see in our subsequent discussion in Section 4.2, the correlation between noise variables can significantly help BSS to identify the correct model. Specifically, we construct an example where the true variables are uncorrelated with the noise variables and show that a high correlation among noise variables helps BSS to identify the correct model. The intuition is that under the presence of correlation, the diversity of the elements in $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$ gets reduced as $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$ becomes small. Thus, BSS needs to search on a comparatively smaller feature space rather than searching over all possible $\binom{p-\widehat{s}}{\widehat{s}-|\mathcal{I}|}$ models, which in turn aids the probability of finding the correct model out of the other candidate ones. Thus, the smaller complexity of $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$ counteracts the adverse effect of correlation to some degree, and it may improve the model recovery performance of BSS.

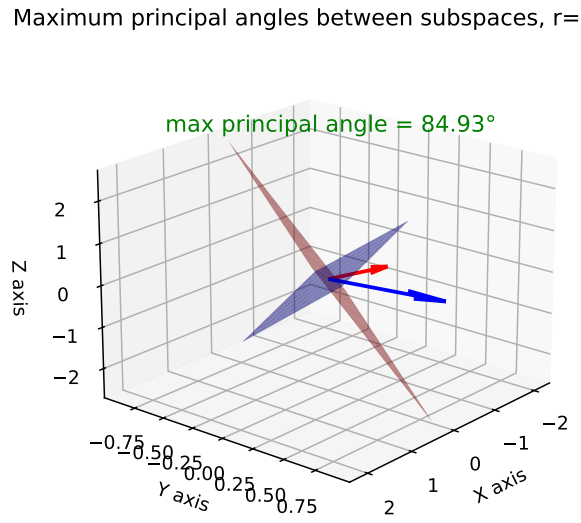


Fig 4: Maximum principal angle between the subspaces $U := \text{span}\{\mathbf{X}_1, \mathbf{X}_2\}$ and $V := \text{span}\{\mathbf{X}_3, \mathbf{X}_4\}$ for $r = 0.99$ under equi-correlated Gaussian design.

However, it is not necessary that the complexity will be small for a highly correlated structure. For example, in the above case, if $\mathcal{S} = \{1, 2\}$, then $\mathcal{G}_{\emptyset}^{(1)}$ might be closer to 1, i.e., the maximum principal angles between the subspaces are closer to 90° as shown in Figure 4. Therefore, the complexity solely depends on the distribution and correlation structure of the design matrix and may behave differently on a case-by-case basis.

Correlation and residualized signals: Now we shift our focus to understanding the behavior of the set of normalized residualized signals denoted by $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$. Recall that for a fixed $\mathcal{I}$, the set $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$ denotes the collection of all the unit vectors $\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}$ (defined in Section 3.2) such that $\mathcal{D} \cap \mathcal{S} = \mathcal{I}$. Similar to $\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}$, the complexity of the set $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$ also depends on the correlation structure among the features. To elaborate more on this, we revisit the example of equi-correlated Gaussian design with correlation parameter r and $\mathcal{S} = \{1\}$. We denote by $\mathbf{P}_j$ the orthogonal projection operator onto the span of $\mathbf{X}_j$, i.e., $\mathbf{P}_j = \mathbf{X}_j\mathbf{X}_j^\top / \|\mathbf{X}_j\|_2^2$. Similar to the previous section, in this case also the set $\mathcal{T}_{\emptyset}^{(1)}$ consists of the scaled residualized signals that take the following form for large n with high probability:

$$\widehat{\boldsymbol{\gamma}}_j = \frac{(\mathbb{I}_n - \mathbf{P}_j)\mathbf{X}_1}{\|(\mathbb{I}_n - \mathbf{P}_j)\mathbf{X}_1\|_2} \approx \frac{\mathbf{X}_1 - r\mathbf{X}_j}{\|\mathbf{X}_1 - r\mathbf{X}_j\|_2}, \quad \text{for all } j \neq 1.$$

Also, note that

$$\widehat{\boldsymbol{\gamma}}_j^\top \widehat{\boldsymbol{\gamma}}_k \approx \frac{1 - 2r^2 + r^3}{1 - r^2} =: f(r).$$

Since, $f(r)$ is a strictly decreasing function on $[0, 1)$, and $\|\widehat{\gamma}_j - \widehat{\gamma}_k\|_2^2 = 2(1 - \widehat{\gamma}_j^\top \widehat{\gamma}_k)$, it follows that $\mathsf{d}_{\mathcal{T}_\varnothing} \geq 1/2$ when r is very close to 1. On the contrary, when $r \approx 0$, the above display suggests that $\mathsf{D}_{\mathcal{T}_\varnothing^{(1)}} \approx 0$, i.e., for uncorrelated design, the complexity $\mathscr{E}_{\mathcal{T}_\varnothing^{(1)}}$ of the set $\mathcal{T}_\varnothing^{(1)}$ is smaller compared to the highly correlated case which is in sharp contrast with the behavior of $\mathscr{E}_{\mathcal{G}_\varnothing^{(1)}}$.

However, it is worth pointing out that the above property of $\mathscr{E}_{\mathcal{T}_\varnothing^{(1)}}$ is very specific to the above considered model. There may exist a correlated structure where higher correlation among noise variables does not increase $\mathscr{E}_{\mathcal{T}_\varnothing^{(1)}}$ (see Section 4.2), and improves the chance of identifying the correct model via BSS. However, understanding such a phenomenon for a more general design could be significantly more challenging.

4. Theoretical properties of BSS

4.1. Model selection consistency of BSS under known sparsity

This section illustrates the interaction between the identifiability margin (4) and the two complexities that characterize the sufficient condition for the exact model recovery. From here on, we assume that the true sparsity is known, i.e., we set $\widehat{s} = s$ in (3), and BSS searches the best model out of all possible models of size s. We now introduce a technical assumption that essentially prevents the noisy features from becoming highly correlated with the true features:

Assumption 1. *The design matrix* $\mathbf{X}$ *enjoys the following property:*

$$\min_{\mathcal{I} \subset \mathcal{S}} \mathscr{E}_{\mathcal{G}_\mathcal{I}^{(s)}} > \{\log(ep)\}^{-1/2}.$$

The above assumption ensures that the noisy features are distinguishable enough from the active features in order for BSS to identify the active features. To see this, consider the case when the noise variables are highly correlated with the true features $\{\mathbf{X}_j\}_{j \in \mathcal{S}}$. In this case, the projection operator $\mathbf{P}_\mathcal{D} - \mathbf{P}_\mathcal{I}$ can be written as $(\mathbb{I}_n - \mathbf{P}_\mathcal{I})\mathbf{P}_\mathcal{D}$ for all $\mathcal{D} \in \mathcal{G}_\mathcal{I}^{(s)}$, whenever $\mathcal{I} \neq \varnothing$. As the features in $\{\mathbf{X}_j : j \in \mathcal{D} \setminus \mathcal{S}\}$ are highly correlated with $\mathbf{X}_\mathcal{I}$, it follows that $\|\mathbf{P}_\mathcal{D} - \mathbf{P}_\mathcal{I}\|_{\text{op}} \approx 0$ and by triangle inequality it follows that $\|\mathbf{P}_\mathcal{D} - \mathbf{P}_{\mathcal{D}'}\|_{\text{op}} \approx 0$ for any two candidate models $\mathcal{D}$ and $\mathcal{D}'$ such that $\mathcal{D} \cap \mathcal{S} = \mathcal{D}' \cap \mathcal{S} = \mathcal{I}$. Thus, Assumption 1 gets rid of such cases by indirectly controlling the correlation between the active features and noisy features. Secondly, the assumption also enforces diversity among the noise variables in the following sense: If the features $\{\mathbf{X}_j : j \notin \mathcal{S}^c\}$ are too similar to each other, then also $\mathscr{E}_{\mathcal{G}_\mathcal{I}^{(s)}}$ shrinks towards 0. Thus, Assumption 1 prevents the noise variables from becoming extremely correlated with each other.

Assumptions with similar spirits are fairly common in the literature on high-dimensional statistics. For example, the well-known Sparse Riesz Condition (SRC) [25] assumes that there exist positive numbers κ_-, κ_+ and $\Psi \geq 1$ such that

$$\kappa_- \leq \frac{\|\mathbf{X}\mathbf{v}\|_2^2}{n} \leq \kappa_+, \quad \text{for all } \mathbf{v} \in \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|_2 = 1, \|\mathbf{u}\|_0 \leq \Psi s\}. \tag{11}$$

The above SRC condition controls the maximum and minimum eigenvalues of all the models of size s, which essentially prevents the features to become extremely correlated with each

other. In comparison, Assumption 1 is much weaker than SRC condition in two aspects. First, unlike the SRC, Assumption 1 imposes conditions only over $(2^s - 2)$ models, whereas SRC imposes conditions on $\Omega((p/s)^{\lfloor \Psi s \rfloor})$ many models. Second, the lower bound requirement in Assumption 1 is rather weak as the bound decays with increasing ambient dimension and allows a higher degree of correlation among the features. In other words, SRC condition (11) implies the condition in Assumption 1, and we formalize this claim below.

Proposition 1. *Let the columns of* $\mathbf{X}$ *be normalized, i.e.,* $\|\mathbf{X}_j\|_2 = \sqrt{n}$. *Also, assume that there exist positive constants* κ_-, κ_+ *such that the SRC condition* (11) *holds with* $\Psi = 2$. *Then the condition in Assumption 1 also holds for large enough p, i.e.,* $\min_{I \subset \mathcal{S}} \mathscr{E}_{\mathcal{G}_I^{(s)}} \geq \kappa_-/\kappa_+ \gg \{\log(ep)\}^{-1/2}$. *Furthermore, the implication in the other direction is not true in general.*

Next, we assume that the noise in model (1) is sub-Gaussian.

Assumption 2. *The noise* $\{\varepsilon_i\}_{i \in [n]}$ *in model* (1) *are i.i.d. mean-zero* σ*-sub-Gaussian noise, i.e.,* $\max_{i \in [n]} \mathbb{E} \exp(t\varepsilon_i) \leq \exp(\sigma^2 t^2/2)$ *for all* $t \in \mathbb{R}$.

Now we are ready to state our main sufficiency result.

Theorem 1 (Sufficiency)**.** *Under Assumption 1 and Assumption 2, there exists a positive universal constant* C_0 *such that for any* $0 \leq \eta < 1$, *whenever the identifiability margin* $\tau_*(s)$ *satisfies*

$$\frac{\tau_*(s)}{\sigma^2} \geq \frac{C_0}{(1-\eta)^2} \left[\max \left\{ \max_{I \subset \mathcal{S}} \mathscr{E}^2_{\mathcal{T}_I^{(s)}}, \max_{I \subset \mathcal{S}} \mathscr{E}^2_{\mathcal{G}_I^{(s)}} \right\} + \sqrt{\frac{\log(es) \vee \log\log(ep)}{\log(ep)}} \right] \frac{\log(ep)}{n}, \tag{12}$$

we have

$$\{\widehat{\mathcal{S}}_{\text{best}}(s)\} \subseteq \left\{ \widehat{\mathcal{S}} : |\widehat{\mathcal{S}}| = s, R_{\widehat{\mathcal{S}}} \leq \min_{\mathcal{D} \in \mathscr{A}_s} R_{\mathcal{D}} + n\eta\tau_*(s) \right\} = \{\mathcal{S}\},$$

with probability at least $1 - O(\{s \vee \log p\}^{-1})$. *In particular, we have* $\mathcal{S} = \arg\min_{\mathcal{D} \in \mathscr{A}_s} R_{\mathcal{D}}$ *with high probability.*

The proof of the above theorem is present Section S1.3 of the supplementary material. The above theorem gives a sufficient condition for BSS to achieve model consistency. The above theorem states that under the margin condition (12) the true model $\mathcal{S}$ is the optimizer of the BSS problem. Furthermore, the parameter η quantifies the magnitude of the sub-optimality gap. For $\eta > 0$, the above theorem shows that $R_{\mathcal{D}} - R_{\mathcal{S}} > n\eta\tau_*(s)$ for any $\mathcal{D} \in \mathscr{A}_s$, i.e, the gap between the optimal RSS value $R_{\mathcal{S}}$ and the next smallest RSS is more pronounced for larger values of η. However, this is more demanding than just the requirement for R_S being the optimal value, and hence the margin condition (12) is more stringent for $\eta > 0$ compared to $\eta = 0$ case.

Next, note that the margin condition (12) involves the identifiability margin $\tau_*(s)$ and the two complexities associated with the sets of residualized signals and spurious projection operators. This condition reveals an interesting interplay between the identifiability margin and the two complexities. To highlight this phenomenon, it is instructive to consider the case when the true model $\mathcal{S} = \{1\}$ and $\mathbf{X}_1$ is orthogonal to the spurious features $\{\mathbf{X}_j\}_{j \neq 1}$. However, the

spurious features are allowed to be extremely correlated to each other. As mentioned in the independent block design example in Section 4.2, in this case, both of the two complexities are small for higher correlation among the spurious features, whereas $\tau_*(s)$ remains roughly unaffected by the strength of correlation. Thus, the margin condition (12) becomes less stringent with increasing strength of correlation, and the performance of BSS should improve. To illustrate this phenomenon, we consider a simulation setup with $p = 2000, n = 500$, and $s = 1$. We generate $\mathbf{X}$ from independent Gaussian block design mentioned in Section 4.2 with the cross-correlation $c = 0$, and $r \in [0, 1)$ being the correlation within the noise variables. Thus, $r = 0$ corresponds to the independent Gaussian design. We set $\boldsymbol{\beta} = (0.1, 0, \ldots, 0)^\top \in \mathbb{R}^p$, and the errors $\{\varepsilon_i\}_{i\in[n]}$ are generated in i.i.d. fashion from $\mathsf{N}(0, 1)$. Finally, the response $\mathbf{y}$ is generated according to model (1). Assuming s is known, we use ABESS [29] as a fast computational surrogate for BSS. The left panel of Figure 5 shows that the mean model recovery rate of ABESS (across 20 independent runs) increases as the correlation between the noise variables increases to 1, which validates the findings in Theorem 1. The right panel of Figure 5 also shows that a similar phenomenon is true even for $s > 1$.

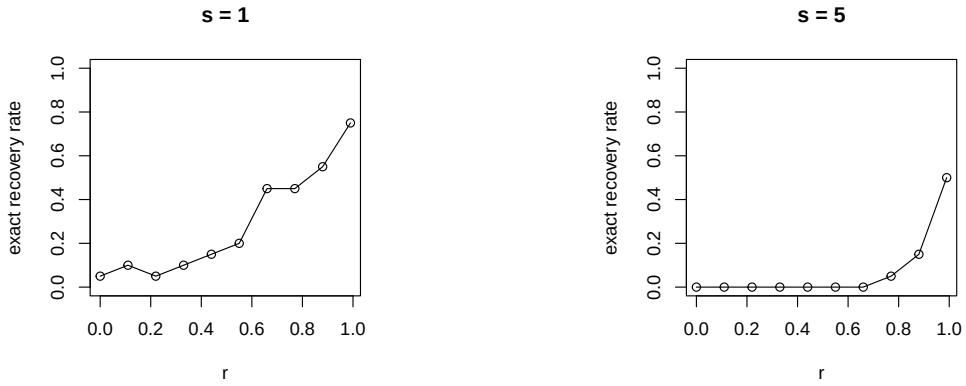


Fig 5: Model recovery rate of ABESS under independent block design.

On the other hand, as mentioned in Remark 1, under equicorrelated model with $\mathcal{S} = \{1\}$ and high correlation, $\mathscr{E}_{\mathcal{T}_\emptyset^{(1)}}$ remains strictly bounded away from 0 and it dominates $\mathscr{E}_{\mathcal{G}_\emptyset^{(1)}}$. However, the identifiability margin $\tau_*(s)$ becomes very small due to the high correlation between the true and noise variables. Hence, the margin condition (12) becomes harder to satisfy with increasing correlation. In the case of independent design, it turns out $\mathscr{E}_{\mathcal{G}_\emptyset^{(1)}}$ is the dominating complexity measure. This is not surprising as under independent design, the features are more spread out in the feature space compared to correlated design, whereas the residualized signals are more concentrated towards a single unit direction, making $\mathscr{E}_{\mathcal{T}_\emptyset^{(1)}}$ smaller compared to $\mathscr{E}_{\mathcal{G}_\emptyset^{(1)}}$.

The above discussion shows that apart from the quantity $\tau_*(s)$, the complexity of residualized signals and the complexity of spurious projection operators also play a decisive role in the margin condition of the best subset selection problem. Specifically, the set with higher complexity characterizes the margin condition in Theorem 1.

We can further represent the condition (12) in terms of the diameter of the sets $\mathcal{T}_{\mathcal{I}}^{(s)}$ and $\mathcal{G}_{\mathcal{I}}^{(s)}$. To see this, recall that $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(s)}} \leq \mathsf{D}_{\mathcal{T}_{\mathcal{I}}^{(s)}}$ and $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(s)}} \leq \mathsf{D}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$ for all $\mathcal{I} \subset \mathcal{S}$. Under the light of this fact, we have the following corollary:

Corollary 1. *Let the conditions in Assumption 1 and Assumption 2 hold. Then there exists a positive universal constant C_0 such that for any $0 \leq \eta < 1$, whenever the identifiability margin $\tau_*(s)$ satisfies*

$$\frac{\tau_*(s)}{\sigma^2} \geq \frac{C_0}{(1-\eta)^2}\left[\max\left\{\max_{\mathcal{I}\subset\mathcal{S}} \mathsf{D}^2_{\mathcal{T}_{\mathcal{I}}^{(s)}}, \max_{\mathcal{I}\subset\mathcal{S}} \mathsf{D}^2_{\mathcal{G}_{\mathcal{I}}^{(s)}}\right\} + \sqrt{\frac{\log(es)\vee\log\log(ep)}{\log(ep)}}\right]\frac{\log(ep)}{n}, \tag{13}$$

we have

$$\{\widehat{\mathcal{S}}_{\text{best}}(s)\} \subseteq \left\{\widehat{\mathcal{S}} : |\widehat{\mathcal{S}}| = s, R_{\widehat{\mathcal{S}}} \leq \min_{\mathcal{D}\in\mathscr{A}_s} R_{\mathcal{D}} + n\eta\tau_*(s)\right\} = \{\mathcal{S}\},$$

with probability at least $1 - O(\{s \vee \log p\}^{-1})$. In particular, we have $\mathcal{S} = \arg\min_{\mathcal{D}\in\mathscr{A}_s} R_{\mathcal{D}}$ with high probability.

Corollary 1 essentially conveys the same message as Theorem 1, only under a slightly stronger margin condition (13). However, in some cases, it could be comparatively easier to give theoretical guarantees on the diameters $\mathsf{D}_{\mathcal{T}_{\mathcal{I}}^{(s)}}, \mathsf{D}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$ rather than their corresponding complexity measures $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(s)}}, \mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$ respectively. In the next section, we will discuss a few illustrative examples to further elaborate on the effects of two complexities.

4.2. Illustrative examples

In this section, we will discuss a few illustrative examples to highlight the effect complexities of the two spaces described in Section 3.2 and Section 3.3.

Block design with a single active feature

Consider the model (1) where the rows of $\mathbf{X}$ are independently generated from p-dimensional multivariate Gaussian distribution with mean-zero and variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & c\mathbf{1}_{p-1}^\top \\ c\mathbf{1}_{p-1} & (1-r)\mathbb{I}_{p-1} + r\mathbf{1}_{p-1}\mathbf{1}_{p-1}^\top \end{pmatrix},$$

where $c \in [0, 0.997], r \in [0, 1)$. We need to further impose a restriction

$$c^2 < r + \frac{1-r}{p-1}$$

to ensure positive definiteness of $\mathbf{\Sigma}$. In this case, we also set the true model $\mathcal{S} = \{1\}$ and the noise variance $\sigma = 1$. Recall that in this case the sets of residualized signals and spurious projection operators are denoted by $\mathcal{T}_{\emptyset}^{(1)}$ and $\mathcal{G}_{\emptyset}^{(1)}$ respectively. Under this setup, we have the following lemma:

Lemma 2. *Assume that* $\log p = o(n)$. *Then under the above setup, there exist universal positive constants* C, L, M *such that the followings are true with* $\varepsilon_{n,p} = C\{(\log p)/n\}^{1/2}$:

(a) For large enough n, p *we have*

$$\mathbb{P}\left[\left\{1+\varepsilon_{n,p}-\frac{(c-\varepsilon_{n,p})^2}{1+\varepsilon_{n,p}}\right\} \geq \frac{\tau_*(1)}{\beta_1^2} \geq \left\{1-\varepsilon_{n,p}-\frac{(c+\varepsilon_{n,p})^2}{1-\varepsilon_{n,p}}\right\}\right] = 1+o(1/p).$$

(b) For large enough n, p *we have*

$$\mathbb{P}\left[\max\left\{\frac{2c^2(1-r)}{1-c^2}-L\varepsilon_{n,p},0\right\} \leq \mathsf{d}^2_{\mathcal{T}^{(1)}_\varnothing} \leq \mathsf{D}^2_{\mathcal{T}^{(1)}_\varnothing} \leq \frac{2c^2(1-r)}{1-c^2}+L\varepsilon_{n,p}\right] = 1+o(1/p).$$

(c) For large enough n, p *we have*

$$\mathbb{P}\left[\max\left\{(1-r^2)-M\varepsilon_{n,p},0\right\} \leq \mathsf{d}^2_{\mathcal{G}^{(1)}_\varnothing} \leq \mathsf{D}^2_{\mathcal{G}^{(}_\varnothing 1)} \leq (1-r^2)+M\varepsilon_{n,p}\right] = 1+o(1/p).$$

From part (b) and (c) of the above lemma, it follows that the complexity $\mathscr{E}_{\mathcal{T}^{(1)}_\varnothing} \approx 0$ when $c = 0$. For any fixed $c > 0$ and $r \in [0, 1)$, we have

$$\mathscr{E}^2_{\mathcal{T}^{(1)}_\varnothing} \sim \frac{2c^2(1-r)}{1-c^2}, \quad \text{and} \quad \mathscr{E}^2_{\mathcal{G}^{(1)}_\varnothing} \sim (1-r^2) \quad \text{for large } n, p. \tag{14}$$

A detailed derivation of the result is present in Section S1.5 of the supplementary material. Left panel of Figure 6 shows the partition of c-r plane based on the dominating complexity. It is worthwhile to note that a high value r, i.e., a high correlation among the noise variables results in a smaller value of the complexity terms in (12). However, Lemma 2(a) suggest that $\tau_*(1)/\beta_1^2 \sim (1-c^2)$, i.e., higher correlation between true and noise variables shrinks the margin quantity $\tau_*(1)$ towards 0. This suggests that a smaller value of c and a higher value r is more favorable to BSS than other possible choices of (r, c). We now discuss these phenomena through some selected examples.

Independent design: In this case $c = r = 0$. In this case (14) suggest that $\mathscr{E}^2_{\mathcal{G}^{(1)}_\varnothing} \approx 1$. On the other hand, we see that $\mathscr{E}^2_{\mathcal{T}^{(1)}_\varnothing} \approx 0$. Thus, the complexity of spurious projections is dominant in this case. Also, in this case, $\tau_*(1) \approx \beta_1^2$ which suggests that higher signal strength results in a better performance in terms of model selection.

Independent block design: In this case, we set $c = 0$ and we vary r in $(0, 1)$. Note that (14) tells that $\mathscr{E}^2_{\mathcal{T}^{(1)}_\varnothing} \approx 0$, and $\mathscr{E}^2_{\mathcal{G}^{(1)}_\varnothing}$ has a decreasing trend with $r \in (0, 1)$. This suggests that the independent block design with a high value of r is more favorable for BSS to identify the true model compared to the independent random design model in the previous example. Finally, noting the fact that $\tau_*(1) \approx \beta_1^2$, we can conclude that for high values of r, the sufficient condition in Theorem 1 becomes less stringent.

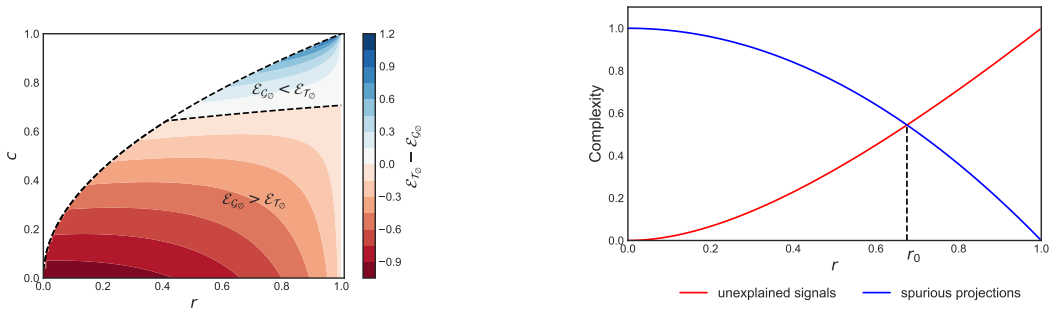


Fig 6: (left) Partition of c-r plane showing dominating regions for the two complexities. The color gradient indicates the value of $\mathscr{E}_{\mathcal{T}_\emptyset^{(1)}} - \mathscr{E}_{\mathcal{G}_\emptyset^{(1)}}$. (right) The plot of two complexities for varying r under equicorrelated design.

Equicorrelated design: Here we set $c = r$ and vary r in the interval $[0, 1)$. Let r_0 be the unique positive solution to the following equation:

$$\frac{2r^2}{1+r} - (1 - r^2) = 0.$$

Calculation shows that $r_0 \approx 0.675$. Using (14), it follows that for $r \in [0, r_0)$ the complexity of spurious projection operators is dominating, i.e., $\mathscr{E}^2_{\mathcal{G}_\emptyset^{(1)}} > \mathscr{E}^2_{\mathcal{T}_\emptyset^{(1)}}$. In contrast, for $r \in (r_0, 1)$, we have the complexity of the residualized signals to be dominating, i.e., $\mathscr{E}^2_{\mathcal{T}_\emptyset^{(1)}} > \mathscr{E}^2_{\mathcal{G}_\emptyset^{(1)}}$. Right panel of Figure 6 indicates the phase transition between the two complexities. Since the identifiability margin $\tau_*(1)$ roughly behaves like $\beta_1^2(1-r^2)$, the margin quantity becomes very small for a high value of r. Hence, for model consistency, we need a high value for β_1^2.

Remark 1. *In the example of equi-correlated design with $\mathcal{S} = \{1\}$, the effect of correlation parameter r on the complexities $\mathscr{E}_{\mathcal{T}_\emptyset^{(1)}}$ and $\mathscr{E}_{\mathcal{G}_\emptyset^{(1)}}$ are complementary to each other. In the case of the set of residualized signals, increasing correlation among the features increases the overall complexity of the set $\mathcal{T}_\emptyset^{(1)}$ and vice versa. In contrast, higher correlation decreases the complexity $\mathscr{E}_{\mathcal{G}_\emptyset^{(1)}}$, thus shrinking the effective size of $\mathcal{G}_\emptyset^{(1)}$. Thus, in this case, the two complexities act as two opposing forces in the margin condition (12).*

4.3. Necessary condition

One question that arises from the preceding discussion is whether the margin condition in Theorem 1 is necessary for model consistency or not. Specifically, it is natural to ask whether the complexities of residualized signals and spurious projections also characterize the necessary margin condition. In this section, we show that a condition very similar to (12) is necessary for model consistency of BSS, which is also governed by a similar margin quantity and complexity measures.

For $j_0 \in \mathcal{S}$, we define the set $\mathscr{C}_{j_0} := \{\mathcal{D} : \mathcal{S} \setminus \mathcal{D} = \{j_0\}, |\mathcal{D}| = s\} \subset \mathscr{A}_{s,1}$. We consider the maximum *leave-one-out* identifiability margin for $j_0 \in \mathcal{S}$ as

$$\widehat{\tau}(s) := \max_{j_0 \in \mathcal{S}} \max_{\mathcal{D} \in \mathscr{C}_{j_0}} \frac{\boldsymbol{\beta}^\top_{\mathcal{S}\setminus\mathcal{D}} \Gamma(\mathcal{D}) \boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}}{|\mathcal{S} \setminus \mathcal{D}|} = \max_{j_0 \in \mathcal{S}} \max_{\mathcal{D} \in \mathscr{C}_{j_0}} \Gamma(\mathcal{D}) \beta^2_{j_0}. \tag{15}$$

Consider the set $\mathcal{I}_0 := \mathcal{S} \setminus \{j_0\}$ for a fixed index $j_0 \in \mathcal{S}$. We capture the complexity of $\mathcal{T}^{(s)}_{\mathcal{I}_0}$ through the following quantity:

$$\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} := \frac{\sup_{\delta>0} \frac{\delta}{2}\sqrt{\log \mathcal{M}(\{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0\cup\{j\}}\}_{j\in\mathcal{S}^c}, \|\cdot\|_2, \delta)}}{\sqrt{\log|\mathcal{T}^{(s)}_{\mathcal{I}_0}|}}. \tag{16}$$

The above display immediately shows that $\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}} \geq \mathsf{d}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}/2$. Also, from the property of packing and covering number, it follows that

$$\mathcal{M}(\{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0\cup\{j\}}\}_{j\in\mathcal{S}^c}, \|\cdot\|_2, \delta) \leq \mathcal{N}(\{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0\cup\{j\}}\}_{j\in\mathcal{S}^c}, \|\cdot\|_2, \delta/2).$$

As $\mathcal{N}(\{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0\cup\{j\}}\}_{j\in\mathcal{S}^c}, \|\cdot\|_2, \delta/2)$ is a decreasing function over $\delta \in (0,\infty)$, we have the following inequality:

$$\sup_{\delta>0}\frac{\delta}{2}\sqrt{\log \mathcal{N}(\{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0\cup\{j\}}\}_{j\in\mathcal{S}^c}, \|\cdot\|_2, \delta/2)} \leq \int_0^\infty \sqrt{\log \mathcal{N}(\{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0\cup\{j\}}\}_{j\in\mathcal{S}^c}, \|\cdot\|_2, \varepsilon)}\, d\varepsilon.$$

The above inequality further shows that $\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} \leq \mathscr{E}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} \leq \mathsf{D}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}$. Hence, similar to $\mathscr{E}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}$, the alternative complexity measure $\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}$ also captures the average separation among the elements in $\mathcal{T}^{(s)}_{\mathcal{I}_0}$.

Next, we focus on the set $\mathcal{G}^{(s)}_{\mathcal{I}_0}$ which is the collection of all the spurious projection operators of the form $\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}_0}$ for all $\mathcal{D} \in \mathscr{C}_{j_0}$. If $\mathcal{D} = \mathcal{I}_0 \cup \{j\}$ for some $j \in \mathcal{S}^c$, then the corresponding spurious projection operator takes the form

$$\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}_0} = \widehat{\mathbf{u}}_j \widehat{\mathbf{u}}_j^\top, \tag{17}$$

where $\widehat{\mathbf{u}}_j$ denotes the unit vector along the residualized feature vector $\mathbf{u}_j := (\mathbb{I}_n - \mathbf{P}_{\mathcal{I}_0})\mathbf{X}_j$. Thus, the above display basically shows that the $\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}_0}$ is the orthogonal projection operator onto the linear span generated by the residualized feature $\mathbf{u}_j$. Similar to (16), we define the complexity measure of $\mathcal{G}^{(s)}_{\mathcal{I}_0}$ as

$$\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}} := \frac{\sup_{\delta>0} \frac{\delta}{2}\sqrt{\log \mathcal{M}(\mathcal{G}^{(s)}_{\mathcal{I}_0}, \|\cdot\|_{\mathrm{op}}, \delta)}}{\sqrt{\log|\mathcal{G}^{(s)}_{\mathcal{I}_0}|}}. \tag{18}$$

By a similar argument, it also follows that $\mathsf{d}_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}/2 \leq \mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}} \leq \mathsf{D}_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}$. Hence, combining the above observation with (17), it also follows that $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}$ captures the angular separation among the elusive features $\{\mathbf{u}_j\}_{j\in\mathcal{S}^c}$. Next, we introduce some technical assumptions that are crucial for our theoretical analysis of the necessity result.

Assumption 3. *The complexities of the $\mathcal{G}^{(s)}_{\mathcal{I}_0}$ and $\mathcal{T}^{(s)}_{\mathcal{I}_0}$ are not too small, i.e.,*

$$\mathscr{E}^{*2}_{\mathcal{G}^{(s)}_{\mathcal{I}_0}} > 16\{\log(ep)\}^{-1}, \quad \textit{and} \quad \mathscr{E}^{*2}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} > 16\{\log(ep)\}^{-1}$$

for all $\mathcal{I}_0 \subset \mathcal{S}$ and $|\mathcal{I}_0| = s-1$.

Assumption 3 combined with the observation (17) essentially tells that the set of elusive features $\{\widehat{\mathbf{u}}_j\}_{j\in\mathcal{S}^c}$ and the scaled spurious signals $\{\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}\}_{\mathcal{D}\in\mathscr{C}_{j_0}}$ are not too identical with each other, as $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}$ and $\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}$ would be typically small otherwise. Thus, Assumption 3 induces diversity in $\mathcal{T}^{(s)}_{\mathcal{I}_0}$ and $\mathcal{G}^{(s)}_{\mathcal{I}_0}$.

Condition 1. *There exists a constant $\alpha \in (0,1)$ such that $\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}/\mathscr{E}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} \in (\alpha, 1)$.*

The condition essentially tells that the set $\mathcal{T}^{(s)}_{\mathcal{I}_0}$ has a somewhat regular geometric shape in the sense that both the lower and upper complexity are of the same order. This essentially implies that minimal separation and maximal separation of the set $\mathcal{T}^{(s)}_{\mathcal{I}_0}$ are of the same order.
.

Now we present our theorem on the necessary condition for model consistency of BSS.

Theorem 2 (Necessity)**.** *Assume $\boldsymbol{\varepsilon} \sim \mathsf{N}(0, \sigma^2\mathbb{I}_n)$, $p > 16e^3$ and $s < p/2$. Also, let the Assumption 3 hold and write $\mathcal{J} = \{\mathcal{I} \subset \mathcal{S} : |\mathcal{I}| = s-1\}$. Then the followings are true:*

(a) If $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}} \notin (\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}, \mathscr{E}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}})$ for all $\mathcal{I}_0 \in \mathcal{J}$, then there exists a universal constant $C_1 > 0$ such that

$$\widehat{\tau}(s) \le C_1 \max\left\{\max_{\mathcal{I}_0\in\mathcal{J}} \mathscr{E}^{*2}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}, \max_{\mathcal{I}_0\in\mathcal{J}} \mathscr{E}^{*2}_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}\right\} \frac{\sigma^2\log(ep)}{n}$$

implies that

$$\mathbb{P}(\widehat{\mathcal{S}}_{\text{best}}(s) \neq \mathcal{S}) \ge \frac{1}{10}.$$

(b) If there exists $\mathcal{I}_\# \in \mathcal{J}$ such that $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_\#}} \in (\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_\#}}, \mathscr{E}_{\mathcal{T}^{(s)}_{\mathcal{I}_\#}})$, then under Condition 1, there exists a constant C_α depending on α, such that

$$\widehat{\tau}(s) \le C_\alpha \max\left\{\max_{\mathcal{I}_0\in\mathcal{J}} \mathscr{E}^{*2}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}, \max_{\mathcal{I}_0\in\mathcal{J}} \mathscr{E}^{*2}_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}\right\} \frac{\sigma^2\log(ep)}{n}$$

implies that

$$\mathbb{P}(\widehat{\mathcal{S}}_{\text{best}}(s) \neq \mathcal{S}) \ge \frac{1}{10}.$$

The detailed proof can be found in Section S1.4 of the supplementary material. The above theorem essentially says that if the maximum leave-one-out margin $\widehat{\tau}(s) \lesssim \sigma^2(\log p)/n$ then the BSS fails to achieve model consistency with positive probability. However, the interesting part of the above theorem is to understand the effect of the term involving complexity measures. Similar to Theorem 1, here also we see that the dominating complexity characterizes the necessary condition for model consistency. However, we reiterate a few major differences between the above theorem and Theorem 1. First, Theorem 1 needs $\tau_*(s)$ to be lower bounded, which is much stronger than the required condition on $\widehat{\tau}(s)$ in Theorem 2. Second, Theorem 2 involves the alternative complexity measures $\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}$ and $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}$, which are typically smaller than the complexity measures used in Theorem 1. Third, the resulting

complexity in Theorem 1 involves the maximum over all possible subsets of $\mathcal{S}$, whereas Theorem 2 involves the maximum only over the subsets of $\mathcal{S}$ of size $(s-1)$. These three facts are the main reasons that the requirement in Theorem 2 is weaker compared to the margin condition (12). Nonetheless, Theorem 2 is still interesting as it shows that the two types of complexities are indeed important quantities to understand the model selection performance of BSS.

Theorem 2 can also be stated in terms of the diameter and minimum separability of the sets $\mathcal{T}_{\mathcal{I}_0}^{(s)}$ and $\mathcal{G}_{\mathcal{I}_0}^{(s)}$. Recall that $\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}^{(s)}} \gtrsim \mathsf{d}_{\mathcal{T}_{\mathcal{I}_0}^{(s)}}$ and $\mathscr{E}^*_{\mathcal{G}_{\mathcal{I}_0}^{(s)}} \gtrsim \mathsf{d}_{\mathcal{G}_{\mathcal{I}_0}^{(s)}}$. Hence, it follows that under the same conditions in Theorem 2, the margin condition

$$\widehat{\tau}(s) \gtrsim \max\left\{ \max_{\mathcal{I}_0 \in \mathcal{J}} \mathsf{d}^2_{\mathcal{T}_{\mathcal{I}_0}^{(s)}}, \max_{\mathcal{I}_0 \in \mathcal{J}} \mathsf{d}^2_{\mathcal{G}_{\mathcal{I}_0}^{(s)}} \right\} \frac{\sigma^2 \log(ep)}{n}$$

is necessary for model consistency of BSS.

5. Experiments

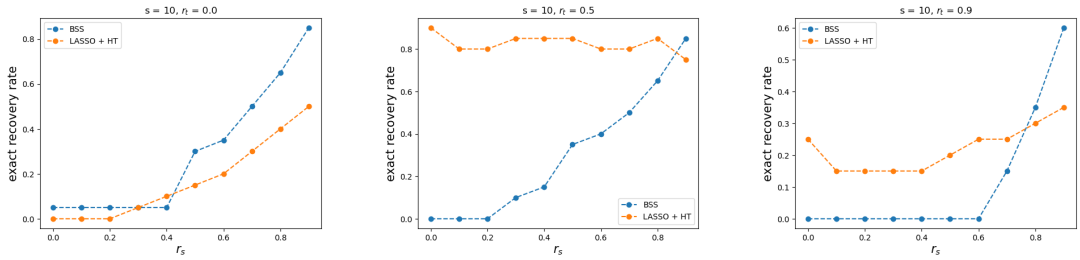
In this section, we will compare the performance of BSS with that of LASSO, one of the most popular tools for model selection. In these experiments, we set $p = 2000, s = 10$ and $n = 500$. We construct the design matrix $\mathbf{X}$ by sampling each row $\mathbf{x}_i \sim N(0, \mathbf{\Sigma})$ for $i \in [n]$, where

$$\mathbf{\Sigma} = \begin{bmatrix} (1-r_t)\mathbb{I}_s + r_t \mathbf{1}_s \mathbf{1}_s^\top & \mathbf{0}_{s\times(p-s)} \\ \mathbf{0}_{(p-s)\times s} & (1-r_s)\mathbb{I}_{p-s} + r_s \mathbf{1}_{p-s}\mathbf{1}_{p-s}^\top \end{bmatrix}$$

and $r_t, r_s \in [0,1]$. We set $\boldsymbol{\beta} \in \mathbb{R}^p$ so that $\beta_j = 0.2 \times \mathbb{1}\{j \le s\}$ for all $j \in [p]$. The responses $\{y_i\}_{i\in[n]}$ are generated according to model (1) with $\varepsilon_i \sim \mathsf{N}(0,1)$. For the experiments, we vary $r_t \in \{0.0, 0.5, 0.9\}$ and $r_s \in \{0.0, 0.1, \ldots, 0.9\}$.

We use ABESS [28] as a computational surrogate for BSS, and we also provide the knowledge of s to the algorithm. For LASSO, we choose the penalty parameter by five-fold cross-validation and then choose the s coordinates with the highest absolute values of the estimated LASSO coefficient, i.e., we perform hard thresholding (HT) operation on the LASSO estimator. In Figure 7, we plot the exact recovery rates of BSS and LASSO + HT for varying choices of r_t and r_s.

In all the cases in Figure 7, we see that the performance of BSS improves as r_s increases to 1. This is in fact consistent with the theoretical results in Theorem 1 as the overall complexity of the spurious signals is likely smaller for high values of r_s, and thus the margin condition is easier to satisfy. In this case, complexity of residualized signals are somewhat unaffected as the cross-correlation between true and spurious signals is 0. However, the performance of LASSO+ HT does not seem to exhibit any particular behavior across different correlation structure. For $r_t = 0$, performance of LASSO + HT, although comparatively worse than BSS, is similar in terms of the trend. However, for $r_t = 0.5$, LASSO + HT is consistently better and somewhat stable. For $r_t = 0.9$, LASSO + HT is generally better than BSS and exact recovery rate increases for $r_s \ge 0.4$. However, its performance is much worse compared to the $r_t = 0.5$ case. These observations indicate that similar complexity theory for LASSO is in fact potentially challenging and evidently requires more future research.



(a) $r_t = 0.0$ (b) $r_t = 0.5$ (c) $r_t = 0.9$

Fig 7: Exact recovery rate of BSS and LASSO + HT for varying choices of r_t and r_s.

6. Conclusion

In this paper, we establish the sufficient and (nearly) necessary conditions for BSS to achieve model consistency in a high-dimensional linear regression setup. Apart from the identifiability margin, we show that the geometric complexity of the residualized signals and spurious projections based on the entropy number and packing numbers also play a crucial role in characterizing the margin condition for model consistency of BSS. In particular, we establish that the dominating complexity among the two plays a decisive role in the margin condition. We also highlight the variation in these complexity measures under different correlation strengths between the features through some simple illustrative examples. Moreover, in the supplementary material, we extend the results in Theorem 1 to the high-dimensional sparse generalized linear models (Section S2). However, it is an open problem to find the analogs of the two complexities in more general settings, e.g., the low-rank matrix regression problem or multi-tasking regression problem.

Supplementary Material

The supplementary material contains the extension of Theorem 1 to the generalized linear model and the proofs of the main results.

References

[1] ADLER, R. J., TAYLOR, J. E. et al. (2007). *Random fields and geometry* **80**. Springer.
[2] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **19** 716–723.
[3] AKAIKE, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* 199–213. Springer.
[4] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields* **113** 301–413.
[5] BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. https://doi.org/10.1214/15-AOS1388

[6] BERTSIMAS, D. and PARYS, B. V. (2020). Sparse high-dimensional regression: exact scalable algorithms and phase transitions. *Ann. Statist.* **48** 300–323. https://doi.org/10.1214/18-AOS1804

[7] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. https://doi.org/10.1214/08-AOS620

[8] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. https://doi.org/10.1198/016214501753382273

[9] FOSTER, D., KARLOFF, H. and THALER, J. (2015). Variable selection is hard. In *Conference on Learning Theory* 696–709. PMLR.

[10] GUO, Y., ZHU, Z. and FAN, J. (2020). Best subset selection is robust against design dependence. *arXiv preprint arXiv:2007.01478*. https://doi.org/10.48550/arXiv.2007.01478

[11] HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statist. Sci.* **35** 579–592. https://doi.org/10.1214/19-STS733

[12] JAIN, P., TEWARI, A. and KAR, P. (2014). On Iterative Hard Thresholding Methods for High-dimensional M-Estimation. In *Advances in Neural Information Processing Systems* **27**. https://proceedings.neurips.cc/paper/2014/hash/218a0aefd1d1a4be65601cc6ddc1520e-Abstract.html.

[13] KRAHMER, F., MENDELSON, S. and RAUHUT, H. (2014). Suprema of chaos processes and the restricted isometry property. *Commun. Pure. Appl. Math.* **67** 1877–1904.

[14] LIFSHITS, M. A. (1995). *Gaussian random functions* **322**. Springer Science & Business Media.

[15] LIU, H. and FOYGEL BARBER, R. (2020). Between hard and soft thresholding: optimal iterative thresholding algorithms. *Inform. Inference: J. IMA* **9** 899–933. https://doi.org/10.1093/imaiai/iaz027

[16] MALLOWS, C. L. (2000). Some comments on Cp. *Technometrics* **42** 87–94.

[17] ROY, S., TEWARI, A. and ZHU, Z. (2022). High-dimensional variable selection with heterogeneous signals: A precise asymptotic perspective. *arXiv preprint arXiv:2201.01508*.

[18] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 461–464.

[19] SHE, Y., SHEN, J. and BARBU, A. (2023). Slow Kill for Big Data Learning. *IEEE Trans. Inform. Theory* **69** 5936-5955. https://doi.org/10.1109/TIT.2023.3273179

[20] TALAGRAND, M. (2005). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

[21] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. Ser. B* **58** 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[22] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3** 1360–1392. https://doi.org/10.1214/09-EJS506

[23] YE, K. and LIM, L.-H. (2016). Schubert varieties and distances between subspaces of different dimensions. *SIAM J. Matrix Anal. Appl.* **37** 1176–1197.

[24] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942.

[25] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. https://doi.org/10.1214/07-AOS520

[26] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593.

[27] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory* 921–948. PMLR https://proceedings.mlr.press/v35/zhang14.html.

[28] ZHU, J., WANG, X., HU, L., HUANG, J., JIANG, K., ZHANG, Y., LIN, S. and ZHU, J. (2022). abess: A Fast Best-Subset Selection Library in Python and R. *Journal of Machine Learning Research* **23** 1–7.

[29] ZHU, J., WEN, C., ZHU, J., ZHANG, H. and WANG, X. (2020). A polynomial algorithm for best-subset selection problem. *Proc. Natl. Acad. Sci. U.S.A.* **117** 33117–33123. https://doi.org/10.1073/pnas.2014241117

Electronic Journal of Statistics
ISSN: 1935-7524

Supplement to "Understanding Best Subset Selection: A Tale of Two C(omplex)ities"

Saptarshi Roy[1], **Ambuj Tewari**[1] **and Ziwei Zhu**[2]

[1]*Department of Statistics, University of Michigan, Ann Arbor, USA*

[2]*Quantitative Research, Radix Trading, Chicago, USA*

Contents

S1. Proof of main results under linear model

S1.1. Proof of Lemma 1

First note that $\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top \Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} = 0 \Leftrightarrow (\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} = 0$. This shows that

$$\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} \in \mathrm{col}(\mathbf{X}_{\mathcal{D}}).$$

Thus, we have $\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} + \mathbf{X}_{\mathcal{S}\cap\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\cap\mathcal{D}} \in \mathrm{col}(\mathbf{X}_{\mathcal{D}})$. This finishes the proof.

S1.2. Proof of Proposition 1

In this section, we will show that the SRC condition (5) is strictly stronger than the condition in Assumption 1. Recall that the features are normalized, i.e., $\left\|\mathbf{X}_j\right\|_2 = \sqrt{n}$ for all $j \in [p]$. Now, we will prove the proposition.

Proof. SRC implies Assumption 1:

For a set $\mathcal{I} \subset \mathcal{S}$, define $\mathscr{A}_{\mathcal{I}} := \{\mathcal{D} \in \mathscr{A}_s : \mathcal{S} \cap \mathcal{D} = \mathcal{I}\}$. Now recall that $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(s)}} \gtrsim \mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$ for large p. Thus, it suffices to show that $\mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(s)}}$ is large for all choices of $\mathcal{I} \subset \mathcal{S}$. Let $\mathcal{D}_1, \mathcal{D}_2 \in \mathscr{A}_{\mathcal{I}}$ and write $\mathcal{M} = \mathcal{D}_1 \cap \mathcal{D}_2$. Let $m = |\mathcal{M}|$ and consider the two subspaces $L_1 = \mathrm{col}(\mathbf{X}_{\mathcal{D}_1}) \cap \mathrm{col}(\mathbf{X}_{\mathcal{M}})^{\perp}$ and $L_2 = \mathrm{col}(\mathbf{X}_{\mathcal{D}_2}) \cap \mathrm{col}(\mathbf{X}_{\mathcal{M}})^{\perp}$. Let $\{\boldsymbol{\xi}_j\}_{j=1}^{m}$ be an orthonormal basis of $\mathcal{M}$. Let $\{\boldsymbol{\alpha}_j\}_{j=1}^{s-m}$ be an orthonormal basis of L_1 and $\{\boldsymbol{\delta}_j\}_{j=1}^{s-m}$ be the orthonormal basis of L_2 such that

$$\theta_j := \angle(\boldsymbol{\alpha}_j, \boldsymbol{\delta}_j), \quad j \in [k],$$

are the principal angles between L_1 and L_2 in decreasing order. Now, we construct the matrix $\mathbf{Z}$ in the following way:

$$\mathbf{Z} = \left[\mathbf{X}_{\mathcal{D}_1 \setminus \mathcal{D}_2} \mid \mathbf{X}_{\mathcal{M}} \mid \mathbf{X}_{\mathcal{D}_2 \setminus \mathcal{D}_1}\right].$$

There exists matrix $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{s-m}$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^m$ such that

$$\boldsymbol{\alpha}_1 = \mathbf{X}_{\mathcal{D}_1 \setminus \mathcal{D}_2}\mathbf{u} + \mathbf{X}_{\mathcal{M}}\mathbf{w}_1 \quad \text{and} \quad \boldsymbol{\delta}_1 = \mathbf{X}_{\mathcal{D}_2 \setminus \mathcal{D}_1}\mathbf{v} + \mathbf{X}_{\mathcal{M}}\mathbf{w}_2.$$

As $\boldsymbol{\alpha}_1 \perp \mathrm{col}(\mathbf{X}_{\mathcal{M}})$, we have

$$1 = \boldsymbol{\alpha}_1^{\top}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^{\top}\mathbf{X}_{\mathcal{D}_1 \setminus \mathcal{D}_2}\mathbf{u} \leq \sqrt{n\kappa_+}\,\|\mathbf{u}\|_2 \Rightarrow \|\mathbf{u}\|_2^2 \geq 1/(n\kappa_+).$$

By a similar argument, we have $\|\mathbf{v}\|_2^2 \geq 1/(n\kappa_+)$. Define the vectors $\boldsymbol{\eta} := (\mathbf{u}^{\top}, (\mathbf{w}_1 - \mathbf{w}_2)^{\top}, \mathbf{v}^{\top})^{\top}$. Hence, $\|\boldsymbol{\eta}\|_2^2 \geq \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 \geq 2/(n\kappa_+)$. Due to SRC condition (4.7), we have

$$\|\mathbf{Z}\boldsymbol{\eta}\|_2^2 \geq n\,\|\boldsymbol{\eta}\|_2^2\,\kappa_- \geq 2(\kappa_-/\kappa_+). \tag{S1.1}$$

$$\begin{aligned} \|\mathbf{Z}\boldsymbol{\eta}\|_2^2 &= \|\boldsymbol{\alpha}_1 - \boldsymbol{\delta}_1\|_2^2 \\ &= 2(1 - \sqrt{1 - \sin^2\theta_1}) \\ &\leq 2\sin\theta_1, \end{aligned} \tag{S1.2}$$

where the last inequality follows from the fact that $1 - x \leq \sqrt{1 - x^2}$ for all $x \in [0, 1]$. Combining (S1.1) and (S1.2), we have

$$\left\|\mathbf{P}_{\mathcal{D}_1} - \mathbf{P}_{\mathcal{D}_2}\right\|_{\mathrm{op}} \geq \frac{\kappa_-}{\kappa_+}.$$

The above display shows that $\mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(s)}} \gtrsim (\kappa_-/\kappa_+) \gg \{\log(ep)\}^{-1/2}$ for all $\mathcal{I} \subset \mathcal{S}$. Hence, the claim follows.

Assumption 1 does not imply SRC:

In this case, assume $\mathcal{S} = \{1\}$ and $\mathbf{e}_j$ be the jth canonical basis in $\mathbb{R}^p$. Under this setup, Assumption 1 becomes

$$\mathscr{E}_{\mathcal{G}_{\emptyset}^{(1)}} > \{\log(ep)\}^{-1/2}. \tag{S1.3}$$

Now assume that

$$\frac{2}{\log(ep)} \leq \frac{\left\|\mathbf{X}_j - \mathbf{X}_{j'}\right\|_2^2}{n} \leq \frac{3}{\log(ep)}, \quad \text{for all } j, j' \in [p].$$

Then, for large p, the condition in (S1.3) holds but SRC fails with the choice of $\mathbf{v} = (\mathbf{e}_j - \mathbf{e}_{j'})/\sqrt{2}$. □

S1.3. Proof of Theorem 1

Recall that $\boldsymbol{\mu} = \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}$, $\boldsymbol{\gamma}_{\mathcal{D}} = n^{-1/2}(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\boldsymbol{\mu}$ and

$$\Gamma(\mathcal{D}) = \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\setminus\mathcal{D},\mathcal{S}\setminus\mathcal{D}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\setminus\mathcal{D},\mathcal{D}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{S}\setminus\mathcal{D}}. \tag{S1.4}$$

Note that for $\mathcal{D} \in \mathscr{A}_{s,k}$ and $0 \leq \eta < 1$ we have the following:

$$\begin{aligned}
n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) &= n^{-1}\{\mathbf{y}^\top(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{y} - \mathbf{y}^\top(\mathbb{I}_n - \mathbf{P}_{\mathcal{S}})\mathbf{y}\} \\
&= n^{-1}\left\{(\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} + \boldsymbol{\varepsilon})^\top(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})(\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} + \boldsymbol{\varepsilon}) - \boldsymbol{\varepsilon}^\top(\mathbb{I}_n - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}\right\} \\
&= \eta\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} \\
&\quad + 2^{-1}(1-\eta)\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} - 2\left\{n^{-1}(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}\right\}^\top(-\boldsymbol{\varepsilon}) \\
&\quad + 2^{-1}(1-\eta)\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} - n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}.
\end{aligned} \tag{S1.5}$$

Also, let $\widetilde{\boldsymbol{\varepsilon}} := (-\boldsymbol{\varepsilon})$. Now, $\mathcal{E}$ be an event under which the following happens:

$$\left\{\widehat{\mathcal{S}} : |\widehat{\mathcal{S}}| = s, \min_{\mathcal{S}\in\mathscr{A}_s} R_{\widehat{\mathcal{S}}} \leq R_{\mathcal{S}} + n\eta\tau_*(s)\right\} = \{\mathcal{S}\}.$$

Define the set $\mathscr{A}_{\mathcal{I}} := \{\mathcal{D} \in \mathscr{A}_s : \mathcal{S} \cap \mathcal{D} = \mathcal{I}\}$. We also set $|\mathcal{I}| = s - k$ for $k \in [s]$. Then we have $\mathscr{A}_{\mathcal{I}} \subset \mathscr{A}_{s,k}$. By union bound we have the following:

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{k=1}^{s} \sum_{\mathcal{I}\subset\mathcal{S}:|\mathcal{I}|=s-k} \mathbb{P}\left\{\min_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) < \eta\tau_*(s)\right\}. \tag{S1.6}$$

Thus, under the light of equation (S1.5) it is sufficient to show the following with high probability:

$$\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^\top\widetilde{\boldsymbol{\varepsilon}} < \frac{n^{1/2}(1-\eta)}{4} \min_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2, \tag{S1.7}$$

$$\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} n^{-1}\left\{\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}\right\} < \frac{1-\eta}{2} \min_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2^2, \tag{S1.8}$$

for every $k \in [s]$. We will analyze the above events separately. We recall the two important sets below:

$$\mathcal{T}_{\mathcal{I}}^{(s)} := \{\widehat{\boldsymbol{\gamma}}_{\mathcal{D}} : \mathcal{D} \in \mathscr{A}_s, \mathcal{D} \cap \mathcal{S} = \mathcal{I}\}, \text{ and } \mathcal{G}_{\mathcal{I}}^{(s)} := \{\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}} : \mathcal{D} \in \mathscr{A}_s, \mathcal{D} \cap \mathcal{S} = \mathcal{I}\}.$$

To reduce notational cluttering, we will drop the s in the superscript, and use $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{G}_{\mathcal{I}}$ to denote the above sets.

Linear term: Let $f_{\mathcal{D}} := \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}}$ and $\|f\| := \max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} f_{\mathcal{D}}$. Since $\mathsf{D}_{\mathcal{T}_{\mathcal{I}}} \leq \sqrt{2}$, Theorem 5.36 of [12] tells that there exists a constant $A_1 > 0$ such that

$$\mathbb{P}\left\{\|f\| \geq A_1\sigma(\mathscr{E}_{\mathcal{T}_{\mathcal{I}}}\sqrt{k\log(ep)} + u)\right\} \leq 3\exp\left(-\frac{u^2}{2}\right), \tag{S1.9}$$

for all $u > 0$. Setting $u = 2c_{\mathcal{T}}\sqrt{k\log(ep)}$ in Equation (S1.9) we get

$$\mathbb{P}(\|f\| \geq A_1\sigma\mathscr{E}_{\mathcal{T}_{\mathcal{I}}}\sqrt{k\log(ep)} + 2A_1c_{\mathcal{T}}\sigma\sqrt{k\log(ep)}) \leq 3(ep)^{-2c_{\mathcal{T}}^2k}. \tag{S1.10}$$

Writing A_1 as c_1, we get

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}} \geq c_1(\mathscr{E}_{\mathcal{T}_{\mathcal{I}}} + 2c_{\mathcal{T}})\sigma\sqrt{k\log(ep)}\right\} \leq 3(ep)^{-2c_{\mathcal{T}}^2k}. \tag{S1.11}$$

Quadratic term: Here we study the quadratic supremum process in Equation (S1.8). First, define the two projection operators $\mathbf{P}_{\mathcal{D}|\mathcal{I}} = \mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}}$ and $\mathbf{P}_{\mathcal{S}|\mathcal{I}} := \mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{I}}$. For any number $c_{\mathcal{G}} \in (\{\log(ep)\}^{-1}, 1)$, by union bound we have,

$$\begin{aligned}
&\mathbb{P}\left\{n^{-1}\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon} > \sigma^2u + \sigma^2c_{\mathcal{G}}u_0\right\} \\
&\mathbb{P}\left\{n^{-1}\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}|\mathcal{I}} - \mathbf{P}_{\mathcal{S}|\mathcal{I}})\boldsymbol{\varepsilon} > \sigma^2u + \sigma^2c_{\mathcal{G}}u_0\right\} \\
&\leq \mathbb{P}\left\{n^{-1}(k\sigma^2 - \boldsymbol{\varepsilon}^{\top}\mathbf{P}_{\mathcal{S}|\mathcal{I}}\boldsymbol{\varepsilon}) > \sigma^2u_0c_{\mathcal{G}}\right\} + \mathbb{P}\left\{n^{-1}\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}}(\boldsymbol{\varepsilon}^{\top}\mathbf{P}_{\mathcal{D}|\mathcal{I}}\boldsymbol{\varepsilon} - k\sigma^2) > \sigma^2u\right\}.
\end{aligned} \tag{S1.12}$$

Also, note that $\mathbb{E}(\boldsymbol{\varepsilon}^{\top}\mathbf{P}_{\mathcal{D}|\mathcal{I}}\boldsymbol{\varepsilon}) \leq k\sigma^2$ and recall that $\mathscr{E}_{\mathcal{G}_{\mathcal{I}}} > \{\log(ep)\}^{-1/2}$ for all $\mathcal{I} \subset \mathcal{S}$. This shows that $\sqrt{k} \leq \mathscr{E}_{\mathcal{G}_{\mathcal{I}}}\sqrt{k\log(ep)}$. Furthermore, by the properties of projection matrices, we have $d_{\mathrm{op}}(\mathcal{G}_{\mathcal{I}}) = 1$ and $d_F(\mathcal{G}_{\mathcal{I}}) = \sqrt{k}$ (defined in Section S4). Also, it follows that the quantities M, V and U (defined in Theorem S4.2) have the following properties:

$$M \leq 2\mathscr{E}_{\mathcal{G}_{\mathcal{I}}}^2 k\log(ep), \quad V \leq 2\sqrt{k\log(ep)}, \quad \text{and} \quad U = 1.$$

Using these facts and Theorem S4.2, we get that there exists a universal positive constants A_2, A_3, such that for $t = A_3c_{\mathcal{G}}k\log(ep)$, we get

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \boldsymbol{\varepsilon}^{\top}\mathbf{P}_{\mathcal{D}|\mathcal{I}}\boldsymbol{\varepsilon} \geq A_2\sigma^2(\mathscr{E}_{\mathcal{G}_{\mathcal{I}}}^2 + c_{\mathcal{G}})k\log(ep)\right\} \leq (ep)^{-2c_{\mathcal{G}}^2k}. \tag{S1.13}$$

Due to Theorem 1.1 of [9], setting $u_0 = k\log(ep)/(2n)$ we can show that there exists an absolute constant $A_4 > 0$ such that

$$\begin{aligned}
\mathbb{P}\left\{n^{-1}\left|\boldsymbol{\varepsilon}^{\top}\mathbf{P}_{\mathcal{S}|\mathcal{I}}\boldsymbol{\varepsilon} - k\sigma^2\right| > \frac{c_{\mathcal{G}}\sigma^2k\log(ep)}{2n}\right\} &\leq 2\exp\left\{-A_4c_{\mathcal{G}}k\log(ep)\right\} \\
&= 2(ep)^{-A_4c_{\mathcal{G}}k},
\end{aligned} \tag{S1.14}$$

Combining Equation (S1.12), (S1.13) and Equation (S1.14) yields

$$\begin{aligned}
&\mathbb{P}\left\{n^{-1}\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon} > c_2(\mathscr{E}_{\mathcal{G}_{\mathcal{I}}}^2 + c_{\mathcal{G}})\sigma^2\frac{k\log(ep)}{n}\right\} \\
&\leq (ep)^{-2c_{\mathcal{G}}^2k} + 2(ep)^{-A_4c_{\mathcal{G}}k},
\end{aligned} \tag{S1.15}$$

where c_2 is a universal constant. Now, if we have

$$\begin{aligned}\tau_*(s) &\triangleq \min_{\mathcal{D}\neq\mathcal{S}} \frac{\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top \Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}}{|\mathcal{S}\setminus\mathcal{D}|} \\ &\geq \frac{64}{(1-\eta)^2}\max\left\{c_1 \max_{\mathcal{I}\subset\mathcal{S}}(\mathscr{E}_{\mathcal{T}_\mathcal{I}} + 2c_\mathcal{T})^2, c_2 \max_{\mathcal{I}\subset\mathcal{S}}(\mathscr{E}^2_{\mathcal{G}_\mathcal{I}} + c_\mathcal{G})\right\}\frac{\sigma^2\log(ep)}{n},\end{aligned} \tag{S1.16}$$

it will ensure that (S1.7) and (S1.8) hold with high probability. Finally, using (S1.11) and (S1.15), we have

$$\begin{aligned}&\mathbb{P}(\mathcal{E}^c) \\ &\leq \sum_{k=1}^{s}\sum_{\mathcal{I}\subset\mathcal{S}:|\mathcal{I}|=s-k}\mathbb{P}\left\{\min_{\mathcal{D}\in\mathscr{A}_\mathcal{I}} n^{-1}(R_\mathcal{D}-R_\mathcal{S}) < \eta\tau_*(s)\right\} \\ &\lesssim \sum_{k=1}^{s}\sum_{\mathcal{I}\subset\mathcal{S}:|\mathcal{I}|=s-k}\left\{(ep)^{-2c_\mathcal{T}^2 k} + (ep)^{-2c_\mathcal{G}^2 k} + (ep)^{-A_4 c_\mathcal{G} k}\right\} \\ &\lesssim \sum_{k=1}^{s}\binom{s}{k}\left\{(ep)^{-2c_\mathcal{T}^2 k} + (ep)^{-2c_\mathcal{G}^2 k} + (ep)^{-A_4 c_\mathcal{G} k}\right\} \\ &\lesssim \sum_{k=1}^{s}(es)^k\left\{(ep)^{-2c_\mathcal{T}^2 k} + (ep)^{-2c_\mathcal{G}^2 k} + (ep)^{-A_4 c_\mathcal{G} k}\right\} \\ &\lesssim \sum_{k=1}^{s}\exp\left[-k\{2c_\mathcal{T}^2\log(ep) - \log(es)\}\right] + \exp[-k\{2c_\mathcal{G}^2\log(ep) - \log(es)\}] \\ &\quad + \exp\left[-k\{A_4 c_\mathcal{G}\log(ep) - \log(es)\}\right].\end{aligned}$$

Now, setting $c_\mathcal{T} = \sqrt{\{\log(es) \vee \log\log(ep)\}/\log(ep)}$ and $c_\mathcal{G} = (2 \vee A_4^{-1})c_\mathcal{T}$ in the above display, and using the identity $(a+b)^2 \leq 2(a^2+b^2)$ we can conclude that the following is sufficient to hold (S1.16):

$$\tau_*(s) \geq \frac{64}{(1-\eta)^2}\max\{8c_1, c_2(2\vee A_4^{-1})\}\left[\max\left\{\max_{\mathcal{I}\subset\mathcal{S}}\mathscr{E}^2_{\mathcal{T}_\mathcal{I}}, \max_{\mathcal{I}\subset\mathcal{S}}\mathscr{E}^2_{\mathcal{G}_\mathcal{I}}\right\} + c_\mathcal{T}\right]\frac{\log(ep)}{n}.$$

Now the result follows by renaming the absolute constant $64\max\{8c_1, c_2(2\vee A_4^{-1})\}$ as C_0.

S1.4. Proof of Theorem 2

Proof. First, we will show that $\|\boldsymbol{\gamma}_\mathcal{D}\|_2$ has to be well bounded away from 0 for every $\mathcal{D} \in \cup_{j_0\in\mathcal{S}}\mathscr{C}_{j_0}$. Again, to reduce notational cluttering, we drop the s in the superscript and use $\mathcal{T}_{\mathcal{I}_0}$ and $\mathcal{G}_{\mathcal{I}_0}$ to denote the sets of scaled residualized signals and spurious projections respectively.

Ruling out the case $\min_{\mathcal{D}\in\cup_{j_0\in\mathcal{S}}\mathscr{C}_{j_0}}\|\boldsymbol{\gamma}_\mathcal{D}\|_2 \leq \sigma/\sqrt{n}$ *:*

Let $\min_{\mathcal{D}\in\cup_{j_0\in\mathcal{S}}\mathscr{C}_{j_0}}\|\boldsymbol{\gamma}_\mathcal{D}\|_2 \leq \sigma/\sqrt{n}$, i.e, there exists $\mathcal{D} \in \mathscr{C}_{j_0}$ for some $j_0 \in \mathcal{S}$ such that

$$\|\boldsymbol{\gamma}_\mathcal{D}\|_2 \leq \sigma/\sqrt{n}.$$

Recall that $\mathbf{y} \sim \mathsf{N}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}, \sigma^2\mathbb{I}_n)$ and define $\mathbf{w} := \mathbf{y} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}$. Hence, we have

$$\begin{aligned}\mathbb{P}\left(R_{\mathcal{D}} - R_{\mathcal{S}} < 0\right) &= \mathbb{P}\left(\left\|\mathbf{y} - \mathbf{P}_{\mathcal{D}}\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}\right\|_2^2 < \left\|\mathbf{y} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}\right\|_2^2\right) \\ &= \mathbb{P}\left(\left\|\mathbf{w} + n^{1/2}\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2^2 < \|\mathbf{w}\|_2^2\right) \\ &\geq \frac{1}{2}\frac{e^{-0.5}}{\sqrt{2\pi}} > 0.1 \quad \text{(By Lemma S5.2)},\end{aligned}$$

i.e., $\mathbb{P}\left(\mathcal{S} \notin \arg\min_{\mathcal{D}:|\mathcal{D}|=s} R_{\mathcal{D}}\right)$ is strictly bounded away from 0. Hence, BSS can not recover the true model. Hence, we rule out this case.

Under the case $\min_{\mathcal{D}\in\cup_{j_0\in\mathcal{S}}\mathscr{C}_{j_0}} \left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2 > \sigma/\sqrt{n}$ *:*

We fix a $j_0 \in \mathcal{S}$.

First decomposition:

$$\begin{aligned}\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) &\leq \min_{\mathcal{D}\in\mathscr{C}_{j_0}} \{\boldsymbol{\beta}^\top_{\mathcal{S}\setminus\mathcal{D}}\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} - 2n^{-1/2}\boldsymbol{\gamma}^\top_{\mathcal{D}}\widetilde{\boldsymbol{\varepsilon}} - n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}\} \\ &\leq \min_{\mathcal{D}\in\mathscr{C}_{j_0}} \left[\left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2 \left\{\left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2 - 2n^{-1/2}\widehat{\boldsymbol{\gamma}}^\top_{\mathcal{D}}\widetilde{\boldsymbol{\varepsilon}}\right\} - n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}_0})\boldsymbol{\varepsilon}\right] \\ &\quad + n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{I}_0})\boldsymbol{\varepsilon} \\ &\leq \min_{\mathcal{D}\in\mathscr{C}_{j_0}} \left[\left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2 \left\{\left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2 - 2n^{-1/2}\widehat{\boldsymbol{\gamma}}^\top_{\mathcal{D}}\widetilde{\boldsymbol{\varepsilon}}\right\}\right] \quad + n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{I}_0})\boldsymbol{\varepsilon}\end{aligned} \tag{S1.17}$$

We start with the quadratic term in the right hand side of the above display. First note that $\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{I}_0})\boldsymbol{\varepsilon}/\sigma^2 = (\widehat{\mathbf{u}}^\top_{j_0}\boldsymbol{\varepsilon})^2/\sigma^2$ follows a chi-squared distribution with degrees of freedom 1. Hence, Markov's inequality shows that

$$\mathbb{P}\left\{n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\mathcal{S}} - \mathbf{P}_{\mathcal{I}_0})\boldsymbol{\varepsilon} > \frac{2\sigma^2}{n}\right\} \leq \frac{1}{2}. \tag{S1.18}$$

Recall that

$$\left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2 \geq \frac{\sigma}{\sqrt{n}}, \quad \text{for all } \mathcal{D} \in \mathscr{C}_{j_0}.$$

Next, by Sudakov's lower bound, we have

$$\mathbb{E}\left(\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \widehat{\boldsymbol{\gamma}}^\top_{\mathcal{D}}\widetilde{\boldsymbol{\varepsilon}}\right) \geq \sigma\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}}\sqrt{\log(p-s)} \geq \frac{\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}}}{\sqrt{2}}\sqrt{\log(ep)}.$$

The last inequality uses the fact that $s < p/2$ and $p > 16e^3$. Again, an application of Borell-TIS inequality yields

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \widehat{\boldsymbol{\gamma}}^\top_{\mathcal{D}}\widetilde{\boldsymbol{\varepsilon}} \geq \sigma\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}}\sqrt{\log(ep)} - c_{\mathcal{T}}\sigma\sqrt{\log(ep)}\right\} \geq 1 - (ep)^{-c^2_{\mathcal{T}}/2}, \tag{S1.19}$$

for any $c_{\mathcal{T}} > 0$. Choosing $c_{\mathcal{T}} = \mathscr{E}^*{}_{\mathcal{T}_{I_0}}(2^{-1/2} - 2^{-1})$, and using the fact that $\mathscr{E}^{*2}_{\mathcal{T}_{I_0}} \geq 16\{\log(ep)\}^{-1}$, we get

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}} \geq \sigma\mathscr{E}^*{}_{\mathcal{T}_{I_0}}\sqrt{\log(ep)}/2\right\} \geq 1 - e^{-1}. \tag{S1.20}$$

Let us define $\widehat{\tau}_{j_0} = \max_{\mathcal{D}\in\mathscr{C}_{j_0}} \Delta(\mathcal{D})\beta_{j_0}^2.$, and by construction we have $\widehat{\tau}(s) = \max_{j_0\in\mathcal{S}} \widehat{\tau}_{j_0}$. If $\widehat{\tau}_{j_0}^{1/2} \leq \sigma\mathscr{E}^*{}_{\mathcal{T}_{I_0}}\sqrt{\log(ep)}/(2n^{1/2})$, then we have

$$\min_{\mathcal{D}\in\mathscr{C}_{j_0}} \left[\|\boldsymbol{\gamma}_{\mathcal{D}}\|_2 \left\{\|\boldsymbol{\gamma}_{\mathcal{D}}\|_2 - 2n^{-1/2}\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}}\right\}\right] \leq -\frac{\sigma^2\mathscr{E}^*{}_{\mathcal{T}_{I_0}}\sqrt{\log(ep)}}{2n}$$

Thus, using (S1.18) and the above display we have

$$\mathbb{P}(\mathcal{S} \notin \arg\min_{\mathcal{D}:|\mathcal{D}|=s} R_{\mathcal{D}}) = \mathbb{P}\left(\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) < 0\right) \geq 1 - \frac{1}{e} - \frac{1}{2} \geq \frac{1}{10},$$

as we have $\mathscr{E}^*{}_{\mathcal{T}_{I_0}}\sqrt{\log(ep)} > 4$ (Assumption 2). Thus the necessary condition turns out to be

$$\widehat{\tau}_{j_0} \geq \frac{\mathscr{E}^{*2}_{\mathcal{T}_{I_0}}}{4}\frac{\sigma^2\log(ep)}{n}. \tag{S1.21}$$

Second decomposition:

We again start with the difference of RSS between a candidate model $\mathcal{D} \in \mathscr{A}_{s,k}$ and the true model $\mathcal{S}$:

$$\begin{aligned}
&n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) \\
&= n^{-1}\{\mathbf{y}^{\top}(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{y} - \mathbf{y}^{\top}(\mathbb{I}_n - \mathbf{P}_{\mathcal{S}})\mathbf{y}\} \\
&= n^{-1}\left\{(\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} + \boldsymbol{\varepsilon})^{\top}(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})(\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} + \boldsymbol{\varepsilon}) - \boldsymbol{\varepsilon}^{\top}(\mathbb{I}_n - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}\right\} \\
&= \boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^{\top}\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} - 2\left\{n^{-1}(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}\right\}^{\top}\widetilde{\boldsymbol{\varepsilon}} - n^{-1}\boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}.
\end{aligned} \tag{S1.22}$$

First of all, in order achieve model consistency, the following is necessary for any $k \in [s]$:

$$\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) > 0. \tag{S1.23}$$

Recall that $\widehat{\tau}_{j_0} = \max_{\mathcal{D}\in\mathscr{C}_{j_0}} \Gamma(\mathcal{D})\beta_{j_0}^2$. Next we note that

$$\begin{aligned}
\min_{\mathcal{D}\in\mathscr{D}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) &\leq \min_{\mathcal{D}\in\mathscr{C}_{j_0}} \{\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^{\top}\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}} - 2n^{-1/2}\boldsymbol{\gamma}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}} - n^{-1}\boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}\} \\
&\leq \widehat{\tau}_{j_0} + 2n^{-1/2}\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \left|\boldsymbol{\gamma}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}}\right| - n^{-1}\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon} \\
&\leq \widehat{\tau}_{j_0} + 2(\widehat{\tau}_{j_0}/n)^{1/2}\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \left|\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}}\right| - n^{-1}\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon}.
\end{aligned} \tag{S1.24}$$

Similar to the proof of Theorem 1, we define $f_{\mathcal{D}} := \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}}$ and $\|f\| := \max_{\mathcal{D}\in\mathscr{C}_{j_0}} f_{\mathcal{D}}$. Hence, we have

$$\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \left|\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\boldsymbol{\varepsilon}}\right| = \max_{\mathcal{D}\in\mathscr{C}_{j_0}} f_{\mathcal{D}} \vee (-f_{\mathcal{D}}) \tag{S1.25}$$

By Borell-TIS inequality [2, Theorem 2.1.1], we have

$$\mathbb{P}\left\{\|f\| - \mathbb{E}(\|f\|) \geq \sigma u\right\} \leq \exp\left(-\frac{u^2}{2}\right),$$

for all $u > 0$. Setting $u = c_{\mathcal{T}}\sqrt{\log(ep)}$ we get

$$\mathbb{P}\left\{\|f\| - \mathbb{E}(\|f\|) \geq c_{\mathcal{T}}\sigma\sqrt{\log(ep)}\right\} \leq (ep)^{-c_{\mathcal{T}}^2/2}.$$

$$\mathbb{E}(\|f\|) \leq 4\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}}\sigma\sqrt{\log(ep)},$$

which ultimately yields

$$\mathbb{P}\left\{\|f\| \geq (4\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}} + c_{\mathcal{T}})\sigma\sqrt{\log(ep)}\right\} \leq (ep)^{-c_{\mathcal{T}}^2/2}.$$

Finally. using (S1.25) we have the following for any $c_{\mathcal{T}} > 0$:

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \left|\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{E}\right| \geq (4\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}} + c_{\mathcal{T}})\sigma\sqrt{\log(ep)}\right\} \leq 2(ep)^{-c_{\mathcal{T}}^2/2}. \tag{S1.26}$$

Next, we will lower bound the quadratic term in Equation (S1.24) with high probability. similar to the proof of Theorem 1, we consider the decomposition

$$\max_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}\boldsymbol{\varepsilon}^{\top}(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\boldsymbol{\varepsilon} = \max_{j\notin\mathcal{S}} n^{-1}\boldsymbol{\varepsilon}^{\top}(\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^{\top} - \widehat{\mathbf{u}}_{j_0}\widehat{\mathbf{u}}_{j_0}^{\top})\boldsymbol{\varepsilon}.$$

For the maximal process we will use Theorem 2.10 of [1]. We begin with the definition of concentration property.

Definition 1 ([1]). *Let Z be random vector in $\mathbb{R}^n$. We say that Z has concentration property with constant K if for every 1-Lipschitz function $\varphi : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}\,|\varphi(Z)| < \infty$ and for every $u > 0$,*

$$\mathbb{P}\left(|\varphi(Z) - \mathbb{E}(\varphi(Z))| \geq u\right) \leq 2\exp(-u^2/K^2).$$

Note that the Gaussian vector $\boldsymbol{\varepsilon}/\sigma$ enjoys concentration property with $K = \sqrt{2}$ [4, Theorem 5.6]. Let $Q_1 := \max_{j\notin\mathcal{S}} \boldsymbol{\varepsilon}^{\top}(\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^{\top} - \widehat{\mathbf{u}}_{j_0}\widehat{\mathbf{u}}_{j_0}^{\top})\boldsymbol{\varepsilon}$. By Theorem 2.10 of [1] we conclude that

$$\mathbb{P}\left\{n^{-1}\left|Q_1 - \mathbb{E}(Q_1)\right| \geq t\sigma^2\right\} \leq 2\exp\left\{-\frac{1}{2}\min\left(\frac{n^2t^2}{16}, \frac{nt}{2}\right)\right\}.$$

Setting $t = 2\delta\log(ep)/n$ in the above equation we get

$$\mathbb{P}\left\{n^{-1}\left|Q_1 - \mathbb{E}(Q_1)\right| \geq 2\delta\sigma^2\log(ep)/n\right\} \leq 2(ep)^{-\frac{\delta}{2}}. \tag{S1.27}$$

Next, we will lower bound the expected value of Q_1. First, note the following:

$$\begin{aligned}\mathbb{E}(Q_1) &= \mathbb{E}\left\{\max_{j\notin\mathcal{S}} \boldsymbol{\varepsilon}^\top(\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^\top - \widehat{\mathbf{u}}_{j_0}\widehat{\mathbf{u}}_{j_0}^\top)\boldsymbol{\varepsilon}\right\} \\ &= \mathbb{E}\left\{\max_{j\notin\mathcal{S}} (\widehat{\mathbf{u}}_j^\top\boldsymbol{\varepsilon})^2\right\} - \sigma^2 \\ &\geq \left\{\mathbb{E}\max_{j\in\mathcal{S}^c} (\widehat{\mathbf{u}}_j^\top\boldsymbol{\varepsilon}) \vee (-\widehat{\mathbf{u}}_j^\top\boldsymbol{\varepsilon})\right\}^2 - \sigma^2\end{aligned} \tag{S1.28}$$

Define the set

$$\mathcal{U}_{\text{sym}} := \{\widehat{\mathbf{u}}_j \ : j\in\mathcal{S}^c\} \cup \{-\widehat{\mathbf{u}}_j \ : j\in\mathcal{S}^c\}.$$

We denote by $\widetilde{\mathbf{u}}_j$ a generic element of $\mathcal{U}_{\text{sym}}$. Thus for any two elements $\widetilde{\mathbf{u}}_j, \widetilde{\mathbf{u}}_k$, we have

$$\left\|\widetilde{\mathbf{u}}_j - \widetilde{\mathbf{u}}_k\right\|_2 \geq \min\left\{\left\|\widehat{\mathbf{u}}_j - \widehat{\mathbf{u}}_k\right\|_2, \left\|\widehat{\mathbf{u}}_j + \widehat{\mathbf{u}}_k\right\|_2\right\} \geq \left\|\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^\top - \widehat{\mathbf{u}}_k\widehat{\mathbf{u}}_k^\top\right\|_{\text{op}}.$$

By Sudakov's lower bound, we have

$$\begin{aligned}\mathbb{E}\max_{j\in\mathcal{S}^c} (\widehat{\mathbf{u}}_j^\top\boldsymbol{\varepsilon}) \vee (-\widehat{\mathbf{u}}_j^\top\boldsymbol{\varepsilon}) &\geq \sup_{\delta>0} \frac{\sigma\delta}{2}\sqrt{\log \mathcal{M}(\mathcal{U}_{\text{sym}}, \|\cdot\|_2, \delta)} \\ &\geq \sup_{\delta>0} \frac{\sigma\delta}{2}\sqrt{\log \mathcal{M}(\{\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^\top\}_{j\in\mathcal{S}^c}, \|\cdot\|_{\text{op}}, \delta)} \\ &\geq \sigma\frac{\mathscr{E}^*_{\mathcal{G}_{I_0}}}{\sqrt{4/3}}\sqrt{\log(ep)}.\end{aligned}$$

The last inequality uses the fact that $p > 16e^3$. Finally, (S1.28) yields

$$\mathbb{E}(Q_1) \geq \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}\log(ep)}{4/3} - \sigma^2.$$

Thus, combined with (S1.27) we finally get

$$\mathbb{P}\left[Q_1 \geq (3/4)\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}\log(ep) - \sigma^2 - 2\delta\sigma^2\log(ep)\right] \geq 1 - 2(ep)^{-\delta/2}.$$

Setting $\delta = \frac{\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{8}$, we get

$$\mathbb{P}\left[Q_1 \geq \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{2}\log(ep) - \sigma^2\right] \geq 1 - 2(ep)^{-\frac{\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{16}}.$$

Thus, finally combining the above with (S1.24) and (S1.26) we get

$$\begin{aligned}&\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}}) \\ &\leq \widehat{\tau}_{j_0} + 2\widehat{\tau}_{j_0}^{1/2}(4\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}} + c_{\mathcal{T}})\sigma\sqrt{\frac{\log(ep)}{n}} - n^{-1}\left(\frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{2}\log(ep) - \sigma^2\right)\end{aligned} \tag{S1.29}$$

with probability at least $1 - 2(ep)^{-c_{\mathcal{T}}^2/2} - 2(ep)^{-\frac{\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{16}}$. Thus, for large p we have

$$\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{D}})$$
$$\leq \widehat{\tau}_{j_0} + 2\widehat{\tau}_{j_0}^{1/2}(4\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}} + c_{\mathcal{T}})\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{4}\frac{\log(ep)}{n}$$

with probability at least $1 - 2(ep)^{c_{\mathcal{T}}^2/2} - 2(ep)^{-\frac{\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{16}}$. Now choose $c_{\mathcal{T}} = 4/\sqrt{\log(ep)}$ and use Assumption 2 to get

$$\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{D}})$$
$$\leq \widehat{\tau}_{j_0} + 2\widehat{\tau}_{j_0}^{1/2}\left(4\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}} + \frac{4}{\sqrt{\log(ep)}}\right)\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{4}\frac{\log(ep)}{n}$$
$$\leq \widehat{\tau}_{j_0} + 8\sqrt{2}\widehat{\tau}_{j_0}^{1/2}\left(\mathscr{E}_{\mathcal{T}_{I_0}} + \frac{1}{\sqrt{2\log(ep)}}\right)\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{4}\frac{\log(ep)}{n}$$
$$\leq \widehat{\tau}_{j_0} + 10\sqrt{2}\widehat{\tau}_{j_0}^{1/2}\mathscr{E}_{\mathcal{T}_{I_0}}\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{4}\frac{\log(ep)}{n}$$

with probability at least $1/5$.

Thus, in light of (S1.23), the following is necessary:

$$\widehat{\tau}_{j_0} \geq \left\{\frac{\sqrt{200\mathscr{E}^2_{\mathcal{T}_{I_0}} + \mathscr{E}^{*2}_{\mathcal{G}_{I_0}}} - 10\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}}}{2}\right\}^2 \frac{\sigma^2\log(ep)}{n}. \tag{S1.30}$$

Case 1: If $\mathscr{E}_{\mathcal{T}_{I_0}} \leq \mathscr{E}^*{}_{\mathcal{G}_{I_0}}$, then the right hand side of (S1.30) is lower bounded by

$$\frac{\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{(\sqrt{201} + 10\sqrt{2})^2}\frac{\sigma^2\log(ep)}{n}.$$

Thus (S1.30) yields the necessary condition

$$\widehat{\tau}_{j_0} \geq \frac{\mathscr{E}^{*2}_{\mathcal{G}_{I_0}}}{(\sqrt{201} + 10\sqrt{2})^2}\frac{\sigma^2\log(ep)}{n}.$$

Combining this with (S1.21) we have the necessary condition to be

$$\widehat{\tau}_{j_0} \geq \tilde{C}_1 \max\{\mathscr{E}^{*2}_{\mathcal{T}_{I_0}}, \mathscr{E}^{*2}_{\mathcal{G}_{I_0}}\}\frac{\sigma^2\log(ep)}{n},$$

for a universal constant $\tilde{C}_1$.

Case 2: If $\mathscr{E}^*{}_{\mathcal{G}_{I_0}} \leq \mathscr{E}^*{}_{\mathcal{T}_{I_0}}$, then using the inequality $\sqrt{1+t} - \sqrt{t} < 1$ for all $t > 0$, we can conclude that the right hand side of (S1.30) is always smaller than

$$\frac{\mathscr{E}^{*2}{}_{\mathcal{G}_{I_0}}}{4} \frac{\sigma^2 \log(ep)}{n},$$

which is further smaller than

$$\frac{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}}{4} \frac{\sigma^2 \log(ep)}{n}.$$

Thus, combining this with (S1.21) we have the necessary condition to be

$$\widehat{\tau}_{j_0} \geq \frac{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}}{4} \frac{\sigma^2 \log(ep)}{n} = \max\{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}, \mathscr{E}^{*2}{}_{\mathcal{G}_{I_0}}\} \frac{\sigma^2 \log(ep)}{4n}.$$

Combining all these cases we finally have the following necessary condition for consistent model selection with $C_1, C_2 > 0$ being some absolute constants:

$$\widehat{\tau}_{j_0} \geq C_1 \max\{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}, \mathscr{E}^{*2}{}_{\mathcal{G}_{I_0}}\} \frac{\sigma^2 \log(ep)}{n}, \quad \text{if } \mathscr{E}^*{}_{\mathcal{G}_{I_0}} \notin (\mathscr{E}^*{}_{\mathcal{T}_{I_0}}, \mathscr{E}_{\mathcal{T}_{I_0}}), \tag{S1.31}$$

or,

$$\widehat{\tau}_{j_0} \geq C_2 \max\left\{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}, \left(\sqrt{200\mathscr{E}^2_{\mathcal{T}_{I_0}} + \mathscr{E}^{*2}{}_{\mathcal{G}_{I_0}}} - 10\sqrt{2}\mathscr{E}_{\mathcal{T}_{I_0}}\right)^2\right\} \frac{\sigma^2 \log(ep)}{n},$$

if $\mathscr{E}^*{}_{\mathcal{G}_{I_0}} \in (\mathscr{E}^*{}_{\mathcal{T}_{I_0}}, \mathscr{E}_{\mathcal{T}_{I_0}})$.

If there exists $I_0 \in \mathcal{J}$ such that $\mathscr{E}^*{}_{\mathcal{T}_{I_0}}/\mathscr{E}_{\mathcal{T}_{I_0}} \in (\alpha, 1)$, then in the preceding case it follows that the last display can be simplified in the following form:

$$\widehat{\tau}_{j_0} \geq C_2 \max\left\{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}, A_\alpha \mathscr{E}^{*2}{}_{\mathcal{G}_{I_0}}\right\} \frac{\sigma^2 \log(ep)}{n}, \quad \text{if } \mathscr{E}^*{}_{\mathcal{G}_{I_0}} \in (\mathscr{E}^*{}_{\mathcal{T}_{I_0}}, \mathscr{E}_{\mathcal{T}_{I_0}}),$$

where

$$A_\alpha = \left(\sqrt{1 + \frac{200}{\alpha^2}} + \frac{10\sqrt{2}}{\alpha}\right)^{-1}.$$

Thus, using the fact that $A_\alpha < 1$ and combining the previous three displays we have the necessary condition to be

$$\widehat{\tau}_{j_0} \geq C_2 A_\alpha \max\left\{\mathscr{E}^{*2}{}_{\mathcal{T}_{I_0}}, \mathscr{E}^{*2}{}_{\mathcal{G}_{I_0}}\right\} \frac{\sigma^2 \log(ep)}{n}. \tag{S1.32}$$

Since, (S1.31) and (S1.32) hold for all choices of j_0 and I_0 depending on whether the $\mathscr{E}^*{}_{\mathcal{G}_{I_0}} \in (\mathscr{E}^*{}_{\mathcal{T}_{I_0}}, \mathscr{E}_{\mathcal{T}_{I_0}})$ is satisfied or not, the claim follows. □

S1.5. Correlated random feature model example (Proof of Lemma 2)

Consider the model (1). We assume that the rows of $\mathbf{X}$ are independently generated from p-dimensional multivariate Gaussian distribution with mean-zero and variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & c\mathbf{1}_{p-1}^\top \\ c\mathbf{1}_{p-1} & (1-r)\mathbb{I}_{p-1} + r\mathbf{1}_{p-1}\mathbf{1}_{p-1}^\top \end{pmatrix},$$

where $c \in [0, 0.997], r \in [0, 1)$ and true model is $\mathcal{S} = \{1\}$. We need to further impose the restriction

$$c^2 < r + \frac{1-r}{p-1}$$

to ensure positive definiteness of Σ. In this case

$$\widehat{\tau} = \beta_1^2 \min_{j \neq 1} \left\{ \frac{\|\mathbf{X}_1\|_2^2}{n} - \frac{(\mathbf{X}_1^\top \mathbf{X}_j/n)^2}{\|\mathbf{X}_j\|_2^2/n} \right\} = \frac{\beta_1^2 \|\mathbf{X}_1\|_2^2}{n} - \beta_1^2 \max_{j \neq 1} \frac{(\mathbf{X}_1^\top \mathbf{X}_j/n)^2}{\|\mathbf{X}_j\|_2^2/n}.$$

We start with providing an upper bound on the margin quantity $\widehat{\tau}$. Using Equation (S5.57) and (S5.58), for any $\varepsilon_{n,p} \in (0, 1)$ we get

$$\mathbb{P}\left(\frac{\|\mathbf{X}_j\|_2^2}{n} \geq 1 + \varepsilon_{n,p} \right) \leq \exp(-n\varepsilon_{n,p}^2/16), \quad \forall j \in [p]. \tag{S1.33}$$

$$\mathbb{P}\left(\frac{\|\mathbf{X}_j\|_2^2}{n} \leq 1 - \varepsilon_{n,p} \right) \leq \exp(-n\varepsilon_{n,p}^2/4), \quad \forall j \in [p]. \tag{S1.34}$$

Using Equation (S1.33) and Equation (S1.34), we also get the following:

$$\mathbb{P}\left(\max_{j \neq 1} \left| \frac{\|\mathbf{X}_j\|_2^2}{n} - 1 \right| \geq \varepsilon_{n,p} \right) \leq 2p \exp(-n\varepsilon_{n,p}^2/16). \tag{S1.35}$$

Let $\mathbf{X}_j = (x_{1,j}, \ldots, x_{n,j})^\top$ for all $j \in [p]$. To this end we recall the *sub-Gaussian norm* $\|\cdot\|_{\psi_2}$ [11, Definiton 2.5.6] and the *sub-exponential norm* $\|\cdot\|_{\psi_1}$ [11, Definition 2.7.5]. Now due to [11, Lemma 2.7.7], we have that $\|x_{u,1}x_{u,j}\|_{\psi_1} \leq \|x_{u,1}\|_{\psi_2} \|x_{u,2}\|_{\psi_2} \leq 4$. Thus, by Berstein's inequality, we have

$$\mathbb{P}\left(\left| \frac{\mathbf{X}_1^\top \mathbf{X}_j}{n} - c \right| > \varepsilon_{n,p} \right) \leq \exp\left(-Cn \min\{\varepsilon_{n,p}, \varepsilon_{n,p}^2\} \right),$$

where $C > 0$ is a universal constant. Thus, we have

$$\mathbb{P}\left(\max_{j \neq 1} \left| \frac{\mathbf{X}_1^\top \mathbf{X}_j}{n} - c \right| > \varepsilon_{n,p} \right) \leq p \exp(-Cn \min\{\varepsilon_{n,p}, \varepsilon_{n,p}^2\}). \tag{S1.36}$$

Combining (S1.34), (S1.35) and (S1.36) we have

$$\begin{aligned}&\mathbb{P}\left[\left\{1+\varepsilon_{n,p}-\frac{(c-\varepsilon_{n,p})^2}{1+\varepsilon_{n,p}}\right\}\geq\frac{\widehat{\tau}}{\beta_1^2}\geq\left\{1-\varepsilon_{n,p}-\frac{(c+\varepsilon_{n,p})^2}{1-\varepsilon_{n,p}}\right\}\right]\\&\geq 1-\exp(-n\varepsilon_{n,p}^2/16)-2p\exp(-n\varepsilon_{n,p}^2/4)-p\exp(-C\varepsilon_{n,p}^2 n)\\&=1+o(1/p),\end{aligned}\tag{S1.37}$$

if $\varepsilon_{n,p}\asymp\{(\log p)/n\}^{1/2}$ and $(\log p)/n$ is small enough. Similarly, due to Bernstein's inequality, it can also be shown that

$$\mathbb{P}\left(\max_{j,k\neq 1}\left|\frac{\mathbf{X}_k^\top\mathbf{X}_j}{n}-r\right|\leq\varepsilon_{n,p}\right)\geq 1-p^2\exp(-Cn\varepsilon_{n,p}^2)=1+o(1/p),\tag{S1.38}$$

where $r\in[0,1)$ and with the same conditions on $\varepsilon_{n,p}$.

Next, we will analyze the geometric quantities. In this case, we have

$$\widehat{\boldsymbol{\gamma}}_j=\frac{\mathbf{X}_1-\frac{\mathbf{X}_j^\top\mathbf{X}_1}{\|\mathbf{X}_j\|_2^2}.\mathbf{X}_j}{\sqrt{\|\mathbf{X}_1\|_2^2-\frac{(\mathbf{X}_1^\top\mathbf{X}_j)^2}{\|\mathbf{X}_j\|_2^2}}}.$$

Note that

$$\|\widehat{\boldsymbol{\gamma}}_j-\widehat{\boldsymbol{\gamma}}_k\|_2^2=2(1-\widehat{\boldsymbol{\gamma}}_j^\top\widehat{\boldsymbol{\gamma}}_k)$$

and

$$\widehat{\boldsymbol{\gamma}}_j^\top\widehat{\boldsymbol{\gamma}}_k=\frac{\|\mathbf{X}_1\|_2^2/n-\frac{(\mathbf{X}_j^\top\mathbf{X}_1/n)^2}{\|\mathbf{X}_j\|_2^2/n}-\frac{(\mathbf{X}_k^\top\mathbf{X}_1/n)^2}{\|\mathbf{X}_k\|_2^2/n}+\frac{(\mathbf{X}_j^\top\mathbf{X}_1/n)(\mathbf{X}_k^\top\mathbf{X}_1/n)(\mathbf{X}_j^\top\mathbf{X}_k/n)}{(\|\mathbf{X}_j\|_2^2/n)(\|\mathbf{X}_k\|_2^2/n)}}{\sqrt{\|\mathbf{X}_1\|_2^2/n-\frac{(\mathbf{X}_1^\top\mathbf{X}_j/n)^2}{\|\mathbf{X}_j\|_2^2/n}}\sqrt{\|\mathbf{X}_1\|_2^2/n-\frac{(\mathbf{X}_1^\top\mathbf{X}_k/n)^2}{\|\mathbf{X}_k\|_2^2/n}}}.$$

Next, we consider the event

$$\mathcal{G}_n:=\left\{\max_{j\in[p]}\left|\frac{\|\mathbf{X}_j\|_2^2}{n}-1\right|\leq\varepsilon_{n,p},\max_{j\neq 1}\left|\frac{\mathbf{X}_1^\top\mathbf{X}_j}{n}-c\right|\leq\varepsilon_{n,p},\max_{j,k\neq 1}\left|\frac{\mathbf{X}_k^\top\mathbf{X}_j}{n}-r\right|\leq\varepsilon_{n,p}\right\}.$$

Due to (S1.33), (S1.34), (S1.36) and (S1.38) we have $\mathbb{P}(\mathcal{G}_n)=1+o(1/p)$. Also, for large n,p, the value of $\varepsilon_{n,p}$ can be chosen such that $\varepsilon_{n,p}<0.001$ so that $c+\varepsilon_{n,p}<0.998$ for all $c\in[0,0.997]$.

Complexity of unexplained signals: Let $\mathbf{u}:=(u_1,u_2,u_3)\in\mathbb{R}^3$ and $\mathbf{t}:=(t_1,t_2,t_2)\in\mathbb{R}^3$. Define the function

$$\Phi(\mathbf{u},\mathbf{t}):=\frac{u_1-(t_1^2/u_2)-(t_2^2/u_3)+(t_1t_2t_3)/(u_2u_3)}{\sqrt{u_1-t_1^2/u_2}\sqrt{u_1-t_2^2/u_3}},$$

where

$$(u_1, u_2, u_3, t_1, t_2, t_3) \in \underbrace{[0.999, 1.001] \times [0.999, 1.001] \times [0.999, 1.001] \times [0, 0.998] \times [0, 0.998] \times [0, 1]}_{:=\mathcal{K}}.$$

It is easy to see that the function Φ is continuously differentiable on the compact set $\mathcal{K}$. Hence, there exists a universal constant $L > 0$ such that

$$|\Phi(\mathbf{u}, \mathbf{t}) - \Phi(\mathbf{u}', \mathbf{t}')| \leq L(\|\mathbf{u} - \mathbf{u}'\|_1 + \|\mathbf{t} - \mathbf{t}'\|_1).$$

Since we have

$$\widehat{\boldsymbol{\gamma}}_j^\top \boldsymbol{\gamma}_k = \Phi\left(\frac{\|\mathbf{X}_1\|_2^2}{n}, \frac{\|\mathbf{X}_j\|_2^2}{n}, \frac{\|\mathbf{X}_K\|_2^2}{n}, \frac{\mathbf{X}_1^\top \mathbf{X}_j}{n}, \frac{\mathbf{X}_1^\top \mathbf{X}_k}{n}, \frac{\mathbf{X}_j^\top \mathbf{X}_k}{n}\right),$$

it follows that on the event $\mathcal{G}_n$, the following holds for all $j, k \in [p] \setminus \{1\}$:

$$\left|\left\|\widehat{\boldsymbol{\gamma}}_j - \widehat{\boldsymbol{\gamma}}_k\right\|_2^2 - \frac{2c^2(1-r)}{1-c^2}\right| \leq 12L\varepsilon_{n,p},$$
$$\Rightarrow \sqrt{\max\left\{\frac{2c^2(1-r)}{1-c^2} - 12L\varepsilon_{n,p}, 0\right\}} \leq \left\|\widehat{\boldsymbol{\gamma}}_j - \widehat{\boldsymbol{\gamma}}_k\right\|_2 \leq \sqrt{\frac{2c^2(1-r)}{1-c^2} + 12L\varepsilon_{n,p}}.$$

Hence, we have

$$\begin{aligned}
&\mathscr{E}_{\mathcal{T}_\varnothing} \\
&= \{\log(ep)\}^{-1/2}\Bigg[\int_0^{\sqrt{\left(\frac{2c^2(1-r)}{1-c^2} - 12L\varepsilon_{n,p}\right)\vee 0}} \sqrt{\log \mathcal{N}(\mathcal{T}_\varnothing, \|\cdot\|_2, \varepsilon)}\, d\varepsilon \\
&\quad + \int_{\sqrt{\left(\frac{2c^2(1-r)}{1-c^2} - 12L\varepsilon_{n,p}\right)\vee 0}}^{\sqrt{\frac{2c^2(1-r)}{1-c^2} + 12L\varepsilon_{n,p}}} \sqrt{\log \mathcal{N}(\mathcal{T}_\varnothing, \|\cdot\|_2, \varepsilon)}\, d\varepsilon\Bigg].
\end{aligned}$$

Applying Lemma S5.4 on the second integral, it follows that

$$\omega_{n,p}\sqrt{\frac{\log p}{\log(ep)}} \leq \mathscr{E}_{\mathcal{T}_\varnothing} \leq \left(\omega_{n,p} + \sqrt{24L\varepsilon_{n,p}}\right)\sqrt{\frac{\log p}{\log(ep)}},$$

where $\omega_{n,p} = \sqrt{\left(\frac{2c^2(1-r)}{1-c^2} - 12L\varepsilon_{n,p}\right) \vee 0}$. Thus, for $c = 0$ we have $0 \leq \mathscr{E}_{\mathcal{T}_\varnothing} \leq \sqrt{24L\varepsilon_{n,p}}$. For any fixed $c > 0$ and $r \in [0, 1)$ we have

$$\mathscr{E}_{\mathcal{T}_\varnothing} \sim \left\{\frac{2c^2(1-r)}{1-c^2}\right\}^{1/2} \quad \text{for large } n, p. \tag{S1.39}$$

Complexity of spurious projections: For $j, k \neq 1$, let $\theta_{j,k}$ denote the angle between $\mathbf{X}_j$ and $\mathbf{X}_k$.

$$\left\|\mathbf{P}_j - \mathbf{P}_k\right\|_{\text{op}} = \sin(\theta_{j,k}) = \sqrt{1 - \cos^2(\theta_{j,k})} = \sqrt{1 - \left(\frac{\mathbf{X}_j^\top \mathbf{X}_k}{\left\|\mathbf{X}_j\right\|_2 \left\|\mathbf{X}_k\right\|_2}\right)^2}.$$

By a similar argument as above, we can conclude that there exists a universal constant $M > 0$ such that on the event $\mathcal{G}_n$ we have,

$$\left|\left\|\mathbf{P}_j - \mathbf{P}_k\right\|_{\text{op}}^2 - (1 - r^2)\right| \leq M\varepsilon_{n,p}, \quad \text{for all } j, k \in [p] \setminus \{1\}.$$

Thus, for any fixed $r \in [0, 1)$ we have

$$\mathscr{E}_{\mathcal{G}_\varnothing} \sim (1 - r^2)^{1/2}. \tag{S1.40}$$

S2. Generalized linear model

In this section, we will focus on the best subset selection problem under generalized linear models (GLM). Similar to the linear regression setup, we will also adopt the fixed design setup in this case. In particular, given the data matrix $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$ we observe the responses $\mathbf{y} := (y_1, \ldots, y_n)^\top$ coming from the distribution

$$f_{\mathbf{x},\boldsymbol{\beta}^*}(y) := h(y) \exp\left\{\frac{y(\mathbf{x}^\top \boldsymbol{\beta}^*) - b(\mathbf{x}^\top \boldsymbol{\beta}^*)}{\phi}\right\} = h(y) \exp\left\{\frac{y\eta - b(\eta)}{\phi}\right\}. \tag{S2.41}$$

Here $\eta = \mathbf{x}^\top \boldsymbol{\beta}^*$ is linear predictor and $\boldsymbol{\beta}^*$ is true parameter with $\|\boldsymbol{\beta}^*\|_0 = s$ and support $\mathcal{S}$. The functions $b : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are known and specific to modeling assumptions. Examples include several well-known models such as

1. Linear regression: Consider the linear regression model $y = \mathbf{x}^\top \boldsymbol{\beta}^* + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$. In this case $h(y) = \exp\{-y^2/(2\sigma^2)\}$ and $b(u) = u^2/2$.
2. Logistic regression: In this model $y \sim \mathsf{Ber}(1/(1 + \exp(-\mathbf{x}^\top \boldsymbol{\beta}^*)))$. Standard calculations show that $h(y) = 1$ and $b(u) = \log(1 + e^u)$.

For the purpose of model selection, we choose the loss function to be the scaled negative log-likelihood function

$$\mathcal{L}(\boldsymbol{\beta}; \{(\mathbf{x}_i, y_i)\}_{i \in [n]}) = \frac{2}{n} \sum_{i \in [n]} \ell(\boldsymbol{\beta}; (\mathbf{x}_i, y_i)),$$

where $\ell(\boldsymbol{\beta}; (\mathbf{x}, y)) = -y(\mathbf{x}^\top \boldsymbol{\beta}) + b(\mathbf{x}^\top \boldsymbol{\beta})$. Furthermore, for a candidate model $\mathcal{D} \in \mathscr{A}_s$ and $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^s$, define the restricted version of the scaled negative log-likelihood function as

$$\mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{\beta}}; \{(\mathbf{x}_{i,\mathcal{D}}, y_i)\}_{i \in [n]}) := (n/2)^{-1} \sum_{i \in [n]} \ell_{\mathcal{D}}(\tilde{\boldsymbol{\beta}}; (\mathbf{x}_{i,\mathcal{D}}, y_i)),$$

where $\ell_{\mathcal{D}}(\tilde{\boldsymbol{\beta}}; (\mathbf{x}_{\mathcal{D}}, y)) := -y(\mathbf{x}_{\mathcal{D}}^\top \tilde{\boldsymbol{\beta}}) + b(\mathbf{x}_{\mathcal{D}}^\top \tilde{\boldsymbol{\beta}})$. Let $\widehat{\boldsymbol{\beta}}_{\mathcal{D}} := \arg\min_{\tilde{\boldsymbol{\beta}}} \mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{\beta}}; \{(\mathbf{x}_{i,\mathcal{D}}, y_i)\}_{i \in [n]})$. Under the oracle knowledge of sparsity s, BSS solves for

$$\widehat{\mathcal{S}}_{\text{best}} = n^{-1} \arg\min_{\mathcal{D}:|\mathcal{D}|=s} \mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}}; \{(\mathbf{x}_{i,\mathcal{D}}, y_i)\}_{i \in [n]}).$$

Next, we will introduce the quantities that capture the degree of separation between the true model $\mathcal{S}$ and a candidate model $\mathcal{D} \in \mathscr{A}_s$ and characterize the identifiability margin for model selection consistency. Let $\mathcal{P}_{\mathcal{D},\tilde{\boldsymbol{\beta}}}$ be probability measure corresponding to the joint density $\prod_{i\in[n]} f_{\mathbf{x}_{i,\mathcal{D}},\tilde{\boldsymbol{\beta}}}(y_i)$ and define

$$\begin{aligned}\Delta_{\mathrm{kl}}(\mathcal{D}) &:= \frac{2\phi}{n} \min_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^s} \mathsf{KL}\left(\mathcal{P}_{\mathcal{S},\boldsymbol{\beta}^*_{\mathcal{S}}} \,\middle\|\, \mathcal{P}_{\mathcal{D},\tilde{\boldsymbol{\beta}}}\right) \\ &= \frac{2}{n}\sum_{i=1}^{n}\left\{(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}^*_{\mathcal{S}}) b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}^*_{\mathcal{S}}) - b(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}^*_{\mathcal{S}})\right\} \\ &\quad - \max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^s} \frac{2}{n}\sum_{i=1}^{n}\left\{(\mathbf{x}_{i,\mathcal{D}}^\top \tilde{\boldsymbol{\beta}}) b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}^*_{\mathcal{S}}) - b(\mathbf{x}_{i,\mathcal{D}}^\top \tilde{\boldsymbol{\beta}})\right\},\end{aligned}$$

where $\mathsf{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence. The above quantity can be thought of as the degree of model separation as it measures the minimum KL-divergence between the likelihood generated by the data under $(\mathcal{S}, \boldsymbol{\beta}^*_{\mathcal{S}})$ and the likelihood generated by $\mathcal{D}$ and all possible choices of $\boldsymbol{\beta} \in \mathbb{R}^p$ with the support in $\mathcal{D}$. Let $\bar{\boldsymbol{\beta}}_{\mathcal{D}}$ be the minimizer of the optimization problem in the above display, i.e.,

$$\bar{\boldsymbol{\beta}}_{\mathcal{D}} := \arg\min_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^s} \mathsf{KL}\left(\mathcal{P}_{\mathcal{S},\boldsymbol{\beta}^*_{\mathcal{S}}} \,\middle\|\, \mathcal{P}_{\mathcal{D},\tilde{\boldsymbol{\beta}}}\right).$$

By definition it follows that $\bar{\boldsymbol{\beta}}_{\mathcal{S}} = \boldsymbol{\beta}^*_{\mathcal{S}}$. Also, note that for $\mathcal{P}_{\mathcal{D},\bar{\boldsymbol{\beta}}_{\mathcal{D}}}$, the *natural parameter* of the density function is $\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}}$. Thus, one can also measure the separation between two models through the mutual distance between the corresponding natural parameters. This motivates the definition of the second measure of separability between the true model $\mathcal{S}$ and candidate model $\mathcal{D}$:

$$\Delta_{\mathrm{par}}(\mathcal{D}) := \frac{\left\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}}\right\|_2^2}{n}.$$

Note that, under the linear regression model with isotropic Gaussian error, both $\Delta_{\mathrm{kl}}(\mathcal{D})$ and $\Delta_{\mathrm{par}}(\mathcal{D})$ becomes equal to the quantity $\boldsymbol{\beta}^\top_{\mathcal{S}\setminus\mathcal{D}}\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}$. To see this, recall that for linear regression model $b(u) = u^2/2$ and the KL-divergence

$$\mathsf{KL}(\mathcal{P}_{\mathcal{S},\boldsymbol{\beta}^*_{\mathcal{S}}} \,\|\, \mathcal{P}_{\mathcal{D},\bar{\boldsymbol{\beta}}_{\mathcal{D}}}) = \left\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}}\right\|_2^2/(2\sigma^2).$$

Thus, from the definition of $\bar{\boldsymbol{\beta}}_{\mathcal{D}}$, it immediately follows that $\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} = \mathbf{P}_{\mathcal{D}}\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}$. Later, we will see that these two notions of distances are equivalent under certain regularity conditions on the link function $b(\cdot)$.

S2.1. Identifiability margin and two complexities

In this section we will introduce the identifiability margin and the two complexities similar the case of linear model. We consider the following identifiability margin:

$$\tilde{\tau}_*(s) := \min_{\mathcal{D}\in\mathscr{A}_s} \frac{\Delta_{\mathrm{kl}}(\mathcal{D})}{|\mathcal{S}\setminus\mathcal{D}|}.$$

We assume that $\tilde{\tau}_*(s) > 0$ to avoid non-identifiability issue. Next, we consider the transformed features as follows:

$$\widetilde{\mathbf{X}}_{\mathcal{D}} = \mathbf{\Lambda}_{\mathcal{D}}^{1/2}\mathbf{X}_{\mathcal{D}},$$

where $\mathbf{\Lambda}_{\mathcal{D}} = \mathbf{diag}(b''(\mathbf{x}_{1,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}}), \ldots, b''(\mathbf{x}_{n,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}}))$. Let $\widetilde{\mathbf{P}}_{\mathcal{D}}$ be orthogonal projection matrices onto the columnspace of $\widetilde{\mathbf{X}}_{\mathcal{D}}$. Let $\widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}}$ be the orthogonal projector onto the columnspace of $[\widetilde{\mathbf{X}}_{\mathcal{D}}]_{\mathcal{I}}$. Now we define the following sets of residualized signals and spurious projections:

$$\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)} = \left\{ \frac{\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}}{\left\|\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}\right\|_2} : \mathcal{D} \in \mathscr{A}_{\mathcal{I}} \right\},$$

$$\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)} = \left\{ \widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}} : \mathcal{D} \in \mathscr{A}_{\mathcal{I}} \right\}.$$

The complexity measures for these two sets are $\mathscr{E}_{\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)}}$ and $\mathscr{E}_{\widetilde{\mathcal{G}}_{\mathcal{I}(s)}}$ respectively, which are defined in the same way as the complexity measures in Section 3.2 and Section 3.3 of the main paper.

S2.2. Main results

In this section we will state the main result analogous to the Theorem 1 in the main paper. We begin with some standard assumption necessary for the theoretical analysis for GLM models.

Assumption S2.1 (Features and parameters)**.** *We assume the following conditions:*

(a) There exists positive constants x_0 and R_0 such that $\max_{i\in[n]} \|\mathbf{x}_i\|_\infty \le x_0$ and $\|\boldsymbol{\beta}^*\|_1 \le R_0$.

(b) There exists a constant $\kappa_0 > 0$ such that

$$\min_{\mathcal{D}\subset[p]:|\mathcal{D}|=s} \lambda_{\min}\left(\mathbf{X}_{\mathcal{D}}^\top\mathbf{X}_{\mathcal{D}}/n\right) \ge \kappa_0.$$

(c) There exists constant $M > 0$ such that

$$\max_{\mathcal{D}\subset[p]:|\mathcal{D}|=s} \left\| \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_{i,\mathcal{D}} \otimes \mathbf{x}_{i,\mathcal{D}} \otimes \mathbf{x}_{i,\mathcal{D}} \right\|_{\text{op}} \le M.$$

(d) There exists a constant $R > 0$ such that $\max_{\mathcal{D}\in\mathscr{A}_s} \max_{i\in[n]} \left|\mathbf{x}_{i,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}}\right| \le x_0 R$.

(e) The design matrix $\mathbf{X}$ enjoys the following property:

$$\min_{\mathcal{I}\subset\mathcal{S}} \mathscr{E}^2_{\tilde{\mathcal{G}}_{\mathcal{I}}^{(s)}} > \{\log(ep)\}^{-1}.$$

Assumption S2.1(a) is very common in high-dimensional literature. Assumption S2.1(b) basically tells that the sparse-eigenvalues of $\mathbf{X}$ are strictly bounded away from 0. Assumption S2.1(c) tells that the third order empirical moment of $\mathbf{X}_{\mathcal{D}}$ is bounded. A stronger version of Assumption S2.1(d) is present in [8, 13], where the authors assume that $\left\|\bar{\boldsymbol{\beta}}_{\mathcal{D}}\right\|_1$ is bounded uniformly over all $\mathcal{D} \in \mathscr{A}_s$. Finally, Assumption S2.1(e) allows diversity among the spurious features. Next, we will assume some technical assumptions on the link function $b(\cdot)$.

Assumption S2.2 ($b(\cdot)$ function)**.** *We assume the following conditions on $b(\cdot)$ function:*

(a) There exists a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any $\eta \in \mathbb{R}$ and any $\omega > 0$, $b''(\eta) \geq \psi(\omega)$ whenever $|\eta| \leq \omega$.

(b) There exists constants $B > 0$ and $\tilde{B} \geq 0$ such that $\|b''\|_\infty \leq B$ and $\|b'''\|_\infty \leq \tilde{B}$.

These assumptions on the link function are pretty common to analyze high-dimensional generalized models. Assumption S2.2(a) basically assumes that $b(\cdot)$ is strongly convex within a compact neighborhood of 0. It is straightforward to check that this assumption is satisfied by standard GLM setups like linear regression and logistic regression. In particular, one can choose $\psi(\omega) = 1$ for linear regression, and $\psi(\omega) = (3 + e^{\omega})^{-1}$ in the case of logistic regression. Furthermore, from (S2.41) it follows that $\mathbb{E}(y) = b'(\mathbf{x}^\top \boldsymbol{\beta}^*)$ and $\mathsf{var}(y) = \phi b''(\mathbf{x}^\top \boldsymbol{\beta}^*) \geq \phi\psi(\omega)$, whenever $|\mathbf{x}^\top \boldsymbol{\beta}^*| \leq \omega$.

Finally, Assumption S2.2(b) tells that the second and third derivatives of $b(\cdot)$ are bounded. This guarantees the first convergence rates of the maximum likelihood estimator. Moreover, this assumption guarantees sub-Gaussianity of y as

$$\begin{aligned}
&\mathbb{E}(\exp\{t(y - b'(\eta)\}) \\
&= e^{-tb'(\eta)} \int_{-\infty}^{\infty} h(y) \exp\left\{\frac{(\eta + \phi t)y - b(\eta)}{\phi}\right\} dy \\
&= \exp\left(\frac{b(\eta + \phi t) - b(\eta) - t\phi b'(\eta)}{\phi}\right) \int_{-\infty}^{\infty} h(y) \exp\left\{\frac{(\eta + \phi t)y - b(\eta + \phi t)}{\phi}\right\} dy \\
&= \exp[\phi^{-1}\{b(\eta + \phi t) - b(\eta) - t\phi b'(\eta)\}] \leq \exp\left(\frac{\phi B t^2}{2}\right).
\end{aligned} \tag{S2.42}$$

Under Assumption S2.2, we can compare between the margin quantities $\Delta_{\mathrm{kl}}(\mathcal{D})$ and $\Delta_{\mathrm{par}}(\mathcal{D})$. To see this, we first focus on $\Delta_{\mathrm{kl}}(\mathcal{D})$. Recall that

$$\begin{aligned}
&\Delta_{\mathrm{kl}}(\mathcal{D}) \\
&= \frac{2}{n}\sum_{i=1}^{n}\left\{(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*) b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*) - b(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*)\right\} - \frac{2}{n}\sum_{i=1}^{n}\left\{(\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}}) b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*) - b(\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\} \\
&= \frac{2}{n}\sum_{i=1}^{n}\left\{b(\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}}) - b(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*) - (\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*) b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} b''\left(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^* + t(\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*)\right)\{\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*\}^2,
\end{aligned}$$

for some $t \in (0, 1)$. Due to Assumption S2.1(a) and Assumption S2.1(d), we get

$$\left|\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^* + t(\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*)\right| \leq x_0(R_0 + R).$$

Finally, strong convexity and smoothness of $b(\cdot)$ (Assumption S2.2(a), S2.2(b)), we have

$$B\Delta_{\mathrm{par}}(\mathcal{D}) \geq \Delta_{\mathrm{kl}}(\mathcal{D}) \geq \psi(x_0 R_0 + x_0 R)\Delta_{\mathrm{par}}(\mathcal{D}). \tag{S2.43}$$

This established the equivalence between $\Delta_{\mathrm{kl}}(\mathcal{D})$ and $\Delta_{\mathrm{par}}(\mathcal{D})$. Now, we present the main below.

Theorem S2.1 (Sufficiency). *Under Assumption 1, there exists a positive constant C depending on $\phi, B, \tilde{B}, x_0, R, R_0, \kappa_0, M$ and the function $\psi(\cdot)$ such that for any $0 \le \eta < 1$, whenever the identifiability margin $\widetilde{\tau}_*(s)$ satisfies*

$$\frac{\widetilde{\tau}_*(s)}{\phi B} \ge \frac{C}{(1-\eta)^2} \max\left\{ \max\left\{ \max_{\mathcal{I}\subset\mathcal{S}} \mathscr{E}^2_{\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)}}, \max_{\mathcal{I}\subset\mathcal{S}} \mathscr{E}^2_{\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}} \right\} + \sqrt{\frac{\log(es)\vee\log\log(ep)}{\log(ep)}}, t^{(1)}_{s,n,p}, t^{(2)}_{s,n,p} \right\} \frac{\log(ep)}{n} \tag{S2.44}$$

for a specified $t^{(1)}_{s,n,p} = O(s\{\log s \vee \log\log p\}/\log p)$ and $t^{(2)}_{s,n,p} = O(\frac{s^2(\log n)^2}{n\log p} + \frac{s^{3/2}(\log n)^{3/2}}{\sqrt{n}\log p})$, we have

$$\left\{ \widehat{\mathcal{S}} : |\widehat{\mathcal{S}}| = s, \min_{\mathcal{S}\in\mathscr{A}_s} \mathcal{L}_{\widehat{\mathcal{S}}}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}}) \le \mathcal{L}_{\mathcal{S}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}) + n\eta\tilde{\tau}_*(s) \right\} = \{\mathcal{S}\},$$

with probability at least $1 - O(\{s\vee\log p\}^{-1} + n^{-7}s\log p)$. In particular, setting $\eta = 0$, we have $\mathcal{S} = \arg\min_{\widehat{\mathcal{S}}\in\mathscr{A}_s} \mathcal{L}_{\widehat{\mathcal{S}}}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{S}}})$ with high probability.

The proof of the above theorem is deferred to Section S3. The above theorem is the generalization of Theorem 1, and (S2.44) also involves the two complexities related to the sets of residualized signals and spurious projection operators. However, condition (S2.44) also involves two extra terms $t^{(1)}_{s,n,p}$ and $t^{(2)}_{s,n,p}$, the exact forms of which can be found in Section S3. It can be shown that both of these terms are exactly 0 for linear models as $\psi \equiv 1, B = 1$ and $\tilde{B} = 0$.

Remark 1. *If $p = \Omega(e^{c_0 n})$ for some universal constant $c_0 > 0$ and $s(\log n)/n \to 0$ as $n \to \infty$, then both $t^{(1)}_{s,n,p}$ and $t^{(2)}_{s,n,p}$ are negligible compared to the complexity term in (S2.44). Hence, in this case, we witness roughly a similar phenomenon involving the two complexities as in the linear model.*

S3. Proof of main results under GLM model

Let $\mathcal{D} \in \mathscr{A}_s$ such that $\mathcal{S}\cap\mathcal{D} = \mathcal{I}$.

Strong convexity

We will start by showing the strong convexity of $\mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{\beta}}; \{\mathbf{x}_i, y_i\}_{i\in[n]})$. For ease of presentation we will just write $\mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{\beta}})$ instead of $\mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{\beta}}; \{\mathbf{x}_i, y_i\}_{i\in[n]})$. Given any $r \in (0, R_0 \wedge R]$ and $\boldsymbol{\Delta} \in \mathbb{B}_1(\mathbf{0}, r)$ define the function

$$\begin{aligned}
\delta\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}} + \boldsymbol{\Delta}; \bar{\boldsymbol{\beta}}_{\mathcal{D}}) &:= \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}} + \boldsymbol{\Delta}) - \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) - \nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})^\top\boldsymbol{\Delta} \\
&= \frac{1}{2}\boldsymbol{\Delta}^\top\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}} + t\boldsymbol{\Delta})\boldsymbol{\Delta} \quad (\text{for some } t \in (0,1)) \\
&= \frac{1}{n}\sum_{i=1}^n b''(\mathbf{x}_{i,\mathcal{D}}^\top(\bar{\boldsymbol{\beta}}_{\mathcal{D}} + t\boldsymbol{\Delta}))(\mathbf{x}_{i,\mathcal{D}}^\top\boldsymbol{\Delta})^2 \\
&\ge \psi(x_0R + x_0r)\kappa_0\|\boldsymbol{\Delta}\|_2^2 \quad (\text{Using Assumption S2.1(a), S2.1(b), S2.1(d)}) \\
&\ge \psi(x_0R + x_0R_0)\kappa_0\|\boldsymbol{\Delta}\|_2^2
\end{aligned}$$

Rate of convergence

Construct an intermediate estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\alpha} = \bar{\boldsymbol{\beta}}_{\mathcal{D}} + \alpha(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})$ where

$$\alpha = \min\left\{1, \frac{r}{\|\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}\|_2}\right\},$$

where r will be chosen later.

Write $\widehat{\boldsymbol{\beta}}_{\mathcal{D},\alpha} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}$ as $\boldsymbol{\Delta}_\alpha$ and note that

$$\psi(x_0R + x_0R_0)\,\|\boldsymbol{\Delta}_\alpha\|_2^2 \le \delta\mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},\alpha}, \bar{\boldsymbol{\beta}}_{\mathcal{D}}) \le -\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})^\top\boldsymbol{\Delta}_\alpha \le \left\|\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_2 \|\boldsymbol{\Delta}_\alpha\|_2\,.$$

Hence we have

$$\|\boldsymbol{\Delta}_\alpha\|_2 \le \frac{\left\|\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_2}{\psi(x_0R + x_0R_0)} \le \frac{\sqrt{s}\left\|\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_\infty}{\psi(x_0R + x_0R_0)}. \tag{S3.45}$$

Now, note that

$$\begin{aligned}
\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) &:= -\frac{2}{n}\sum_{i\in[n]}\{y_i - b'(\mathbf{x}_{i,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}})\}\mathbf{x}_{i,\mathcal{D}} \\
&= -\frac{2}{n}\sum_{i\in[n]}\{y_i - b'(\mathbf{x}_{i,\mathcal{S}}^\top\boldsymbol{\beta}^*_{\mathcal{S}})\}\mathbf{x}_{i,\mathcal{D}} - \underbrace{\frac{2}{n}\sum_{i\in[n]}\{b'(\mathbf{x}_{i,\mathcal{S}}^\top\boldsymbol{\beta}^*_{\mathcal{S}}) - b'(\mathbf{x}_{i,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}})\}\mathbf{x}_{i,\mathcal{D}}}_{=0} \\
&= -\frac{2(\phi B)^{1/2}}{n}\sum_{i\in[n]}\underbrace{\frac{\{y_i - b'(\mathbf{x}_{i,\mathcal{S}}^\top\boldsymbol{\beta}^*_{\mathcal{S}})\}}{(\phi B)^{1/2}}}_{:=\epsilon_i}\mathbf{x}_{i,\mathcal{D}}
\end{aligned}$$

Note that $\mathbb{E}\{\exp(\lambda\epsilon_i[\mathbf{x}_{i,\mathcal{D}}]_j) \le \exp(\lambda^2x_0^2/2)\}$, i.e., $\epsilon_i[\mathbf{x}_{i,\mathcal{D}}]_j$ is sub-Gaussian with parameter x_0. Hence, by an application of union bound and Hoeffding's inequality we have

$$\mathbb{P}\left(\left\|\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_\infty \ge 2tx_0(\phi B)^{1/2}\right) \le 2s\exp\left(-\frac{nt^2}{2}\right). \tag{S3.46}$$

Setting $t = 4(\log n/n)^{1/2}$ in (S3.46) we get

$$\mathbb{P}\left(\left\|\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_\infty \ge 8x_0(\phi B)^{1/2}\sqrt{\frac{\log n}{n}}\right) \le \frac{2}{n^7}. \tag{S3.47}$$

Using the above fact and (S3.45) we finally get that with probability at least $1 - 2n^{-7}$ the following holds:

$$\|\boldsymbol{\Delta}_\alpha\|_2 \le \frac{8x_0(\phi B)^{1/2}}{\psi(x_0R + x_0R_0)}\sqrt{\frac{s\log n}{n}}.$$

Now we set $r = \frac{9x_0(\phi B)^{1/2}}{\psi(x_0R)}\sqrt{\frac{s\log n}{n}}$. Hence, we have $\|\boldsymbol{\Delta}_\alpha\|_2 < r$, i.e., $\|\boldsymbol{\Delta}\|_2 < r$. This shows that

$$\mathbb{P}\left(\max_{\mathcal{D}\in\mathscr{A}_s}\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}\right\|_2 > \frac{9x_0(\phi B)^{1/2}}{\psi(x_0R + x_0R_0)}\sqrt{\frac{s\log n}{n}}\right) \le \frac{4s\log p}{n^7}. \tag{S3.48}$$

By a similar argument, it can be shown that

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\right\|_2 > \frac{9x_0(\phi B)^{1/2}}{\psi(x_0R + x_0R_0)}\sqrt{\frac{s\log n}{n}}\right) \leq \frac{2}{n^7}. \tag{S3.49}$$

Expansion of likelihood estimate

Now that we have determined the rate of estimation, we can now write $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ in terms of $\bar{\boldsymbol{\beta}}_{\mathcal{D}}$. To see this, note that

$$\mathbf{0} = \nabla\mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}}) = \nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \mathbf{R}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})^{\otimes 2},$$

where $\mathbf{R}_{\mathcal{D}} = (1/2)\nabla^3\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}} + t_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}))$ for some $t_{\mathcal{D}} \in (0,1)$. Thus, we have

$$\widehat{\boldsymbol{\beta}}_{\mathcal{D}} = \bar{\boldsymbol{\beta}}_{\mathcal{D}} - [\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}\left(\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \mathbf{R}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})^{\otimes 2}\right) \tag{S3.50}$$

Higher order Taylor's expansion of loss function

Now we are ready to analyze the loss functions. We do so by expanding the Taylor series of the loss function. Write $\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}$ ad $\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}$. Then, using (S3.50) we have

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}}) \\
&= \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})^\top\widehat{\boldsymbol{\Delta}}_{\mathcal{D}} + \frac{1}{2}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\widehat{\boldsymbol{\Delta}}_{\mathcal{D}} + (1/3)\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top\tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\otimes\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}) \\
&= \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) - \nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})^\top[\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}\left(\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})^{\otimes 2}\right) \\
&+ \frac{1}{2}\left(\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})^{\otimes 2}\right)^\top[\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}\left(\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})^{\otimes 2}\right) \\
&+ \frac{1}{2}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top\tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\otimes\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}) \\
&= \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) - \frac{1}{2}\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})^\top[\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}\nabla\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) \\
&+ \frac{1}{2}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2})^\top[\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2}) + (1/3)\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top\tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\otimes\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}) \\
&= \frac{2}{n}\sum_{i\in[n]}\{-y_i(\mathbf{x}_{i,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + b(\mathbf{x}_{i,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}})\} \\
&- \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}}))^\top\mathbf{X}_{\mathcal{D}}(\widetilde{\mathbf{X}}_{\mathcal{D}}^\top\widetilde{\mathbf{X}}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}})) \\
&\quad + \frac{1}{2}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2})^\top[\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2}) + (1/3)\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top\tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\otimes\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}) \\
&= -\frac{2}{n}\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})^\top\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} + \frac{2}{n}\sum_{i\in[n]}b(\mathbf{x}_{i,\mathcal{D}}^\top\bar{\boldsymbol{\beta}}_{\mathcal{D}}) - \frac{2}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} \\
&\quad - \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top\mathbf{X}_{\mathcal{D}}(\widetilde{\mathbf{X}}_{\mathcal{D}}^\top\widetilde{\mathbf{X}}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \\
&\quad + \frac{1}{2}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2})^\top[\nabla^2\mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2}) + (1/3)\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top\tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\otimes\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}),
\end{aligned}$$

where $\tilde{\mathbf{R}}_{\mathcal{D}} = (1/2)\nabla^3 \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}} + \tilde{t}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}))$ for some $\tilde{t}_{\mathcal{D}} \in (0,1)$.

Thus, we have the following:

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}}) - \mathcal{L}_{\mathcal{S}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}) \\
&= \Delta_{\mathrm{kl}}(\mathcal{D}) \underbrace{- \frac{2}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top (\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})}_{\text{linear term}} \\
&\quad \underbrace{- \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \left\{ \mathbf{X}_{\mathcal{D}}(\widetilde{\mathbf{X}}_{\mathcal{D}}^\top \widetilde{\mathbf{X}}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top - \mathbf{X}_{\mathcal{S}}(\widetilde{\mathbf{X}}_{\mathcal{S}}^\top \widetilde{\mathbf{X}}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top \right\} (\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))}_{\text{quadratic term}} \\
&\quad + \frac{1}{2}(\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2})^\top [\nabla^2 \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1} (\tilde{\mathbf{R}}_{\mathcal{D}}\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^{\otimes 2}) + (1/3)\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top \tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}} \otimes \widehat{\boldsymbol{\Delta}}_{\mathcal{D}}) \\
&\quad - \frac{1}{2}(\tilde{\mathbf{R}}_{\mathcal{S}}\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}^{\otimes 2})^\top [\nabla^2 \mathcal{L}_{\mathcal{S}}(\boldsymbol{\beta}^*_{\mathcal{S}})]^{-1} (\tilde{\mathbf{R}}_{\mathcal{S}}\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}^{\otimes 2}) - (1/3)\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}^\top \tilde{\mathbf{R}}_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_{\mathcal{D}} \otimes \widehat{\boldsymbol{\Delta}}_{\mathcal{D}}).
\end{aligned}
$$

Write $\psi_* = \min\{\psi(x_0 R), \psi(x_0 R_0)\}$. By definition of $\widetilde{\mathbf{X}}_{\mathcal{D}}$, we have

$$
\widetilde{\mathbf{X}}_{\mathcal{D}}^\top \widetilde{\mathbf{X}}_{\mathcal{D}} = \mathbf{X}_{\mathcal{D}}^\top \boldsymbol{\Lambda}_{\mathcal{D}} \mathbf{X}_{\mathcal{D}} \succeq \psi_* \mathbf{X}_{\mathcal{D}}^\top \mathbf{X}_{\mathcal{D}}, \quad \text{where} \quad \boldsymbol{\Lambda}_{\mathcal{D}} = \mathbf{diag}(b''(\mathbf{x}_{1,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}}), \ldots, b''(\mathbf{x}_{n,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}})).
$$

Hence, we have $(\widetilde{\mathbf{X}}_{\mathcal{D}}^\top \widetilde{\mathbf{X}}_{\mathcal{D}})^{-1} \preceq \frac{1}{\psi_*}(\mathbf{X}_{\mathcal{D}}^\top \mathbf{X}_{\mathcal{D}})^{-1} \Rightarrow \mathbf{X}_{\mathcal{D}}(\widetilde{\mathbf{X}}_{\mathcal{D}}^\top \widetilde{\mathbf{X}}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top \preceq \frac{1}{\psi_*}\mathbf{P}_{\mathcal{D}}$. Similarly, $\mathbf{X}_{\mathcal{S}}(\widetilde{\mathbf{X}}_{\mathcal{S}}^\top \widetilde{\mathbf{X}}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}} \succeq \frac{1}{B}\mathbf{P}_{\mathcal{S}}$. Also, recall that $\widetilde{\mathbf{P}}_{\mathcal{D}} = \widetilde{\mathbf{X}}_{\mathcal{D}}(\widetilde{\mathbf{X}}_{\mathcal{D}}^\top \widetilde{\mathbf{X}}_{\mathcal{D}})^{-1}\widetilde{\mathbf{X}}_{\mathcal{D}}^\top$ for any $\mathcal{D} \in \mathscr{A}_s \cup \{\mathcal{S}\}$. Let $\widetilde{\mathbf{P}}_{I|\mathcal{D}}$ be the orthogonal projector onto the $\mathrm{col}([\widetilde{\mathbf{X}}_{\mathcal{D}}]_I)$. Using these facts, for any $\eta \in [0,1)$, the difference between the two losses can be lower bounded as follows:

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}}) - \mathcal{L}_{\mathcal{S}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}) \\
&\geq \eta \Delta_{\mathrm{kl}}(\mathcal{D}) \\
&\quad + 2^{-1}(1-\eta)\Delta_{\mathrm{kl}}(\mathcal{D}) - \frac{2}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top (\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}) \\
&\quad + 2^{-1}(1-\eta) - \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}(\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{I|\mathcal{D}})\boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \\
&\quad \underbrace{+ \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \boldsymbol{\Lambda}_{\mathcal{S}}^{-1/2}(\widetilde{\mathbf{P}}_{\mathcal{S}} - \widetilde{\mathbf{P}}_{I|\mathcal{S}})\boldsymbol{\Lambda}_{\mathcal{S}}^{-1/2}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))}_{\geq 0} \\
&\quad - \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}\widetilde{\mathbf{P}}_{I|\mathcal{D}}\boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \\
&\quad + \frac{1}{n}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \boldsymbol{\Lambda}_{\mathcal{S}}^{-1/2}\widetilde{\mathbf{P}}_{I|\mathcal{S}}\boldsymbol{\Lambda}_{\mathcal{S}}^{-1/2}(\mathbf{y} - \boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \\
&\quad - \frac{\tilde{B}^2 M^2}{4\kappa_0 \psi_*}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\right\|_2^4 - \frac{\tilde{B}^2 M^2}{4\kappa_0 \psi_*}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\right\|_2^4 - \frac{\tilde{B}M}{6}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{D}}\right\|_2^3 - \frac{\tilde{B}M}{6}\left\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\right\|_2^3.
\end{aligned}
\tag{S3.51}
$$

Now recall that

$$
\widetilde{\mathcal{T}}_I^{(s)} = \left\{ \frac{\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}}{\left\|\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}\right\|_2} : \mathcal{D} \in \mathscr{A}_I \right\}
$$

$$
\widetilde{\mathcal{G}}_I^{(s)} = \left\{ \widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{I|\mathcal{D}} : \mathcal{D} \in \mathscr{A}_I \right\}
$$

Now we will handle the linear and the quadratic terms separately. We assume that $|\mathcal{I}| = s - k$, where $1 \le k \le s$.

Analysis of likelihood lower bound

Analysis of linear term

To analyze the linear term we will use the deviation bound for the supremum of the sub-Gaussian process. In particular, we will use Theorem 5.36 of [12]. First, note that the diameter $\mathsf{diam}(\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)}) \le \sqrt{2}$ and recall $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, where $\epsilon_i = \frac{\{y_i - b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*)\}}{(\phi B)^{1/2}}$. Due to 1-sub-Gaussianity, we have $\max_{i\in[n]} \mathsf{var}(\epsilon_i) \le \sigma_\epsilon^2$ for some universal constant $\sigma_\epsilon > 0$. Also, recall that

$$\widehat{\boldsymbol{r}}_{\mathcal{D}} = \frac{\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^*}{\left\|\mathbf{X}_{\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^*\right\|_2}.$$

Thus, using the aforementioned theorem we get

$$\begin{aligned}&\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \widehat{\boldsymbol{r}}_{\mathcal{D}}^\top \boldsymbol{\epsilon} \ge A_1(\mathscr{E}_{\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)}}\sqrt{k\log(ep)} + \sqrt{2k\{\log(es)\vee\log\log(ep)\}})\right\} \\ &\le 3\{(es)\vee\log(ep)\}^{-2k},\end{aligned} \tag{S3.52}$$

for some universal constant $A_1 > 0$.

Analysis of quadratic terms

Now, we focus on the quadratic terms. Define the random vector $\boldsymbol{\xi}_{\mathcal{D}} := (\xi_{1,\mathcal{D}}, \ldots, \xi_{n,\mathcal{D}})$, where

$$\xi_i = \frac{\{y_i - b'(\mathbf{x}_{i,\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^*)\}}{(\phi B \psi_*^{-1})^{1/2}\sqrt{b''(\mathbf{x}_{i,\mathcal{D}}^\top \bar{\boldsymbol{\beta}}_{\mathcal{D}})}}, \quad i \in [n].$$

Note that $\{\xi_{i,\mathcal{D}}\}_{i\in[n]}$ are independent 1-sub-Gaussian. We will study the random quantity $Q_{\mathscr{A}_{\mathcal{I}}} := \max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \boldsymbol{\xi}_{\mathcal{D}}^\top(\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}})\boldsymbol{\xi}_{\mathcal{D}}$. Let us assume that $\max_{i\in[n]} \mathsf{var}(\xi_{i,\mathcal{D}}) = \sigma_{\mathcal{D}}^2$. First, we note that $\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}}$ is a projection matrix of rank k and hence it is idempotent. Also note that $\mathbb{E}\left\{\boldsymbol{\xi}_{\mathcal{D}}^\top(\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}})\boldsymbol{\xi}_{\mathcal{D}}\right\} = \mathsf{tr}\left\{(\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}})\mathbb{E}(\boldsymbol{\xi}_{\mathcal{D}}\boldsymbol{\xi}_{\mathcal{D}}^\top)\right\} = k\sigma_{\mathcal{D}}^2 \le k\sigma_0^2$, where σ_0^2 is a universal constant. Now, to bound $Q_{\mathscr{A}_s}$ we will use Theorem S4.2. By the properties of projection matrices, we have $d_{\mathrm{op}}(\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}) = 1$ and $d_F(\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}) = \sqrt{k}$. Hence, equipped with Assumption S2.1(e), the quantities M, V and U (defined in Theorem S4.2) has the following properties:

$$M \le 2\mathscr{E}_{\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}}^2 k\log(ep), \quad V \le 2\sqrt{k\log(ep)}, \quad \text{and} \quad U = 1.$$

Due to Theorem S4.2, there exists a universal positive constant A_3, such that for

$$t = A_3 k\sqrt{\log(ep)\{\log(es)\vee\log\log(ep)\}},$$

we get

$$\mathbb{P}\left(C_{\mathscr{A}_{\mathcal{I}}}(\boldsymbol{\xi}_{\mathcal{D}}) \geq A_2 \mathscr{C}^2_{\widetilde{\mathcal{G}}^{(s)}_{\mathcal{I}}} k\log(ep) + A_3 k\sqrt{\log(ep)\{\log(es)\vee\log\log(ep)\}}\right)$$
$$\leq \{(es)\vee\log(ep)\}^{-2k},$$

for a universal positive constant A_2. As we have

$$\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \mathbb{E}\{\boldsymbol{\xi}^\top(\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}})\boldsymbol{\xi}\} \leq k\sigma_0^2 \leq k\sigma_0^2 \mathscr{C}^2_{\widetilde{\mathcal{G}}^{(s)}_{\mathcal{I}}}\log(ep),$$

it follows that

$$\mathbb{P}\left(Q_{\mathscr{A}_s} \leq A_4 \mathscr{C}^2_{\widetilde{\mathcal{G}}^{(s)}_{\mathcal{I}}} k\log(ep) + A_3 k\sqrt{\log(ep)\{\log(es)\vee\log\log(ep)\}}\right)$$
$$\leq \{(es)\vee\log(ep)\}^{-2k}, \tag{S3.53}$$

where A_4 is a universal positive constant.

Next, by construction, we have

$$\begin{aligned} &n^{-1}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}\widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}}\boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \\ &= n^{-1}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \mathbf{X}_{\mathcal{I}}(\widetilde{\mathbf{X}}_{\mathcal{I}}^\top\widetilde{\mathbf{X}}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}^\top(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \\ &\preceq n^{-1}\psi_*^{-1}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \mathbf{P}_{\mathcal{I}}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})). \end{aligned}$$

Similarly,

$$\frac{1}{n}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}\widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}}\boldsymbol{\Lambda}_{\mathcal{D}}^{-1/2}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})) \succeq \frac{1}{nB}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}))^\top \mathbf{P}_{\mathcal{I}}(\mathbf{y}-\boldsymbol{\rho}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})).$$

By Theorem 1 of [9], there exists a universal constant $A_5 > 0$ such that

$$\mathbb{P}\left\{\left|\boldsymbol{\epsilon}^\top\mathbf{P}_{\mathcal{I}}\boldsymbol{\epsilon} - (s-k)\sigma_\epsilon^2\right| \geq t\right\} \leq 2\exp\left[-A_5\min\left\{\frac{t^2}{s-k}, t\right\}\right].$$

For $t = 2A_5^{-1}s\{\log(es)\vee\log\log(ep)\}$ and a universal constant $A_6 > 0$, we get

$$\mathbb{P}\left[\boldsymbol{\epsilon}^\top\mathbf{P}_{\mathcal{I}}\boldsymbol{\epsilon} \geq A_6 s\max\{\log(es), \log\log(ep)\}\right] \leq \{\log(ep)\}^{-2s} \quad \text{for all } \mathcal{D}\in\mathscr{A}_s\cup\{\mathcal{S}\}. \tag{S3.54}$$

Final 0-1 error bound

Define the event

$$\Omega = \left\{\max_{\mathcal{D}\in\mathscr{A}_s\cup\{\mathcal{S}\}} \left\|\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}\right\|_2 \leq \frac{9x_0(\phi B)^{1/2}}{\psi(x_0R + x_0R_0)}\sqrt{\frac{s\log n}{n}}\right\}$$

.

By (S3.48) and (S3.49) we have $\mathbb{P}(\Omega^c) \lesssim (s\log p)/n^7$. Also, let $\mathcal{E}$ be an event inside of which the assertion of the theorem holds. It follows that

$$\begin{aligned} \mathbb{P}(\mathcal{E}^c) &= \mathbb{P}(\mathcal{E}^c\cap\Omega) + \mathbb{P}(\mathcal{E}^c\cap\Omega^c) \\ &\lesssim \left[\sum_{k=1}^{s}\sum_{\mathcal{I}\subset\mathcal{S}:|\mathcal{I}|=s-k} \mathbb{P}\left(\min_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \mathcal{L}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}}) - \mathcal{L}_{\mathcal{S}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}) < \eta\widetilde{\tau}_*(s)\right)\right] + \frac{s\log p}{n^7} \end{aligned}$$

Now, assume the follwing

$$\widetilde{\tau}_{\text{glm}}(s) := \min_{\mathcal{D}\neq\mathcal{S},|\mathcal{D}|=s} \min\left\{\frac{\Delta_{\text{kl}}^2(\mathcal{D})}{\Delta_{\text{par}}(\mathcal{D})\,|\mathcal{D}\setminus\mathcal{S}|}, \frac{\Delta_{\text{kl}}(\mathcal{D})}{|\mathcal{D}\setminus\mathcal{S}|}\right\}$$
$$\gtrsim \frac{(1\vee\psi_*^{-1})}{(1-\eta)^2}\max\left\{\mathsf{Comp}_1, \mathsf{Comp}_2, t^{(1)}_{s,n,p}, t^{(2)}_{n,s,p}\right\}\frac{(\phi B)\log(ep)}{n},$$

where

$$\mathsf{Comp}_1 = \left(\mathscr{E}_{\widetilde{\mathcal{T}}_I^{(s)}} + \sqrt{\frac{\log(es)\vee\log\log(ep)}{\log(ep)}}\right)^2,$$

$$\mathsf{Comp}_2 = \left(\mathscr{E}^2_{\widetilde{\mathcal{G}}_I^{(s)}} + \sqrt{\frac{\log(es)\vee\log\log(ep)}{\log(ep)}}\right),$$

$$t^{(1)}_{s,n,p} := (\psi_*^{-1} - B^{-1})\frac{s\{\log(es)\vee\log\log(ep)\}}{\log(ep)},$$

$$t^{(2)}_{s,n,p} := \left(\frac{\tilde{B}^2M^2x_0^4\phi^2B^2}{\kappa_0\psi_*\psi_{**}^4}\right)\frac{s^2(\log n)^2}{n\log p} + \left(\frac{\tilde{B}Mx_0^3\phi^{3/2}B^{3/2}}{6}\right)\frac{s^{3/2}(\log n)^{3/2}}{\sqrt{n}\log p},$$

where $\psi_{**} = \psi(x_0R + x_0R_0)$. Now, recall the property (S2.43) of $\Delta_{\text{kl}}(\mathcal{D})$. Thus, for the aforementioned condition to hold for $\widetilde{\tau}_{\text{glm}}(s)$, it is sufficient to have

$$\widetilde{\tau}_*(s) := \min_{\mathcal{D}\neq\mathcal{S},|\mathcal{D}|=s}\frac{\Delta_{\text{kl}}(\mathcal{D})}{|\mathcal{S}\setminus\mathcal{D}|}$$
$$\gtrsim \frac{(\phi B)(1\vee\psi_*)^{-1}}{(\psi_{**}\wedge 1)(1-\eta)^2}\max\left\{\mathsf{Comp}_1, \mathsf{Comp}_2, t^{(1)}_{s,n,p}, t^{(2)}_{n,s,p}\right\}\frac{\log(ep)}{n}$$

Under the above inequality and due to (S3.52), (S3.53), and (S3.54), we can finally conclude

$$\mathbb{P}(\mathcal{E}^c) \lesssim \sum_{k=1}^{s}\binom{s}{k}\{(es)\vee\log(ep)\}^{-2k} + \frac{s\log p}{n^7}$$
$$\lesssim \frac{1}{(s\vee\log p)} + \frac{s\log p}{n^7}.$$

S4. Quadratic chaos process

Let $\mathcal{A}$ be a set of $m\times n$ matrices and $\boldsymbol{\xi}$ be a 1-sub-Gaussian random vector. The random variable of interest is

$$C_{\mathcal{A}}(\boldsymbol{\xi}) := \sup_{A\in\mathcal{A}}\left|\|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|A\boldsymbol{\xi}\|_2^2\right|.$$

This quantity is studied by [6] and [3]. In the literature of empirical process, this is known as order-2 sub-Gaussian chaos. Before we present the main result for $C_{\mathcal{A}}(\boldsymbol{\xi})$, we introduce some useful definitions.

Definition 2. *For a metric space (T, d), an admissible sequence of T is a collection of subsets of T, $\{T_r : r \geq 0\}$, such that for every $r \geq 0$, $|T_r| \leq 2^{2^r}$ and $|T_0| = 1$. For $\alpha \geq 1$, define the γ_α functional by*

$$\gamma_\alpha(T, d) := \inf \sup_{t \in T} \sum_{r=0}^{\infty} 2^{r/\alpha} d(t, T_r).$$

The γ_α functional can be bounded in terms of the covering numbers $\mathcal{N}(T, d, \epsilon)$ by the well-known Dudley's integral (see [10]). A more specific formulation for the γ_2 functional of a set of matrices $\mathcal{A}$ endowed with the operator norm, the scenario which we will focus on in this article, is

$$\gamma_2(\mathcal{A}, \|\cdot\|_{\text{op}}) \leq \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{A}, \|\cdot\|_{\text{op}}, \epsilon)}\, d\epsilon.$$

We also define the two quantities $d_{\text{op}}(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_{\text{op}}$ and $d_F(\mathcal{A}) := \sup_{A \in \mathcal{A}} \sqrt{\text{tr}(A^\top A)}$.

Now, we present the main deviation bound for $C_{\mathcal{A}}(\boldsymbol{\xi})$.

Theorem S4.2 (Theorem 1 of [3])**.** *Let $\mathcal{A}$ be a set of $m \times n$ matrices and $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_n)^\top$ be a random vector with independent 1-sub-Gaussian entries. Let*

$$\begin{aligned} M &= \gamma_2(\mathcal{A}, \|\cdot\|)_{\text{op}} \left\{\gamma_2(\mathcal{A}, \|\cdot\|)_{\text{op}} + d_F(\mathcal{A})\right\}, \\ V &= d_{\text{op}}(\mathcal{A}) \left\{\gamma_2(\mathcal{A}, \|\cdot\|)_{\text{op}} + d_F(\mathcal{A})\right\}, \\ U &= d_{\text{op}}(\mathcal{A}). \end{aligned}$$

Then, for $t > 0$,

$$\mathbb{P}\left(C_{\mathcal{A}}(\boldsymbol{\xi}) \geq c_1 M + t\right) \leq 2 \exp\left(-c_2 \min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}\right),$$

where c_1, c_2 are universal positive constants.

S5. Technical lemmas

Lemma S5.1 (Equation (9) in [5])**.** *Let $\Phi(\cdot)$ denote the cumulative distribution function of standard Gaussian distribution. Then for all $x \geq 0$, the following inequalities are true:*

$$\left(\frac{x}{1+x^2}\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \leq 1 - \Phi(x) \leq \left(\frac{1}{x}\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Lemma S5.2. *Let $\mathbf{w} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ and $\boldsymbol{\mu} \in \mathbb{R}^n \setminus \{0\}$ such that $\|\boldsymbol{\mu}\|_2 \leq \sigma\delta$. Then*

$$\mathbb{P}(\|\mathbf{w} + \boldsymbol{\mu}\|_2^2 < \|\mathbf{w}\|_2^2) \geq \frac{\delta}{1+\delta^2} \frac{e^{-\delta^2/2}}{\sqrt{2\pi}}.$$

Proof. By straightforward algebra, it follows that

$$\mathsf{p}_0 := \mathbb{P}(\|\mathbf{w} + \boldsymbol{\mu}\|_2^2 < \|\mathbf{w}\|_2^2) = \mathbb{P}(\boldsymbol{\mu}^\top \mathbf{w} + \|\boldsymbol{\mu}\|_2^2 < 0).$$

Note that $\boldsymbol{\mu}^\top \mathbf{w} \overset{\mathrm{d}}{=} \|\boldsymbol{\mu}\|_2 \, w$, where $w \sim \mathsf{N}(0, \sigma^2)$. Hence, due to Lemma S5.1 we have

$$\begin{aligned} \mathsf{p}_0 &= \mathbb{P}(w < -\|\boldsymbol{\mu}\|_2) \\ &= \mathbb{P}(w > \|\boldsymbol{\mu}\|_2) \\ &\geq \mathbb{P}(w > \sigma\delta) \\ &\geq \frac{\delta}{1+\delta^2} \frac{e^{-\delta^2/2}}{\sqrt{2\pi}}. \end{aligned}$$

□

Lemma S5.3 (Lemma 1 in [7])**.** *Let W be chi-squared random variable with degrees of freedom m. Then, we have the following large-deviation inequalities for all $x > 0$*

$$\mathbb{P}(W - m > 2\sqrt{mx} + 2x) \leq \exp(-x), \quad \textit{and} \tag{S5.55}$$

$$\mathbb{P}(W - m < -2\sqrt{mx}) \leq \exp(-x). \tag{S5.56}$$

If we set $x = mu$ in Equation (S5.55) for $u > 0$, then we get

$$\mathbb{P}\left(\frac{W}{m} - 1 \geq 2\sqrt{u} + 2u\right) \leq \exp(-mu).$$

Note that for $u < 1$, we have $2\sqrt{u} + 2u < 4\sqrt{u}$. Thus, setting $u = \delta^2/16$ for any $\delta < 1$, we get

$$\mathbb{P}\left(\frac{W}{m} - 1 \geq \delta\right) \leq \exp(-m\delta^2/16). \tag{S5.57}$$

Similarly, setting $x = mu$ and $u = \delta^2/4$ in Equation (S5.56), we get

$$\mathbb{P}\left(\frac{W}{m} - 1 \leq -\delta\right) \leq \exp(-m\delta^2/4). \tag{S5.58}$$

Lemma S5.4. *Let $\delta \in (0, \infty)$. Then for any $x > 0$ the following inequality holds:*

$$0 < \sqrt{x+\delta} - \sqrt{(x-\delta) \vee 0} \leq \sqrt{2\delta}.$$

Proof. It is obvious that $f_\delta(x) := \sqrt{x+\delta} - \sqrt{(x-\delta) \vee 0} > 0$. Now for the other inequality, we will consider two cases:

Case 1: $x \leq \delta$ In this case $f_\delta(x) = \sqrt{x+\delta} \leq \sqrt{2\delta}$.

Case 2: $x > \delta$ In this case we have

$$f'_\delta(x) = \frac{1}{2}\left(\frac{1}{\sqrt{x+\delta}} - \frac{1}{\sqrt{x-\delta}}\right) < 0 \quad \text{for all } x > \delta.$$

Hence $f_\delta(x) \leq f_\delta(\delta) = \sqrt{2\delta}$. □

References

[1] ADAMCZAK, R. (2015). A note on the Hanson-Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.* **20** 1–13.

[2] ADLER, R. J., TAYLOR, J. E. et al. (2007). *Random fields and geometry* **80**. Springer.

[3] BANERJEE, A., GU, Q., SIVAKUMAR, V. and WU, S. Z. (2019). Random quadratic forms with dependence: Applications to restricted isometry and beyond. *Advances in Neural Information Processing Systems* **32**.

[4] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[5] GORDON, R. D. (1941). Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statist.* **12** 364–366.

[6] KRAHMER, F., MENDELSON, S. and RAUHUT, H. (2014). Suprema of chaos processes and the restricted isometry property. *Commun. Pure. Appl. Math.* **67** 1877–1904.

[7] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* 1302–1338.

[8] PIJYAN, A., ZHENG, Q., HONG, H. G. and LI, Y. (2020). Consistent Estimation of Generalized Linear Models with High Dimensional Predictors via Stepwise Regression. *Entropy* **22**. https://doi.org/10.3390/e22090965

[9] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.* **18** 1–9. https://doi.org/10.1214/ECP.v18-2865

[10] TALAGRAND, M. (2005). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

[11] VERSHYNIN, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. https://doi.org/10.1017/9781108231596

[12] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: a non-asymptotic viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. https://doi.org/10.1017/9781108627771

[13] ZHENG, Q., HONG, H. G. and LI, Y. (2020). Building generalized linear models with ultrahigh dimensional features: A sequentially conditional approach. *Biometrics* **76** 47–60.