Statistics in the Modern Era: High Dimensions, Decision-Making, and Privacy

by

Saptarshi Roy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2024

Doctoral Committee:

Professor Ambuj Tewari, Chair
Professor Stilian A. Stoev
Professor Martin J. Strauss
Dr. Ziwei Zhu

Saptarshi Roy
roysapta@umich.edu
ORCID iD: 0000-0003-4183-205X

DEDICATION

Dedicated to my parents for their complete and
selfless devotion towards the success of their son.

ACKNOWLEDGEMENTS

I would like to begin by thanking my PhD advisor Ambuj Tewari without whom this thesis wouldn't have been possible. Through the course of these five years, Ambuj has not only been an exceptional advisor but also an inspirational role model for a student like me who is still in the nascent stage of his research career and overall life. There hasn't been a single instance where Ambuj was unable to provide me with a glimmer of hope through one-on-one conversations during the critical stages of my academic or career life. His positive attitude and sincere desire to assist a struggling researcher, like myself, have been the driving factors behind my success during my PhD journey. Apart from that, Ambuj has always given me the freedom to explore new research areas that have helped me to mature as an independent researcher. His enthusiasm and genuine respect for my research interests have truly made this journey a memorable one. I would say, it has been an honor to be his student as I have witnessed a truly meaningful student-advisor relationship getting its perfect shape over the last five years. I would also like to thank Ziwei, who has also played an important role in directing me towards the right direction. Collaboration with him has taught me a lot of lessons that are essential for succeeding as an independent researcher.

Apart from these, a large part of the enjoyment during my PhD life came from my exceptionally good peers, who later became good friends. I am in debt to the whole CASI lab for being part of this whole journey. They have made this hurdle many times easier to conquer. I would like to thank Vinod and Unique for all those afternoon strolls filled with philosophical debates. Although, jokingly we called those "Time wasting sessions", in retrospect, I believe that missing out on those moments would have been a real "Waste of good times". I am grateful to Yash and Pan for encouraging me to join the lifting group. This decision has indeed helped me to become more disciplined and stoic, which I believe is very important to push through difficult times. I am also grateful to the current second year Statistics PhD students who have completely changed the face of the department and made it a more fun place to work. I would also like to thank my roommate Sunrit. Our occasional jamming sessions during the pandemic era provided much-needed peace of mind.

Above all, my family has been the most important pillar of strength, carrying me through the most difficult times. Whatever I have become today, it is only because of their selfless

support and sacrifices. Lastly, I would like to acknowledge my partner Aesha, who kept the fire in me alive during the turmoil times in my life. Aesha, you have done your best, as best as one can do from another part of the world. You have always been the place of my return, you have always been my home.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

High dimensional data analysis has become increasingly frequent and important in diverse fields of sciences, engineering, genomics, and machine learning (ML), and it has quite evidently spawned new complexities concerning modern problems ranging from variable selection, distributed learning, and computational efficiency, data privacy, and online decision-making. To address these modern emerging statistical problems in data science, my research focuses on the intersection of high-dimensional statistics and the above modern ML problems.

The first chapter studies the problem of exact support recovery for high-dimensional sparse linear regression when the signals are weak, rare, and possibly heterogeneous. Specifically, we broaden the theoretical understanding of model selection accuracy of best subset selection (BSS) and marginal screening (MS) under independent Gaussian design. Furthermore, to overcome the computational bottleneck of BSS, we also propose an efficient two-stage algorithm called "Estimate Then Screen" (ETS) which shares exactly the same asymptotic optimality in terms of exact recovery as BSS.

The second chapter follows up on the work of the first chapter by considering correlated features. In this chapter, we also study the model selection accuracy of BSS. We show that apart from the separation margin between the true and noise variables, the complexity of residualized signals and projections of spurious features also play intricate roles in characterizing the model consistency of BSS. In this chapter, we demonstrate the interplay between the margin separation and the two complexities through a simple margin condition which further helps to understand the theoretical properties of BSS.

In the third chapter, we propose a differentially private algorithm for model selection under the high-dimensional sparse regression setup. We adopt the well-known exponential mechanism for designing a sampling scheme that can identify the true set of features under desirable conditions on the signal. In fact, under low privacy regime, we show that the minimum signal strength requirement exactly matches the requirement under the non-private setting. Moreover, to achieve computational expediency over the intractable exponential mechanism, we design a Metropolis-Hastings chain that quickly mixes to the target distribution to generate private estimates of the model.

In the final chapter, we propose a novel Thompson sampling algorithm for the high-

dimensional sparse linear contextual bandit. We specifically use a sparsity-inducing prior for Thompson sampling that exploits the low dimensional structure of the problem, and we theoretically show that our algorithm enjoys desirable regret bound. Furthermore, for computational speed-up, we adopt a variational inference framework and demonstrate the superior performance of our algorithm over its competitors both for simulated and real data.

CHAPTER 1

Introduction

High-dimensional data has become increasingly abundant in contemporary machine learning (ML) problems across diverse fields of science including genomics, engineering, and computer vision. However, the rapid growth of data and modern artificial intelligence (AI) technologies have also presented new challenges related to data privacy, computational scalability, and reliable decision-making. To resolve these important problems through the advancement of responsible AI research and open science (Vicente-Saez and Martinez-Fuentes, 2018), two key aspects of modern data science, my research has primarily focused on the intersection of high-dimensional statistics and emerging ML problems in the areas of computation, model selection, differential privacy, and online decision-making.

Over the last decade, differential privacy, online learning, and algorithmic fairness have been at the forefront of ML research. However, statistically principled investigations of these frameworks for high-dimensional data are still relatively rare. In the era of big data, this is particularly concerning as it questions the reliability of existing ML frameworks that are being used in our everyday lives, thereby hindering the development of open science that advocates the sharing of scientific discoveries to all classes of community.

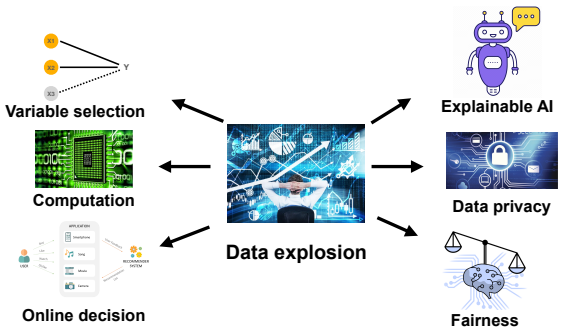


Figure 1.1: Modern problems in high-dimension.

In my dissertation, I have made contributions to particularly address the issues related to responsible ML research and open science by developing novel methodologies along with theoretical justifications. Therefore, my dissertation paves the way to underpin modern AI applications ranging from genetics and healthcare to computer vision and engineering. My thesis touches upon the following important topics:

- Chapter 2: Scalable feature selection with optimal performance.
- Chapter 3: New theoretical insights on best subset selection (BSS).
- Chapter 4: Privacy protection in model selection.
- Chapter 5: Efficient Thompson sampling for high-dimensional bandits.

One recurrent theme in all of the above chapters is understanding the sparsity structure of the underlying model. This is an important component of almost all high-dimensional learning to battle the curse of dimensionality. In particular, it is imperative to avoid the polynomial dependence of the ambient dimension in the learning rate, and there have been different strategies to achieve this goal based on the different methodologies. Now, we will discuss some of the strategies that have been used to induce sparsity in the learning mechanisms.

1.1 Sparsity inducing techniques

1.1.1 Frequentist methods

Over the past few decades, statisticians have developed a plethora of sparsity-inducing techniques including LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and adaptive LASSO (Huang et al., 2008). Essentially all these methods try to achieve a sparse estimate of the parameter of interest by inducing a penalization term in the objective. For example, the well-known LASSO method uses ℓ_1-penalty with the objective function to produce sparse solutions. To be precise, in the case of high-dimensional linear regression, LASSO minimizes the following loss function:

$$L(\boldsymbol{\theta}) := \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 ,$$

where $\mathbf{X}$ and $\mathbf{y}$ are the design matrix and response vector respectively.

However, all these approaches were developed as computational surrogates to the ℓ_0-constrained problems where one solves the following optimization problem:

$$\min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_0 \leq \widehat{s}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 , \tag{1.1}$$

where $\widehat{s}$ is the sparsity budget. This is known as the best subset selection (BSS) problem. However, in general, the BSS problem is NP-hard, which is one of the reasons that statisticians have been reluctant to use BSS. However, recent computational advancements in Bertsimas et al. (2016); Bertsimas and Parys (2020) have rekindled the interest in ℓ_0

constrained problems both from practical and theoretical perspectives. In these works, the authors developed a mixed-integer-optimization (MIO) framework for solving problem (1.1) using commercial MIO solvers like GUROBI, CPLEX, and MOSEK. However, these methods do not scale very well with the ambient dimension. Therefore, significant effort has been given to propose methods that have better performance with the scaling of the ambient dimension. See Chapter 2 for more detailed discussion on this. However, there is another family of first-order methods that also enjoys good performance under high-dimensional regression settings. The well-known iterative hard thresholding (IHT) (Jain et al., 2014; Liu and Foygel Barber, 2020) falls under this category which is essentially a projected gradient descent method. Typically, such methods are way faster than the MIO solvers in high-dimensional regimes.

1.1.2 Bayesian methods

In the Bayesian setting, the most important component of the learning method is to choose a proper prior over the parameter space. If there is prior knowledge about the structure of the parameter, then one can choose an appropriate prior on the parameter space to invoke that structure. In the sparse regression case, one has to consider sparsity-inducing prior. Popular choices are slab-and-spike prior (George and McCulloch, 1993a; Mitchell and Beauchamp, 1988) and horseshoe prior (Carvalho et al., 2010a; Piironen and Vehtari, 2017).

In the slab-and-spike prior approach, estimating the p-dimensional sparse parameter $\boldsymbol{\beta}$ begins with placing independent slab-and-spike prior on each component $\boldsymbol{\beta}_j$ for $j \in [p]$. A particularly appealing spike-and-slab prior has been

$$\pi(\boldsymbol{\beta}_j \mid \gamma_j, \lambda) = (1-\gamma_j)\delta_0 + \gamma_j \psi(\boldsymbol{\beta}_j \mid \lambda), \quad \gamma_j \in [0,1] \tag{1.2}$$

where δ_0 is the Dirac distribution at zero and $\psi(\cdot \mid \lambda)$ is an absolutely continuous distribution with hyperparameter λ. If γ_j's are small, then prior (1.2) will induce more sparsity in the model. Such prior structure can be alternatively viewed as placing prior on the space of models and then placing prior $\psi(\cdot \mid \lambda)$ on the components of the selected model. Therefore, posterior computation in such models leads to numerical bottlenecks as it takes account of an exponentially large number of model candidates.

On the other hand, the horseshoe prior (Carvalho et al., 2010b), which is a continuous shrinkage prior, takes a different approach: instead of placing prior directly on the model, it models directly the posterior inclusion probability $\mathbb{P}(\boldsymbol{\beta}_j \neq 0 \mid D)$, where D is the observed data. To illustrate this point, consider the normal means model: $y_i \mid \theta_i \sim \mathsf{N}(\theta_i, \sigma^2)$. If the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is known to be sparse then one can construct the following prior on

$\boldsymbol{\theta}$:

$$\theta_i \mid (\lambda_i, \tau) \sim \mathsf{N}(0, \lambda_i^2\tau^2),\ \lambda_i^2 \sim \text{Half-Cauchy}(0,1), i = 1, \ldots, n.$$

Here the parameter τ plays the role of the global shrinkage parameter. Under this model, the posterior mean of θ_i turns out to be $\mathbb{E}(\theta_i \mid y_i) = \{1 - \mathbb{E}(\kappa_i \mid y_i)\}y_i$ where $\kappa_i = 1/(1+\lambda_i^2\tau^2)$. Therefore, $\kappa_i \approx 0$ yields virtually no shrinkage on θ_i. On the contrary $\kappa_i \approx 1$ indicates high shrinkage over θ_i.

1.2 Minimum signal strength condition

A very important concept that will be revisited again and again throughout this thesis is the minimum signal strength condition. Consider the linear model $y = \mathbf{x}^\top\boldsymbol{\beta} + \varepsilon$, where ε is random noise. Also, assume that $\boldsymbol{\beta}$ is sparse. Under such a model. it is well known that each non-zero component of $\boldsymbol{\beta}$ needs to be large in absolute value to be identified by a statistical procedure. In other words $\beta_{min} := \min_{j:\boldsymbol{\beta}_j \neq 0} \left|\boldsymbol{\beta}_j\right|$ needs to be large enough so that a suitable method can distinguish them from zero. Otherwise, it is possible that no method can recover the non-zero support of $\boldsymbol{\beta}$. For the past few decades, there has been ample research in this direction (Rad, 2011; Wainwright, 2009a,b; Fletcher et al., 2009; Genovese et al., 2012; Ndaoud and Tsybakov, 2020). There is also a long line of works that shows certain phase transition phenomena on β_{min} under the sparse regression setting. A detailed discussion on this can be found in Ndaoud and Tsybakov (2020) and Chapter 2.

1.3 Thesis organization

Next, I provide brief descriptions of the individual chapters.

Chapter 2: Scalable feature selection with optimal performance. Feature selection is an important component of feature engineering in modern AI applications that aims to find better representations of predictors (i.e., features) to improve the predictive performance of ML models. In addition, feature selection is often used as a dimension reduction scheme in various scientific fields ranging from genetics (Ang et al., 2015; Liu et al., 2018) to signal processing (Lu et al., 2016; Alotaiby et al., 2015), which leads to more efficient and explainable statistical models compared to other complicated neural network models. However, the BSS approach (Miller, 2002), a classical feature selection method, has eluded ML community for a long time mainly due to its computational inefficiency despite some of the recent computational advancements (Bertsimas et al., 2016) and promising empirical findings (Hastie et al., 2020). In our work Roy et al. (2022), *we theoretically justify the*

superior performance of BSS through a novel analytical framework, thereby broadening the current understanding of BSS. Moreover, we propose a new computationally efficient feature selection method that enjoys the same optimal performance as BSS in practice. Therefore, our proposed algorithm can be used as a more reliable and efficient dimension reduction scheme for much larger datasets, contributing to the promotion of responsible AI and open science in contemporary data science practices.

Chapter 3: New theoretical insights on best subset selection. Over the past decades, the BSS approach has received considerable attention due to its wide applicability and intuitive nature. For example, in genetics research, BSS approaches have become the "gold standard" for feature screening processes (Guo et al., 2023; Kong et al., 2023; Matsumoto et al., 2023) to achieve more interpretable and transparent predictions. However, this necessitates the need for awareness about the possible pitfalls of BSS in the ML community from a safe and responsible ML research point of view. In our work Roy et al. (2023), *we are the earliest to theoretically explore and point out the intricate influence of feature correlation on the performance of BSS using a novel topological argument. Through our research, we explicitly identify two fundamental geometric quantities that relate to the correlation structure of the underlying feature space and help us understand the quality of the selected features via BSS.* As a consequence, our theoretical findings can be used in practice to qualitatively assess the outcome of the BSS approach using the correlation structure of the design, even before its deployment. Therefore, our research takes the initiative to raise awareness about the possible drawbacks of BSS and pave the way towards safe ML research.

Chapter 4: Privacy protection in model selection. The rapid growth of AI and the abundance of data collection from edge devices like cell phones, personal computers, and smartwatches, have put the privacy of personal data at risk. Therefore, differential privacy (DP) Dwork (2006), a mathematical framework that guarantees data privacy protection by ensuring similar output irrespective of the presence or absence of an individual in the database, has emerged as one of the leading research areas in the landscape of modern AI (Dwork, 2008; Steil et al., 2019; Wei et al., 2020; Kim et al., 2019). Although, DP algorithms have been used in several ML problems including risk minimization (Jain and Thakurta, 2014; Kasiviswanathan and Jin, 2016; Wang et al., 2017), density estimation (Wasserman and Zhou, 2010), and hypothesis testing (Dwork et al., 2015c), theoretical investigation of model selection under the DP framework remained somewhat scarce for high-dimensional data. This is concerning as model selection methods are heavily used in high-dimensional genetic data containing sensitive information that may compromise patient's

privacy, thereby hindering data sharing and delaying scientific advancements. In our work Roy and Tewari (2023), *we propose a computationally efficient DP algorithm for model selection that enjoys superior utility compared to the existing benchmarks. In particular, we adopt the well-known Metropolis-Hastings algorithm that generates a private estimate of the model (set of influential features) within polynomial time in the problem parameters.* Therefore, our research provides the first private algorithm for model selection that provably achieves high utility along with computational efficiency, allowing efficient sharing of scientific discoveries to a broader community to practice open science.

Chapter 4: Efficient Thompson sampling for high-dimensional bandits. Contextual multi-armed bandit (MAB) has been one of the most active areas of interest in diverse fields of research ranging from clinical trials (Villar et al., 2015), personalized healthcare (Tewari and Murphy, 2017) to the autonomous driving (Ferdowsi et al., 2019). In such

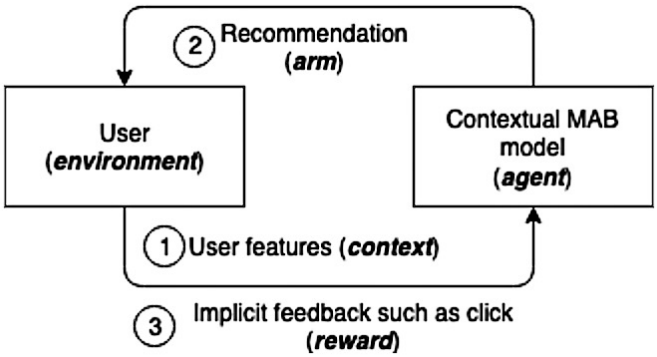


Figure 1.2: A schematic diagram of bandit.

applications, the learner sequentially takes actions based on their past rewards and high-dimensional features (See Figure 1.2) to gradually improve the decision-making process. However, quick and accurate decision-making is of utmost importance in such applications due to safety concerns. Thompson sampling (TS), a widely used heuristics, is a more effective algorithmic framework than greedy MAB algorithms in many such applications (Chapelle and Li, 2011; Kaufmann et al., 2012) under low-dimensional settings. However, a similar TS framework was lacking for high-dimensional bandit problems. In our work (Chakraborty et al., 2023), *we provide the first TS algorithm for high-dimensional sparse linear bandit that uses a novel sparsity inducing prior distribution to achieve an almost dimension-free regret bound guarantee including minimax optimal performance in many practical regimes. In addition, we design an efficient Variational Bayes framework that achieves computational expedi-*

ency under the high-dimensional setting. Our algorithm achieves superior performance than previously existing benchmark algorithms, which we demonstrate by an online cancer identification task on the well-known Gravier's breast cancer data (Gravier et al., 2010). Therefore, our proposed algorithm enjoys quicker and more accurate decision-making, thereby advancing trustworthy AI research.

CHAPTER 2

High-Dimensional Variable Selection with Heterogeneous Signals: A Precise Asymptotic Perspective

In this chapter, we study the problem of exact support recovery for high-dimensional sparse linear regression under independent Gaussian design when the signals are weak, rare, and possibly heterogeneous. Under a suitable scaling of the sample size and signal sparsity, we fix the minimum signal magnitude at the information-theoretic optimal rate and investigate the asymptotic selection accuracy of best subset selection (BSS) and marginal screening (MS) procedures. We show that despite the ideal setup, somewhat surprisingly, marginal screening can fail to achieve exact recovery with probability converging to one in the presence of heterogeneous signals, whereas BSS enjoys model consistency whenever the minimum signal strength is above the information-theoretic threshold. To mitigate the computational intractability of BSS, we also propose an efficient two-stage algorithmic framework called ETS (Estimate Then Screen) comprised of an estimation step and gradient coordinate screening step, and under the same scaling assumption on sample size and sparsity, we show that ETS achieves model consistency under the same information-theoretic optimal requirement on the minimum signal strength as BSS. Finally, we present a simulation study comparing ETS with LASSO and marginal screening. The numerical results agree with our asymptotic theory even for realistic values of the sample size, dimension, and sparsity.

2.1 Introduction

Consider n independent observations $(\mathbf{x}_i, y_i)_{i\in[n]}$ of a random pair $(\mathbf{x}, y)$ drawn from the following linear regression model:

$$\begin{aligned}(\mathbf{x}, w) &\sim \mathcal{P}_x \times \mathcal{P}_w,\\ y &= \mathbf{x}^\top \boldsymbol{\beta} + w,\end{aligned} \tag{2.1}$$

where $\mathcal{P}_x$ is the p-dimensional isotropic Gaussian distribution $\mathsf{N}_p(0, \mathbb{I}_p)$, and $\mathcal{P}_w$ is the standard Gaussian distribution on $\mathbb{R}$. In matrix notation, the observations can be represented as

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

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top, \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top$. We consider the standard *high-dimensional* setup where $n < p$, and possibly $n \ll p$, and the vector $\boldsymbol{\beta}$ is unknown but 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. We denote by $\mathbb{P}_{\boldsymbol{\beta}_0}(\cdot)$ and $\mathbb{E}_{\boldsymbol{\beta}_0}(\cdot)$ the probability measure and the expectation with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ respectively. In this chapter, we focus on the variable selection problem, i.e., identifying the active set $\mathcal{S}_{\boldsymbol{\beta}} := \{j : \beta_j \neq 0\}$. We primarily use the 0-1 loss, i.e., $\mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}} \neq \mathcal{S}_{\boldsymbol{\beta}})$, to assess the quality of a model estimator $\widehat{\mathcal{S}}$.

The isotropic Gaussian design has been widely used to conduct precise analysis of variable selection procedures (Fletcher et al., 2009; Genovese et al., 2012; Ndaoud and Tsybakov, 2020; Su et al., 2017; Kowshik and Polyanskiy, 2021). Specifically, these works either derive the necessary and sufficient condition for exact model recovery (Fletcher et al., 2009; Aeron et al., 2010; Rad, 2011; Jin et al., 2011; Akçakaya and Tarokh, 2009), or establish tight asymptotic bounds of model selection error (Genovese et al., 2012; Ji et al., 2012; Su et al., 2017). The isotropic Gaussian design is also used in compressed sensing to generate a measurement matrix (Candès et al., 2006; Candes and Tao, 2006; Donoho, 2006) so that one can sense the sparse signals with few measurements of the high-dimensional signal. Scarlett and Cevher (2016); Wang et al. (2010) considered the variable selection problem and studied the information-theoretic limit of support recovery under non-Gaussian setup. It is worth emphasizing that these works also assumed that the entries of $\mathbf{X}$ are independent and identically distributed, which is also the setup of this chapter.

weak and rare signals: Recently there has been growing interest in the variable selection problem in the presence of *weak* and *rare* signal regimes (Genovese et al., 2012; Ji et al., 2012) where the active signals are highly sparse with very low magnitude of the or-

der $O(\sqrt{(\log p)/n})$, which is known to be the information-theoretic optimal rate necessary to achieve model consistency. This regime is ubiquitous in modern data analytics such as those in Genome-Wide Association Study (GWAS). There the genes that exhibit detectable association with the trait of interest can be extremely few with weak effects (Consortium et al., 2007; Marttinen et al., 2013). Moreover, the number of subjects n typically ranges in thousands, while the number of features p can range from tens of thousands to hundreds of thousands. Such a high dimension further adds to the difficulty of identifying weak signals. Weak and rare signals also arise in multi-user detection problems (Arias-Castro et al., 2011) where one typically uses linear model of the form (2.1). The signal received from user j is $\beta_j \mathbf{X}_j$. Thus $\beta_j = 0$ means that jth user is not sending any signal. It is a common practice to model the mixing matrix $\mathbf{X}$ as random with i.i.d. entries. Under the presence of strong noise, one might be interested in knowing whether information is being transmitted or not. Typically, in some applications, it is reasonable to assume that a very few numbers of users are sending signals. Also, due to strong noise environment, the signals become quite weak, making them harder to detect. Therefore, from an application point of view, understanding variable selection in *weak* and *rare* signal regimes is crucial. Despite its importance, typically most of the popular methods such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Huang et al., 2008) have been extensively analyzed in terms of 0-1 loss when the signals are uniformly strong (Zhao and Yu, 2006; Guo et al., 2015; Zhang and Huang, 2008; Huang et al., 2008; Zheng et al., 2014) in the sense that

$$a := \min_{j:\beta_j \neq 0} |\beta_j| \gg \left(\frac{\log p}{n}\right)^{1/2}.$$

However, Wainwright (2009b) established a sharp phase transition for LASSO in terms of exact recovery under a general combination of (n, p, s). Under some regularity conditions on the design matrix, the author shows that $a \gtrsim \{(\log p)/n\}^{1/2}$ is necessary and sufficient for the model consistency of LASSO in terms of 0-1 loss. Zhang and Huang (2008) proposed MC+ method based on *minimax concave penalty*, which also achieves model consistency under the optimal rate for a. Therefore, these works can accommodate weak and rare signal regimes, and it is important to emphasize that the analyses presented in them are *non-asymptotic*. Other works on weak and rare signal regimes include Genovese et al. (2012); Ji et al. (2012), and Jin et al. (2014) that establish sharp *asymptotic* requirement on the signal strength.

Signal heterogeneity and marginal screening: Besides the weakness and rarity of signals, heterogeneity in the signal *strength* is another important feature of modern data applications that has not yet received sufficient attention. Roughly, *heterogeneity* in the

signal allows the magnitude of the active β_j's to differ in an arbitrary fashion, whereas *homogeneity* restricts the magnitude of the active signals to be in the same order. One limitation of the existing literature on variable selection in the *weak* and *rare* signals regime is that it typically assumes that the true signals are homogeneous (Genovese et al., 2012; Ji et al., 2012; Jin et al., 2014). Ji et al. (2012) refer to this setup as the *Asymptotically Rare and Weak* (ARW) signal regime. Many popular approaches have been shown to enjoy satisfactory variable selection properties under the ARW regime. For instance, Genovese et al. (2012) showed that both LASSO and marginal screening enjoy model consistency in terms of Hamming loss under independent random design. Ji et al. (2012) and Jin et al. (2014) investigated the same problem under sparsely correlated design. They proposed two-stage screen and clean algorithms that also exhibit model consistency in terms of Hamming loss. However, their theory heavily relies on homogeneous signals and does not extend to the heterogeneous case that is of interest to us. In reality, the ARW setup seldom occurs: the signals almost always have different strengths (Li et al., 2019).

To underscore the contrasting effects of homogeneous and heterogeneous signals in terms of exact model recovery, we study the variable selection property of marginal screening (see Section 2.3). We show that under the presence of strong heterogeneity in the signal, marginal screening fails to recover the exact model with probability converging to 1, whereas under homogeneous signal it can recover the exact model asymptotically (Genovese et al., 2012). It turns out that due to heterogeneity, the spurious correlations become large and create impediment to selecting the exact model. In correlated design, a different problem known as *unfaithfulness* (Wasserman and Roeder, 2009; Robins et al., 2003) prevents marginal screening from achieving model consistency. Specifically, due to "correlation cancellation", the marginal correlation between $\mathbf{y}$ and $\mathbf{X}_j$ becomes negligible even when β_j is large and this ultimately leads to *false negatives*. In this chapter, we study an independent random design model in which correlation cancellation does not occur. Instead, we identify a different source of the problem under the presence of signal heterogeneity that affects the exact variable selection performance of marginal screening. Varying effect of signal heterogeneity in variable selection was also identified in the case of LASSO by Su et al. (2017) and Wang et al. (2022b) for i.i.d. Gaussian design under a special asymptotic setting. In particular, Su et al. (2017) studied the tradeoff between power and type-I error of LASSO and showed that strong heterogeneity in the signal helps to reduce the false discovery in the LASSO path. The same effect was also analyzed in more detail in Wang et al. (2022b). These works use approximate message passing (AMP) theory to obtain the exact asymptotic behavior of the LASSO estimator in terms of variable selection and show that it is unable to achieve model consistency under *linear sparsity* regime.

Computational advancements: On the computational side, modern methods like LASSO, SCAD, MC+ were initially motivated as alternatives to Best Subset Selection (BSS). BSS is in general an NP-hard optimization problem and was believed to be practically intractable even for p as small as 30. Thanks to recent advancements in algorithms and hardware, the optimal solution to the BSS problem can now be computed, sometimes with approximations, for some practical settings. Jain et al. (2014) 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. Liu and Foygel Barber (2020) 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. Bertsimas et al. (2016) viewed the BSS problem through the lens of mixed integer optimization (MIO) and showed that for n in 100s and p in 1000s, the MIO algorithm can obtain a near optimal solution reasonably fast. Bertsimas and Parys (2020) developed a new cutting plane method that solves to provable optimality the Tikhonov-regularized (Tikhonov, 1943) BSS problem for n and p in the 100,000s. Xie and Deng (2020) considered solving the Tikhonov-regularized BSS via mixed integer second-order cone formulation and the largest problem instance they considered has $p \sim 10^3$. Most recently, Hazimeh et al. (2022) developed a Branch-and-Bound method that solves the ℓ_0/ℓ_2-regularized BSS problem for $p \sim 10^7$. A recent work (Zhu et al., 2020) proposed an iterative splicing method called Adaptive Best Subset Selection (ABESS) to solve the BSS problem. They also showed that ABESS enjoys both statistical accuracy and polynomial computational complexity when the design matrix satisfies sparse Reisz condition and minimum signal strength is of order $\Omega\{(s \log p \log\log n/n)^{1/2}\}$.

Given these recent advances in solving BSS, there has been growing acknowledgment that BSS enjoys significant statistical superiority over the aforementioned alternative methods. Bertsimas et al. (2016) and Bertsimas and Parys (2020) numerically demonstrated higher predictive power and lower false discovery rate (FDR) respectively of the BSS solution compared to LASSO. Guo et al. (2020) and Zhu and Wu (2021) reported that the approximate BSS solutions provided by IHT have much fewer false discoveries than LASSO, SCAD, and SIS, especially in the presence of highly correlated design. They also theoretically showed that the model selection behavior of BSS does not explicitly depend on the restricted eigenvalue condition for the design (Bickel et al., 2009; Van De Geer and Bühlmann, 2009), a condition which appears unavoidable (assuming a standard computational complexity conjecture) for any polynomial-time method (Zhang et al., 2014). This suggests that BSS is robust against design collinearity in terms of model selection.

Main contribution: In this chapter, we mainly focus on the *precise* asymptotic bound, i.e., the bound with the optimal constant for the minimum signal strength that allows BSS to achieve model consistency. Under a specific asymptotic setup, we show that BSS achieves asymptotic exact recovery of the true model once the minimum signal strength parameter is above the information-theoretic lower bound, meaning that BSS is optimal in terms of the requirement on the signal strength. In contrast, previous works such as Aeron et al. (2010); Wainwright (2009a); Rad (2011) analyze BSS from a sample complexity point of view: they show that BSS can achieve model consistency under the optimal rate of the sample complexity and different asymptotic regimes. Later Ndaoud and Tsybakov (2020) showed the existence of a polynomial-time method that achieves model consistency under the same sufficient condition on n as BSS for i.i.d. Gaussian design. For general Gaussian design, Wainwright (2009a) showed a similar result for BSS. However, the analyses of all these works are non-asymptotic, and thus cannot yield the optimal constant in the minimum signal strength. Next, we summarize the main contributions of in this chapter below:

1. Under suitable asymptotic scaling of the sample size and signal sparsity, we present a sharp and novel asymptotic analysis of BSS that leverages results from Fan et al. (2018) and achieves the optimal constant in minimum signal strength required for model consistency of BSS (Theorem 2.4.1).

2. To establish the sharpness of Theorem 2.4.1, we first present a novel impossibility result of BSS (Theorem 2.4.3) at the information-theoretic boundary which also relies on asymptotic analysis of BSS. This result together with the impossibility result in Wang et al. (2010), establishes the sharp optimality of the constant obtained in Theorem 2.4.1.

3. To achieve computational expediency, we propose a computationally tractable two-stage algorithmic framework that also enjoys model consistency under the same condition on the minimum signal strength as BSS (Theorem 2.5.2).

The rest of the chapter is organized as follows. Section 2.2 introduces the Asymptotically Ultra-Rare and Weak Minimum signal (AURWM) regime that accommodates heterogeneous signal strengths. In Section 2.3, we point out a potential adverse effect of signal heterogeneity on model selection under AURWM regime. For demonstration, we present an illustrative example of marginal screening to substantiate our claim. In Section 2.4, we derive an impossibility result for BSS under the AURWM regime and show that BSS is optimal in terms of the requirement on the minimum signal strength. In Section 2.5, we propose the aforementioned two-stage algorithmic framework and establish its optimality property in terms

of the constant in minimum signal strength. Finally, in Section 2.6, we carry out simulation studies and numerically demonstrate the superiority of our method over other competing methods.

2.2 Ultra rare and weak minimum signal regime

In this section, we focus on a specific asymptotic setup that allows *heterogeneity* among the sparse signals in high dimension. Throughout our paper, we consider the following signal class:

$$\mathcal{M}_s^a := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 = s, \min_{j:\beta_j \neq 0} |\beta_j| \geq a\}.$$

Here a denotes the minimum signal strength of $\boldsymbol{\beta}$. Note that the signal class $\mathcal{M}_s^a$ only imposes a lower bound for the minimum signal strength and thus allows arbitrarily large magnitudes across the true signals. This implicitly accommodates heterogeneity in the signal, which is in sharp contrast with the homogeneous signal setup considered by Genovese et al. (2012).

Now we are in a position to introduce the *Asymptotically Ultra Rare and Weak Minimum signal* regime (AURWM), in which we mainly consider the signal class above with

$$a = \left(\frac{2r \log p}{n}\right)^{1/2} \quad \text{and} \quad s = O(\log p), \tag{2.2}$$

where the parameter r controls the magnitude of the minimum signal strength. As we will see in Section 2.4, the model consistency of BSS will depend on the value of r. Besides, we set the sample size n as

$$n = \lfloor p^k \rfloor, \quad 0 < k < 1.$$

The condition $s \lesssim \log p$ characterizes the ultra-rarity of the signals, which is common in genetic studies such as GWAS (Yang et al., 2020). Unless stated otherwise, from now on our statistical analysis follows the scalings of n, p, s, a in this AURWM regime. We say a support estimator $\widehat{\mathcal{S}}$ achieves *asymptotic consistent recovery* in the AURWM regime if

$$\lim_{p\to\infty} \sup_{\beta \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}} \neq \mathcal{S}_{\boldsymbol{\beta}}) = 0. \tag{2.3}$$

This paper mainly focuses on the criterion (2.3) to measure the quality of exact recovery performance for an estimator $\widehat{\mathcal{S}}$.

It is also worth mentioning that a relevant but different asymptotic setup is studied by Genovese et al. (2012) and Ji et al. (2012). There the authors assumed a Bayesian model such that all the signals are independent and identically distributed and that the sparsity

$s \sim p^{1-\vartheta}$ for some $\vartheta \in (0,1)$. Under such a setup they obtained asymptotically tight phase transition boundaries with respect to Bayesian Hamming risk, which partitions the r-ϑ plane into three regions: (a) Region of exact recovery, (b) Region of almost recovery, (c) Region of no recovery. We skip the details of these results for brevity. The major differences between their setup and ours are twofold: (1) They essentially assume homogeneous signals; (2) They assume s to grow in a polynomial fashion with respect to p.

2.3 Marginal screening under heterogeneous signal

Marginal screening (MS) is one of the most widely used variable selection methods in practice. It selects the variables with the top absolute marginal correlation with the response. Formally, for any $j \in [p]$, write $\mu_j := \mathbf{X}_j^\top \mathbf{y}/n$. Given any possibly data-driven threshold $\tau(\mathbf{X}, \mathbf{y})$, define the marginal screening estimator as follows:

$$\widehat{\mathcal{S}}_\tau := \{j \in [p] : |\mu_j| \geq \tau(\mathbf{X}, \mathbf{y})\}. \tag{2.4}$$

Note that μ_j is essentially equivalent to the marginal correlation between $\mathbf{X}_j$ and $\mathbf{y}$ because of the isotropic nature of $\mathbf{X}$. Marginal screening has been applied in various fields for feature selection and dimension reduction, including biomedicine (Huang et al., 2019; Lu, 2005; Leisenring et al., 1997), survival data analysis (Hong et al., 2018; Li et al., 2016), economics and econometrics (Wang et al., 2022a; Huang et al., 2014).

Besides the broad applications, marginal screening has been shown to enjoy some desirable statistical properties. Fan and Lv (2008) established the sure screening property of marginal screening under an ultra-high dimensional setup, which serves as a theoretical justification for MS to be used for dimension reduction in many applications. Later, Genovese et al. (2012) showed that MS enjoys the minimax optimal rate under Hamming loss with homogeneous signals. Nevertheless, as mentioned in Section 2.1, precise asymptotic characterization of the 0-1 loss of MS remains fairly underexplored under high dimension, especially in the presence of heterogeneity in signal strength.

2.3.1 Failure of MS in the AURWM regime

In this section, we study the 0-1 risk of the MS estimator. Define $\mathcal{T} := \{\widehat{\mathcal{S}}_\tau \mid \tau : \mathbb{R}^{n\times p} \times \mathbb{R}^n \to \mathbb{R}_+\}$, which is the class of all possible marginal screening estimators. Under the AURWM regime, we show that MS fails to achieve exact model recovery in the minimax sense.

Theorem 2.3.1. *Under the AURWM regime with $n = \lfloor p^k \rfloor$ for some $k \in (0,1)$, none of the MS estimators of the form (2.4) can achieve asymptotic exact recovery, i.e.,*

$$\lim_{p\to\infty} \inf_{\widehat{\mathcal{S}}_\tau \in \mathcal{T}} \sup_{\boldsymbol{\beta} \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}}_\tau \neq \mathcal{S}_{\boldsymbol{\beta}}) = 1.$$

To understand the main message of this theorem, it is instructive to compare it with the parallel result in Genovese et al. (2012) with homogeneous signal. Specifically, Genovese et al. (2012) considers a Bayesian setup where all the signal coefficients are independent and identically distributed Bernoulli random variables (up to a universal constant). Under the AURWM regime, $s = O(\log p)$, which implies that $\vartheta = 1$ in Theorem 10 of Genovese et al. (2012). Then Theorem 10 in Genovese et al. (2012) says that when $r > 1$, MS enjoys consistency in terms of Hamming risk and thus 0-1 risk too. In contrast, when we broaden the signal class to $\mathcal{M}_s^a$ that embraces possibly heterogeneous signals, the same model consistency fails to hold anymore for MS as shown in Theorem 2.3.1. This comparison clearly reveals the curse of signal heterogeneity on MS. However, the above impossibility result does not contradict Theorem 2 in Fletcher et al. (2009). The result therein states that asymptotically $r > (1 + \|\beta\|_2^2)$ is sufficient for the model consistency of MS. However, in the AURWM regime, r can be smaller than $1 + \|\beta\|_2^2$, which would violate the previous condition. In fact, the proof of the above theorem essentially relies on constructing a sequence of signal patterns that violates the condition $r > (1 + \|\beta\|_2^2)$ asymptotically. Thus, in a way, the proof techniques of Theorem 2.3.1 show that $r > (1 + \|\beta\|_2^2)$ is also necessary for MS to achieve model consistency in the AURWM regime. Hence, this establishes the sharpness of Theorem 2 of Fletcher et al. (2009), at least in AURWM regime.

To see how signal heterogeneity hurts MS, for any $j \in [p]$, write μ_j as

$$\mu_j = (\beta_j/n)\,\|\mathbf{X}_j\|_2^2 + \mathbf{X}_j^\top \Big(\sum_{\ell \neq j} \mathbf{X}_\ell \beta_\ell + \mathbf{w}\Big)/n =: \mu_j^{(1)} + \mu_j^{(2)}. \tag{2.5}$$

Here $\mu_j^{(1)} = n^{-1}\beta_j \|\mathbf{X}_j\|_2^2$ represents the marginal contribution from β_j to μ_j, and $\mu_j^{(2)}$ represents the random error of μ_j due to the cross covariance between $\mathbf{X}_j$ and the other signals and noise. Suppose there are spiky signals among $\{\beta_\ell\}_{\ell \neq j}$. Though $\mathbb{E}(\mu_j^{(2)}) = 0$ regardless of the magnitude of β_j, the spiky signals may incur large variance of $\mu_j^{(2)}$ and overwhelm the magnitude of $\mu_j^{(1)}$, which is the essential indicator of the significance of β_j. Consequently, for weak signals, one cannot tell if β_j is a true variable based on only μ_j in the presence of

spiky signals. For further understanding of this phenomenon, it is instructive to note that

$$\mu_j \mid \mathbf{X}_j \sim \mathsf{N}\left(\frac{1}{n}\left\|\mathbf{X}_j\right\|_2^2 \beta_j, \frac{1}{n^2}\left\|\mathbf{X}_j\right\|_2^2\left(1+\sum_{\ell \neq j} \beta_\ell^2\right)\right).$$

If the non-zero components of $\boldsymbol{\beta}$ are arbitrarily large, then the variance of the above Gaussian distribution becomes extremely large. Therefore, even if $\beta_j = 0$, there is a high probability that the correlation μ_j will take larger values. Hence, there is a chance that the true variables associated with weak signals would lose to a noise variable, thereby ultimately leading to false discovery. To rigorously show these claims, we construct a specific example as mentioned before and we study the asymptotic limits of $\max_{j \notin \mathcal{S}_{\boldsymbol{\beta}}} \mu_j^{(2)}$ and μ_{j_0}, where j_0 denotes the index of a weak signal. While the asymptotic analysis of μ_{j_0} is rather straightforward, we borrow some non-trivial results from Fan et al. (2018) to obtain the asymptotic properties of $\max_{j \notin \mathcal{S}_{\boldsymbol{\beta}}} \mu_j^{(2)}$. Details of the proof can be found in Section A of the supplementary material.

In contrast, the AMP line of works on LASSO in Su et al. (2017) and Wang et al. (2022b) show that under a certain asymptotic regime, signal heterogeneity helps LASSO in terms of variable selection. Specifically, under i.i.d. Gaussian design and *linear sparsity* regime (i.e. $s/p \to \alpha$ for some constant $\alpha \in (0,1)$), Wang et al. (2022b) show that higher signal heterogeneity delays the inclusion of false variables in the LASSO solution path whereas, under signal homogeneity, false discovery occurs in a much earlier stage in the solution path. The effect is somewhat opposite to what we discussed for MS. The reason perhaps lies in the fact that LASSO tries to select the features that are highly correlated with the *shrinkage noise* (see Su et al. (2017)), whereas, MS tries to select the features that have a higher correlation with the response. In the case of LASSO, higher signal heterogeneity makes the magnitudes of the correlations between features and shrinkage noise more distinguishable compared to a homogeneous signal pattern, thus preventing early false discovery in the first case. In the case of MS, higher signal heterogeneity increases the variance of $\mu_j^{(2)}$, which essentially dwarfs the influence of weak signals and leads to false discovery. However, these two phenomena are not directly comparable as the asymptotic settings are different for the two cases. In fact, when $s = O(\log p)$, the effect of shrinkage noise is much smaller (see Section 3.2 in Su et al. (2017)) and such phenomenon does not occur for LASSO.

2.4 Best subset selection

Now we shift our focus to BSS, one of the most classical variable selection approaches. With the oracle knowledge of true sparsity s, BSS solves for

$$\widehat{\boldsymbol{\beta}}_{\text{best}} \in \arg\min_{\mathbf{b}\in\mathbb{R}^p, \|\mathbf{b}\|_0=s} n^{-1} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 .$$

Define $\mathbf{P}_{\mathcal{D}} := \mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^\top \mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top$, which is the orthogonal projection operator onto the column space of $\mathbf{X}_{\mathcal{D}}$. The BSS above can be alternatively viewed as solving for

$$\widehat{\mathcal{S}}_{\text{best}} := \arg\min_{\mathcal{D}\subseteq[p]:|\mathcal{D}|=s} n^{-1}\mathbf{y}^\top(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{y} = \arg\max_{\mathcal{D}\subseteq[p]:|\mathcal{D}|=s} n^{-1}\mathbf{y}^\top\mathbf{P}_{\mathcal{D}}\mathbf{y}. \tag{2.6}$$

Using a union bound as in Wainwright (2009a) or Guo et al. (2020), one can show that there exists a universal positive constant φ (approximately equal to 0.618) such that whenever $r > 4/(1-\varphi)$, BSS achieves model consistency, i.e.,

$$\lim_{p\to\infty} \sup_{\beta\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}}_{\text{best}} \neq \mathcal{S}_{\boldsymbol{\beta}}) = 0.$$

We emphasize that the requirement on r here is more stringent than needed: we will show that BSS achieves model consistency whenever $r > 1$. Moreover, a direct application of Theorem 1 of Wang et al. (2010) tells us that $r \geq 1$ is necessary for model consistency in AURWM regime. Therefore, our result (Theorem 2.4.1) essentially resolves the gap between the existing necessary ($r \geq 1$) and sufficient ($r > 4/(1-\varphi)$) conditions on the parameter r by showing that $r > 1$ is sufficient.

2.4.1 Exact support recovery of BSS

In the following theorem, we show that $r > 1$ is sufficient for BSS to achieve asymptotic exact recovery. Recall that $n = \lfloor p^k \rfloor$ with $0 < k < 1$.

Theorem 2.4.1. *Let $r > 1$ and write $\delta = r - 1$. Then there exists a universal positive constant C_0 such that whenever*

$$s < C_0 \min\left\{2k, \frac{\delta^2}{\{(1+0.75\delta)^{1/2} + (1+0.5\delta)^{1/2}\}^2}\right\} \log p,$$

we have

$$\lim_{p\to\infty} \sup_{\beta\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}}_{\text{best}} \neq \mathcal{S}_{\boldsymbol{\beta}}) = 0.$$

In order for BSS to achieve model consistency, we need to ensure that the maximum spurious correlation, i.e., the correlation between the spurious variables and the response, is well controlled so that the best subset does not involve any false discovery. One important ingredient of our analysis is the asymptotic distribution of the maximum spurious correlation due to Fan et al. (2018), based on which we can derive the sharp constant in the minimum

signal strength for BSS to be model-consistent. It is worth emphasizing that pursuing the exact asymptotic distributions is crucial to obtain constant-sharp results; typically, standard non-asymptotic analysis can only yield optimal rates rather than optimal constants. Detailed proof can be found in Section B.1 of the supplementary material.

Note also that Theorem 2.4.1 requires s to grow slowly. Given that we have at least $\binom{p-s}{s}$ spurious models and that this number increases with respect to s when s is small, a larger s implies a higher maximum spurious correlation due to randomness and thus a thinner chance for the best subset to remain the true model.

Remark 2.4.2. *The result of Theorem 2.4.1 can be extended to the sub-Gaussian case. In particular, if the coordinates of* $\mathbf{x}$ *follow i.i.d. distribution with mean-zero and unite variance, and* w *is also distributed as a mean-zero sub-Gaussian distribution with unit variance and independently from* $\mathbf{x}$*, then BSS is model consistent under a similar condition on sparsity* s*. Details can be found in Section A.1.2 of the supplementary material.*

In the next section, we show that $r > 1$ is the weakest possible requirement on the minimum signal strength for BSS to achieve asymptotic model consistency.

2.4.2 Necessary condition for model consistency of BSS

In this section, we establish that under the AURWM regime, it is impossible for BSS to achieve model consistency if $r \leq 1$, i.e., $r > 1$ is necessary for BSS to exactly recover the true support of β. To begin with, Theorem 1 of Fletcher et al. (2009) in our setting yields that whenever $r < 1$, i.e., $r < 1 - \delta_0$ for some $\delta_0 \in (0, 1)$, the 0-1 loss of BSS approaches to 1 as p grows to infinity. This immediately shows that if $r < 1$, BSS is not model consistent. However, to the best of our knowledge, there is no existing research that has explored the compatibility of the condition $r = 1$ with the model consistency of BSS. In the following theorem, we answer this question negatively.

Theorem 2.4.3. *Under AURWM regime* (2.2) *with* $r = 1$ *and* $n = \lfloor p^k \rfloor$ *for some* $k \in (0, 1)$*, BSS is unable to achieve model consistency, i.e.,*

$$\lim_{p\to\infty} \sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}(\widehat{\mathcal{S}}_{\text{best}} \neq \mathcal{S}_{\boldsymbol{\beta}}) > \frac{1}{10}.$$

The above theorem together with Theorem 1 of Fletcher et al. (2009) shows that $r > 1$ is necessary for model consistency of BSS which is an improvement over the prior existing necessary condition $r \geq 1$. Therefore, These results along with Theorem 2.4.1, provide a *complete characterization* of model consistency of BSS in terms of the magnitude of r and resolve the existing theoretical gap at $r = 1$.

It is worth mentioning that a more general information-theoretic impossibility result is true for the regime $r < 1$. In other words, if $r < 1$, all methods (including BSS) fail to achieve model consistency. This claim is a direct consequence of Theorem 1 of Wang et al. (2010) that states that the minimax 0-1 loss is bounded away from 0 as p diverges to infinity, i.e.,

$$\lim_{p\to\infty} \inf_{\widehat{\mathcal{S}}} \sup_{\beta\in\mathcal{M}_s^a} \mathbb{P}_\beta(\widehat{\mathcal{S}} \neq \mathcal{S}_{\boldsymbol{\beta}}) \geq c,$$

where $c > 0$ is a universal constant and the infimum is taken over the class of all possible measurable functions $\widehat{\mathcal{S}} : (X, Y) \to \{\mathcal{D} \subseteq [p] : |\mathcal{D}| = s\}$. The above inequality suggests that $r \geq 1$ is a necessary condition for exact support recovery. Combining this with Theorem 2.4.1 and Theorem 2.4.3, we can see that BSS is *almost* optimal in terms of the requirement on the constant r in minimum signal strength to achieve model consistency. More discussion on this can be found in Section A.1.3 and Section A.1.4 of the supplementary material.

Next, we briefly discuss the intuition behind the Theorem 2.4.3. In the regime $r = 1$, the main difficulty for BSS arises from the fact that it gets confused between $\mathcal{S}_{\boldsymbol{\beta}}$ and its closest competitors. To be precise, let j_0 denote the index of a weak signal, i.e., $\beta_{j_0} = \{(2r \log p)/n\}^{1/2}$ with $r = 1$. Due to the weak magnitude of β_{j_0}, it becomes indistinguishable from 0 and as a result, BSS confuses $\mathcal{S}_{\boldsymbol{\beta}}$ with other candidate models $\{\mathcal{D} \subset [p] : \mathcal{S}_{\boldsymbol{\beta}} \setminus \mathcal{D} = \{j_0\}, |\mathcal{D}| = s\}$ of size s that differ only at j_0 with non-negligible probability. To prove Theorem 2.4.3, we also analyze the asymptotic distribution of an appropriate maximum spurious correlation statistics using results from Fan et al. (2018). We point the readers to Section A.1.3 of the supplementary material for further details of the proof.

Comparison with previous literature: As pointed out before, there is a sharp contrast between the above results and the results in the previous works like Wainwright (2009a); Rad (2011); Aeron et al. (2010), where the authors study the necessary and sufficient conditions for model consistency of BSS in terms of sample complexity under different asymptotic regimes. For example, under *strong-noise* regime, Aeron et al. (2010) showed that the necessary and sufficient conditions for model consistency in terms of 0-1 loss are given by $n = \Omega(s \log(p/s))$ and $a^2 = \Omega(\log(p - s))$, and BSS is optimal in the sense that it achieves exact recovery under these conditions. For the *fixed noise-variance* regime, the results are different. Firstly, Wang et al. (2010) showed that the following condition is necessary for any method to achieve exact recovery:

$$n = \Omega\left(\frac{s\log(p/s)}{\log(1+sa^2)} \vee \frac{\log(p-s)}{\log(1+a^2)}\right), \tag{2.7}$$

where $u \vee v := \max\{u, v\}$. Under the restriction that $a = O(1)$ and $a = \Omega(1/\sqrt{s})$, which represents *strong-signal* regime, Rad (2011) showed that BSS achieves model consistency under the necessary condition (2.7). In the general case, that is with no assumption on the joint behavior of (n, p, s, a), Wainwright (2009a) established that $n = \Omega(\max\{s \log(p/s), a^{-2} \log(p - s)\})$ is a sufficient condition for model consistency of BSS. One can check that the previous condition matches with the condition in (2.7) under the weak signal regime $a = O(1/\sqrt{s})$. This indicates that BSS is also optimal in this regime in terms of sample complexity. It is interesting to note that the AURWM regime (2.2) also falls under this regime as $a = O(\sqrt{(\log p)/n}) \ll 1/\sqrt{s}$, and $n \asymp p^k$. However, all of these results fail to capture the precise dependence on a in terms of sharp requirement on the constant r.

2.5 Achieving information-theoretic optimality with computational efficiency

Despite the optimality of BSS in terms of model selection, its NP-hardness seriously restricts its practical applicability. To address the computational issue, we propose a two-stage algorithm framework called ETS (Estimate then Screen) that combines an estimation step with a follow-up coordinate screening step. Under this framework, one has the flexibility to use any sensible algorithm in the first stage that outputs an estimate with a good estimation guarantee for $\boldsymbol{\beta}$. For example, one choice could be the well-known *iterative hard thresholding* (IHT) algorithm (Blumensath and Davies, 2009) which is a computational surrogate for BSS and enjoys a desirable estimation guarantee (Jain et al., 2014). Other choices may include algorithms like *pathwise calibrated sparse shooting algorithm* (PICASSO) or *prox-gradient homotopy* (PGH) method that are known to produce good approximate solutions for LASSO (see Zhao et al. (2018); Xiao and Zhang (2013)) in high-dimensional setup. We show that in the AURWM regime, ETS enjoys the same selection optimality as BSS in terms of the requirement on the minimum signal strength, i.e., ETS asymptotically achieves model consistency whenever r is greater than the information-theoretic threshold 1, which is also the the optimal requirement for BSS to achieve exact recovery.

The above framework is similar to the methodology introduced in Ndaoud and Tsybakov (2020). In that paper, the authors used the square-root SLOPE estimator (Bogdan et al., 2015) for the estimation step, and under i.i.d. Gaussian design they showed that their algorithm achieves model consistency under the same *sample complexity* as BSS. However, they do not study optimal dependence on r, which is the main focus of this chapter.

2.5.1 The ETS algorithm

In this section, we introduce our ETS algorithm (Algorithm 1) in detail. Given a partition parameter $0 < \gamma < 1$, ETS first splits the full sample $(\mathbf{x}_i, y_i)_{i\in[n]}$ into two subsamples D_1, D_2 of respective sizes $n_1 = \lfloor \gamma n \rfloor$ and $n_2 = n - n_1$. Then ETS performs two main steps on these two sub-samples respectively:

1. Given an objective function $f_{n_1}(\cdot; D_1)$ and a constraint set $\mathcal{C} \subseteq \mathbb{R}^p$, in the estimation step, ETS procures a close approximation to $\boldsymbol{\beta}$ by solving for an approximate solution to the optimization problem

$$\text{minimize}_{\theta\in\mathcal{C}} f_{n_1}(\theta; D_1) \tag{2.8}$$

   via a suitable iterative algorithm $\mathcal{A}(\cdot,\cdot)$ that takes the objective function $f_{n_1}(\cdot; D_1)$ and a set of tuning parameters $\mathcal{T}_{\mathcal{A}}$ as inputs. In particular, in this step, ETS outputs an estimator $\widehat{\boldsymbol{\beta}} := \mathcal{A}(f_{n_1}(\cdot; D_1), \mathcal{T}_{\mathcal{A}})$ of the true signal vector $\boldsymbol{\beta}$.

2. In the second step, ETS performs a coordinatewise screening based on D_2 and $\widehat{\boldsymbol{\beta}}$ to select the true variables.

Algorithm 1: ETS

Input: Data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, objective function $f_{n_1}(\cdot; D_1)$, partition parameter γ , threshold parameter ς;

1. Randomly partition the whole dataset D into two disjoint subsets $D_1 = (\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ and $D_2 = (\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$;

2.Apply the algorithm $\mathcal{A}$ to compute an approximate solution $\widehat{\boldsymbol{\beta}}$ of the optimization problem (2.8) ;

3. Construct the statistics $\{\Delta_i\}_{i=1}^p$ and thresholds $\{\kappa_\varsigma(\mathbf{X}_i^{(2)})\}_{i=1}^p$ using (2.9)-(2.10);

4. Finally compute the selector $\hat{\eta}(\mathbf{X}, \mathbf{y})$;

Output: The selector $\hat{\eta}(\mathbf{X}, \mathbf{y})$.

To elaborate more on the method, for $\ell \in \{1, 2\}$, let $\mathbf{X}^{(\ell)} \in \mathbb{R}^{n_\ell \times p}$ and $\mathbf{y}^{(\ell)} \in \mathbb{R}^{n_\ell}$ denote the design matrix and the response vector of the ℓth sub-sample respectively. ETS computes $\hat{\beta}$ based on the first sub-sample $D_1 := (\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ by finding an approximate solution the optimization problem (2.8) via algorithm $\mathcal{A}$. In practice, there could be several choices for both the objective function $f_{n_1}(\cdot; D_1)$ and the algorithm $\mathcal{A}$. For example, one of the most common choices is to consider the ℓ_0-constrained squared-error loss, i.e., $f_{n_1}(\boldsymbol{\theta}; D_1) = n_1^{-1}\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\theta}\|_2^2$ with $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_0 \le s\}$. In this case, a natural choice for $\mathcal{A}$ is the IHT algorithm which is basically a projected gradient descent method. Another popular

choice for the objective function is the well-known ℓ_1-regularized LASSO objective function $f_{n_1}(\boldsymbol{\theta}; \lambda, D_1) = n_1^{-1}\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$ with $\mathcal{C} = \mathbb{R}^p$ and one can choose either PICASSO, PGH or the composite gradient method proposed in Agarwal et al. (2012) as the algorithm $\mathcal{A}$. Besides these, another choice could be to solve the square-root LASSO problem (Bogdan et al., 2015) via proximal-gradient descent algorithm proposed in Li et al. (2020).

Next comes the screening step of ETS. For each $i \in [p]$, define

$$\Delta_i := \frac{\mathbf{X}_i^{(2)\top}\left(\mathbf{y}^{(2)} - \sum_{j\neq i}\mathbf{X}_j^{(2)}\hat{\beta}_j\right)}{\|\mathbf{X}_i^{(2)}\|_2} \tag{2.9}$$

and

$$\kappa_\varsigma(u) := \frac{a\|u\|_2}{2} + \frac{\varsigma^2 \log p}{a\|u\|_2}, \quad \forall u \in \mathbb{R}^{n_2}, \tag{2.10}$$

where $\varsigma > 0$ is specified later. ETS selects the ith variable if and only if $|\Delta_i| > \kappa_\varsigma(\mathbf{X}_i^{(2)})$. To see why we can screen variables based on $\{\Delta_i\}_{i\in[p]}$, note that

$$\Delta_i = \beta_i\|\mathbf{X}_i^{(2)}\|_2 + \frac{\mathbf{X}_i^{(2)\top}\left(\sum_{j\neq i}\mathbf{X}_j^{(2)}(\beta_j - \hat{\beta}_j) + \mathbf{w}\right)}{\|\mathbf{X}_i^{(2)}\|_2}. \tag{2.11}$$

A straightforward argument shows that conditioned on D_1 and $\mathbf{X}_i^{(2)}$, Δ_i is distributed as:

$$\Delta_i \,\big|\, \left(D_1, \mathbf{X}_i^{(2)}\right) \stackrel{\mathrm{d}}{=} \beta_i\left\|\mathbf{X}_i^{(2)}\right\|_2 + \left\{1 + \sum_{j\neq i}(\beta_j - \hat{\beta}_j)^2\right\}^{1/2} g_i,$$

where $g_i \sim \mathsf{N}(0,1)$ and is independent of $\mathbf{X}_i^{(2)}$. If estimation method performs well in the sense that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ is small, then for all $i \in \mathcal{S}_{\boldsymbol{\beta}}$, $\beta_i\|\mathbf{X}_i^{(2)}\|_2$ becomes the dominant term in Δ_i. In contrast, for all $i \notin \mathcal{S}_{\boldsymbol{\beta}}$, $\beta_i\|\mathbf{X}_i^{(2)}\|_2 = 0$ and we thus expect Δ_i to be small. This suggests the existence of a threshold $t(\cdot)$ on $(\Delta_i)_{i\in[p]}$ that distinguishes the true support $\mathcal{S}_{\boldsymbol{\beta}}$ from the irrelevant variables. We follow Ndaoud and Tsybakov (2020) to choose the threshold function in (2.10), which is shown to be a reasonable choice to identify the true variables.

For each $i \in [p]$, define $\hat{\eta}_i(\mathbf{X}, \mathbf{y}) := \mathbb{1}\{|\Delta_i| > \kappa_\varsigma(\mathbf{X}_i^{(2)})\}$ and write

$$\hat{\eta}(\mathbf{X}, \mathbf{y}) := (\hat{\eta}_1(\mathbf{X}, \mathbf{y}), \ldots, \hat{\eta}_p(\mathbf{X}, \mathbf{y}))^\top.$$

The selector $\hat{\eta}(\mathbf{X}, \mathbf{y})$ is the final estimate of the support $\mathcal{S}_{\boldsymbol{\beta}}$ produced by the ETS algorithm. Algorithm 1 shows the detailed steps of the ETS algorithm.

2.5.1.1 Model consistency of ETS

In this section, we establish theoretical guarantees for ETS-IHT. First, we introduce a technical assumption that concerns how fast algorithm $\mathcal{A}$ can generate a good approximation of the true signal β.

Assumption 2.5.1. *The following holds with a probability converging to 1 as p diverges to infinity:*

For any given tolerance level $\epsilon > 0$, there exists a suitable set of deterministic tuning parameters $\mathcal{T}_{\mathcal{A}}$ such that the algorithm $\mathcal{A}$ requires no more than $T(\epsilon, p, \boldsymbol{\beta})$ iterations to produce a solution $\widehat{\boldsymbol{\beta}} := \mathcal{A}(f_{n_1}(\cdot; D_1), \mathcal{T}_{\mathcal{A}})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq \epsilon$.

The above assumption essentially tells that the $T(\epsilon, p, \boldsymbol{\beta})$th iterate of algorithm $\mathcal{A}$ is already ϵ-close to $\boldsymbol{\beta}$ in squared ℓ_2-distance with high probability for large enough p. If ϵ is small, then $\widehat{\boldsymbol{\beta}}$ is a good estimate of $\boldsymbol{\beta}$ and we can use it in the screening step to select the variables. Typically, as ϵ decreases towards 0, the iteration counts $T(\epsilon, p, \boldsymbol{\beta})$ increases to infinity as higher accuracy generally demands more computation. However, in many examples, as we will see in Section 2.5.1.2, $T(\epsilon, p, \boldsymbol{\beta})$ depends only poly-logarithmically on ϵ^{-1}, which alleviates the computational cost.

Next, we define the binary decoder of the true support $\mathcal{S}_{\boldsymbol{\beta}}$ as $\eta_{\boldsymbol{\beta}} := (\mathbb{1}\{\beta_1 \neq 0\}, \ldots, \mathbb{1}\{\beta_p \neq 0\})^\top$. The following theorem shows that ETS can achieve exact recovery under suitable choices of tuning parameters.

Theorem 2.5.2. *Assume the condition in Assumption 2.5.1 hold and the sample size $n = \lfloor p^k \rfloor$ for some $k \in (0, 1)$. Let $r > 1$ and write $\delta = r - 1$. Then, under AURWM regime (2.2), there exist universal positive constants A_1, A_2 such that with overall iteration count no more than $T(A_1\delta, p, \boldsymbol{\beta})$ for algorithm $\mathcal{A}$, $\gamma \in (0, \delta/(8 + 8\delta))$ and $\varsigma = (1 + A_2\delta)^{1/2}$, we have that $\lim_{p\to\infty} \sup_{\boldsymbol{\beta}\in\mathcal{M}_a^s} \mathbb{P}_\beta(\hat{\eta} \neq \eta_{\boldsymbol{\beta}}) = 0$.*

Note that as the signal strength parameter r approaches the information-theoretic boundary, i.e., as δ approaches 0, ETS may require more iterations to achieve model consistency as $T(A_1\delta, p, \boldsymbol{\beta})$ generally increases as δ decreases to 0. This is not surprising: intuitively, weaker signals are harder to identify than strong ones.

Besides, ETS does not require the knowledge of the true sparsity s, but requires the knowledge of a in the second stage for accurate screening. If the true sparsity s is known, then we can enforce ETS to select exactly s features as follows: Let $|\Delta|_{(m)}$ denote the mth largest value of $\{|\Delta_i|\}_{i\in[p]}$. For each $i \in [p]$, define

$$\hat{\eta}_i(\mathbf{X}, \mathbf{y}; s) = \mathbb{1}\{|\Delta_i| \geq |\Delta|_{(s)}\}. \tag{2.12}$$

Hence $\hat{\eta}(s) := (\hat{\eta}_1(s), \ldots, \hat{\eta}_p(s))^\top$ selects exactly s features and the knowledge of a is not required in this case. The following corollary shows that under the same conditions of Theorem 2.5.2, $\hat{\eta}(s)$ achieves model consistency.

Corollary 2.5.3. *Assume the condition in Assumption 2.5.1 holds and the sample size $n = \lfloor p^k \rfloor$ for some $k \in (0, 1)$. Let A_1 be the same universal constant as in Theorem 2.5.2, $r > 1$ and write $\delta = r - 1$. Then, under AURWM regime (2.2), with overall iteration count no more than $T(A_1\delta, p, \boldsymbol{\beta})$ for algorithm $\mathcal{A}$ and $\gamma \in (0, \delta/(8 + 8\delta))$, we have that $\lim_{p\to\infty} \sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(\hat{\eta}(s) \neq \eta_{\boldsymbol{\beta}}) = 0$.*

Remark 2.5.4. *The algorithm can be made adaptive to a in some certain regime of r. In particular, if there exists a known positive constant δ_* such that $r > 1 + \delta_*$, then a threshold as in (2.10) can be constructed without the knowledge of a or r so that ETS still enjoys model consistency. In this case, δ_* can be arbitrarily small and as long as δ_* is known, an adaptive choice of threshold exists.*

Detailed proofs of Theorem 2.5.2, Corollary 2.5.3 and Remark 2.5.4 can be found in Section A.2 of the supplementary materials. Next, we will discuss some concrete examples of ETS methods that enjoy model consistency.

2.5.1.2 Examples of ETS methods

In this section, will present a few examples of ETS methods. In particular, we will consider the ETS methods with different choices for the base algorithm $\mathcal{A}$: (1) the IHT algorithm which solves the ℓ_0-constrained optimization problem, (2) the PICASSO and PGH algorithm which solves the ℓ_1-regularized optimization problem. We will show that Assumption 2.5.1 is met in all these cases and we will explicitly derive the dependence of $T(\epsilon, p, \boldsymbol{\beta})$ on $(\epsilon, p, \boldsymbol{\beta})$. Hence, this will automatically establish the model consistency of these three variants of the ETS method due to the result in Theorem 2.5.2. For clarity, depending on the algorithm used in the estimation step, we will refer to these methods as ETS-IHT, ETS-PICASSO, and ETS-PGH. To be self-contained, we describe the steps of IHT in Algorithm 2. However, we do not add the description of PICASSO and PGH as those are too involved to add in this paper. Detailed description of PICASSO and PGH can be found in Zhao et al. (2018) and Xiao and Zhang (2013) respectively. We remind the readers that throughout the discussion in this section, we will consider the AURWM regime defined in (2.2) with sample size $n = \lfloor p^k \rfloor$ for some $k \in (0, 1)$. More details and proofs related to the examples can be found in Section A.2.4 of the supplementary material.

Solving ℓ_0-constrained problem: As discussed in Section 2.5.1, In this case, the optimization problem (2.8) takes the form

$$\text{minimize}_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_0 \leq s} n_1^{-1} \left\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\theta}\right\|_2^2. \tag{2.13}$$

We consider the ETS-IHT in this case which uses IHT (Algorithm 2) to obtain an approximate solution to the above optimization problem. In this case $\mathcal{T}_{\mathcal{A}} = \{\hat{s}, h\}$, where $\hat{s}$ is the sparsity level and h is the gradient step-size. Following the discussion of Section 4 in Jain et al. (2014), in particular, using Theorem 3 of that paper we have that the final output $\widehat{\boldsymbol{\beta}}$ of IHT satisfies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq \epsilon$ with probability converging to 1, when $\hat{s} = 2592s, h \leq 8/27$ and $T(\epsilon, p, \boldsymbol{\beta}) = O(\log p + \log((1 + \|\boldsymbol{\beta}\|_\infty)/\epsilon))$. Hence, the conditions in Assumption 2.5.1 hold. Moreover, Theorem 3 of Jain et al. (2014) suggests that if $f_{n_1}(\widehat{\boldsymbol{\beta}}; \mathcal{D}_1) - \min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_0 \leq s} f_{n_1}(\boldsymbol{\theta}; D_1) \leq (\epsilon/16)$, then for large values of p, we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq \epsilon$. Hence, it is enough to output an estimator $\widehat{\boldsymbol{\beta}}$ which incurs a sub-optimality gap of the order $O(\epsilon)$.

Algorithm 2: IHT

Input: Objective function f, sparsity level $\hat{s}$, step size h ;
$\boldsymbol{\beta}^{(0)} = 0$;
$t = 0$;
while *not converged* **do**
  $\boldsymbol{\beta}^{(t+1)} = P_{\hat{s}}^0(\boldsymbol{\beta}^{(t)} - h\nabla_\theta f(\boldsymbol{\beta}^t))$, where $P_{\hat{s}}^0(\mathbf{v}) = \arg\min_{\mathbf{z}:\|\mathbf{z}\|_0 = \hat{s}} \|\mathbf{v} - \mathbf{z}\|_2$;
  $t \leftarrow t + 1$
end
Output: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$.

Solving ℓ_1-regularized problem: In this case, the objective function is

$$f_{n_1}(\boldsymbol{\theta}; \lambda, D_1) = n_1^{-1} \left\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\theta}\right\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 ,$$

and $\mathcal{C} = \mathbb{R}^p$. For brevity of discussion, we only consider PICASSO and PGH as the candidate methods for solving the above optimization problem. We omit the details of the tuning parameters for these algorithms in this paper, but details of those can be found in Zhao et al. (2018) and Xiao and Zhang (2013) respectively.

- **ETS-PICASSO**: For ETS-PICASSO, Theorem 3.12 of Zhao et al. (2018) yields that with $T(\epsilon, p, \boldsymbol{\beta}) \lesssim (\log p)^3(\log p + \log \|\boldsymbol{\beta}\|_\infty)(\log\log p + \log(\epsilon^{-1} \vee C_1))^2$ and the regularization parameter $\lambda = C_2\{(\log p)/n_1\}^{1/2}$ for appropriate absolute constants $C_1, C_2 > 0$, the approximate solution $\hat{\beta}$ has the property $f_{n_1}(\widehat{\boldsymbol{\beta}}; \lambda, \mathcal{D}_1) - f_{n_1}(\widehat{\boldsymbol{\beta}}_L; \lambda, \mathcal{D}_1) = O(\epsilon)$ with

probability converging to 1, where

$$\widehat{\boldsymbol{\beta}}_L := \arg\min_\theta f_{n_1}(\boldsymbol{\theta}; \lambda, D_1).$$

Then, the strong convexity property of the Gram matrix $\mathbf{X}^{(1)\top}\mathbf{X}^{(1)}/n_1$, and the good estimation property of $\widehat{\boldsymbol{\beta}}_L$ yields that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \le \epsilon$.

- (ETS-PGH): For ETS-PGH, Theorem 3.2 of Xiao and Zhang (2013) yields that with $T(\epsilon, p, \boldsymbol{\beta}) = O((\log p + \log \|\boldsymbol{\beta}\|_\infty) \log\log p + \log(\epsilon^{-1} \vee \tilde{C}_1))$ and $\lambda = \tilde{C}_2\{(\log p)/n_1\}^{1/2}$ for appropriate absolute constants $\tilde{C}_1, \tilde{C}_2 > 0$, the approximate solution $\widehat{\boldsymbol{\beta}}$ satisfies $f_{n_1}(\widehat{\boldsymbol{\beta}}; \lambda, D_1) - f_{n_1}(\widehat{\boldsymbol{\beta}}_L; \lambda, D_1) = O(\epsilon)$ with probability converging to 1. Then, again by the strong convexity property of the Gram matrix $\mathbf{X}^{(1)\top}\mathbf{X}^{(1)}/n_1$, and the good estimation property of $\hat{\boldsymbol{\beta}}_L$, it follows that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \le \epsilon$.

It is worth mentioning that the choice of PICASSO or PGH is not special for solving the ℓ_1-regularized problem. As long as the base algorithm $\mathcal{A}$ outputs $\widehat{\boldsymbol{\beta}}$ which enjoys a sub-optimality gap of the order $O(\epsilon)$ in the functional value, it follows that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \le \epsilon$. We formalize this result in the next proposition.

Proposition 2.5.5. *Consider the AURWM regime in* (2.2) *and let the sample size* $n = \lfloor p^k \rfloor$ *for some* $k \in (0, 1)$. *Then, there exists a positive universal constant* C_3 *such that for all* $\epsilon \in (0, C_3)$, *the following holds with probability at least* $1 - 3p^{-0.5}$ *for large enough* p *and* $\lambda = 8\{(\log p)/n_1\}^{1/2}$*:*

$$f_{n_1}(\widehat{\boldsymbol{\beta}}; \lambda, D_1) - f_{n_1}(\widehat{\boldsymbol{\beta}}_L; \lambda, D_1) \le \epsilon/C_3 \quad \textit{implies} \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \le \epsilon.$$

The proof of the above result is deferred to Section A.2.4 of the supplementary material. The above proposition basically shows that in the case of solving the LASSO problem, if $\widehat{\boldsymbol{\beta}}$ can be produced efficiently, then an optimality-gap of $C_3^{-1}\epsilon$ in the functional value is enough to guarantee that $\widehat{\boldsymbol{\beta}}$ falls inside the $\epsilon^{1/2}$ neighborhood of the true parameter $\boldsymbol{\beta}$, i.e., the conditions in Assumption 2.5.1 hold with probability at least $1 - O(p^{-0.5})$. Hence, this provides us the flexibility to use any sensible algorithm for the ℓ_1-regularized problem such as the composite gradient method proposed in Agarwal et al. (2012).

2.5.2 Discussion on information-theoretic optimality, statistical accuracy and computational efficiency

In the previous section, we have shown that ETS achieves the model consistency under the same information-theoretic optimal requirement on r with computational expediency. How-

ever, the computational efficiency of ETS heavily depends on the magnitude of r. Theorem 2.5.2 suggests that as r approaches the information-theoretic boundary 1, the demand on the number of iterations in the estimation step increases. It could be possible that the computational load of ETS surpasses the computational load of BSS when r is extremely close to 1 for a fixed ambient dimension. Hence, even though ETS is able to recover the weak signals *asymptotically* (as n approaches infinity) under the optimal requirement on r, it may suffer from high computational costs. Moreover, as r approaches 1, it turns out that the decaying rate of the error probability worsens. This fact can be verified from the rates obtained in the proof of Theorem 2.5.2 in the supplementary material and we do not include those in the main theorem for conciseness. This suggests that weak signals also hurt the statistical power or accuracy of the ETS methods. Hence, both statistical accuracy and computational efficiency suffer as the signals get weaker.

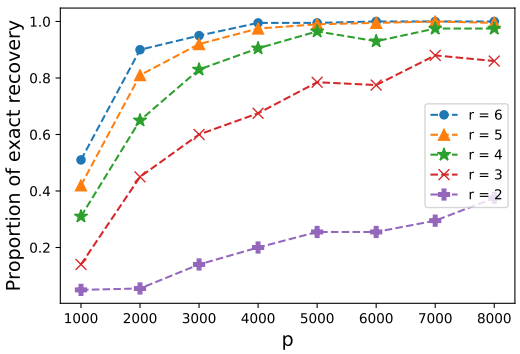
2.6 Numerical experiments

In this section, we first numerically investigate the probability for MS to achieve exact recovery of the true model with growing ambient dimension p under both homogeneous and heterogeneous signal setups. Our results show that while MS exhibits model consistency under the homogeneous signal regime, it completely fails to do so under the heterogeneous signal regime, which is consistent with Theorem 2.3.1. We then conduct simulation experiments to demonstrate the superiority of ETS methods over competing methods including LASSO and MS as signal strength grows or signal heterogeneity grows. For ETS methods, we only include ETS-IHT and ETS-PICASSO methods. For ETS-PICASSO, we used `picasso` package in R which uses the PICASSO method for solving the LASSO problem. To this end, we mention that we do not numerically compare *exact* BSS in this section mainly due to computational issues. In most of our simulation setups, we consider p in thousands and exact BSS suffers from high computational costs in such regimes, which is also a limitation of the commercial solver Gurobi (Hastie et al., 2020). Bertsimas and Parys (2020); Hazimeh et al. (2022); Xie and Deng (2020) have made efforts to overcome this computational bottleneck by considering different methods for solving approximate versions of BSS. In particular, they all consider different regularized versions of BSS which is beyond the scope of this paper, and hence we do not include those in the numerical experiments. Instead, we focus on LASSO and ETS, both of which are two different computational surrogates of the BSS problem.

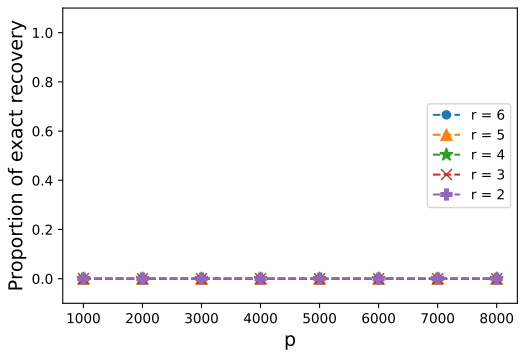
2.6.1 Exact recovery performance of MS

In Figure 2.1, we demonstrate the asymptotics of MS under both homogeneous and heterogeneous signal patterns. We consider $p \in \{1000, 2000, \ldots, 8000\}$ and signal strength parameter $r \in \{2, 3, 4, 5, 6\}$. We set $s = \lfloor 2 \log p \rfloor$ and $n = \lfloor p^{0.9} \rfloor$. We let $\tau(X, Y)$ in (2.4) be equal to the sth largest value of $\{|\mu_1|, \ldots, |\mu_p|\}$, so that MS always chooses a model of size s. For the homogeneous signal setup, we consider β with $\|\boldsymbol{\beta}\|_0 = s$ and $\beta_j = a$ for all $j \in \mathcal{S}_{\boldsymbol{\beta}}$, where a is defined in (2.2). This implies that the SNR varies between 0.19 and 2.15 across different choices of (r, p). For the heterogeneous signal setup, we consider $\boldsymbol{\beta}$ with $(s-1)$ active coordinates equal to a and one "spiky" coordinate equal to $\{10 - (s-1)a^2\}^{1/2}$. This ensures that the SNR is fixed at 10 for all choices of r, p.

Figure 2.1(a) shows that under homogeneous signal MS is able to recover the exact model with probability converging to 1 as p grows. In contrast, Figure 2.1(b) shows that under heterogeneous signal MS never achieves exact model recovery: plots for all values of r are at level 0. Such a contrast corroborates Theorem 2.3.1: signal spikes can give rise to substantial spurious correlation and jeopardize the accuracy of MS.



(a) Proportion of exact recovery under weak homogeneous signal. ($0.19 \leq \text{SNR} \leq 2.15$)



(b) Proportion of exact recovery under heterogeneous signal. ($\text{SNR} = 10$)

Figure 2.1: Asymptotics of MS with growing dimension p.

2.6.2 Effect of growing signal strength

Here we numerically compare the probability of exact support recovery of ETS with those of LASSO and MS as signal strength parameter r grows. We investigate both homogeneous and heterogeneous signal patterns. We set $p = 2000$, $s \in \{13, 52\}$ and $n = \lfloor p^{0.9} \rfloor = 935$. We set signal strength parameter $r \in \{1.5, 2, 2.5, \ldots, 9\}$ in (2.2). The support $\mathcal{S}$ is chosen

uniformly over all the size-s subsets of $[p]$, and each support coordinate of β is chosen as follows:

$$\beta_j = (1 - b_j)(1 + Z_j^2/n)^{1/2} a + b_j r^{1/2} \quad \forall j \in \mathcal{S},$$

where $(Z_j)_{j \in \mathcal{S}} \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, and where $(b_j)_{j \in \mathcal{S}} \overset{i.i.d.}{\sim} \mathsf{Ber}(\pi)$ with $\pi \in \{0, 0.2\}$. $\pi = 0$ corresponds to the homogeneous signal pattern, and $\pi = 0.2$ corresponds to the heterogeneous signal pattern, where spiky signals are present with probability 0.2. Each entry x_{ij} of the design matrix X is generated independently from $\mathsf{N}(0,1)$.

In this experiment, we grant all the approaches with the knowledge of s, so that the comparison is fair. Using this oracle knowledge, we only look at the solutions of the aforementioned three methods with sparsity exactly equal to s. Specifically, for LASSO, we look at the solution path and select the model of size exactly equal to s. For MS, we just select the top s variables corresponding to the largest absolute values of μ's. For ETS, we do not split data for estimation and screening separately; instead, we use the full data in both steps. Specifically, we replace $\mathbf{X}^{(1)}$ and $\mathbf{y}^{(1)}$ with $\mathbf{X}$ and $\mathbf{y}$ respectively in (2.13) and replace $\mathbf{X}^{(2)}$ and $\mathbf{y}^{(2)}$ with $\mathbf{X}$ and $\mathbf{y}$ respectively in (2.9). We set gradient step size $h = 0.5$ in IHT. We choose projection size $\hat{s}$ by cross-validation in terms of mean squared prediction error. Lastly, for selecting exactly s features we use (2.12) in the screening stage of ETS. It is worthwhile to mention that from an application point of view, incorporating data splitting in ETS is not necessary as we are only interested in identifying the active signals, which is akin to point estimation. Also, given the fact that $n \ll p$ in high dimensional regime, using full sample in both the estimation and screening step delivers greater sample efficiency and provides better inference.

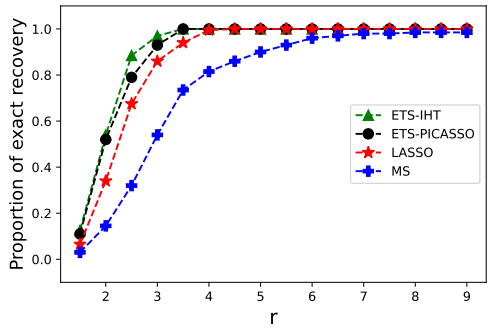
Next, for each choice of r, we run LASSO, MS, and ETS over 200 independent Monte Carlo experiments to compute the empirical probability of exact recovery. Figure 2.2 presents the results. We make the following important observations:

1. All three methods enjoy a higher chance of exact support recovery as the signal strength grows.

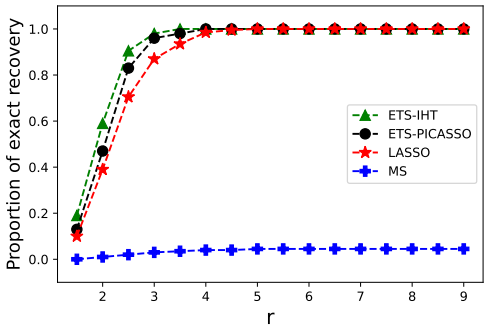
2. MS completely fails to achieve exact support recovery when s becomes large (compare panels (a) and (c)) or the signal becomes heterogeneous (compare panels (a) and (b)).

3. LASSO and ETS algorithms are insensitive to the heterogeneity of the signal. However, LASSO suffers from larger sparsity, while ETS algorithms are much more robust against it.

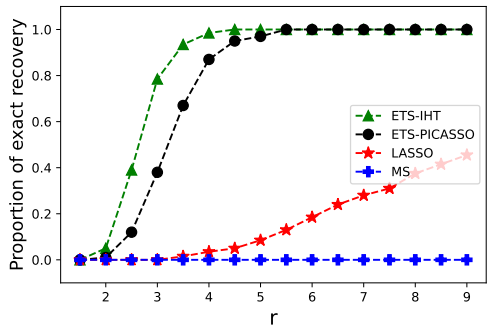
4. Overall, ETS is the best among all the three methods in terms of exact support recovery. However, ETS-IHT is somewhat better than ETS-PICASSO, and the difference



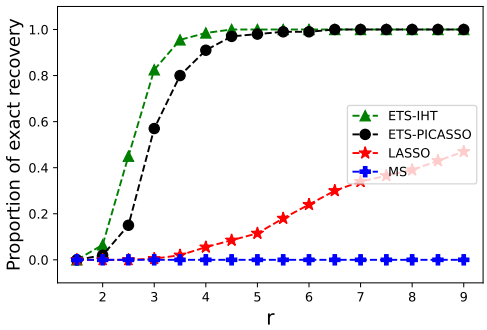
(a) Homogeneous signal pattern ($\pi = 0, s = 13$)



(b) Heterogeneous signal pattern ($\pi = 0.2, s = 13$)



(c) Homogeneous signal pattern ($\pi = 0, s = 52$)



(d) Heterogeneous signal pattern ($\pi = 0.2, s = 52$)

Figure 2.2: Plot of the proportion of exact recovery for varying r. (Gaussian case)

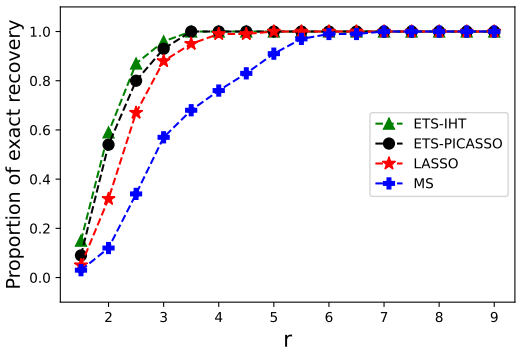
between their performance is more prominent when s is large.

Remark 2.6.1. *Genovese et al. (2012) established model consistency of LASSO under rare and weak signal regime for i.i.d. Gaussian design under a different asymptotic setting and the scaling $s = O(\log p)$ is a special case of their setting. However, they assume a homogeneous signal, and our simulation results in Figure 2.2(a) concur with their theoretical findings. However, the performance of LASSO degrades significantly for larger sparsity, even under homogeneous signal (see Figure 2.2(b)), which perhaps shows the limitations of the results in Genovese et al. (2012) for realistic values of p and s. In contrast, Wainwright (2009b) obtains the sharpest possible results for model consistency of LASSO under a very general setting. To be precise, under i.i.d. Gaussian design and for general combination of (n, p, s), the paper shows that LASSO achieves model consistency for $a = \Omega(\lambda)$, where λ is the regularization parameter and $\lambda \gtrsim \{(\log p)/n\}^{1/2}$. Hence, it is also valid for rare and weak signal regimes and also accommodates heterogeneity in the signal when $\lambda \asymp \{(\log p)/n\}^{1/2}$. However, those are*

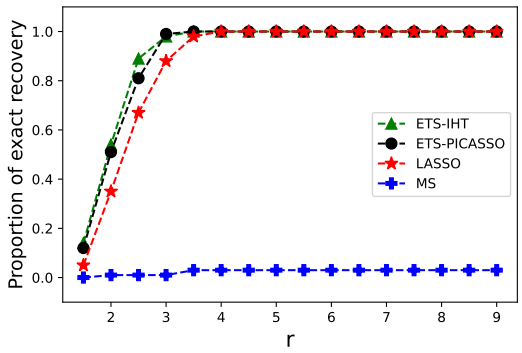
tight only up to multiplicative constants. Moreover, as pointed out in Section B of Wainwright (2009b), the model consistency of LASSO depends on whether or not the following is achieved:

$$n > s\log(p-s) + \lambda^{-2}\log(p-s).$$

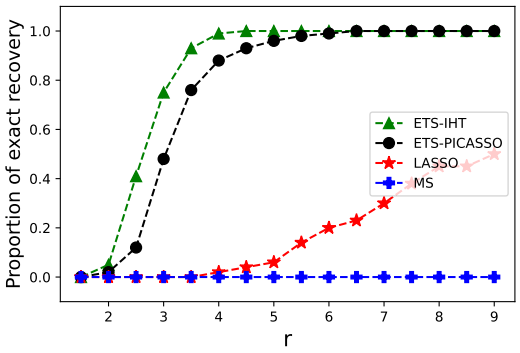
The above condition is harder to satisfy if s becomes large keeping other parameters fixed, which could be a possible explanation for the phenomenon observed in the third point of the prior observations.



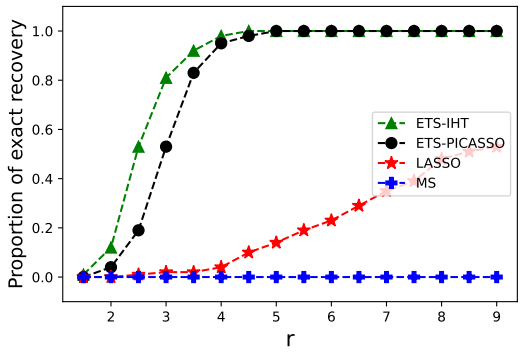
(a) Homogeneous signal pattern ($\pi = 0, s = 13$)



(b) Heterogeneous signal pattern ($\pi = 0.2, s = 13$)



(c) Homogeneous signal pattern ($\pi = 0, s = 52$)



(d) Heterogeneous signal pattern ($\pi = 0.2, s = 52$)

Figure 2.3: Plot of the proportion of exact recovery for varying r. (non-Gaussian case)

To demonstrate the effectiveness of the ETS framework beyond Gaussian design, we also evaluate the performance of ETS-IHT and ETS-PICASSO under the case where the entries of X are generated from $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$ in an i.i.d. fashion with the same choices of (n, p, s) as in the Gaussian example. Based on Figure 2.3, in this case, ETS also enjoys superior per-

formance than LASSO and MS. Moreover, we observe a similar effect of signal heterogeneity and sparsity on the model recovery rate of the methods. This provides empirical evidence in favor of the good model selection performance of ETS under sub-Gaussian design.

2.6.3 Effect of growing heterogeneity

In this numerical experiment, we study the effect of growing heterogeneity on ETS-IHT, LASSO and MS. We set $p = 2000, s = 13$, $n = \lfloor p^{0.9} \rfloor = 935$ and $r \in \{2, 6\}$ in (2.2). Next, we introduce $\mathsf{n}_{\text{spike}}$, the number of "spiky" signals in $\mathcal{S}$. We vary $\mathsf{n}_{\text{spike}}$ in $\{0\} \cup [6]$. The case $\mathsf{n}_{\text{spike}} = 0$ corresponds to the homogeneous signal setup where the true signals are set as a uniformly. For $\mathsf{n}_{\text{spike}} > 0$, we randomly set $(s - \mathsf{n}_{\text{spike}})$ signals in $\mathcal{S}$ to be equal to a and the remaining signals to be equal to a_{spike}, which is defined as

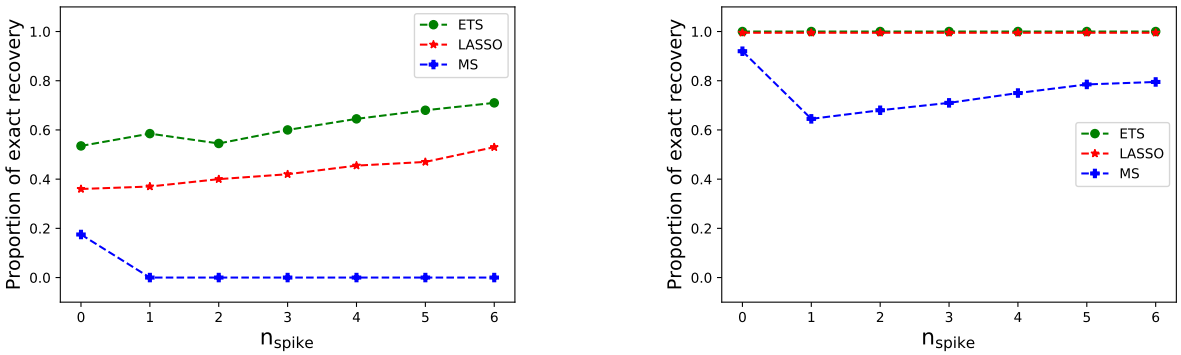
$$a_{\text{spike}} := \left\{ \frac{(2 - sa^2)}{\mathsf{n}_{\text{spike}}} + a^2 \right\}^{1/2}.$$

Such a choice of a_{spike} ensures that the SNR always equals 2 whenever $\mathsf{n}_{\text{spike}} > 0$. We perform ETS-IHT, LASSO, and MS over 200 Monte Carlo simulations for each choice of r and $\mathsf{n}_{\text{spike}}$ to obtain the empirical probability of exact support recovery. Similarly to the previous sections, we assume that the true sparsity s is known and we apply the three methods in the same fashion as before.

Figure 2.4 shows again the detrimental effect of heterogeneity on MS in terms of exact recovery. In both panels we see a significant drop in the proportion of exact recovery for MS when $\mathsf{n}_{\text{spike}}$ changes from 0 to 1. This is consistent with the theory in Section 2.3. However, in Figure 2.4(b) we see that the proportion of exact recovery is slowly increasing as $\mathsf{n}_{\text{spike}}$ grows from 1 to 6. This is because as $\mathsf{n}_{\text{spike}}$ increases, a_{spike} monotonically decreases, so that the signals become more homogeneous. MS is then able to recover the exact model more frequently. We do not see a similar phenomenon in Figure 2.4(a) because a_{spike} is too large. Another important observation is that while ETS-IHT and LASSO are both performing nearly perfectly when $r = 6$, ETS-IHT significantly outperforms both LASSO and MS when $r = 2$. Therefore, ETS-IHT is again the overall winner.

2.7 Conclusion

In this chapter, we studied the exact support recovery in high-dimensional sparse linear regression with independent Gaussian design. We focus on the AURWM regime that not only accommodates *rare* and *weak* signals as the ARW regime does but also allows *heterogeneity*



(a) $r = 2$ (b) $r = 6$

Figure 2.4: Plot of proportion of exact recovery with varying $\mathsf{n}_{\text{spike}}$.

in the signal strength. Our first theoretical result (Theorem 2.3.1) shows that marginal screening fails to achieve exact support recovery under the AURWM regime. The main reason is that the presence of "spiky" signals increases the maximum spurious marginal correlation, thereby blinding the marginal screening procedure to weak signals. Therefore, one needs to be cautious with usage of marginal screening for variable selection in practice.

In contrast, we show that BSS is robust to signal spikes and is able to achieve model consistency under the AURWM regime with the optimal requirement on signal strength (Theorem 2.4.1, 2.4.3). The primary reason behind this is that unlike MS, BSS takes into account multiple features simultaneously and thus selects variables based on their capability of fitting the residualized responses given the other variables rather than the responses themselves. Therefore, spiky signals do not affect BSS: They are very likely to be in plausible candidate models in the first place and their effect on the response has been removed in the residualization procedure. Given the recent computational advancements in solving BSS, our positive result on BSS makes it more appealing from an application point of view.

However, it is worth mentioning that even with modern advances in optimization, BSS suffers from high computational costs when the ambient dimension is extremely high. To address this issue, we propose a computationally tractable two-stage method ETS that delivers essentially the same optimal exact recovery performance as BSS (Theorem 2.5.2). Similar to BSS, ETS seeks for the features that exhibit high explanation power for the residuals from the model that excludes these features themselves (see (2.9)). Therefore, ETS is robust to spiky signals. This fact together with the slowly growing sparsity condition in (2.2) yields the optimal exact recovery accuracy of ETS.

Our work naturally raises several important questions for future research. One question

is whether similar optimality results hold for BSS and ETS when the sparsity s grows faster than $\log p$. The same question can also be asked for *correlated* random design. Another direction of our interest is studying the problem of exact recovery in a *distributed* setting where data are stored at different places and communication between them is restricted.

CHAPTER 3

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

In the previous chapter, we extensively studied the properties of marginal screening and BSS when entries of the design matrix are generated from i.i.d. Gaussian distribution. However, in reality, it is often the case that the features are highly correlated with each other. For example, in GWAS studies, the features carrying genetic information are highly correlated with other genetic features. In such cases, multi-collinearity adds a further layer of difficulty on top of high dimensionality of the data. However, recent simulation studies in Hastie et al. (2020) shows that BSS achieves superior performance in terms of model selection compared to its computational surrogates like LASSO, SCAD, and MCP. Recently, Guo et al. (2020) theoretically validated this fact by showing that the model selection performance of BSS depends on a certain *identifiability margin* that is robust to the design dependence, unlike its computational surrogates such as LASSO, SCAD, MCP, etc.

To this end, in this chapter, we consider the problem of best subset selection under a high-dimensional sparse linear regression model and add a new geometric perspective to the BSS problem that unravels a deeper understanding of the model selection problem. In particular, we show that the topological properties of the feature space are fundamental to understanding the model selection property of BSS, i.e., the complexities of the *residualized signals* and *spurious projections* (see Section 3.3.4) play fundamental roles in characterizing the margin condition for model consistency of BSS. Furthermore, 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).

3.1 Introduction

Similar to the previous chapter, we consider n observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ following the linear model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + w_i, \quad i \in \{1, \ldots, n\}, \tag{3.1}$$

where $\{\mathbf{x}_i\}_{i\in[n]}$ are *fixed* p-dimensional feature vectors, $\{w_i\}_{i\in[n]}$ are i.i.d. *mean-zero* σ-sub-Gaussian noise, i.e., $\mathbb{E}\exp(\lambda w_i) \leq \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$ and $i \in [n]$, 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

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

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$, and $\mathbf{w} = (w_1, \ldots, w_n)^\top$. 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. We focus on the variable selection problem, i.e., identifying the active set $\mathcal{S} := \{j : \beta_j \neq 0\}$ under the 0-1 loss.

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 (Akaike, 1974, 1998), BIC (Schwarz, 1978), Mallow's C_p (Mallows, 2000) 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 (Barron et al., 1999; Zhang and Zhang, 2012), such ℓ_0 regularized methods are known to suffer from huge computational bottleneck (Foster et al., 2015). This motivated a whole generation of statisticians to develop alternative penalization methods such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), MC+ (Zhang, 2010), and many others that have both strong statistical guarantees and computational expediency.

However, as discussed in Chapter 2, after recent computational advancements in solving BSS (Bertsimas et al., 2016; Bertsimas and Parys, 2020; Zhu et al., 2020), there has been growing acknowledgment that BSS enjoys significant statistical superiority over its computational surrogates and has inevitably motivated statisticians to investigate the variable selection properties of BSS. For example, through extensive simulations, Hastie et al. (2020)

shows that BSS performs better than LASSO in a high signal-to-noise ratio regime in terms of the prediction risk. On the theoretical side, Guo et al. (2020) showed that the model selection behavior of BSS does not explicitly depend on the restricted eigenvalue condition for the design (Bickel et al., 2009; Van De Geer and Bühlmann, 2009), a condition which appears unavoidable (assuming a standard computational complexity conjecture) for any polynomial-time method (Zhang et al., 2014). Specifically, they show that BSS is robust to design collinearity. Under a particular asymptotic regime and independent design, Roy et al. (2022) further established information-theoretic optimality of BSS in terms of precise constants for the signal strength parameter under weak and heterogeneous signal regimes.

Main contribution: In this chapter, 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 (Guo et al., 2020), 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 chapter is organized as follows. In Section 4.3 we discuss the preliminaries of BSS. Section 3.3 is devoted to the discussion of the key quantities, i.e., identifiability margin and the two complexities. In particular, Section 3.3.1-3.3.3 carefully introduce the notion of identifiability margin discussed in Guo et al. (2020) and the two novel complexity measures. In Section 3.3.4, we build intuition for understanding the effect of these two complexities with varying correlation. In Section 3.4, we present both sufficient (Section 3.4.1) and necessary (Section 3.4.3) 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.

3.2 Best subset selection

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

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

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 $\mathbf{X}_{\mathcal{D}} := (\mathbf{X}_j; j \in \mathcal{D})$. Let $\mathbf{P}_{\mathcal{D}} := \mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^\top \mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top$ be orthogonal projection operator onto the column space of $\mathbf{X}_{\mathcal{D}}$. Also, 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

$$\widehat{\mathcal{S}}_{\text{best}}(\widehat{s}) := \arg\min_{\mathcal{D}\subseteq[p]:|\mathcal{D}|\le\widehat{s}} R_{\mathcal{D}}. \tag{3.2}$$

Given any candidate model $\mathcal{D} \subset [p]$, we can rewrite the model (3.1) as

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

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

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

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

Then, it follows that $n^{-1}\left\|(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})\mathbf{X}_{\mathcal{S}\setminus\mathcal{D}}\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}\right\|_2^2 = \boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top\Gamma(\mathcal{D})\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}$. 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 3.2.1. *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.

3.3 Identifiability margin and two complexities

3.3.1 Identifiability margin

The discussion in Section 4.3 motivates us to define the following *identifiability margin*:

$$\tau_*(\widehat{s}) := \min_{\mathcal{D} \in \mathscr{A}_{\widehat{s}}} \frac{\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}^\top \Gamma(\mathcal{D}) \boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}}{|\mathcal{D} \setminus \mathcal{S}|}. \tag{3.3}$$

If we define $\mathscr{A}_{\widehat{s},k} := \{\mathcal{D} \in \mathscr{A}_{\widehat{s}} : |\mathcal{D} \setminus \mathcal{S}| = k\}$, 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, Guo et al. (2020) showed that under the condition

$$\tau_*(s) \gtrsim \sigma^2 \frac{\log p}{n}, \tag{3.4}$$

BSS is able to achieve model consistency. In general, Condition (4.6) is less restrictive than the well known β-min condition which demands

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

To see this, let $\hat{\lambda}_m := \min_{\mathcal{D} \in \mathscr{A}_s} \lambda_{\min}(\Gamma(\mathcal{D}))$, and note that $\tau_*(s) \geq \hat{\lambda}_m a^2$. Thus a sufficient condition for (4.6) to hold is $a \gtrsim \sigma\{\log p/(n\hat{\lambda}_m)\}^{1/2}$. In comparison, Zhang and Zhang (2012) showed that the ℓ_0-regularized least square estimator is able to achieve model consistency

when $a \gtrsim \sigma\{\log p/(n\kappa_-)\}^{1/2}$, 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 Guo et al. (2020). From now on, unless otherwise mentioned, we will assume that $\tau_*(\widehat{s}) > 0$ to avoid the non-identifiability issue as pointed out in Lemma 3.2.1.

Next, we will shift focus on the underlying geometric structures of two spaces that govern the difficulty of the BSS problem (4.4). We essentially 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.

3.3.2 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

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

and the corresponding unit vector $\widehat{\boldsymbol{\gamma}}_{\mathcal{D}} := \boldsymbol{\gamma}_{\mathcal{D}}/\left\|\boldsymbol{\gamma}_{\mathcal{D}}\right\|_2$. 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 (4.6) 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. To capture this notion of separation within the vectors $\{\boldsymbol{\gamma}_{\mathcal{D}}\}_{\mathcal{D}\in\mathscr{A}_{\widehat{s}}}$, we also need to capture the spatial alignment of their corresponding unit vectors $\{\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}\}_{\mathcal{D}\in\mathscr{A}_{\widehat{s}}}$. 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

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

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

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

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* (Talagrand, 2005), which often comes up in controlling the expectation of the supremum of Gaussian processes (Krahmer et al., 2014; Lifshits, 1995; Adler et al., 2007). In this chapter, 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:

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

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

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

This shows that the quantity $\mathscr{E}_{\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}}$ roughly captures the average separation of the set $\mathcal{T}_{\mathcal{I}}^{(\widehat{s})}$. 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.

3.3.3 Complexity of spurious projections

In this section, we will introduce the space of projection operators that also controls 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

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

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 L_1, L_2 we look at their *maximum sin-theta distance*:

$$\mathrm{d}(L_1, L_2) := \|\mathbf{\Pi}_{L_1} - \mathbf{\Pi}_{L_2}\|_{\mathrm{op}},$$

where $\mathbf{\Pi}_{L_1}, \mathbf{\Pi}_{L_2}$ are the orthogonal projection operators of L_1, L_2 respectively. It turns out that $\mathrm{d}(L_1, L_2)$ evaluates the trigonometric sine function at the maximum principal angle between the subspaces L_1 and L_2. We point the readers to Ye and Lim (2016) for a more detailed discussion on this topic.

Under this distance, we define the scaled entropy integral as

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

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 3.3.2, it can be shown that $\mathsf{d}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} \leq \mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}} \leq \mathsf{D}_{\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}}$, where

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

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 $\left|\mathcal{G}_{\mathcal{I}}^{(\widehat{s})}\right| = \binom{p-\widehat{s}}{\widehat{s}-|\mathcal{I}|}$ 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.

3.3.4 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 spurious features. If the spurious features are highly correlated then it is expected that these subspaces are 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 (3.1) follows i.i.d. mean-zero Gaussian distribution with covariance matrix

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

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 $\mathcal{I} = \varnothing$. Under this setup, we have $\mathcal{G}_{\varnothing}^{(1)} = \{\mathbf{X}_j\mathbf{X}_j^\top / \|\mathbf{X}_j\|_2^2 : j \notin \mathcal{S}\}$ 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 $\mathscr{E}_{\mathcal{G}_{\varnothing}^{(1)}}$ 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

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

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 $\mathscr{E}_{\mathcal{G}_{\varnothing}^{(1)}}$ 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 3.4.2, 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.

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 3.3.2) 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 $\mathcal{T}_{\varnothing}^{(1)}$ 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 $\left\|\widehat{\boldsymbol{\gamma}}_j - \widehat{\boldsymbol{\gamma}}_k\right\|_2^2 = 2(1 - \widehat{\boldsymbol{\gamma}}_j^\top \widehat{\boldsymbol{\gamma}}_k)$, 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 3.4.2), 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.

3.4 Theoretical properties of BSS

3.4.1 Model selection consistency of BSS under known sparsity

This section illustrates the interaction between the identifiability margin (4.5) 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 (4.4), 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 3.4.1. *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 3.4.1 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 3.4.1 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) Zhang and Huang (2008) assumes that there exist positive numbers κ_-, κ_+ and $\Psi \geq 1$ such that

$$\kappa_- \leq \frac{\|\mathbf{X}\mathbf{v}\|_2^2}{n} \leq \kappa_+, \quad \text{for all } \mathbf{v} \in \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|_2 = 1, \|\mathbf{u}\|_0 \leq \Psi s\}. \tag{3.5}$$

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

each other. In comparison, Assumption 3.4.1 is much weaker than SRC condition in two aspects. First, unlike the SRC, Assumption 3.4.1 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 3.4.1 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 (3.5) implies the condition in Assumption 3.4.1, and we formalize this claim in the following proposition.

Proposition 3.4.2. *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* (3.5) *holds with* $\Psi = 2$. *Then the condition in Assumption 3.4.1 also holds for large enough* p, *i.e.,* $\min_{\mathcal{I} \subset \mathcal{S}} \mathscr{E}_{\mathcal{G}_{\mathcal{I}}^{(s)}} \geq \kappa_-/\kappa_+ \gg \{\log(ep)\}^{-1/2}$. *Furthermore, the implication in the other direction is not true in general.*

Now we are ready to state our main sufficiency result.

Theorem 3.4.3 (Sufficiency)**.** *Under Assumption 3.4.1, there exists a positive universal constant* C_0 *such that for any* $0 \leq \eta < 1$, *whenever the identifiability margin* $\tau_*(s)$ *satisfies*

$$\frac{\tau_*(s)}{\sigma^2} \geq \frac{C_0}{(1-\eta)^2} \left[\max \left\{ \max_{\mathcal{I} \subset \mathcal{S}} \mathscr{E}^2_{\mathcal{T}_{\mathcal{I}}^{(s)}}, \max_{\mathcal{I} \subset \mathcal{S}} \mathscr{E}^2_{\mathcal{G}_{\mathcal{I}}^{(s)}} \right\} + \sqrt{\frac{\log(es) \vee \log\log(ep)}{\log(ep)}} \right] \frac{\log(ep)}{n}, \tag{3.6}$$

we have

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

with probability at least $1 - O(\{s \vee \log p\}^{-1})$. *In particular, setting* $\eta = 0$, *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 in Section B.1.3 of the supplementary material. The above theorem gives a sufficient condition for BSS to achieve model consistency.

Moreover, note that the margin condition (3.6) 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 $\{\mathbf{X}_j\}_{j \neq 1}$ but the spurious features may be extremely correlated to each other. As mentioned in the independent block design example in Section 3.4.2, 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 (3.6) 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 3.4.2 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 $\{w_i\}_{i\in[n]}$ are generated in i.i.d. fashion from $\mathsf{N}(0, 1)$. Finally, the response $\mathbf{y}$ is generated according to model (3.1). Assuming s is known, we use ABESS (Zhu et al., 2020) as a fast computational surrogate for BSS. The left panel of Figure 3.1 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 3.4.3. The right panel of Figure 3.1 also shows that a similar phenomenon is true even for $s > 1$.

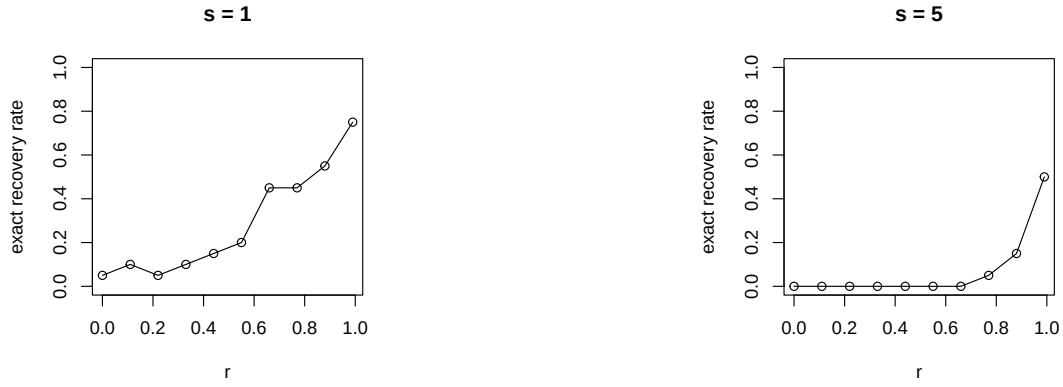


Figure 3.1: Model recovery rate of ABESS under independent block design.

On the other hand, as mentioned in Remark 3.4.6, under equicorrelated model with $\mathcal{S} = \{1\}$ and high correlation, $\mathscr{E}_{\mathcal{T}^{(1)}_{\varnothing}}$ remains strictly bounded away from 0 and it dominates $\mathscr{E}_{\mathcal{G}^{(1)}_{\varnothing}}$. However, the identifiability margin $\tau_*(s)$ becomes very small due to the high correlation between the true and noise variables. Hence, the margin condition (3.6) becomes harder to satisfy with increasing correlation. In the case of independent design, it turns out $\mathscr{E}_{\mathcal{G}^{(1)}_{\varnothing}}$ 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 $\mathscr{E}_{\mathcal{T}^{(1)}_{\varnothing}}$ smaller compared to $\mathscr{E}_{\mathcal{G}^{(1)}_{\varnothing}}$.

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 3.4.3.

We can further represent the condition (3.6) 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 3.4.4. *Let the condition in Assumption 3.4.1 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*

$$\frac{\tau_*(s)}{\sigma^2} \geq \frac{C_0}{(1-\eta)^2}\left[\max\left\{\max_{\mathcal{I}\subset\mathcal{S}} \mathsf{D}^2_{\mathcal{T}_{\mathcal{I}}^{(s)}}, \max_{\mathcal{I}\subset\mathcal{S}} \mathsf{D}^2_{\mathcal{G}_{\mathcal{I}}^{(s)}}\right\} + \sqrt{\frac{\log(es)\vee\log\log(ep)}{\log(ep)}}\right]\frac{\log(ep)}{n}, \tag{3.7}$$

we have

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

with probability at least $1 - O(\{s \vee \log p\}^{-1})$. In particular, setting $\eta = 0$, we have $\mathcal{S} = \arg\min_{\mathcal{D}\in\mathscr{A}_s} R_{\mathcal{D}}$ with high probability.

Corollary 3.4.4 essentially conveys the same message as Theorem 3.4.3, only under a slightly stronger margin condition (3.7). 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. Now in the next section, we will discuss a few illustrative examples to further elaborate on the effects of two complexities.

3.4.2 Illustrative examples

In this section, we will discuss a few illustrative examples to highlight the effect complexities of the two spaces described in Section 3.3.2 and Section 3.3.3.

Block design with a single active feature

Consider the model (3.1) 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 $\mathcal{T}_{\varnothing}^{(1)}$ and $\mathcal{G}_{\varnothing}^{(1)}$ respectively. Under this setup, we have the following lemma:

Lemma 3.4.5. *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*

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

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

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

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

$$\mathscr{E}^2_{\mathcal{T}_{\varnothing}^{(1)}} \sim \frac{2c^2(1-r)}{1-c^2}, \quad \text{and} \quad \mathscr{E}^2_{\mathcal{G}_{\varnothing}^{(1)}} \sim (1-r^2) \quad \text{for large } n, p. \tag{3.8}$$

A detailed derivation of the result is present in Section B.1.5 of the supplementary material. Left panel of Figure 3.2 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 (3.6). However, Lemma 3.4.5(a) suggest

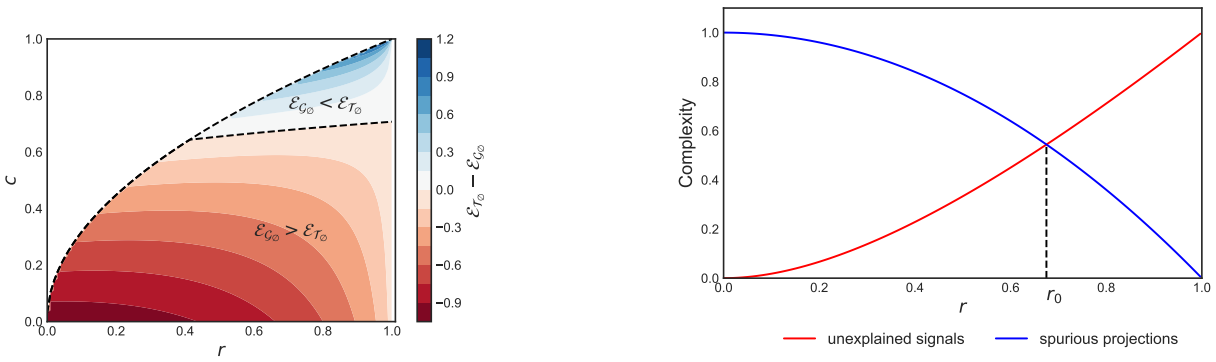


Figure 3.2: (left) Partition of c-r plane showing dominating regions for the two complexities. The color gradient indicates the value of $\mathscr{E}_{\mathcal{T}_\varnothing^{(1)}} - \mathscr{E}_{\mathcal{G}_\varnothing^{(1)}}$. (right) The plot of two complexities for varying r under equicorrelated design.

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 (3.8) suggest that $\mathscr{E}^2_{\mathcal{G}_\varnothing^{(1)}} \approx 1$. On the other hand, we see that $\mathscr{E}^2_{\mathcal{T}_\varnothing^{(1)}} \approx 0$. 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 (3.8) tells that $\mathscr{E}^2_{\mathcal{T}_\varnothing^{(1)}} \approx 0$, and $\mathscr{E}^2_{\mathcal{G}_\varnothing^{(1)}}$ 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 3.4.3 becomes less stringent.

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 (3.8), it follows that for $r \in [0, r_0)$ the complexity of spurious projection operators is dominating, i.e., $\mathscr{E}^2_{\mathcal{G}^{(1)}_\varnothing} > \mathscr{E}^2_{\mathcal{T}^{(1)}_\varnothing}$. In contrast, for $r \in (r_0, 1)$, we have the complexity of the residualized signals to be dominating, i.e., $\mathscr{E}^2_{\mathcal{T}^{(1)}_\varnothing} > \mathscr{E}^2_{\mathcal{G}^{(1)}_\varnothing}$. Right panel of Figure 3.2 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 3.4.6. *In the example of equi-correlated design with $\mathcal{S} = \{1\}$, the effect of correlation parameter r on the complexities $\mathscr{E}_{\mathcal{T}^{(1)}_\varnothing}$ and $\mathscr{E}_{\mathcal{G}^{(1)}_\varnothing}$ 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 $\mathcal{T}^{(1)}_\varnothing$ and vice versa. In contrast, higher correlation decreases the complexity $\mathscr{E}_{\mathcal{G}^{(1)}_\varnothing}$, thus shrinking the effective size of $\mathcal{G}^{(1)}_\varnothing$. Thus, in this case, the two complexities act as two opposing forces in the margin condition* (3.6).

3.4.3 Necessary condition

One question that arises from the preceding discussion is whether the margin condition in Theorem 3.4.3 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 (3.6) is essential 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

$$\widehat{\tau}(s) := \max_{j_0 \in \mathcal{S}} \max_{\mathcal{D} \in \mathscr{C}_{j_0}} \frac{\boldsymbol{\beta}^\top_{\mathcal{S}\setminus\mathcal{D}} \Gamma(\mathcal{D}) \boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{D}}}{|\mathcal{D} \setminus \mathcal{S}|} = \max_{j_0 \in \mathcal{S}} \max_{\mathcal{D} \in \mathscr{C}_{j_0}} \Gamma(\mathcal{D}) \beta^2_{j_0}. \tag{3.9}$$

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:

$$\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} := \frac{\sup_{\delta > 0} \frac{\delta}{2} \sqrt{\log \mathcal{M}(\delta, \{\widehat{\boldsymbol{\gamma}}_{\mathcal{I}_0 \cup \{j\}}\}_{j \in \mathcal{S}^c}, \|\cdot\|_2)}}{\sqrt{\log |\mathcal{T}^{(s)}_{\mathcal{I}_0}|}}. \tag{3.10}$$

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

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

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

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

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

$$\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{I}_0} = \widehat{\mathbf{u}}_j \widehat{\mathbf{u}}_j^\top, \tag{3.11}$$

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 (3.10), we define the complexity measure of $\mathcal{G}^{(s)}_{\mathcal{I}_0}$ as

$$\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}} := \frac{\sup_{\delta>0} \frac{\delta}{2}\sqrt{\log \mathcal{M}(\delta, \mathcal{G}^{(s)}_{\mathcal{I}_0}, \|\cdot\|_{\mathrm{op}})}}{\sqrt{\log |\mathcal{G}^{(s)}_{\mathcal{I}_0}|}}. \tag{3.12}$$

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 (3.11), 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 3.4.7. *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 3.4.7 combined with the observation (3.11) 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 3.4.7

induces diversity in $\mathcal{T}_{\mathcal{I}_0}^{(s)}$ and $\mathcal{G}_{\mathcal{I}_0}^{(s)}$.

Condition 3.4.8. *There exists a constant $\alpha \in (0,1)$ such that $\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}^{(s)}}/\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}^{(s)}} \in (\alpha, 1)$.*

The condition essentially tells that the set $\mathcal{T}_{\mathcal{I}_0}^{(s)}$ 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 $\mathcal{T}_{\mathcal{I}_0}^{(s)}$ are of the same order.

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

Theorem 3.4.9 (Necessity)**.** *Assume $\mathbf{w} \sim \mathsf{N}(0, \sigma^2 \mathbb{I}_n)$, $p > 16e^3$ and $s < p/2$. Also, let the Assumption 3.4.7 hold and write $\mathcal{J} = \{\mathcal{I} \subset \mathcal{S} : |\mathcal{I}| = s-1\}$. Then the following are true:*

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

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

implies that

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

(b) If there exists $\mathcal{I}_\# \in \mathcal{J}$ such that $\mathscr{E}^*_{\mathcal{G}_{\mathcal{I}_\#}^{(s)}} \in (\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_\#}^{(s)}}, \mathscr{E}_{\mathcal{T}_{\mathcal{I}_\#}^{(s)}})$, then under Condition 3.4.8, there exists a constant C_α depending on α, such that

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

implies that

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

The detailed proof can be found in Section 3.4.3 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 3.4.3, 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 3.4.3. First, Theorem 3.4.3 needs $\tau_*(s)$ to be lower bounded, which is much stronger than the required condition on $\widehat{\tau}(s)$ in Theorem 3.4.9. Second, Theorem 3.4.9 involves the alternative complexity measures $\mathscr{E}^*_{\mathcal{T}_{\mathcal{I}_0}^{(s)}}$

and $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}$, which are typically smaller than the complexity measures used in Theorem 3.4.3. Third, the resulting complexity in Theorem 3.4.3 involves the maximum over all possible subsets of $\mathcal{S}$, whereas Theorem 3.4.9 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 3.4.9 is weaker compared to the margin condition (3.6). Nonetheless, Theorem 3.4.9 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 3.4.9 can also be stated in terms of the diameter and minimum separability of the sets $\mathcal{T}^{(s)}_{\mathcal{I}_0}$ and $\mathcal{G}^{(s)}_{\mathcal{I}_0}$. Recall that $\mathscr{E}^*_{\mathcal{T}^{(s)}_{\mathcal{I}_0}} \gtrsim \mathsf{d}_{\mathcal{T}^{(s)}_{\mathcal{I}_0}}$ and $\mathscr{E}^*_{\mathcal{G}^{(s)}_{\mathcal{I}_0}} \gtrsim \mathsf{d}_{\mathcal{G}^{(s)}_{\mathcal{I}_0}}$. Hence, it follows that under the same conditions in Theorem 3.4.9, the margin condition

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

is necessary for model consistency of BSS.

3.5 Extension to GLM

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

$$f_{\mathbf{x},\boldsymbol{\beta}^*}(y) := h(y)\exp\left\{\frac{y(\mathbf{x}^\top\boldsymbol{\beta}^*) - b(\mathbf{x}^\top\boldsymbol{\beta}^*)}{\phi}\right\} = h(y)\exp\left\{\frac{y\eta - b(\eta)}{\phi}\right\}. \tag{3.13}$$

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 $y = \mathbf{x}^\top\boldsymbol{\beta}^* + w$, where $w \sim \mathsf{N}(0, \sigma^2)$. 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 $\mathcal{L}_\mathcal{D}(\tilde{\boldsymbol{\beta}}; \{(\mathbf{x}_{i,\mathcal{D}}, y_i)\}_{i\in[n]}) := (n/2)^{-1}\sum_{i\in[n]} \ell_\mathcal{D}(\tilde{\boldsymbol{\beta}}; (\mathbf{x}_{i,\mathcal{D}}, y_i))$ with $\ell_\mathcal{D}(\tilde{\boldsymbol{\beta}}; (\mathbf{x}_\mathcal{D}, y)) := -y(\mathbf{x}_\mathcal{D}^\top\tilde{\boldsymbol{\beta}}) + b(\mathbf{x}_\mathcal{D}^\top\tilde{\boldsymbol{\beta}})$. Let $\widehat{\boldsymbol{\beta}}_\mathcal{D}$ be the unique minimizer of $\mathcal{L}_\mathcal{D}(\tilde{\boldsymbol{\beta}}; \{(\mathbf{x}_{i,\mathcal{D}}, y_i)\}_{i\in[n]})$. 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

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

The above quantity can be thought of as the degree of model separation as it measures the minimum KL-distance 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 $\mathcal{D}$:

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

Note that, under the linear regression model with isotropic Gaussian error, both $\Delta_{\text{kl}}(\mathcal{D})$

and $\Delta_{\text{par}}(\mathcal{D})$ 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 $b(\cdot)$.

3.5.1 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:

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

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

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

where $\boldsymbol{\Lambda}_{\mathcal{D}} = \mathbf{diag}(b''(\mathbf{x}^{\top}_{1,\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}}), \ldots, b''(\mathbf{x}^{\top}_{n,\mathcal{D}}\bar{\boldsymbol{\beta}}_{\mathcal{D}}))$. 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:

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

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

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

3.5.2 Main results

In this section, we will state the main result analogous to the Theorem 3.4.3. We begin with some standard assumptions necessary for the theoretical analysis of GLM models.

Assumption 3.5.1 (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

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

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

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

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

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

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

Assumption 3.5.1(a) is very common in high-dimensional literature. Assumption 3.5.1(b) basically tells that the sparse-eigenvalues of $\mathbf{X}$ are strictly bounded away from 0. Assumption3.5.1(c) tells that the third order empirical moment of $\mathbf{X}_{\mathcal{D}}$ is bounded. A stronger version of Assumption 3.5.1(d) is present in Pijyan et al. (2020); Zheng et al. (2020), 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 3.5.1(e) allows diversity among the spurious features. Next, we will assume some technical assumptions on the link function $b(\cdot)$.

Assumption 3.5.2 ($b(\cdot)$ 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 in analyzing high-dimensional generalized models. Assumption 3.5.2(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 logisitc 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 (3.13) 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 $\left|\mathbf{x}^\top\boldsymbol{\beta}^*\right| \leq \omega$.

Finally, Assumption 3.5.2(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

$$\begin{aligned}
&\mathbb{E}(\exp\{t(y-b'(\eta)\}) \\
&= e^{-tb'(\eta)} \int_{-\infty}^{\infty} h(y) \exp\left\{\frac{(\eta+\phi t)y - b(\eta)}{\phi}\right\} dy \\
&= \exp\left(\frac{b(\eta+\phi t) - b(\eta) - t\phi b'(\eta)}{\phi}\right) \int_{-\infty}^{\infty} h(y) \exp\left\{\frac{(\eta+\phi t)y - b(\eta+\phi t)}{\phi}\right\} dy \\
&= \exp[\phi^{-1}\{b(\eta+\phi t) - b(\eta) - t\phi b'(\eta)\}] \leq \exp\left(\frac{\phi B t^2}{2}\right).
\end{aligned} \tag{3.14}$$

Under Assumption 3.5.2, we can compare between the margin quantities $\Delta_{\text{kl}}(\mathcal{D})$ and $\Delta_{\text{par}}(\mathcal{D})$. To see this, we first focus on $\Delta_{\text{kl}}(\mathcal{D})$. Recall that

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

where $t \in (0,1)$. Due to Assumption 3.5.1(a) and Assumption 3.5.1(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 3.5.2(a), 3.5.2(b)), we have

$$B\Delta_{\text{par}}(\mathcal{D}) \geq \Delta_{\text{kl}}(\mathcal{D}) \geq \psi(x_0 R_0 + x_0 R)\Delta_{\text{par}}(\mathcal{D}). \tag{3.15}$$

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

Theorem 3.5.3 (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 $0 \leq \eta < 1$,*

whenever the identifiability margin $\widetilde{\tau}_*(s)$ *satisfies*

$$\frac{\widetilde{\tau}_*(s)}{\phi B} \geq \frac{C}{(1-\eta)^2}\left[\max\left\{\max_{\mathcal{I}\subset\mathcal{S}} \mathscr{E}^2_{\widetilde{\mathcal{T}}^{(s)}_{\mathcal{I}}}, \max_{\mathcal{I}\subset\mathcal{S}} \mathscr{E}^2_{\widetilde{\mathcal{G}}^{(s)}_{\mathcal{I}}}\right\} + \sqrt{\frac{\log(es)\vee\log\log(ep)}{\log(ep)}}, t^{(1)}_{s,n,p}, t^{(2)}_{s,n,p}\right]\frac{\log(ep)}{n} \tag{3.16}$$

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*

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

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 B.2. The above theorem is the generalization of Theorem 1, and (3.16) also involves the two complexities related to the sets of residualized signals and spurious projection operators. However, condition (3.16) 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 B.2. 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 3.5.4. *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* (3.16)*. Hence, in this case, we witness roughly a similar phenomenon involving the two complexities as in the linear model.*

3.6 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 3.4.3 to the high-dimensional sparse generalized linear models. To be precise, we identified that a margin quantity based

on the KL distance between the true distribution and a candidate distribution governs the model selection performance of BSS. Moreover, we showed that in the GLM case, two complexity measures of the transformed feature space are fundamental to understanding the quality of model selection od BSS. 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.

CHAPTER 4

On the Computational Complexity of Private High-dimensional Model Selection

The rapid development of AI technologies has pushed the frontiers of data processing power. Large amount data are constantly being collected and processed in various ML domains ranging from engineering, computer vision to genetics and neuroimaging. This poses serious concerns related to privacy protection of sensitive user data. That is why data-privacy has emerged as one of the important aspects in the landscape of modern AI applications, and differential privacy (DP) has become a popular mathematical framework to analyze privacy loss in modern ML tasks.

In this chapter, we consider the problem of model selection in a high-dimensional sparse linear regression model under privacy constraints. We propose a differentially private best subset selection method with strong utility properties by adopting the well-known exponential mechanism for selecting the best model. We propose an efficient Metropolis-Hastings algorithm and establish that it enjoys polynomial mixing time to its stationary distribution. Furthermore, we also establish approximate differential privacy for the estimates of the mixed Metropolis-Hastings chain. Finally, we perform some illustrative experiments that show the strong utility of our algorithm.

4.1 Introduction

In this chapter, we consider the problem of *private model selection* in high-dimensional sparse regression. Once again, to clarify the mathematical model, we consider n observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ following the linear model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + w_i, \quad i \in \{1, \ldots, n\}, \tag{4.1}$$

where $\{\mathbf{x}_i\}_{i\in[n]}$ are *fixed* p-dimensional feature vectors, $\{w_i\}_{i\in[n]}$ are i.i.d. *mean-zero* σ-sub-Gaussian noise, i.e., $\mathbb{E}\exp(\lambda w_i) \leq \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$ and $i \in [n]$, 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

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

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, and $\mathbf{w} = (w_1, \ldots, w_n)^\top$. The main goal, in this case, is to identify the support $\gamma^* := \{j : \beta_j \neq 0\}$ without violating the user's privacy.

Despite these theoretical and computational advancements related to BSS (see Chapter 2 and Chapter 3), to the best of our knowledge, there is no computationally efficient private algorithmic framework for BSS for high-dimensional sparse regression setup (4.1). This is especially surprising as the private model selection is important in many contemporary applications involving sensitive data including genetics (He and Lin, 2011), neuroimaging (Mwangi et al., 2014), and computer vision Zhang et al. (2018). One major reason for this could be the lack of DP mechanisms for MIO problems which restricts us from exploiting the MIO formulation of BSS introduced in Bertsimas et al. (2016). Secondly, the apparent computational burden stemming from the requirement of exponentially large numbers of search queries in private BSS has eluded the majority of the machine learning and statistics community.

Main contribution: In this chapter, we address the latter issue by mainly focusing on the utility and computational complexity of BSS under privacy constraints. To be specific, we make the following contributions listed below:

1. We adopt the exponential mechanism (McSherry and Talwar, 2007) to design a DP BSS algorithm, and we establish its good statistical or utility guarantee under high-privacy regime whenever $\beta_{min} := \min_{j\in\gamma^*} |\beta_j| \gtrsim \sigma\{(s\log p)/n\}^{1/2}$.
2. Under the low-privacy regime, we show that accurate model recovery is possible whenever $\beta_{min} \gtrsim \sigma\{(\log p)/n\}^{1/2}$, which is the minimax optimal β_{min} requirement for model recovery under non-private setting. Therefore, this paper points out an inflection phenomenon in the signal strength requirement for the model consistency across different privacy regimes.
3. In addition, we design an MCMC chain that converges to its stationary distribution that matches the sampling distribution in the exponential mechanism. As a consequence, the model estimator generated by the MCMC also enjoys (approximate) DP. Furthermore, under certain regularity conditions on the design, we show that the

MCMC chain enjoys a polynomial mixing time in (n, p, s) to the stationary distribution with a good utility guarantee.

In summary, this chapter proposes a DP version of BSS that generates a private model estimator of γ^* with strong model recovery property within polynomial time in the problem parameters n, p, s. In the next section, we will discuss some prior related works on DP model selection and discuss some of their limitations.

4.1.1 Related Works

In the past decades, there has been a considerable amount of work studying differentially private sparse regression problems. However, most of these works focus either on empirical risk minimization (Jain and Thakurta, 2014; Talwar et al., 2015; Kasiviswanathan and Jin, 2016; Wang et al., 2017) or establishing ℓ_2-consistency rate (Wang and Xu, 2019; Cai et al., 2021) which are not directly related to the task of model selection. To the best of our knowledge, there are only three works considering the problem of variable selection in sparse regression problems under the DP framework: Kifer et al. (2012), Thakurta and Smith (2013), and Lei et al. (2018). Table 4.1 shows a clear comparison between this work and all of the prior works. Kifer et al. (2012) proposed two algorithms under sparse regression setting. One of them is based on the exponential mechanism, which is known to be computationally inefficient. However, they do not analyze the algorithm under the model selection framework. Moreover, for the privacy analysis, they assume that the loss functions are bounded over the space of sparse vectors, which is generally not true for model (4.1). In comparison, our paper provides a solid model recovery guarantee (Theorem 4.3.5) for a similar exponential mechanism without using the bounded loss assumption. Furthermore, under a slightly stronger assumption, we design a computationally efficient MCMC algorithm that also enjoys desirable utility similar to the exponential mechanism (Theorem 4.4.3) under DP framework. The other algorithm in Kifer et al. (2012) is based on the resample-and-aggregate framework (Nissim et al., 2007; Smith, 2011). Although computationally efficient, this method requires sub-optimal β_{min} condition compared to Theorem 4.3.5. In Thakurta and Smith (2013), the authors introduced two concepts of stability for LASSO and proposed two PTR-based (propose-test-release) algorithms for variable selection. However, these methods have nontrivial probabilities of outputting the null (no result), which is undesirable in practice. Also, the support recovery probabilities for these methods do not approach 1 with a growing sample size which could be problematic. In Lei et al. (2018), the authors proposed to use the Akaike information criterion or Bayesian information criterion coupled with the exponential mechanism to choose the proper model. However, the runtime of this algorithm is exponen-

tial and also requires stronger β_{min} condition. As mentioned earlier, in this paper, we show that our proposed MCMC algorithm is both computationally efficient and produces approximate DP estimates of γ^* with a strong utility guarantee under a better β_{min} condition.

Table 4.1: Comparison of DP model selection methods.

Paper	Method	β_{min} cond.	failure prob. $\to 0$	runtime
Kifer et al. (2012)	Exp-Mech	NA	NA	exp
	Lasso + Samp-Agg	$\Omega(\sqrt{\frac{s \log p}{n^{1/2}}})$	yes	poly
Thakurta and Smith (2013)	Lasso + Sub-samp. stability	$\Omega(\sqrt{\frac{s \log p}{n\varepsilon}})$	no	poly
	Lasso + Pert. stability	$\Omega(\max\{\sqrt{\frac{s \log p}{n}}, \frac{\varepsilon s^{3/2}}{n}\})$	no	poly
Lei et al. (2018)	Exp-Mech	$\Omega(\sqrt{\max\{1, \frac{s}{\varepsilon}\}\frac{s \log n}{n}})$	yes	exp
This paper	Exp-Mech	$\Omega(\sqrt{\max\{1, \frac{s}{\varepsilon}\}\frac{\log p}{n}})$	yes	exp
	Approx. Exp-Mech via MCMC	$\Omega(\sqrt{\max\{1, \frac{s}{\varepsilon}\}\frac{\log p}{n}})$	yes	poly

4.1.2 Chapter Organization and Notations

The remainder of the chapter is organized as follows. In Section 4.2, we introduce the notion of differential privacy and briefly discuss some standard DP mechanisms. Section 4.3 is devoted to a brief discussion on BSS and the development of the differentially private BSS algorithm. In particular, Section 3.3.1 discusses the notion of identifiability margin introduced in Guo et al. (2015), and Section 4.3.1 presents the exponential mechanism along with its utility guarantee (Theorem 4.3.5). Next, in Section 4.4, we provide some background on MCMC and mixing time followed by the main mixing time result (Theorem 4.4.3) and the result of the approximate DP property of MCMC samples (Corollary 4.4.4). Finally, in Section 4.5, we perform some numerical experiments on synthetic data to demonstrate the usefulness of our proposed algorithm. The concluding remarks are provided in Section 4.6. The proofs of the main results are relegated to the appendix sections.

4.2 Differential Privacy

Differential privacy requires the output of a randomized procedure to be robust with respect to a small perturbation in the input dataset, i.e., an attacker can hardly recover the presence or absence of a particular individual in the dataset based on the output only. It is important

to note that differential privacy is a property of the randomized procedure, rather than the output obtained.

4.2.1 Preliminaries

In this section, we will formalize the notion of differential privacy. Consider a dataset $D := \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$ consisting of n datapoints in the sample space $\mathcal{Z}$. A *randomized* algorithm $\mathcal{A}$ maps the dataset D to $\mathcal{A}(D) \in \mathcal{O}$, an output space. Thus, $\mathcal{A}(D)$ is a random variable on the output space $\mathcal{O}$.

For any two datasets D and D', we say they are *neighbors* if $|D \Delta D'| = 1$. We can now formally introduce the definition of differential privacy.

Definition 4.2.1 ((ε, δ)-DP, Dwork (2006))**.** *Given the privacy parameters $(\varepsilon, \delta) \in \mathbb{R}^+ \times \mathbb{R}^+$, a randomized algorithm $\mathcal{A}(\cdot)$ is said to satisfy the (ε, δ)-DP property if*

$$\mathbb{P}(\mathcal{A}(D) \in \mathcal{K}) \leq e^{\varepsilon} \mathbb{P}(\mathcal{A}(D') \in \mathcal{K}) + \delta \tag{4.2}$$

for any measurable event $\mathcal{K} \in$ range$(\mathcal{A})$ and for any pair of neighboring datasets D and D'.

In the above definition, the probability is only with respect to the randomness of the algorithm $\mathcal{A}(\cdot)$, and it does not impose any condition on the distribution of D or D'. If both ε and δ are small, then Definition 4.2.1 essentially entails that distribution of $\mathcal{A}(D)$ and $\mathcal{A}(D')$ are essentially indistinguishable from each other for any choices of neighboring datasets D and D'. This guarantees strong privacy against an attacker by masking the presence or absence of a particular individual in the dataset. As a special case, when $\delta = 0$, the notion of DP in Definition 4.2.1 is known as the *pure differential privacy*.

4.2.2 Privacy Mechanisms

For any DP procedure, a specific randomized procedure $\mathcal{A}$ must be designed that takes a database $D \in \mathcal{Z}^n$ as input and returns an element of the output space $\mathcal{O}$ while satisfying the condition in (4.2). Several approaches exist that are generic enough to be adaptable to different tasks, and which often serve as building blocks for more complex ones. A few popular examples include the Laplace mechanism (Dwork et al., 2006b), Gaussian mechanism (Dwork et al., 2006a), and Exponential mechanism (McSherry and Talwar, 2007). We discuss each of them in detail below.

Laplace mechanism: Let $f : \mathcal{Z}^n \to \mathbb{R}$ be a non-private mechanism. *Laplace mechanism* preserves privacy by perturbing the output $f(\mathcal{D})$ with noise generated from the Laplace dis-

tribution $\mathsf{Lap}(\lambda)$, whose density is $(2\lambda)^{-1}\exp(-|x|/\lambda)$. The scale $\lambda > 0$ should be calibrated to the sensitivity of the statistic f, defined as follows:

$$\Delta f := \sup_{D,D' : D,\, D' \text{ are neighbors}} |f(D) - f(D')| .$$

In particular, for any non-private mechanism f with global sensitivity $\Delta f < \infty$, we have the following result:

Lemma 4.2.2 ((Dwork et al., 2006b)). *Laplace mechanism $\mathcal{A}_L(\mathcal{D})$ that outputs*

$$\mathcal{A}_L(\mathcal{D}) = f(\mathcal{D}) + Z$$

preserves $(\varepsilon, 0)$-differential privacy, where $Z \sim \mathsf{Lap}(\Delta f/\varepsilon)$.

Intuitively, sensitivity quantifies the effect of any individual in the dataset on the outcome of the analysis. In this mechanism, Laplace noise with a magnitude proportional to the sensitivity has the effect of masking the characteristics of any individual, thereby preserving privacy.

Gaussian mechanism: Let $f : \mathcal{Z}^n \to \mathbb{R}^d$ be a non-private mapping that takes a dataset D of size n as input and outputs a vector. *Gaussian mechanism* preserves privacy by perturbing the output $f(\mathcal{D})$ with noise generated from the Gaussian distribution $\mathsf{N}(\mathbf{0}, \sigma^2\mathbb{I}_d)$. The noise scale $\sigma > 0$ should be calibrated to the ℓ_2-sensitivity of the statistic f, defined as follows:

$$\Delta_2(f) := \sup_{D,D' : D,\, D' \text{ are neighbors}} \|f(D) - f(D')\|_2 .$$

In particular, for any deterministic mapping f with global sensitivity $\Delta_2(f) < \infty$, we have the following result:

Lemma 4.2.3 (Dwork et al. (2014)). *Gaussian mechanism $\mathcal{A}_G(\mathcal{D})$ that outputs*

$$\mathcal{A}_G(\mathcal{D}) = f(\mathcal{D}) + Z$$

preserves (ε, δ)-differential privacy, where $Z \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbb{I}_d)$ with $\sigma = \sqrt{2\log(1.25/\delta)}\Delta_2(f)/\varepsilon$.

Exponential mechanism: The exponential mechanism is designed for discrete output space, Suppose $\mathscr{S} = \{\alpha_i : i \in \mathcal{I}\}$ for some index set $\mathcal{I}$, and let $u : \mathscr{S} \times \mathcal{Z}^n \to \mathbb{R}$ be score function that measures the quality of $\alpha \in \mathscr{S}$. Denote by Δu the global sensitivity of the

score function u, i.e.

$$\Delta u := \max_{\alpha \in \mathscr{S}} \max_{D,\, D' \text{ are neighbors}} |u(\alpha, D) - u(\alpha, D')| .$$

The score function $u(\cdot, \cdot)$ is called *data monotone* if the addition of a data record can either increase (decrease) or remain the same with any outcome, e.g., $u(\alpha, D) \leq u(\alpha, D \cup \{z\})$. Next, we have the following result.

Lemma 4.2.4 ((Durfee and Rogers, 2019)). *Exponential mechanism $\mathcal{A}_E(D)$ that outputs samples from the probability distribution*

$$\mathbb{P}(\mathcal{A}_E(D) = \alpha) \propto \exp\left\{ \frac{\varepsilon u(\alpha, D)}{\Delta u} \right\} \tag{4.3}$$

preserves $(2\varepsilon, 0)$-differential privacy. If $u(\cdot, \cdot)$ is data monotone, then we have $(\varepsilon, 0)$-differential privacy.

In general, if the $\mathscr{S}$ is too large the sampling from the distribution could be computationally inefficient. However, in the prequel, we show that the special structure of the linear model (4.1) allows us to design an MCMC chain that can generate approximate samples *efficiently* from an appropriate distribution similar to (4.3) for privately solving BSS.

4.3 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 $\widehat{\boldsymbol{\beta}}_{\text{best}}(\widehat{s}) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p, \|\boldsymbol{\theta}\|_0 \leq \widehat{s}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$. For model selection purposes, we can choose the best fitting model to be $\widehat{\gamma}_{\text{best}}(\widehat{s}) := \{j : [\widehat{\boldsymbol{\beta}}_{\text{best}}(\widehat{s})]_j \neq 0\}$. For a subset $\gamma \subseteq [p]$, define the matrix $\mathbf{X}_\gamma := (\mathbf{X}_j; j \in \gamma)$. Let $\boldsymbol{\Phi}_\gamma := \mathbf{X}_\gamma (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^\top$ be orthogonal projection operator onto the column space of $\mathbf{X}_\gamma$. Also, define the corresponding residual sum of squares (RSS) for model γ as $L_\gamma(\mathbf{y}, \mathbf{X}) := \mathbf{y}^\top (\mathbb{I}_n - \boldsymbol{\Phi}_\gamma)\mathbf{y}$. With this notation, the $\widehat{\gamma}_{\text{best}}(\widehat{s})$ can be alternatively written as

$$\widehat{\gamma}_{\text{best}}(\widehat{s}) := \arg\min_{\gamma \subseteq [p] : |\gamma| \leq \widehat{s}} L_\gamma(\mathbf{y}, \mathbf{X}). \tag{4.4}$$

Let $\mathbf{X}_\gamma$ be the matrix comprised of only the columns of $\mathbf{X}$ with indices in γ, and $\boldsymbol{\Phi}_\gamma$ denotes the orthogonal projection matrix onto the column space of $\mathbf{X}_\gamma$. In addition, let $\widehat{\boldsymbol{\Sigma}} := n^{-1}\mathbf{X}^\top \mathbf{X}$ be the sample covariance matrix and for any two sets $\gamma_1, \gamma_2 \subset [p]$, $\widehat{\boldsymbol{\Sigma}}_{\gamma_1, \gamma_2}$ denotes the submatrix of $\boldsymbol{\Sigma}$ with row indices in γ_1 and column indices in γ_2. Finally, define the collection $\mathscr{A}_{\widehat{s}} := \{\gamma \subset [p] : \gamma \neq \gamma^*, |\gamma| = \widehat{s}\}$, and for $\gamma \in \mathscr{A}_{\widehat{s}}$ write

$\Gamma(\gamma) = \widehat{\mathbf{\Sigma}}_{\gamma^*\setminus\gamma,\gamma^*\setminus\gamma} - \widehat{\mathbf{\Sigma}}_{\gamma^*\setminus\gamma,\gamma}\widehat{\mathbf{\Sigma}}^{-1}_{\gamma,\gamma}\widehat{\mathbf{\Sigma}}_{\gamma,\gamma^*\setminus\gamma}$. Then, it follows that $\boldsymbol{\beta}^\top_{\gamma^*\setminus\gamma}\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}$ is equal to the *residualized* signal strength $n^{-1}\|(\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}\|_2^2$. Therefore, $\boldsymbol{\beta}^\top_{\gamma^*\setminus\gamma}\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}$ quantifies the separation between γ and the true model γ^*. Ideally, a larger value of the quantity will help BSS to discriminate between γ^* and any other candidate model γ. More details on this can be found in Roy et al. (2023). Now we are ready to introduce the identifiability margin that characterizes the *model discriminative power* of BSS. We recall the *identifiablity margin*:

$$\mathfrak{m}_*(\widehat{s}) := \min_{\gamma\in\mathscr{A}_{\widehat{s}}} \frac{\boldsymbol{\beta}^\top_{\gamma^*\setminus\gamma}\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}}{|\gamma \setminus \gamma^*|}. \tag{4.5}$$

As mentioned in Chapter 3, the quantity $\mathfrak{m}_*(\widehat{s})$ captures the degree of separation between the true model γ^* and any candidate model $\gamma \neq \gamma^*$. It turns out $\mathfrak{m}_*(\widehat{s})$ is a pivotal quantity to understand the model recovery quality of BSS. For example, Guo et al. (2020) showed that under the knowledge of true sparsity, i.e., when $\widehat{s} = s$, the condition

$$\mathfrak{m}_*(s) \gtrsim \sigma^2\frac{\log p}{n}, \tag{4.6}$$

is sufficient for BSS to achieve model consistency. If we define $\lambda_* = \min_{\gamma\in\mathscr{A}_s}\lambda_{min}(\Gamma(\gamma))$, then it suffices to have $\min_{j\in\gamma^*}|\beta_i| \gtrsim \sigma\{(\log p)/(n\lambda_*)\}^{1/2}$ in order to satisfy condition (4.6).

4.3.1 Differentially Private BSS and Utility Analysis

In order to privatize the optimization problem in (4.4), we will adopt the exponential mechanism discussed in Section 4.2.2. In particular, for $K > 0$, we consider the score function

$$u_K(\gamma; \mathbf{X}, \mathbf{y}) := -\min_{\boldsymbol{\theta}\in\mathbb{R}^s:\|\boldsymbol{\theta}\|_1\leq K} \|\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\theta}\|_2^2$$

, and for a given privacy budget $\varepsilon > 0$, we sample $\gamma \in \mathscr{A}_{\widehat{s}}$ from the distribution

$$\pi(\gamma) \propto \exp\left\{\frac{\varepsilon u_K(\gamma; \mathbf{X}, \mathbf{y})}{\Delta u_K}\right\}\mathbb{1}(\gamma \in \mathscr{A}_{\widehat{s}} \cup \{\gamma^*\}). \tag{4.7}$$

As we are concerned with the exact recovery γ^*, we assume $\widehat{s} = s$. The above algorithm is essentially the same as Algorithm 4 in Kifer et al. (2012); however, they do not introduce the extra ℓ_1 constraint on the parameter space. Instead, their algorithm needs the loss-term $(y - \mathbf{x}_\gamma^\top\boldsymbol{\theta})^2$ to be bounded by a constant for every possible choice of $\mathbf{x}, y, \gamma$ and $\boldsymbol{\theta}$. This assumption is not true in general for the squared error loss, and to remedy this issue, we introduce the extra ℓ_1 constraint in the score function. This is a common truncation strategy that is used to guarantee worst-case sensitivity bound and similar methods also have been

adopted in Lei et al. (2018) and Cai et al. (2021) to construct private estimators. Next, we present the following lemma that shows the data-monotonicity of the proposed score function.

Lemma 4.3.1. *The score function $u_K(\gamma;\cdot)$ in (4.7) is data monotone.*

Therefore, Lemma 4.2.4 automatically guarantees that the above procedure is $(\varepsilon, 0)$-DP. However, in practice, we need an explicit form for Δu_K to carry out the sampling method, and it is also needed to analyze the utility guarantee of the exponential mechanism. To provide a concrete upper bound on the global sensitivity of $u_K(\cdot;\cdot)$, we make the following boundedness assumption on the database:

Assumption 4.3.2. *There exists positive constants r, x_{max} such that $\sup_{y\in\mathcal{Y}}|y| \leq r, \sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_\infty \leq x_{\mathsf{max}}$.*

Under this assumption, the following lemma provides an upper bound on the global sensitivity of the score function.

Lemma 4.3.3 (Sensitivity bound and DP)**.** *Under Assumption 4.3.2, the global sensitivity Δu_K is bounded by $\Delta_K := (r + x_{\mathsf{max}}K)^2$. Therefore, the exponential mechanism (4.7) with Δu_K replaced by Δ_K satisfies $(\varepsilon, 0)$-DP.*

The above lemma provides an upper bound on the global sensitivity of the score function rather than finding the exact value of it. Therefore, to guarantee $(\varepsilon, 0)$-DP property of exponential mechanism, it suffices to use the upper bound of Δu_K in (4.7). Now we will shift towards the utility analysis of the proposed exponential mechanism. First, we require some technical assumptions.

Assumption 4.3.4. *We assume the following hold:*

(a) There exists positive constants r, x_{max} and b_{max} such that $\sup_{y\in\mathcal{Y}}|y| \leq r, \sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_\infty \leq x_{\mathsf{max}}$ and $\|\boldsymbol{\beta}\|_1 \leq b_{\mathsf{max}}$.

(b) There exists positive constants $\kappa_{,}\kappa_+$ such that

$$\kappa_- \leq \lambda_{\min}\left(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma/n\right) \leq \lambda_{\max}\left(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma/n\right) \leq \kappa_+, \tag{4.8}$$

for all $\gamma \in \mathscr{A}_s \cup \{\gamma^*\}$.

(c) The true sparsity level s follows the inequality $s \leq \frac{n}{\log p}$.

Assumption 4.3.4(a) tells that the true parameter $\boldsymbol{\beta}$ lies inside a ℓ_1-ball. Similar boundedness assumptions are fairly standard in privacy literature (Wang, 2018; Lei et al., 2018; Cai et al., 2021). Assumption 4.3.4(b) is a well-known assumption in the high-dimensional literature (Zhang and Huang, 2008; Huang et al., 2018; Meinshausen and Yu, 2009) which is known as the Sparse Riesz Condition (SRC). Finally, Assumption 4.3.4(c) essentially assumes that the $s = o(n)$, i.e., sparsity grows with a sufficiently small rate compared to the sample size n.

Theorem 4.3.5 (Utility gurantee)**.** *Let the conditions in Assumption 4.3.2 and Assumption 4.3.4 hold. Set* $K \geq \{(\kappa_+/\kappa_-)b_{\mathsf{max}} + (8x_{\mathsf{max}}/\kappa_-)\sigma\}\sqrt{s}$. *Then, under the data generative model* (4.1)*, there exist universal positive constants* c_1, C_1 *such that whenever*

$$\mathbb{m}_*(s) \geq C_1\sigma^2 \max\left\{1, \frac{\Delta_K}{\varepsilon\sigma^2}\right\}\frac{\log p}{n}, \tag{4.9}$$

with probability at least $1 - c_1 p^{-2}$ *we have* $\pi(\gamma^*) \geq 1 - p^{-2}$.

Theorem 4.3.5 essentially says that whenever the identifiability margin is large enough, the exponential mechanism outputs the true model γ^* with high probability. Note that $\Delta_K/\sigma^2 = \Omega(s)$. In the low privacy regime, i.e., for $\varepsilon > \Delta_K/\sigma^2$ we only require $\mathbb{m}_*(s) \gtrsim \sigma^2(\log p)/n$ to achieve model consistency and this matches with the optimal rate for model consistency of non-private BSS. In contrast, in a high privacy regime, i.e., for $\varepsilon < \Delta_K/\sigma^2$, Condition (4.9) essentially demands $\mathbb{m}_*(s) \gtrsim \sigma^2(s\log p)/(n\varepsilon)$ to achieve model consistency. Thus, in a high privacy regime, we pay an extra factor of (s/ε) in the margin requirement.

Remark 4.3.6. *The failure probability in Theorem 4.3.5 can be improved to* $O(p^{-M})$ *for any arbitrary integer* $M > 2$. *However, we have to pay a cost in the universal constant* C_1 *in terms of a multiplicative constant larger than 1.*

Remark 4.3.7. *Under Assumption 4.3.4(b), it follows that* $\lambda_* \geq \kappa_-$. *Therefore, it suffices to have* $\min_{j\in\gamma^*}\beta_j^2 \geq \left(\frac{C_1\sigma^2}{\kappa_-}\right)\max\{1, \Delta_K/(\varepsilon\sigma^2)\}\frac{\log p}{n}$ *in order to hold condition* (4.9). *Therefore, in high-privacy regime, our method requires* $\min_{j\in\gamma^*}|\beta_j| \gtrsim \sigma\{(s\log p)/(n\varepsilon\kappa_-)\}^{1/2}$. *In contrast, under the low-privacy regime, we retrieve the optimal requirement* $\min_{j\in\gamma^*}|\beta_j| \gtrsim \sigma\{(\log p)/(n\kappa_-)\}^{1/2}$.

4.4 Efficient Sampling through MCMC

In this section, we will propose an efficient sampling method to generate approximate samples from the distribution (4.7). One of the challenges of sampling methods in high-dimension is

their high computational complexity. For example, the distribution in (4.7) places mass on all $\binom{p}{s}$ subsets of $[p]$, and it is practically infeasible to sample γ from the distribution as we have to essentially explore over an exponentially large space. This motivates us to resort to sampling techniques based on MCMC, through which we aim to obtain approximate samples from the distribution in (4.7). Past works on MCMC algorithms for Bayesian variable selection can be divided into two main classes – Gibbs sampler (George and McCulloch, 1993b; Ishwaran and Rao, 2005; Narisetty et al., 2018) and Metropolis-Hastings (Hans et al., 2007; Lamnisos et al., 2013). In this chapter, we focus on a particular form of Metropolis-Hastings updates.

In general terms, Metropolis-Hastings random walk is an iterative and local-move based method involving three steps:

1. Given the current state γ, construct a neighborhood $\mathcal{N}(\gamma)$ of proposal states.

2. Choose a new state $\gamma' \in \mathcal{N}(\gamma)$ according to some proposal distribution $\mathbf{F}(\gamma, \cdot)$ over the neighborhood $\mathcal{N}(\gamma)$.

3. Move to the new state γ' with probability $\mathbf{R}(\gamma, \gamma')$, and stay in the original state γ with probability $1 - \mathbf{R}(\gamma, \gamma')$, where the acceptance probability is given by

$$\mathbf{R}(\gamma, \gamma') = \min\left\{1, \frac{\pi(\gamma')\mathbf{F}(\gamma', \gamma)}{\pi(\gamma)\mathbf{F}(\gamma, \gamma')}\right\},$$

   where $\pi(\cdot)$ is same as in Equation (4.7).

This procedure generates a Markov chain for any choice of the neighborhood structure $\mathcal{N}(\gamma)$ with the following transition probability:

$$\mathbf{P}_{\mathsf{MH}}(\gamma, \gamma') = \begin{cases} \mathbf{F}(\gamma, \gamma')\mathbf{R}(\gamma, \gamma'), & \text{if } \gamma' \in \mathcal{N}(\gamma), \\ 1 - \sum_{\gamma' \neq \gamma} \mathbf{P}_{\mathsf{MH}}(\gamma, \gamma'), & \text{if } \gamma' = \gamma, \\ 0, & \text{otherwise.} \end{cases}$$

The specific form of Metropolis-Hastings update analyzed in this chapter is obtained by following the *double swap update* scheme to update γ.

Double swap update: Let $\gamma \in \mathscr{A}_s \cup \{\gamma^*\}$ be the initial state. Choose an index pair $(k, \ell) \in \gamma \times \gamma^c$ *uniformly* at random. Construct the new state γ' by setting $\gamma' = \gamma \cup \{\ell\} \setminus \{k\}$.

The above scheme can be viewed as a general Metropolis-Hastings update scheme when $\mathcal{N}(\gamma)$ is the collection of all models γ' which can be obtained by swapping two distinct

coordinates of γ and γ^c respectively. Thus, letting $d_H(\gamma, \gamma') = |\gamma \setminus \gamma'| + |\gamma' \setminus \gamma|$ denote the Hamming distance between γ and γ', the neighborhood is given by $\mathcal{N}(\gamma) = \{\gamma' \mid d_H(\gamma, \gamma') = 2, \exists\ (k, \ell) \in \gamma \times \gamma^c \text{ such that } \gamma' = \gamma \cup \{\ell\} \setminus \{k\}\}$.

With the above definition, the transition matrix of the previously described Metropolis-Hastings scheme can be written as follows:

$$\mathbf{P}_{\mathsf{MH}}(\gamma, \gamma') = \begin{cases} \frac{1}{|\gamma||\gamma^c|} \min\{1, \frac{\pi(\gamma')}{\pi(\gamma)}\}, & \text{if } \gamma' \in \mathcal{N}(\gamma), \\ 1 - \sum_{\gamma' \neq \gamma} \mathbf{P}_{\mathsf{MH}}(\gamma, \gamma'), & \text{if } \gamma' = \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{4.10}$$

4.4.1 Mixing Time and Approximate DP

Let $\mathcal{C}$ be a Markov chain on the discrete space $\mathscr{S}$ with a transition probability matrix $\mathbf{P} \in \mathbb{R}^{|\mathscr{S}| \times |\mathscr{S}|}$ with stationary distribution ν. Throughout our discussion, we assume that $\mathcal{C}$ is reversible, i.e., it satisfies the balanced condition $\nu(\gamma)\mathbf{P}(\gamma, \gamma') = \nu(\gamma')\mathbf{P}(\gamma', \gamma)$ for all $\gamma, \gamma' \in \mathscr{S}$. Note that the previously described transition matrix $\mathbf{P}_{\mathsf{MH}}$ in (4.10) satisfies the reversibility condition. It is convenient to identify a reversible chain with a weighted undirected graph G on the vertex set $\mathscr{S}$, where two vertices γ and γ' are connected if and only if the edge weight $\mathbf{Q}(\gamma, \gamma') := \nu(\gamma)\mathbf{P}(\gamma, \gamma')$ is strictly positive. For $\gamma \in \mathscr{S}$ and any subset $S \subseteq \mathscr{S}$, we write $\mathbf{P}(\gamma, S) = \sum_{\gamma' \in S} \mathbf{P}(\gamma, \gamma')$. If γ is the initial state of the chain, then the total variation distance to the stationary distribution after t iterations is

$$\Delta_\gamma(t) = \left\| \mathbf{P}^t(\gamma, \cdot) - \nu(\cdot) \right\|_{\mathsf{TV}} := \max_{S \subset \mathscr{S}} \left| \mathbf{P}^t(\gamma, S) - \nu(S) \right|.$$

The η-mixing time is given by

$$\tau_\eta := \max_{\gamma \in \mathscr{S}} \min\{t \in \mathbb{N} \mid \Delta_\gamma(t') \leq \eta \text{ for all } t' \geq t\}, \tag{4.11}$$

which measures the number of iterations needed for the chain to be within distance $\eta \in (0, 1)$ of the stationary distribution.

Privacy of MCMC estimator: Now, we will show that once the MH chain in (4.10) has mixed with its stationary distribution $\pi(\cdot)$ defined in (4.7), the model estimators at each iteration will enjoy approximate DP. To fix the notation, let γ_t be the tth iteration of the MH chain in (4.10). Then, we have the following useful lemma:

Lemma 4.4.1. *The model estimator γ_{τ_η} is (ε, δ)-DP with $\delta = \eta(1 + e^\varepsilon)$.*

The above lemma shows that smaller η entails a better privacy guarantee for a fixed level ε as δ decreases with η. Therefore, allowing more mixing of the chain will provide better privacy protection. However, this raises a concern about how long a practitioner must wait until the chain archives η-mixing. In particular, it is important to understand how τ_η scales in the difficulty parameters of the problem, for example, the dimension of the parameter space and sample size. In our case, we are interested in the covariate dimension p, sample size n, sparsity s, and the privacy parameter ε. In the next section, we will show that the chain with transition matrix (4.10) enjoys rapid mixing, meaning that the mixing time τ_η grows at most at a polynomial rate in p, s and the sample size n.

4.4.2 Rapid Mixing of MCMC and Approximate DP

We now turn to develop sufficient conditions for Metropolis-Hastings scheme (4.10) to be rapidly mixing. To this end, we make a technical assumption on the design matrix. Essentially, the following assumption controls the amount of correlation between active features and spurious features.

Assumption 4.4.2. *For every $\gamma' \in \mathscr{A}_s \setminus \{\gamma^*\}$, there exists $k \notin \gamma^* \cup \gamma'$ such that*

$$\max_{j \in \gamma^* \setminus \gamma'} \frac{\left|\mathbf{X}_j^\top (\mathbb{I}_n - \mathbf{\Phi}_{\gamma'}) \mathbf{X}_k\right|}{\left\|(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'}) \mathbf{X}_k\right\|_2} \leq b_{\mathsf{max}}^{-1} \sqrt{(\kappa_- C_1 \sigma^2/2) \log p},$$

where C_1 is the same universal positive constant as in Theorem 4.3.5.

First, note that Assumption 4.4.2 basically controls the length of the projection of the feature $\mathbf{X}_j$ on the unit vector $\mathbf{u}_k := (\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k / \left\|(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k\right\|_2$. Therefore, the above inequality restricts the correlation between an active feature $\mathbf{X}_j \boldsymbol{\beta}_{\gamma^*}$ and the spurious scaled feature $\mathbf{u}_k$ from being too large. To this end, we emphasize that stronger assumptions on model correlation (on top of the SRC condition) are common in literature for establishing the computational efficiency of Bayesian variable selection methods involving MH algorithm. For example, to show the computational efficiency MH algorithm under Zellner's g-prior, Yang et al. (2016) assumes

$$\max_{\gamma : |\gamma| \leq s_0} \left\|(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^\top \mathbf{X}_{\gamma^* \setminus \gamma}\right\|_{\mathrm{op}}^2 = O\left(\frac{n}{s \log p}\right), \tag{4.12}$$

where s_0 (larger than s) is a specific tuning parameter of their algorithm that controls the model size. The assumption in the above display is akin to the well-known irrepresentability condition Zhao and Yu (2006) which is a very strong assumption on the design. On a high level, at any given current state γ, Assumption 4.4.2 or Condition (4.12) helps to identify

a good local move towards the true model γ^* in the MH algorithm via deletion of the least influential covariate in γ. Now, we present our main result for the mixing time of MCMC.

Theorem 4.4.3 (Rapid mixing time)**.** *Let the conditions in Assumption 4.3.2, Assumption 4.3.4 and Assumption 4.4.2 hold. Then, under the data generative model* (4.1)*, there exists a universal constant $C_1' > 0$ such that under the margin condition*

$$\mathfrak{m}_*(s) \geq C_1' \sigma^2 \max\left\{1, \frac{\Delta_K}{(\kappa_- \wedge 1)\varepsilon\sigma^2}\right\} \frac{\log p}{n}, \tag{4.13}$$

there exist universal positive constants c_2, C_2 such that the mixing time τ_η of the MCMC chain (4.10) *enjoys the following with probability at least $1 - c_2 p^{-2}$:*

$$\tau_\eta \leq C_2 p s^2 \left\{n\varepsilon\Psi^{-1}\kappa_+ b_{\mathsf{max}}^2 + \log(1/\eta)\right\}, \tag{4.14}$$

where $\Psi = \{r + (\kappa_+/\kappa_-)b_{\mathsf{max}}x_{\mathsf{max}} + (\sigma/\kappa_-)x_{\mathsf{max}}^2\}^2$.

First note that the margin condition (4.13) is slightly stronger than the margin condition in Theorem 4.3.5. Under that condition, the above theorem shows that the η-mixing time of the MCMC algorithm designed for approximate sampling from the distribution (4.7) grows at a polynomial rate in (n, p, s). Recall that according to the previous definition (4.11) of the mixing time, Theorem 4.4.3 characterizes the *worst-case* mixing time, meaning the number of iterations when starting from the worst possible initialization. If we start with a good initial state — for example, the true model γ^* would be an ideal though impractical choice - then we can remove the n term in the upper bound in (4.14). Therefore, the bound in (4.14) can be thought of as the worst-case number of iterations required in the burn-in period of the MCMC algorithm. Furthermore, it is important to point out that Assumption 4.4.2 is only needed to ensure the "quick" mixing time of the MCMC chain. It is possible to relax this assumption, however, in that case, the MCMC chain is not guaranteed to mix under polynomial time. Nonetheless, given enough iterations, the chain will indeed converge to the distribution (4.7) as MH algorithm always generates an ergodic chain that eventually mixes to its stationary distribution.

It is interesting to note that Theorem 4.4.3 suggests that in a large ε regime, the chain mixes slower compared to the small ε regime. The main reason for this is that Theorem 4.4.3 only relies on worst-case analysis. The intuition is the following: When ε is very large, then the target distribution is essentially fully concentrated on γ^* (assuming the score for γ^* is highest). Now, the current analysis of Theorem 4.4.3 does not assume any condition on the initial state of the MCMC chain. It treats the initial state γ_0 as if it is chosen in a completely random manner, i.e., it is the worst case. From this point of view, it is hard for a completely

uninformative distribution to converge to a target distribution that is concentrated on a single subset (very informative), and resulting in a longer mixing time. Finally, Theorem 4.4.3 leads to the following corollary:

Corollary 4.4.4. *Let π_t denote the distribution of the tth iterate γ_t of the MCMC scheme (4.10). Then, under the conditions of Theorem 4.3.5 and Theorem 4.4.3, there exists a universal constant $c_3 > 0$ such that for any fixed iteration t such that $t \geq C_2 ps^2 \{n\varepsilon\Psi^{-1}\kappa_+ b_{\mathsf{max}}^2 + \log(1/\eta)\}$, we have $\pi_t(\gamma^*) \geq 1 - \eta - p^{-2}$ with probability at least $1 - c_3 p^{-2}$.*

The above corollary is useful in the sense that it provides a quantitative choice of η that yields high utility of the estimator γ_t. For example, if we set $\eta = p^{-2}$ and $\varepsilon = O(1)$, then for any $t \gtrsim ps^2(n\varepsilon + \log p)$ the resulting sample γ_t will match γ^* with probability $1 - c_3 p^{-2}$.

Remark 4.4.5. *Similar to Remark 4.3.6, the failure probability in Theorem 4.4.3 and Corollary 4.4.4 can be improved to $O(p^{-M})$ for arbitrary large $M > 2$, but at the cost of paying higher values for the absolute constants C_1' and C_2.*

4.5 Numerical experiments

In this section, we will conduct some illustrative simulations[1]. To compare the quality of the DP model estimator, we compare `F-score` (Hastie et al., 2020) of the estimated model with that of the true model γ^* and the BSS estimator. As the actual BSS is computationally infeasible, we use the adaptive best subset selection (ABESS) algorithm (Zhu et al., 2022) as a computational surrogate to BSS. Throughout this section, we assume that the true sparsity s is known, i.e., we provide the knowledge of s to the algorithm.

We consider a random design matrix, formed by choosing each entry from the distribution Uniform$(-1, 1)$ in i.i.d. fashion. In detail, we set $n = 900, p = 2000$, and the sparsity level $s = 4$. we generate the entries of the noise $\mathbf{w}$ independently from Uniform$(-0.1, 0.1)$, and consider the liner model (4.1). We choose the design vector $\boldsymbol{\beta}$ with true sparsity $s = 4$ and the support set $\gamma^* = \{j : 1 \leq j \leq 4\}$. We set all the signal strength to be equal, taking the following two forms: (i) **Strong signal:** $\beta_j = 2\{(s \log p)/n\}^{1/2}$, and (ii) **Weak signal:** $\beta_j = 2\{(\log p)/n\}^{1/2}$ for all $j \in \gamma^*$.

Under these setups, we consider the privacy parameter $\varepsilon \in \{0.5, 1, 3, 5, 10\}$ which are acceptable choices of ε (Near and Darais, 2022). Moreover, similar (or larger) choices of ε are common in various applications including US census study (Garfinkel, 2022), socio-economic study (Rogers et al., 2020), and industrial applications (Apple, 2017; Young et al.,

[1]Codes are available online: Github link.

Table 4.2: Comparison of mean `F-score`'s across chains for $K = 2$. $(*)$ denotes that the chain has mixed reasonably.

Privacy	`F-score`	
	Strong signal	Weak signal
$\varepsilon = 0.5$	**0.025**	0.00
$\varepsilon = 1$	**0.15**	0.05
$\varepsilon = 3$	**1.00***	0.15
$\varepsilon = 5$	**1.00***	0.40
$\varepsilon = 10$	**1.00***	**1.00***
Non-private	**1.00**	**1.00**

2020). For the Metropolis-Hastings random walk, we vary $K \in \{0.5, 2, 3, 3.5\}$ and initialize 10 independent Markov chains from random initializations and record the `F-score` of the last iteration. We also track the qualities of the model through its explanatory power for convergence diagnostics. In particular, we calculate the scale factor $R_\gamma := \mathbf{y}^\top \mathbf{\Phi}_\gamma \mathbf{y} / \|\mathbf{y}\|_2^2$ for each model update along the random walk and compare those with $R_{\widehat{\gamma}_{\text{best}}}$ to heuristically gauge the quality of mixing. More details and a set of comprehensive plots can be found in Appendix C.1 where we also discuss more about the effect of ε and K on the utility. For $K = 2$, Table 4.2 shows that `F-score` increases as ε increases both in the cases of strong and weak signals. In fact, for $\varepsilon \geq 3$, the performance of the algorithm is on par with the non-private BSS. This is consistent with the inflection phenomenon pointed out in Theorem 4.3.5 and Corollary 4.4.4. Furthermore, as expected, we see that for a fixed ε, the `F-score` is generally higher in the strong signal case.

4.6 Conclusion

In this paper, we study the variable selection performance of BSS under the differential privacy constraint. In order to achieve (pure) differential privacy, we adopt the exponential mechanism and establish its high utility guarantee in terms of exact model recovery. Furthermore, to overcome the computational bottleneck of the sampling step in the exponential mechanism, we design a Metropolis-Hastings random walk that provably mixes with the stationary distribution within a mixing time of the polynomial order in (n, p, s). We also show that the samples from the Metropolis-Hastings random walk also enjoy approximate DP with a high utility guarantee. Finally, we carry out a few illustrative simulation experiments to demonstrate the good performance of our algorithm. In summary, as discussed in Section

4.1.1, we establish both high utility and efficient computational guarantee for our model selection algorithm under privacy constraints, which is in sharp contrast with the previous works in DP model selection literature.

To this end, we also point out some of the open problems and future directions that naturally arise from this research. One limitation, of our main result Theorem 4.3.5 is that it required the condition $\min_{j\in\gamma^*} |\beta_j| = \Omega(\sqrt{(s\log p)/n})$ in high-privacy regime. It is still an open question whether the extra $\sqrt{s}$ factor is necessary for model selection. Future research along this line could focus on solving BSS through DP mixed integer optimization (MIO). This would mean an important contribution in this field as commercial solvers like GUROBI or MOSEK would be capable of solving the BSS problem with high computational efficiency using a general DP framework via objective perturbation.

CHAPTER 5

Thompson Sampling for High-Dimensional Sparse Linear Contextual Bandits

Online learning has emerged as one of the important fields of research in machine learning, with applications ranging from recommendation systems and robotics to personalized medicine and clinical trials. In such applications, the learner needs to take quick and accurate actions. However, in the modern era of big data, both processing of the data and learning tasks have become increasingly difficult due to the curse of dimensionality. One special case of the online learning problem is the contextual linear bandit problem where each action is associated with a context that carries information about the users. For example, it could be the user's search history, and based on this context information the bandit learner (the browser) can recommend the next set of actions that will maximize the reward (the clicks) for the learner. This problem has been studied extensively both in low and high-dimensional settings. However, exploration of the Thompson sampling algorithm in high-dimensional settings remained somewhat limited.

In this chapter [1], we consider the stochastic linear contextual bandit problem with high-dimensional features. We analyze the Thompson sampling algorithm using special classes of sparsity-inducing priors (e.g., spike-and-slab) to model the unknown parameter and provide a nearly optimal upper bound on the expected cumulative regret. To the best of our knowledge, this is the first work that provides theoretical guarantees of Thompson sampling in high-dimensional and sparse contextual bandits. For faster computation, we use variational inference instead of Markov Chain Monte Carlo (MCMC) to approximate the posterior distribution. Extensive simulations demonstrate the improved performance of our proposed algorithm over existing ones.

[1] First authorship of the paper is shared between Saptarshi Roy and Sunrit Chakraborty.

5.1 Introduction

Sequential decision-making, including bandits problems and reinforcement learning, has been one of the most active areas of research in machine learning. It formalizes the idea of selecting actions based on current knowledge to optimize some long-term reward over sequentially collected data. On the other hand, the abundance of personalized information allows the learner to make decisions while incorporating this contextual information, a setup that is mathematically formalized as contextual bandits. Moreover, in the big data era, the personal information used as contexts often has a much larger size, which can be modeled by viewing the contexts as high-dimensional vectors. Examples of such models cover internet marketing and treatment assignments in personalized medicine, among many others.

A particularly interesting special case of the contextual bandit problem is the linear contextual bandit problem, where the expected reward is a linear function of the features (Abe et al., 2003; Auer, 2002). Under this setting, Dani et al. (2008), Chu et al. (2011) and Abbasi-Yadkori et al. (2011) showed polynomial dependence of the cumulative regret on ambient dimension d and time horizon T in low dimensional case. Specifically, Dani et al. (2008) and Abbasi-Yadkori et al. (2011) proved a regret upper bound scaling as $O(d\sqrt{T})$, while Chu et al. (2011) showed a regret upper bound of the order $O(\sqrt{dT})$. It is worthwhile to mention that all of the aforementioned algorithms fall under a certain class of algorithms known as upper confidence bound (UCB) type algorithms that rely on the construction of a specific confidence set for the unknown parameter. In contrast, Thompson Sampling (TS) maintains uncertainty about the unknown parameter in the form of a posterior distribution. The first TS algorithm under this setting was proposed by Agrawal and Goyal (2013) where they established a regret bound of the order $O(d^2\delta^{-1}\sqrt{T^{1+\delta}})$ for any $\delta \in (0, 1)$.

There is also a large body of work present in high-dimensional sparse linear contextual bandit setup, where the reward only depends on a small subset of features of the observed contexts. This area has recently attracted considerable attention due to its abundance in modern reinforcement learning applications (e.g, clinical trials, personalized recommendation systems, etc.) and has quite naturally spawned theoretical research in this direction. Some of the important references include Bastani and Bayati (2020); Wang et al. (2018); Hao et al. (2020); Chen et al. (2022); Ariu et al. (2022); Kim and Paik (2019); Li et al. (2022); Oh et al. (2021); Li et al. (2021) among others. A more detailed discussion on existing literature in the high dimensional bandit field is provided in Section 5.3.3. However, there has been very limited work dedicated to analyzing TS algorithms in high-dimensional sparse bandit setups. Hao et al. (2021) proposed a sparse information-directed sampling (IDS) algorithm which under a special case reduces to a TS algorithm based on a spike-and-slab Gaussian-Laplace

prior. However, the regret bound of IDS scales polynomially in d, which is sub-optimal in the high-dimensional regime. In related work, Gilton and Willett (2017) proposed a linear TS algorithm based on a relevance vector machine (RVM) which again suffers from sub-optimal dependence on d.

In this chapter, we specifically focus on the high-dimensional sparse linear contextual bandit (SLCB) setup and propose a TS algorithm based on a sparsity-inducing prior that enjoys almost dimension-independent regret bound. While TS algorithms have been known to empirically perform better than optimism-based algorithms (Chapelle and Li, 2011; Kaufmann et al., 2012), theoretical understanding of these is challenging due to the complex dependence structure of the bandit environment. Moreover, posterior sampling in high-dimensional regression (which is the crucial step for TS in high-dimensional SLCB), using MCMC, generally suffers from computational bottleneck. Our work overcomes all these challenges and makes the following contributions:

1. We use the sparsity inducing prior proposed in Castillo et al. (2015) for posterior sampling and establish posterior contraction result for *non-i.i.d. observations* coming from bandit environment and for a wide class of noise distributions.
2. Using the posterior contraction result, we establish an almost *dimension free* regret bound for our proposed TS algorithm under different arm-separation regimes parameterized by ω. The algorithm enjoys minimax optimal performance for $\omega \in [0, 1)$. To the best of our knowledge, this is the first work that proposes a novel TS algorithm with desirable regret guarantees in high-dimensional and sparse SLCB setup.
3. Our algorithm does *not* need the knowledge of model sparsity level, unlike other algorithms such as LASSO-bandit, MCP-bandit, ESTC, etc.
4. Finally, the prior allows us to design a computationally efficient TS algorithm based on Variational Bayes.

The rest of the chapter is organized as follows. In Section 5.2, we introduce the problem formally and discuss the assumptions and prior distribution. In Section 5.3, we present the crucial posterior contraction result and the main regret bound for our proposed algorithm. In Section 5.4, we discuss the challenges of drawing samples from the posterior in such problems and present a faster alternative relying on variational inference. In Section 5.5, we present simulation studies under different setups comparing our proposed method with existing algorithms. Detailed proofs of results and technical lemmas are deferred to the appendix.

5.2 Problem Formulation

We consider a linear stochastic contextual bandit with K arms. At time $t \in [T]$, context vectors $\{x_i(t)\}_{i\in[K]}$ are revealed for every arm i. We assume $x_i(t) \in \mathbb{R}^d$ for all $i \in [K], t \in [T]$ and for every i, $\{x_i(t)\}_{t\in[T]}$ are i.i.d. from some distribution $\mathcal{P}_i$. At every time step t, an action $a_t \in [K]$ is chosen by the learner and a reward $r(t)$ is generated according to the following linear model:

$$r(t) = x_{a_t}(t)^\top \beta^* + \epsilon(t) \tag{5.1}$$

where $\beta^* \in \mathbb{R}^d$ is the unknown true signal and $\{\epsilon(t)\}_{t\in[T]}$ are independent sub-Gaussian random noise, also independent of all the other processes. We assume that the true parameter β^* is s^*–sparse, i.e., $\|\beta^*\|_0 = s^*$. We denote by S^* the true support of β^*, i.e., $S^* = \{j : \beta_j^* \neq 0\}$.

The goal is to design a sequential decision-making policy π that maximizes the expected cumulative reward over the time horizon. To formalize the notion, we define the history $\mathcal{H}_t$ up to time t as follows:

$$\mathcal{H}_t := \left\{(a_\tau, r(\tau), \{x_i(\tau)\}_{i\in[K]}) : \tau \in [t]\right\},$$

and an admissible policy π generates a sequence of random variables $a_1, a_2, \ldots$ taking values in $[K]$ such that a_t is measurable with respect to the σ-algebra generated by the previous feature vectors from each arm, observed rewards of the chosen arms till the previous round and the current feature vectors, i.e., measurable with respect to the filtration $\mathcal{F}_t := \sigma\left(x_{a_\tau}(\tau), r(\tau), x_i(t); \tau \in [t-1], i \in [K]\right)$.

Thus, an algorithm for contextual bandits is a policy π, which at every round t, chooses an action (arm) a_t based on history $\mathcal{H}_{t-1}$ and current contexts. We note that although contexts of the previous round corresponding to arms that were not chosen are in $\mathcal{F}_t$, however, they do not provide useful information on the parameter, since we do not observe rewards corresponding to them under the bandit feedback, and hence are not included in the history $\mathcal{H}_t$. To measure the quality of performance, we compare it with the oracle policy π^* which uses the knowledge of the true β^* to choose the optimal action $a_t^* := \arg\max_{i\in[K]} x_i(t)^\top \beta^*$. Define $\Delta_i(t)$ to be the difference between the mean rewards of the optimal arm and ith arm at time t, i.e., $\Delta_i(t) = x_{a_t^*}(t)^\top \beta^* - x_i(t)^\top \beta^*$. Note that under the random-design assumption, a_t^* is also random. Then the regret at time t is defined as $\text{regret}(t) = \Delta_{a_t}(t)$ and the objective of the learner is to minimize the total regret till time T, defined as $R(T) = \sum_{t\in[T]} \text{regret}(t)$. We also define the matrix $X_t := (x_{a_1}(1), \ldots, x_{a_t}(t))^\top$. The time horizon T is finite but

possibly unknown, but much smaller compared to the ambient dimension of the parameter, i.e. $d \gg T$. Hao et al. (2021) refers to this regime as "data-poor" regime; such a regime adds an extra layer of hardness on top of the difficulty incurred by the sparse structure of β. We also assume that K is fixed and much smaller compared to both d and T.

5.2.1 Assumptions

In this section, we discuss the assumptions of our model.

Definition 5.2.1 (Sparse Riesz Condition (SRC))**.** *Let M be a $d \times d$ positive semi-definite matrix. The maximum and minimum sparse eigenvalues of M with parameter $s \in [d]$ are defined as follows:*

$$\phi_{\min}(s; M) := \inf_{\delta:\delta\neq 0, \|\delta\|_0 \leq s} \frac{\delta^\top M \delta}{\|\delta\|_2^2},$$
$$\phi_{\max}(s; M) := \sup_{\delta:\delta\neq 0, \|\delta\|_0 \leq s} \frac{\delta^\top M \delta}{\|\delta\|_2^2}.$$

We say M satisfies the SRC if $0 < \phi_{\min}(s, M) \leq \phi_{\max}(s; M) < \infty$.

Now we are ready to state the assumptions on the context distributions, which are as follows:

Assumption 5.2.2 (Assumptions on Context Distributions)**.** *We assume that*

(a) For some constant $x_{\mathsf{max}} \in \mathbb{R}^+$, we have that for all $i \in [K]$, $\mathcal{P}_i(\|x\|_\infty \leq x_{\mathsf{max}}) = 1$.

(b) For all arms $i \in [K]$, the distribution $\mathcal{P}_i$ is sub-Gaussian, i.e., there exists a constant $\vartheta > 0$ such that $\max_{i\in[K]} \|x_i(t)\|_{\psi_2} \leq \vartheta$ for all $t \in [T]$.

(c) There exists a constant $\xi \in \mathbb{R}^+$ such that for each $u \in \mathbb{S}^{d-1} \cap \{v \in \mathbb{R}^d : \|v\|_0 \leq Cs^*\}$ and $h \in \mathbb{R}^+$ $\mathcal{P}_i(\langle x, u\rangle^2 \leq h) \leq \xi h$, for all $i \in [K]$, ,where $C \in (2, \infty)$.

(d) The matrix $\Sigma_i := \mathbb{E}_{x\sim\mathcal{P}_i}[xx^\top]$ has bounded maximum sparse eigenvalue, i.e., $\phi_{\max}(Cs^*, \Sigma_i) \leq \phi_u < \infty$, for all $i \in [K]$, where C is the same constant as in part (c).

Assumption 5.2.2(a) basically tells that the contexts are bounded; such assumptions are standard in the bandit literature to obtain results on regret bound that are independent of the scaling of the contexts (or parameter). Assumption 5.2.2(b) says that all the arm-contexts are generated from sub-Gaussian distributions with a common bound of the order $O(\vartheta^2)$ on the proxy-variance, for all time point t. This is indeed a very mild assumption on the context distribution and a broad class of distributions enjoys such property. For

example, truncated multivariate normal distribution with covariance matrix $\mathbb{I}_d$, where the truncation is over the set $\{u \in \mathbb{R}^d : \|u\|_\infty \leq 1\}$ is a valid distribution for the contexts. In comparison, most of the previous literature such as Kim and Paik (2019); Oh et al. (2021); Li et al. (2022) assume that $\|x_i(t)\|_2 \leq L$, for some constant $L > 0$. This condition automatically implies that Assumption 5.2.2(b) holds with $\vartheta = (L/\log 2)^{1/2}$. As a result, the theory in Kim and Paik (2019); Oh et al. (2021); Li et al. (2022) can not accommodate the aforementioned truncated multivariate normal distribution as in this case $\|x_i(t)\|_2 = \Theta(\sqrt{d})$ and their analysis yields $O(\sqrt{d})$ dependence in the regret bound. Assumption 5.2.2(c) talks about anti-concentration condition that plays a critical role in controlling the estimation accuracy of β^*. This condition is also assumed by Li et al. (2021) and a variant (diverse covariates) of this condition is assumed in Ren and Zhou (2020). Intuitively, this condition prohibits the context features to fall along a singular direction. When u is not constrained to be sparse, this condition implies that the distribution of the contexts is not supported on a lower dimensional sub-space, allowing diversity and in the contexts and leading to inherent exploration. Assumption 5.2.2(c) captures this notion using the weaker condition where u is only sparse. Existing works like Oh et al. (2021); Kim and Paik (2019) try to capture this notion via compatibility/RE condition which is a somewhat stronger assumption and cannot be easily checked in practice. Assumption 5.2.2(c) is more interpretable - under the mere existence of bounded density for $u^\top x_i(t)$ (for all sparse u), the condition holds. Moreover, The entire Assumption 5.2.2 does not need any existence of pdf, whereas Oh et al. (2021); Ariu et al. (2022) need the relaxed symmetry assumption which requires the existence of pdf. Lastly, from the discussion in Section 2.3 of Ren and Zhou (2020), it follows that minimum sparse eigenvalue assumption and relaxed symmetry assumption (both used in Ariu et al. (2022)) implies diverse covariate property when $K = 2$. This suggests that Assumption 5.2.2(b) is very mild. Lastly, Assumption 5.2.2(d) imposes an upper bound on the maximum sparse eigenvalue of Σ_i which is a common assumption in high-dimensional literature (Zhang and Huang, 2008; Zhang, 2010).

Next, we come to the assumptions on the true parameter β^*

Assumption 5.2.3 (Assumptions on the true parameter)**.** *We assume the followings:*

(a) Sparsity and Soft-sparsity: There exist positive constants $s^* \in \mathbb{N}$ and $b_{\mathsf{max}} \in \mathbb{R}^+$ such that $\|\beta^*\|_0 = s^*$ and $\|\beta^*\|_1 \leq b_{\mathsf{max}}$.

(b) Margin condition: There exists positive constants Δ_*, A and $\omega \in [0, \infty]$, such that for $h \in \left[A\sqrt{\log(d)/T}, \Delta_*\right]$ and for all $t \in [T]$,

$$\mathbb{P}\left(x_{a_t^*}(t)^\top \beta^* \leq \max_{i \neq a_t^*} x_i(t)^\top \beta^* + h\right) \leq \left(\frac{h}{\Delta^*}\right)^\omega.$$

The first part of the assumption requires boundedness of the true parameter β^* to make the final regret bound scale free. Such an assumption is also standard in bandit literature (Bastani and Bayati, 2020; Abbasi-Yadkori et al., 2011).

The second part of the assumption imposes a margin condition on the arm distributions. Essentially, this assumption controls the probability of the optimal arm falling into h-neighborhood of the sub-optimal arms. As ω increases, the margin condition becomes stronger as the sub-optimal arms are less likely to fall close to the optimal arms. As a result, it becomes easier for any bandit policy to distinguish the optimal arm. As an illustration, consider the two extreme cases $\omega = \infty$ and $\omega = 0$. The $\omega = \infty$ case tells that there is a deterministic gap between rewards corresponding to the optimal arm and sub-optimal arms. This is the same as the "gap assumption" in Abbasi-Yadkori et al. (2011). Thus, quite evidently it is easy for any bandit policy to recognize the optimal arm. This phenomenon is reflected in the regret bound of Theorem 5 in Abbasi-Yadkori et al. (2011), where the regret depends on the time horizon T only though poly-logarithmic terms. In contrast, $\omega = 0$ corresponds to the case when there is no apriori information about the separation between the arms, and as a consequence, we pay the price in regret bound by a $\sqrt{T}$ term (Hao et al., 2021; Agrawal and Goyal, 2013; Chu et al., 2011).

The margin condition with $\omega = 1$ has been assumed in Goldenshluger and Zeevi (2013); Bastani and Bayati (2020); Wang et al. (2018) and will be satisfied when the density of $x_i(t)^\top \beta^*$ is uniformly bounded for all $i \in [K]$. Li et al. (2021) also discusses an example where the margin condition holds for different values of ω.

The final assumption is on the noise variables:

Assumption 5.2.4 (Assumption on Noise)**.** *We assume that the random variables $\{\epsilon(t)\}_{t\in[T]}$ are independent and also independent of the other processes and each one is $\sigma-$Sub-Gaussian, i.e., $\mathbb{E}[e^{a\epsilon(t)}] \leq e^{\sigma^2 a^2/2}$ for all $t \in [T]$ and $a \in \mathbb{R}$.*

Various families of distribution satisfy such a requirement, including normal distribution and bounded distributions, which are commonly chosen noise distributions. Note that such a requirement automatically implies that for every $t \in [T]$, $\mathbb{E}[\epsilon(t)] = 0$ and $\text{Var}[\epsilon(t)] \leq \sigma^2$.

5.2.2 Thompson Sampling and Prior

We discuss the basics of Thompson sampling and introduce the specific structure of the prior that we use and analyze. Typically, we place a prior Π on the unknown parameter (β in our case) along with a *specified likelihood model* on the data, and do the following: while taking action, we draw a sample from the posterior distribution of the parameter given the data and use that as the proxy for the unknown parameter value, hence in our case at time t,

we draw a sample $\hat{\beta}_t \sim \Pi(\beta|\mathcal{H}_{t-1})$ and choose $a_t = \arg\max_i x_i(t)^\top \hat{\beta}_t$ as the action. While simple enough to describe, Thompson sampling has been difficult to analyze theoretically, particularly because of the complex dependence between the observations due to the bandit structure. The choice of prior plays a crucial role, as we shall see, in providing the correct exploration-exploitation trade-off. In the high-dimensional sparse case that we are dealing with, this choice is specifically important since we do not wish to have a linear dependence on the dimension d in our regret bound - which would be incurred if we use the normal prior-likelihood setup of Agrawal and Goyal (2013), which analyzes Thompson sampling in contextual bandits.

While there is a rich literature on Bayesian priors for high dimensional regression, including horseshoe priors and slab-and-spike priors among others, we shall be using the complexity prior introduced in Castillo et al. (2015). Specifically, we consider a prior Π on β that first selects a dimension s from a prior π_d on the set $[d]$, next a random subset $S \subset [d]$ of size $|S| = s$ and finally, given S, a set of nonzero values $\beta_S := \{\beta_i : i \in S\}$ from a prior density g_S on $\mathbb{R}^S$. Formally, the prior on (S, β) can be written as

$$(S, \beta) \mapsto \pi_d(|S|) \frac{1}{\binom{d}{|S|}} g_S(\beta_S) \delta_0(\beta_{S^c}), \tag{5.2}$$

where the term $\delta_0(\beta_{S^c})$ refers to coordinates β_{S^c} being set to 0. Moreover, we choose g_S as a product of Laplace densities on $\mathbb{R}$ with parameter λ/σ, i.e., $\beta_i \mapsto (2\sigma)^{-1}\lambda \exp(-\lambda|\beta_i|/\sigma)$ for all $i \in S$. Note that, here we assume that the noise level σ is known. In practice, one can add another level of hierarchy by setting a prior on σ but in this chapter we do not pursue that direction.

The prior π_d plays the role of expressing the sparsity of the parameter. This is in contrast to other priors like product of independent Laplace densities over the coordinates (typically known as Bayesian LASSO), where the Laplace parameter plays the role of shrinking the coefficients towards 0. However, in our case, the scale parameter λ of the Laplace does not have this role and we assume that during the tth round we use $\lambda \in [(5/3)\overline{\lambda}_t, 2\overline{\lambda}_t]$, where $\overline{\lambda}_t \asymp \sqrt{t \log d}$, which is the usual order of the regularization parameter used in the LASSO.

The choice of the prior π_d is very critical; it should downweight big models, but at the same time give enough mass to the true model. Following Castillo et al. (2015), we assume that there are constants $A_1, A_2, A_3, A_4 > 0$ such that $\forall\, s \in [d]$

$$A_1 d^{-A_3} \pi_d(s-1) \leq \pi_d(s) \leq A_2 d^{-A_4} \pi_d(s-1). \tag{5.3}$$

Complexity priors of the form $\pi_d(s) \propto c^{-s} d^{-as}$ for constants a, c satisfy the above require-

ment. Moreover, slab and spike priors of the form $(1-r)\delta_0 + r\mathsf{Lap}(\lambda/\sigma)$ independently over the coordinates satisfy the requirement with hyperprior on r being $\mathsf{Beta}(1, d^u)$.

Finally, we specify the data likelihood that is crucial for the TS algorithm. At each time point $t \in [T]$, given the observations coming from model (5.1), we model the $\{\epsilon(\tau)\}_{\tau \le t}$ as i.i.d. $\mathsf{N}(0, \sigma^2)$. We emphasize that this Gaussian assumption is *only required for likelihood modeling* and our main results hold under any true error distribution satisfying Assumption 5.2.4. The same strategy is also used in Agrawal and Goyal (2013) for the LinTS algorithm in low-dimensional setting.

5.3 Main Results

5.3.1 Posterior contraction

Now, we present an informal version of the main posterior contraction result for the estimation of β. A more detailed version of the result with exact rates, along with the measure theoretic details, is in Appendix D.3.

Theorem 5.3.1 (Informal). *Write* $\mathbf{r}_t = (r(1), \ldots, r(t))^\top$, *and let the Assumption 5.2.2–5.2.4 hold with* $C = \Theta(\phi_u \vartheta^2 \xi K \log K)$, *and* $K \ge 2, d \ge T$. *With* $\lambda \asymp x_{\mathsf{max}}(t \log d)^{1/2}$ *and* $\varepsilon_{t,d,s} = s^*\{(\log d + \log t)/t\}\}^{1/2}$, *the following holds as* $t \to \infty$*:*

$$\mathbb{E}_{\mathbf{r}_t} \Pi \left(\|\beta - \beta^*\|_1 \gtrsim \sigma \varepsilon_{t,d,s} \,\middle|\, \mathbf{r}_t, X_t \right) \xrightarrow{a.s.} 0.$$

The above result is similar to Theorem 3 in Castillo et al. (2015) under classical linear regression setup with i.i.d. observations and Gaussian noise. However, we generalize their result under bandit setup and sub-Gaussian noise by carefully controlling the correlation between noise and observed contexts, which is crucial for our regret analysis.

5.3.2 Algorithm and regret bound

In this section, we introduce the Thomson sampling algorithm for high-dimensional contextual bandit, a pseudo-code for which is provided below in Algorithm 3. Similar to the Thompson sampling algorithm in Agrawal and Goyal (2013), in the tth round Algorithm 3 sets the a specific prior on β and updates it sequentially based on the observed rewards and contexts. In particular, it chooses the prior described in (5.2) with an appropriate choice of round-specific prior scaling λ_t and updates the posterior using the observed rewards and

contexts until $(t-1)$th round. Then a sample is generated from the posterior and an arm a_t is chosen greedily based on the generated sample.

Now, we show that the Thompson sampling algorithm achieves desirable regret upper bound.

In this section we introduce the Thomson sampling algorithm for high-dimensional contextual bandit, a pseudo-code for which is provided below in Algorithm 3. Similar to the Thompson sampling algorithm in Agrawal and Goyal (2013), in the tth round Algorithm 3 sets the a specific prior on β and updates it sequentially based on the observed rewards and contexts. In particular, it chooses the prior described in (5.2) with an appropriate choice of round-specific prior scaling λ_t and updates the posterior using the observed rewards and contexts until $(t-1)$th round. Then a sample is generated from the posterior and an arm a_t is chosen greedily based on the generated sample.

Now, we show that the Thompson sampling algorithm achieves desirable regret upper bound.

Theorem 5.3.2. *Let the Assumption 5.2.2–5.2.4 hold with $C = \Theta(\phi_u \vartheta^2 \xi K \log K)$, and $K \geq 2, d \geq T$. Define the quantity $\kappa(\xi, \vartheta, K) := \min\{(4c_3 K \xi \vartheta^2)^{-1}, 1/2\}$ where c_3 is a universal positive constant. Also, set the prior scaling λ_t as follows:*

$$(5/3)\overline{\lambda}_t \leq \lambda_t \leq 2\overline{\lambda}_t, \quad \overline{\lambda}_t = x_{\mathsf{max}}\sqrt{2t(\log d + \log t)}.$$

Then there exists a universal constant $C_0 > 0$ such that we have the following regret bound for Algorithm 3:

$$\mathbb{E}\{R(T)\} \lesssim I_b + I_\omega,$$

where,

$$I_b = \left\{ \frac{b_{\mathsf{max}} x_{\mathsf{max}} \phi_u \vartheta^2 \xi (K \log K)}{\min\{\kappa^2(\xi, \vartheta, K), \log K\}} \right\} s^* \log(Kd),$$

$$I_\omega = \begin{cases} \Phi^{1+\omega} \left(\frac{s^{*1+\omega} (\log d)^{\frac{1+\omega}{2}} T^{\frac{1-\omega}{2}}}{\Delta_*^\omega} \right), & \text{for } \omega \in [0, 1), \\ \Phi^2 \left(\frac{s^{*2} [\log d + \log T] \log T}{\Delta_*} \right), & \text{for } \omega = 1, \\ \frac{\Phi^2}{(\omega - 1)} \left(\frac{s^{*2} [\log d + \log T]}{\Delta_*} \right), & \text{for } \omega \in (1, \infty) \\ \Phi^2 \left(\frac{s^{*2} [\log d + \log T]}{\Delta_*} \right), & \text{for } \omega = \infty, \end{cases}$$

and $\Phi = \sigma x_{\mathsf{max}}^2 \xi K \left(2 + 40 A_4^{-1} + C_0 K \xi x_{\mathsf{max}}^2 A_4^{-1}\right)$.

Discussion on the above result: The regret bound provided by Theorem 5.3.2 shows that the regret of the algorithm grows poly-logarithmically in d, i.e., $\mathbb{E}\{R(T)\} = O((\log d)^{\frac{1+\omega}{2}})$, when $\omega \in [0, 1)$; logarithmically in d, i.e., $O(\log d)$ when $\omega \in [1, \infty]$. Meanwhile, the expected cumulative regret depends polynomially in T, i.e., $\mathbb{E}\{R(T)\} = O(T^{\frac{1-\omega}{2}})$ when $\omega \in [0, 1)$; ploy-logarithmically in T; i.e., $\mathbb{E}\{R(T)\} = O((\log T)^2)$, when $\omega = 1$. In $\omega \in (1, \infty]$ regime, the expected cumulative regret depends poly-logarithmically in both the time horizon T and ambient dimension d. As $T \ll d$, the expected regret ultimately scales as $O(\log d)$. Comparing our upper bound result with minimax regret lower bound established in Theorem 1 of Li et al. (2021), it follows that our algorithm enjoys optimal dependence on both ambient dimension d and time-horizon T when $\omega \in [0, 1)$. In $\omega = 1$ region, the regret upper bound in the above theorem is optimal up to a $O(\log T)$ term. To the best of our knowledge, there does not exist any result on minimax lower bound in the regime $\omega > 1$ in the high-dimensional linear contextual bandit literature. It is worth mentioning that this is an upper bound on the expected (frequentist) regret, as compared to Bayesian regret which is often considered for Thompson sampling based algorithms.

Algorithm 3: TS algorithm

$\mathcal{H}_0 = \varnothing$.
for $t = 1, \cdots, T$ **do**
  if $t \leq 1$ **then**
    Choose action a_t uniformly over $[K]$.
  end if
  if $t > 1$ **then**
    Set $\overline{\lambda}_t = x_{\max}\sqrt{2t(\log d + \log t)}$ and choose $\lambda_t \in (5\overline{\lambda}_t/3, 2\overline{\lambda}_t)$.
    Generate sample $\tilde{\beta}_t \sim \Pi(\cdot \mid \mathcal{H}_{t-1})$ with prior Π in (5.2)-(5.3), $\lambda = \lambda_t$ and Gaussian likelihood.
    Play arm: $a_t = \arg\max_{i \in [K]} x_i(t)^\top \tilde{\beta}_t$.
  end if
  Observe reward $r(t)$.
  Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(a_t, r_{a_t}(t), x_{a_t}(t))\}$.
end for

Intuitively, the initial term I_b in regret upper bound in Theorem 5.3.2 describes the regret caused by the "burn-in" period of exploring the space of contexts and it does not contribute to the asymptotic regret growth. Note that we consider running Thompson sampling from the very beginning, without an explicit random exploration phase, in contrast to most of the existing algorithms; the distinction between the burn-in phase and the subsequent phase is only a construct of our theoretical analysis. Furthermore, the constant Δ_* plays the role of gap parameter which commonly appears in a problem-dependent regret bound (Abbasi-Yadkori et al., 2011). Note that, for $\omega = 0$, we get a problem-independent regret bound of

the order $O(s^*\sqrt{T}\log d)$. The appearance of the $\sqrt{T}$, term is not surprising, as the condition $\omega = 0$ poses no prior knowledge on the arm-separability, Thus, in the worst case, the context vectors may fall into each other, making the bandit environment harder to learn. In contrast, as ω increases the optimal arm becomes more distinguishable than the sub-optimal arms and the bandit environment becomes easier to learn. As a result, the effect of the time horizon becomes less and less severe as ω increases. In particular, when $\omega \in [1, \infty]$, the time horizon does not affect the asymptotic growth of the regret bound. Finally, as we mainly focus on the case when the number of arms is very small, the quantity Φ roughly has an inflating effect of $O(1)$ on the regret bound.

Sketch of the proof of Theorem 5.3.2: While a self-contained and detailed proof of the above result is given in the Appendix, here we go through the main steps and ideas of the proof. The proof is broadly divided into 3 parts for clarity:

(i) In Section D.2.1 we will first show that the estimated covariance matrix $\widehat{\Sigma}_t := X_t^\top X_t/t$ enjoys SRC condition with high probability for sufficiently large t. In our analysis, we carefully decouple this complex dependent structure and exploit the special temporal dependence structure of the bandit environment to establish SRC property of $\widehat{\Sigma}_t$.

(ii) Next, in Section D.2.2 we will establish a compatibility condition for the matrix $\widehat{\Sigma}_t$. We use a Transfer Lemma (Lemma D.2.9) which essentially translates the uniform lower bound on $\phi_{\min}(Cs^*, \widehat{\Sigma}_t)$ to a certain compatibility number.

(iii) Finally, in Section D.2.3, under the compatibility condition we use the posterior contraction result in Theorem 5.3.1 to give bound on the per round regret $\Delta_{a_t}(t)$.

5.3.3 Comparison with existing literature

Over the past few years, the problem of high dimensional stochastic linear contextual bandit has attracted significant attention and has quite evidently generated a large body of work in this field under different problem settings. However, there are mainly two types of settings that have been considered in high-dimensional linear bandit literature: **(S1)** Each arm has different parameters β_i^* for $i \in [K]$ and only one context vector $x(t)$ is generated at every time point t, **(S2)** K different contexts $x_i(t)$ are generated for each arm $i \in [K]$ at every time point t and all of the arms have one common parameter β^*, which is also the setting of this chapter.

There has been an ample amount of work in both of these settings. To mention a few, Wang et al. (2018); Bastani and Bayati (2015); Wang and Cheng (2020) consider the setting in **(S1)**, whereas Kim and Paik (2019); Li et al. (2021); Oh et al. (2021) consider the setting

Table 5.1: This table compares the regret bounds and working assumptions of this chapter with existing works under different SLCB settings. We focus four most important assumptions: (1) 'Margin' - similar to Assumption 5.2.2(b) with $\omega \in \{0, 1\}$, (2)'Comp/RE'- Compatibility or RE condition, (3) 'ℓ_2-bound'- boundedness of contexts in ℓ_2-norm, (4) 'Pdf exst'- existence of pdf. ✓ symbol indicates that the corresponding condition is assumed in the chapter. ✓$(\star)$ symbol indicates that Chen et al. (2022) assumes that the coordinates of the contexts are i.i.d and the second moments are lower bounded, which is typically much stronger than compatibility or RE condition.

Setting	Paper	Regret Bound	Margin	Comp/RE	ℓ_2-bound	Pdf exists
(S1)	Bastani and Bayati (2015)	$O(s^{*2}[\log d + \log T]^2)$	✓	✓		
	Wang et al. (2018)	$O(s^{*2}[\log d + \log T] \log T)$	✓	✓		
	Wang and Cheng (2020)	$O(s^{*2}[\log d + \log T]^2)$	✓	✓		
(S2)	Kim and Paik (2019)	$O(s^*\sqrt{T}\log(dT))$		✓	✓	
	Oh et al. (2021)	$O(\sqrt{s^*T\log(dT)})$		✓	✓	✓
	Li et al. (2021)	$O(s^*\sqrt{T\log d})$				
		$O(s^{*2}[\log d + \log T]\log T)$	✓			
	Li et al. (2022)	$O(s^{*1/3}T^{2/3}\sqrt{\log(dT)})$		✓	✓	
	Ariu et al. (2022)	$O(s^{*2}\log d + \sqrt{s^*T})$		✓		✓
		$O(s^{*2}\log d + s^*\log T)$	✓	✓		✓
	Chen et al. (2022)	$O(s^*\sqrt{T}\log^2(Td))$		✓$(\star)$		
	This chapter	$O(s^*\sqrt{T\log d})$				
		$O(s^{*2}[\log d + \log T]\log T)$	✓			

in **(S2)**. It is worth mentioning that most of these works assume very strong compatibility or restricted eigenvalue (RE) conditions on the feature distribution, which is in general hard to check in real-world applications. Instead, in this chapter we show that TS algorithm enjoys desirable regret bound under much weaker and easily interpretable assumptions on the feature distribution. There is also a parallel line of work that considers the set of features or contexts to be infinite but fixed (Hao et al., 2020, 2021; Jang et al., 2022), which is in sharp

contrast to the setting considered in this chapter. Moreover, in the setup of these works, the optimal arm remains the same for every round. On the other hand, in our setting, due to the randomness of the observed contexts, the optimal arm does not necessarily remain the same in every round. Lastly, Hao et al. (2021) and Hao et al. (2022) study the properties of information-directed sampling and provide a guarantee for Bayesian regret, which is much weaker than the result in Theorem 5.3.2.

Now we compare the results and assumptions of this chapter with existing literature in SLCB setting. Table 5.1 shows the comparison of regret bound and working assumptions of different papers under both **(S1)** and **(S2)** settings. Under the setup of **(S1)**, Bastani and Bayati (2015) and Wang et al. (2018) proposed the LASSO-bandit algorithm and MCP bandit algorithm respectively, and under the margin condition $\omega = 1$, they established a regret bound of the order $\widetilde{O}((s^* \log T)^2)$[2]. However, Theorem 5.3.2 accommodates a broader range of ω and our method does not need the knowledge of ω, but enjoys the same regret upper bound for $\omega = 1$. Moreover, unlike LASSO-bandit or MCP-bandit, our method does not require forced sampling which could be expensive in certain marketing applications.

Under the setting in **(S2)**, Kim and Paik (2019) and Ariu et al. (2022) proposed Doubly-robust LASSO and Threshold LASSO bandit algorithms respectively. Under strong compatibility or RE condition they established $\widetilde{O}(\sqrt{T})$ regret bound. With margin condition $\omega = 1$, Ariu et al. (2022) improved their bound to $\widetilde{O}(\log T)$. Li et al. (2022) proposed "Explore Structure then Commit" framework and established regret bound of the order $\widetilde{O}(T^{2/3})$. However, all of these algorithms require the knowledge of true sparsity s^*, and as mentioned before, also need some type of compatibility/RE conditions or some other strong conditions on the covariance structure and density functions of the contexts. In comparison, our theory does not assume any strong compatibility/RE condition on the context distribution and the TS algorithm also does not require the knowledge of true sparsity but still enjoys better or comparable regret bound. In some recent works, Oh et al. (2021); Li et al. (2021) also proposed LASSO-based algorithms which do not require the knowledge of true sparsity s^*. It is worth mentioning that under a similar set of assumptions as in this chapter, Li et al. (2021) showed that the LASSO-L1 confidence ball algorithm enjoys similar regret bounds as in Theorem 5.3.2 for different values of ω. However, the Sparsity Agnostic LASSO algorithm proposed in Oh et al. (2021) needs strong RE and balanced covariance assumptions. Lastly, Chen et al. (2022) recently proposed Sparse LinUCB and SupLinUCB algorithm which relies on best subset selection and showed that it enjoys $\widetilde{O}(\sqrt{T})$ regret bound. However, they assume that the contexts are sub-Gaussian with independent coordinates, which is far stronger than the compatibility condition and even unrealistic in most real-world applications.

[2] $\widetilde{O}(\cdot)$ hides the logarithmic dependence on d or T.

5.4 Computation

In this section, we discuss the computational challenges and how these are overcome by using Variational Bayes (VB). While priors as (5.2) have been shown to perform well, both empirically and in theory, the discrete model selection component of the prior makes it challenging to allow computation and inference on the posterior. For $\beta \in \mathbb{R}^d$, inference using the slab and spike prior requires a combinatorial search over 2^d possible models, which in the case of high dimension is computationally infeasible. Fast algorithms are known only in the special diagonal design case and traditional Markov Chain Monte Carlo methods have very slow mixing in such high dimensional cases. Thus, following Ray and Szabó (2021) we use Variational Bayes to make computations faster. Specifically, in the sampling step of Algorithm 3, we consider the VB approximation of the posterior $\Pi(\cdot \mid \mathcal{H}_{t-1})$ arising from slab and spike prior with $\mathsf{Lap}(\lambda/\sigma)$ slab in the mean-field family

$$\left\{ \bigotimes_{j=1}^{d} [\gamma_j \mathsf{N}(\mu_j, \sigma_j^2) + (1-\gamma_j)\delta_0] : (\mu_j, \sigma_j, \gamma_j) \in \mathcal{R} \right\},$$

where $\mathcal{R} = \mathbb{R} \times \mathbb{R}^+ \times [0,1]$. We use the `sparsevb` package (Clara et al., 2021) in R to use the Coordinate Ascent Variational Inference (CAVI) algorithm proposed in Ray and Szabó (2021) to obtain the VB posterior. This makes the Thompson sampling algorithm much faster as one can efficiently obtain samples from the VB posterior due to its structure. The details of the algorithm for the Variational Bayes Thompson Sampling (VBTS)[3] are in the appendix (see Section D.5).

5.5 Numerical Experiments

In both simulations and real data experiments, we present results corresponding to $\lambda_t = 1$ for all $t \in [T]$. Recall that Theorem 5.3.2 suggests that in tth round $\lambda_t \asymp \sqrt{t \log d}$ is a reasonable choice for the exact Thompson sampling algorithm. However, in practice, we noticed that such choices of λ_t lead to numerical instability. Some recent findings in Ray and Szabó (2021) suggest that λ_t in the order of $O(\sqrt{t \log d}/s^*)$ should be an appropriate choice, which is smaller than the predicted order of λ_t in our main theorem. Motivated by this, we also present the simulation results for synthetic data experiments with $\lambda_t = \lambda_* \sqrt{t}$ for $\lambda_* \in \{0.2, 0.3, 0.4, 0.5\}$ in Section 5.5.3. We found the performance of VBTS to be robust with respect to the choice of the tuning parameter λ_t.

[3]Codes are available online: `Github link`.

5.5.1 Synthetic data

In this section, we illustrate the performance of the VBTS algorithm on a simulated data set. As a benchmark, we consider, DR-LASSO (Kim and Paik, 2019), LASSO-L1 confidence ball algorithm (Li et al., 2021), ESTC (Li et al., 2022), sparsity agnostic (SA) LASSO (Oh et al., 2021), thresholded (TH) LASSO (Ariu et al., 2022), and TS algorithm based on Bayesian LASSO (Park and Casella, 2008) (BLASSO TS) to compare the performance of VBTS (Algorithm **??**). In this section, we only include the methods that are designed for the high-dimensional linear contextual bandit. Simulation results for LinUCB (Abbasi-Yadkori et al., 2011) and LinTS (Agrawal and Goyal, 2013) can be found in Appendix D.1.

Equicorrelated (EC) structure

We set the number of arms $K = 10$ and we generate the context vectors $\{x_i(t)\}_{i=1}^K$ from multivariate d-dimensional Gaussian distribution $\mathsf{N}_d(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j| \wedge 1}$ and $\rho = 0.3$. We consider $d = 1000$ and the sparsity $s^* = 5$. We choose the set of active indices S^* uniformly over all the subsets of $[d]$ of size s^*. Next, for each choice of d, we consider two types generating scheme for β:

- **Setup 1:** $\{U_i\}_{i \in S^*} \overset{i.i.d.}{\sim} \text{Uniform}(0.3, 1)$ and set $\beta_j = U_j (\sum_{\ell \in S^*} U_\ell^2)^{-1/2} \mathbb{1}(j \in S^*)$.
- **Setup 2:** $\{Z_i\}_{i \in S^*} \overset{i.i.d.}{\sim} \mathsf{N}(0, 1)$ and set $\beta_j = Z_j (\sum_{\ell \in S^*} Z_\ell^2)^{-1/2} \mathbb{1}(j \in S^*)$.

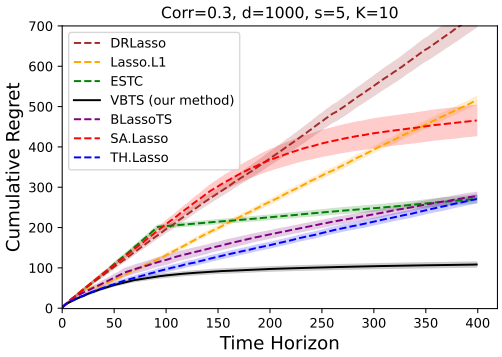
We run 40 independent simulations and plot the mean cumulative regret with 95% confidence band in Figure 5.1a-5.1b. In all the setups, we see that VBTS outperforms its competitors by a wide margin.

Autoregressive (AR) structure

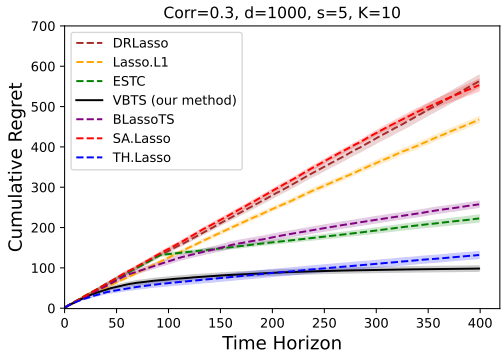
We consider the same setups as in the EC structure above, with the exception of the context distribution. Here we generate the context vectors $\{x_i(t)\}_{i=1}^K$ from $\mathsf{N}_d(\mathbf{0}, \Sigma)$ where $\Sigma_{ij} = \phi^{|i-j|}$ and $\phi = 0.3$. VBTS also enjoys superior empirical performance under this setup (see Figure 5.1c-5.1d). Table 5.2 shows the mean execution time (across Setup 1 and 2) of all TS algorithms for both EC and AR structure simulations. Among the class of TS algorithms, VBTS outperforms its other competing algorithms.

5.5.2 Real data - `gravier` Breast Carcinoma Data

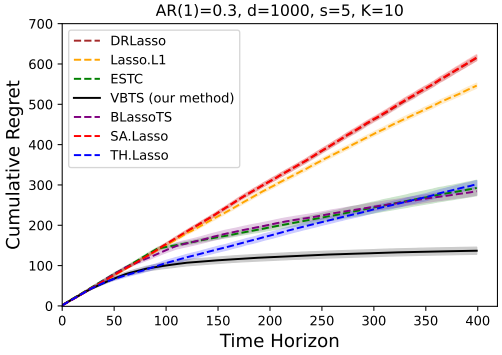
We consider breast cancer data `gravier` (`microarray` package in `R`) for 168 patients to predict metastasis of breast carcinoma based on 2905 gene expressions (bacterial artificial



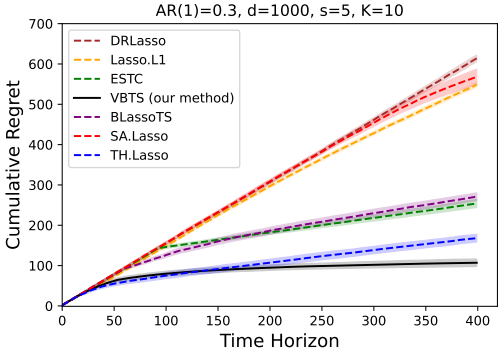
(a) EC (Setup 1)



(b) EC (Setup 2)



(c) AR (Setup 1)



(d) AR (Setup 2)

Figure 5.1: Cumulative regret of competing algorithms.

chromosome or BAC array). (Gravier et al., 2010) considered small, invasive ductal carcinomas without axillary lymph node involvement (T1T2N0) to predict metastasis of small node-negative breast carcinoma. Using comparative genomic hybridization arrays, they examined 168 patients over a five-year period. The 111 patients with no event after diagnosis were labeled good (class 0), and the 57 patients with early metastasis were labeled poor (class 1). The 2905 gene expression levels were normalized with a $\log_2$ transformation.

Similar to Kuzborskij et al. (2019); Chen et al. (2021), in our experimental setup we convert the breast cancer classification problem into 2-armed contextual bandit problem as follows: Given the `gravier` data set with 2 classes, we first set Class 1 as the target class. In each round, the environment randomly draws one sample from each class and composes a set of contexts of 2 samples. The learner chooses one sample and observes the reward following a logit model. In particular, we model the reward as

$$r(t) := \log\left\{\frac{\mathbb{P}(\text{Selected class} = 1)}{\mathbb{P}(\text{Selected class} = 0)}\right\} = x_{a_t}^\top \beta^* + \epsilon(t),$$

Table 5.2: Time comparison among the competing algorithms.

Type	Algorithm	Mean time of execution (seconds)	
		Equicorrelated	Auto-regressive
TS	LinTS	1344.39	1346.46
	BLASSO TS	1511.68	1455.53
	VBTS	**29.33**	**27.65**

where $a_t \in \{1, 2\}$ is the selected arm at round t. Thus, small cumulative regret insinuates that the learner is able to differentiate the positive patients eventually. Such concepts can be used for constructing online classifiers to differentiate carcinoma metastasis from healthy patients based on gene expression data. However, in practice, we can not measure the regret defined in (5.2), unless we have the knowledge of β^*. To resolve this issue, we first fit a logit model on the whole `gravier` data set and consider the estimated $\hat{\beta}$ as the ground truth and report the expected regret with respect to the estimated β. As reported in Gravier et al. (2010), 24 (out of 2905) BACs showed statistically significant difference (comparing Cy3/Cy5 values) between the two groups, motivating the use of a sparse logit model in our case. The estimated β^* in our sparse logistic model on the dataset had a sparsity of 18 using the dataset. In addition to this, we also treat the estimated noise variance from the fitted logit model as the true noise variance of the error induced by the environment in each round.

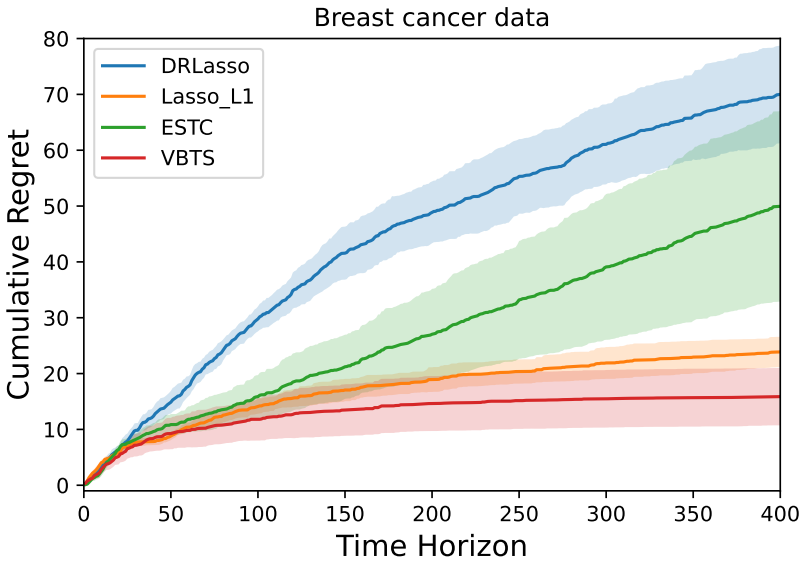


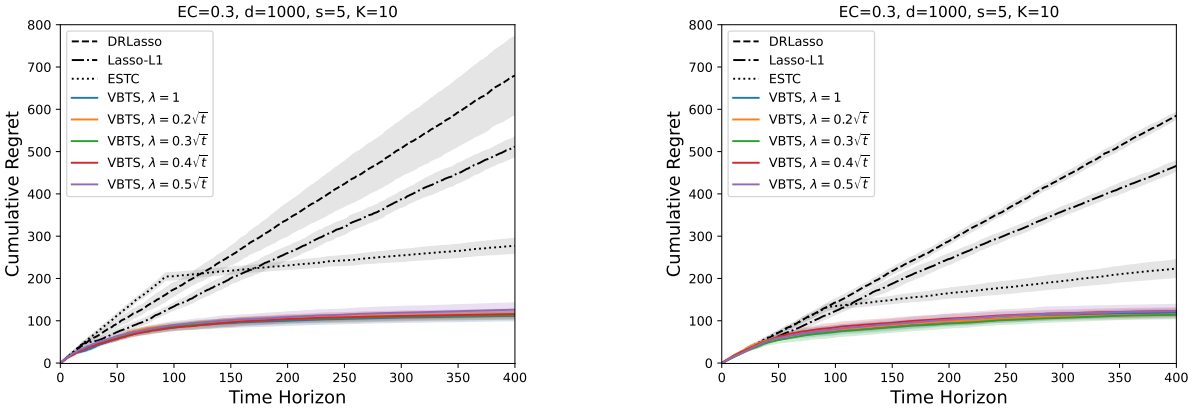
Figure 5.2: Cumulative regret plot for breast cancer data set.

Table 5.3: Classification accuracy of competing algorithms.

Algorithm	DR-LASSO	LASSO-L1	ESTC	**VBTS**
Accuracy(%)	65.63	81.20	73.32	**81.88**

5.5.3 Siumlation for different choices of λ

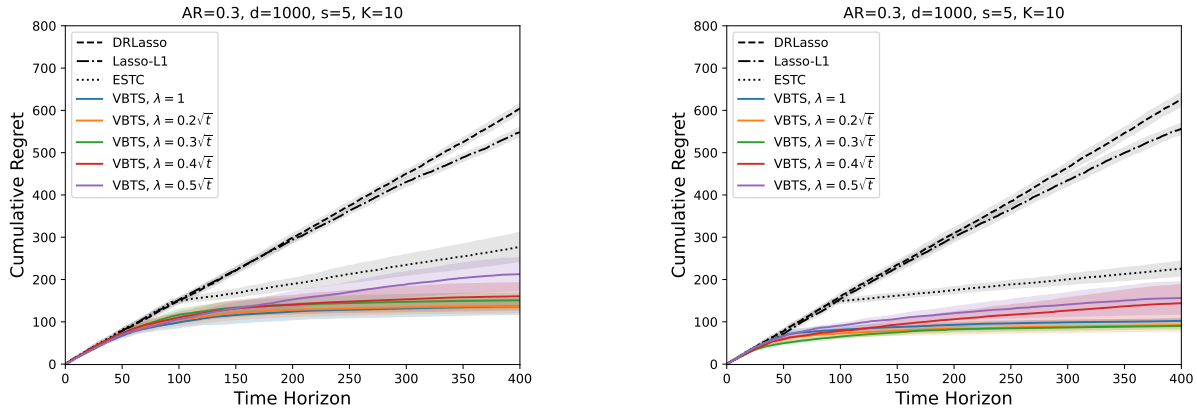
As discussed in the first paragraph of Section 5.5, for each of these simulation settings, we tried a few choices for the tuning parameter λ_t. In addition to the default choice of $\lambda_t = 1$ (for all time points t), we also explored the performance of the algorithm under growing λ, as required by our theoretical results. In particular, we tried $\lambda_t = \lambda_* \sqrt{t}$ for $\lambda_* \in \{0.2, 0.3, 0.4, 0.5\}$. For comparison, we only kept the faster optimism based methods DRLasso, Lasso-L1 and ESTC. We found the results to be roughly robust to the choice of this tuning parameter. The results are summarized in Figure 5.3 and Figure 5.4 below. However, we found that larger values of λ_* lead to numerical issues, we conjecture that this is an artifact of the variational Bayes approximation, rather than the prior itself. For our simulation settings, the choice $\sqrt{t \log d}/s^* \approx 0.5\sqrt{t}$ and hence by the findings in Ray and Szabó (2021), values of λ_* higher than this may yield inaccurate Variational Bayes estimation.



(a) Setup 1

(b) Setup 2

Figure 5.3: Regret bound for equi-correlated design for different tuning parameter choices



(a) Setup 1

(b) Setup 2

Figure 5.4: Regret bound for auto-regressive design for different tuning parameter choices

5.6 Conclusion

In this chapter, we consider the stochastic linear contextual bandit problem with high-dimensional sparse features and a fixed number of arms. We propose a Thompson sampling algorithm for this problem by placing a suitable *sparsity-inducing* prior on the unknown parameter to induce sparsity. We also develop a crucial posterior contraction result for *non-i.i.d.* data that allows us to obtain an almost *dimension independent* regret bound for our proposed algorithm. We explicitly point out the dependences on d and T for different arm-separation regimes parameterized by ω, which is also minimax optimal for $\omega \in [0, 1)$. Moreover, the choice of prior allows us to devise a Variational Bayes algorithm that enjoys computational expediency over traditional MCMC. We demonstrate the superior performance of our algorithm through extensive simulation studies. We finally perform an experiment on the `gravier` dataset, for which our method performs better compared to other existing algorithms.

Now we point the readers toward some of the natural research directions that we plan to cover in our future works. The regret analysis, similar to most of the recent works in high dimensional contextual bandits, relies on upper bounding the regret through estimation of the parameter, i.e., we rely on the estimation of β^* to be able to provide meaningful regret bound. However, this should not be required - as an example, consider the case where the first coordinate of $x_i(t)$ is 0 for all $i \in [K], t \in [T]$. Then the first coordinate of β^* is not estimable, however, this does not pose any problem to designing a sensible policy since this coordinate does not appear in the regret. Unfortunately, Assumption 5.2.2(c) is not satisfied for such degeneracy in the contexts and as a result, it would require a modified analysis of the regret

bound. Secondly, we underscore the fact that in our setup we adopt the Variational Bayes framework only to sidestep the computational hurdles of MCMC arising from a myriad of challenges such as slow mixing times of the chains, lack of easy implementation, etc. However, in high-dimensional regression setup Yang et al. (2016) has proposed Metropolis-Hastings algorithms based on truncated sparsity priors that do not meet the above roadblocks. It could be very well possible that some other prior structure will allow us to design more efficient MCMC algorithms with faster mixing times in the high-dimensional SLCB setup along with theoretical guarantees.

CHAPTER 6

Summary and Future Works

In this thesis, we address the modern emerging problems at the intersection of contemporary ML and high-dimensional statistics. In particular, we studied the problem of feature selection both from statistical and computational perspectives. In fact, we propose computationally efficient methods that enjoy the same statistical properties as BSS under the weak signal regime. Moreover, we provide a novel theoretical perspective on the model selection property of BSS that unravels the effect of the intrinsic topological structure of the design. In addition, we theoretically study the model selection performance of BSS under the setting of differential privacy. More importantly, we proposed a computationally efficient Metropolis-Hastings method that enjoys polynomial mixing time to its stationary distribution and produces a DP model estimator with a high utility guarantee. Therefore, our algorithm enjoys the best of three worlds: privacy, utility, and computation. Finally, we also propose a novel Thompson sampling algorithm for high-dimensional sparse linear contextual bandit that has regret scaling poly-logarithmically with the ambient dimension.

Future research vision and goals

As a researcher in the big-data era, my long-term vision is to promote the practice of "open science" (Vicente-Saez and Martinez-Fuentes, 2018) across diverse fields of ML and make reproducible scientific findings accessible to broader communities through relevant research. However, there are several pressing challenges: (i) the recent data explosion has necessitated statistical learning under distributed frameworks where large-scale data are distributed across several or even a large number of sites, restricting seamless communication between data sites, (ii) the paucity of scalable algorithms in various ML domains that can process and analyze complex data to produce reproducible results without excessive time and resource consumption, and (iii) alongside the escalation in data scale, there is a continuous accumulation of sensitive information in diverse formats, which also restricts data sharing across different studies. Hence, the need to construct scalable and privacy-preserving algorithmic

frameworks has become imperative to encourage data sharing in the landscape of modern AI and domain science applications. To address these emerging concerns, I plan to focus my future research at the juncture of the following areas:

Objective 1: Private mixed integer programming and large scale data sharing
Integer programming (IP) and mixed integer programming (MIP) are mathematical optimization problems where all or some of the decision variables are restricted to be integers. IPs and MIPs are critical for solving discrete and combinatorial optimization problems in various applications, including production planning, scheduling, and vehicle routing (Pochet and Wolsey, 2006; Malandraki and Daskin, 1992). In many of these applications, the privacy protection of the user is of utmost interest (Atmaca et al., 2021; Tong et al., 2017). However, currently, there does not exist any user-friendly private algorithmic framework that can leverage the computational power of commercial MIP solvers such as GUROBI, CPLEX, and MOSEK. This is a huge bottleneck in industrial-level applications where efficient and accurate algorithms are important requirements. To address this issue, I plan to do research related to developing DP frameworks that would provide seamless integration of modern MIP solvers to facilitate efficient and high-quality solutions to modern problems. For example, an immediate relevant statistical application could be to obtain an approximate DP solution to the BSS problem using the MIP formulation introduced in Bertsimas et al. (2016). This methodology could be used in statistical genetics to make quick and meaningful discoveries that can be shared with other studies to advance the scientific understanding of the field while maintaining user privacy.

Objective 2: Designing robust algorithm for reproducible results
Modern AI algorithms hold considerable promise in various ML domains encompassing healthcare, computer vision, and engineering. However, their dissemination and adoption have been slow, owing partially to unpredictable AI model performance once deployed in real-world applications such as digital healthcare, that involve extremely high-dimensional data. A recent study in Berisha et al. (2021) points out the "dataset blind spot", a phenomenon due to the curse of dimensionality, as the potential reason behind the failure of the AI algorithms - algorithms that achieve high performance in their training phases suffers much higher error rates when deployed for use. Similar issues regarding false discoveries have been pointed out in Dwork et al. (2015a,b), and the authors propose a robust algorithm based on perturbation that leverages the idea of DP to learn about the training data as a whole, i.e., the data distribution instead of individual entries. I plan to propose a similar approach for general high-dimensional analysis that will facilitate a robust and scalable algorithmic framework. In particular, I plan to use the privatized version of the low-dimensional

sufficient statistics for the training phase which needs relatively less amount of noise. This will allow us to produce reproducible and reliable results in a computationally efficient way which is an important requirement in open science.

Objective 3: Distributed learning in neuroimaging using summary statistics
Neuroimaging data are often complex and high-dimensional, and they are typically distributed across different sites or hospitals. Therefore, neuroimaging analysis require advanced ML methods to enhance precision, facilitating detection of subtle brain changes critical for understanding neurological disorders. However, there are two conflicting challenges: on the one hand, increasing sample size is necessary for neuroimaging research because its effect size is known to be small (Marek et al., 2022), and it is becoming more common to integrate neuroimaging studies from multiple sites; and on the other hand, there are privacy related challenges (White et al., 2022) that need to be addressed before such an integration, which causes unwanted delays in research and hinders reproducibility of scientific studies. Therefore, it is important to build AI-assisted statistical framework for promoting "open science" in neuroimaging research, enabling the integration of multimodal data and fostering a more comprehensive understanding of the brain. To achieve this goal, I intend to put forth summary-statistics methodologies rooted in high-dimensional ML techniques that inherently safeguard data privacy, as third parties are not granted access to sensitive neuroimaging data. Consequently, this approach will facilitate the aggregation of multiple summary statistics within the distributed framework, thereby amplifying the statistical power of the methodology for detecting subtle changes in the brain structure with better sample efficiency under more economic budget constraints. In particular, I will propose summary statistics akin to polygenic risk scores, as seen in the field of statistical genetics, that can be aggregated under the distributed framework to accurately forecast the likelihood of brain disorders using high-dimensional multivariate ML methodologies.

APPENDIX A

Appendix for Chapter 2

A.1 Proof of Theorem 2.3.1

Consider a MS procedure $\widehat{\mathcal{S}}_\tau \in \mathcal{T}$. Now, there are mainly two steps of the proof:

1. Upper bound the probability of recovery in terms of the probability of an event depending only on $\max_{j\in\mathcal{S}^c}|\mu_j|$ and $\min_{j\in\mathcal{S}}|\mu_j|$.

2. Find the asymptotic limits of the above random variables and find the limiting probability of the aforementioned event.

To start with note that

$$\mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}}_\tau = \mathcal{S}_{\boldsymbol{\beta}}) = \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{j\in\mathcal{S}^c}|\mu_j| < \tau(\mathbf{X},\mathbf{y}) \le \min_{j\in\mathcal{S}}|\mu_j|\right) \le \mathbb{P}\left(\max_{j\in\mathcal{S}^c}|\mu_j| < \min_{j\in\mathcal{S}}|\mu_j|\right).$$

Recall that for for $j \in [p]$ we have

$$\mu_j = \begin{cases} \beta_j \|\mathbf{X}_j\|_2^2/n + \omega_j \|\mathbf{X}_j\|_2 g_j/n & \text{if } j \in \mathcal{S}_{\boldsymbol{\beta}} \\ \omega_j \|\mathbf{X}_j\|_2 g_j/n & \text{if } j \notin \mathcal{S}_{\boldsymbol{\beta}}, \end{cases}$$

where $\omega_j^2 = 1 + \sum_{k\neq j}\beta_k^2$ and $g_j = \mathbf{X}_j^\top(\sum_{k\neq j}\mathbf{X}_k\beta_k + \mathbf{w})/(\omega_j\|\mathbf{X}_j\|_2) \sim \mathsf{N}(0,1)$. Thus we have

$$\mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}}_\tau = \mathcal{S}) \le \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{j\in S^c}\omega_j\|\mathbf{X}_j\|_2|g_j|/(2n\log p)^{1/2} < \min_{j\in S}|\beta_j\|\mathbf{X}_j\|_2^2 + \omega_j\|\mathbf{X}_j\|_2 g_j|/(2n\log p)^{1/2}\right). \tag{A.1}$$

The right-hand side of Equation (A.1) does not depend on $\widehat{\mathcal{S}}_\tau$ hence the above inequality is valid uniformly over the class $\mathcal{T}$. Now choose a sequence $\{c_p\}_{p=1}^\infty$ such that $\lim_{p\to\infty} c_p^2/r \ge 1$. Next construct a sequence of $\boldsymbol{\beta}^{(p)}$ in the following manner:

- Consider the set $\mathcal{S}_0 = \{1,\ldots,s\} \subseteq [p]$ with $s = O(\log p)$.

- Set $\beta_1^{(p)} = c_p$. For all other $i \in \mathcal{S}_0 \setminus \{1\}$ set $\beta_i^{(p)} = a = (2r(\log p)/n)^{1/2}$.
- Set $\beta_i^{(p)} = 0$ if $i \notin \mathcal{S}_0$.

In this setup we have $\omega_j \sim (1+c_p^2)^{1/2}$ for all $j \neq 1$. Now fix $k_0 \in \mathcal{S}_0 \setminus \{1\}$ (say $k_0 = 2$). From Equation (A.1) it can be concluded that

$$\begin{aligned}
&\sup_{\widehat{\mathcal{S}}_\tau \in \mathcal{T}} \mathbb{P}_{\boldsymbol{\beta}^{(p)}}(\widehat{\mathcal{S}}_\tau = \mathcal{S}_0) \\
&\leq \mathbb{P}_{\boldsymbol{\beta}^{(p)}} \left(\max_{j \in S^c} \omega_j \left\| \mathbf{X}_j \right\|_2 |g_j| / (2n \log p)^{1/2} < |\boldsymbol{\beta}_{k_0} \left\| \mathbf{X}_{k_0} \right\|_2^2 + \omega_{k_0} \left\| \mathbf{X}_{k_0} \right\|_2 g_{k_0}| / (2n \log p)^{1/2} \right).
\end{aligned}$$

Also note that $\omega_j > (1+c_p^2)^{1/2}$ for all $j \in \mathcal{S}_0^c$. Using these facts, we get

$$\frac{1}{(1+c_p^2)^{1/2}} \max_{j \in \mathcal{S}_0^c} \omega_j \frac{\left\| \mathbf{X}_j \right\|_2 |g_j|}{(2n \log p)^{1/2}} \geq \min_{j \in \mathcal{S}_0^c} \frac{\left\| \mathbf{X}_j \right\|_2}{n^{1/2}} \max_{j \in \mathcal{S}_0^c} \frac{|g_j|}{(2 \log p)^{1/2}}.$$

Now, note that for $j \in \mathcal{S}_0^c$, we have all $\omega_j^2 = 1 + \|\boldsymbol{\beta}\|_2^2$ and

$$g_j = \frac{\mathbf{X}_j^\top \mathbf{z}}{\left\| \mathbf{X}_j \right\|_2},$$

where $\mathbf{z} = (\sum_{k \in \mathcal{S}_0} \mathbf{X}_k \beta_k + \mathbf{w})/\omega_j \sim \mathsf{N}_n(0, \mathbb{I}_n)$ and it is independent of $\{\mathbf{X}_j\}_{j \in \mathcal{S}_0^c}$. Thus, $g_j^2 = \|\mathbf{P}_j \mathbf{z}\|_2^2$, where $\mathbf{P}_j$ is the orthogonal projection operator onto the subspace span$\{\mathbf{X}_j\}$. This shows that the random quantity $\max_{j \in \mathcal{S}_0^c} g_j^2$ is the scaled version of maximum spurious correlation (defined in Section 7.2 of Fan et al. (2018)) between $\{\mathbf{X}_j\}_{j \in \mathcal{S}_0^c}$ and the noise $\mathbf{z}$ with sparsity level 1, i.e., $\max_{j \in \mathcal{S}_0^c} g_j^2 = n \widehat{R}_n^2(1, p-s)$, where

$$\widehat{R}_n(1, p-s) := \sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_2 = 1, \|\boldsymbol{\alpha}\|_0 = 1} \frac{1}{n} \sum_{i=1}^n \frac{\boldsymbol{\alpha}^\top (\mathbf{z}_i \mathbf{x}_{i, \mathcal{S}_0^c})}{(\boldsymbol{\alpha}^\top \widehat{\boldsymbol{\Sigma}}_{n, \mathcal{S}_0^c} \boldsymbol{\alpha})^{1/2}},$$

with $\widehat{\boldsymbol{\Sigma}}_{n, \mathcal{S}_0^c} = n^{-1} \sum_{i=1}^n \mathbf{x}_{i, \mathcal{S}_0^c} \mathbf{x}_{i, \mathcal{S}_0^c}^\top$. Thus, following the arguments of Fan et al. (2018), in particular, using Theorem 3.1 and Remark 3.3 of Fan et al. (2018) we have,

$$\max_{j \in \mathcal{S}_0^c} \frac{|g_j|}{(2 \log p)^{1/2}} \xrightarrow{p} 1.$$

Using the above fact and Lemma 3 from Fletcher et al. (2009), we get

$$\frac{1}{(1+c_p^2)^{1/2}} \max_{j \in \mathcal{S}_0^c} \omega_j \frac{\left\| \mathbf{X}_j \right\|_2 |g_j|}{(2n \log p)^{1/2}} \geq \min_{j \in \mathcal{S}_0^c} \frac{\left\| \mathbf{X}_j \right\|_2}{n^{1/2}} \max_{j \in \mathcal{S}_0^c} \frac{|g_j|}{(2 \log p)^{1/2}} \xrightarrow{p} 1.$$

Now, recall that $\boldsymbol{\beta}_{k_0} = \{2r(\log p)/n\}^{1/2}$ and and define $u_r := \frac{1}{2}(1 + \frac{r^{1/2}}{\sqrt{1+r}}) < 1$ .Thus we have the following:

$$\mathbb{P}(|\boldsymbol{\beta}_{k_0} \|\mathbf{X}_{k_0}\|_2^2 + \omega_{k_0} \|\mathbf{X}_{k_0}\|_2 g_{k_0}|/\{(1 + c_p^2)(2n \log p)\}^{1/2} < u_r) \to 1.$$

This tells that,

$$\lim_{p\to\infty} \sup_{\widehat{\mathcal{S}}_\tau \in \mathcal{T}} \inf_{\boldsymbol{\beta} \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}}_\tau = \mathcal{S}_{\boldsymbol{\beta}}) \leq \lim_{p\to\infty} \sup_{\widehat{\mathcal{S}}_\tau \in \mathcal{T}} \mathbb{P}_{\boldsymbol{\beta}^{(p)}}(\widehat{\mathcal{S}}_\tau = \mathcal{S}_0) = 0.$$

This finishes the proof.

A.1.1 Proof of Theorem 2.4.1

In this proof, we reparametrize δ by $8\delta_0$ for algebraic convenience. The main result can be salvaged by back substituting $8\delta_0$ by δ in all the main equations in this section. Also, for brevity of notation, in this proof we use $\widehat{\mathcal{S}}$ and $\mathcal{S}$ to denote that oracle BSS estimator and $\mathcal{S}_{\boldsymbol{\beta}}$ respectively. We highlight the three main steps of the proof:

1. Convert the BSS problem in the problem of selecting the model with maximum spurious correlation.

2. Use results from Fan et al. (2018) to find the asymptotic distribution of the maximum spurious correlation statistics.

3. Use the asymptotic distribution along with non-asymptotic concentration inequalities to upper bound the error probability.

Recall that BSS is defined as

$$\widehat{\mathcal{S}} = \arg\max_{\mathcal{D},|\mathcal{D}|=s} \|\mathbf{P}_{\mathcal{D}} Y\|_2^2 = \arg\min_{\mathcal{D}:|\mathcal{D}|=s} \|(\mathbb{I}_n - \mathbf{P}_{\mathcal{D}})Y\|_2^2 .$$

Thus from the above definition we have the following equality:

$$\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) = \mathbb{P}\left(\|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2 < \max_{\mathcal{D}\neq\mathcal{S}} \|\mathbf{P}_{\mathcal{D}}\mathbf{y}\|_2^2 \right).$$

Now we will try to understand how the quantity $\|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2$ behaves asymptotically. First it is easy to see that $\mathbf{P}_{\mathcal{S}}\mathbf{y} = \sum_{j\in\mathcal{S}} \mathbf{X}_j\beta_j + \mathbf{P}_{\mathcal{S}}\mathbf{w}$. Note that $\sum_{j\in\mathcal{S}} \mathbf{X}_j\boldsymbol{\beta}_j = \|\boldsymbol{\beta}\|_2 \tilde{\boldsymbol{\varepsilon}}$ where

$\tilde{\boldsymbol{\varepsilon}} \sim \mathsf{N}_n(0, \mathbb{I}_n)$ and independent of the noise z. Hence we up with the following:

$$\begin{aligned}\|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2 &= \Big\| \sum_{j\in S} \mathbf{X}_j \boldsymbol{\beta}_j \Big\|_2^2 + 2\mathbf{w}^\top \mathbf{P}_{\mathcal{S}} \left(\sum_{j\in S} \mathbf{X}_j \boldsymbol{\beta}_j \right) + \mathbf{w}^\top \mathbf{P}_{\mathcal{S}} \mathbf{w} \\ &= \Big\| \sum_{j\in \mathcal{S}} \mathbf{X}_j \boldsymbol{\beta}_j \Big\|_2^2 + 2\mathbf{w}^\top \left(\sum_{j\in S} \mathbf{X}_j \boldsymbol{\beta}_j \right) + \mathbf{w}^\top \mathbf{P}_{\mathcal{S}} \mathbf{w} \\ &= \|\boldsymbol{\beta}\|_2^2 \, \|\tilde{\boldsymbol{\varepsilon}}\|_2^2 + 2 \, \|\boldsymbol{\beta}\|_2 \, \mathbf{w}^\top \tilde{\boldsymbol{\varepsilon}} + \mathbf{w}^\top \mathbf{P}_{\mathcal{S}} \mathbf{w}.\end{aligned}$$

Recall that $\left|\boldsymbol{\beta}_j\right| \geq \{2r(\log p)/n\}^{1/2}$ for all $j \in \mathcal{S}$. This is presumably the hardest setup as increasing signal strength can only decrease the error probability. Then $\|\boldsymbol{\beta}\|_2^2 \geq (2rs \log p)/n$. also note that $\mathbf{w}^\top \mathbf{P}_{\mathcal{S}} \mathbf{w} \sim \chi_s^2$. Hence we have,

$$\frac{\|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2}{s \log p} \geq 2r \frac{\|\tilde{\boldsymbol{\varepsilon}}\|_2^2}{n} + 2 \left(\frac{2r}{s \log p} \right)^{1/2} \frac{\tilde{\boldsymbol{\varepsilon}}^\top \mathbf{w}}{n^{1/2}} + \frac{\mathbf{w}^\top \mathbf{P}_{\mathcal{S}} \mathbf{w}}{s \log p} \xrightarrow{\text{p}} 2r.$$

Thus $\lim_{p\to\infty} \mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) \leq \mathbb{P}\left(2r \leq \limsup_{p\to\infty} \max_{\mathcal{D}\neq\mathcal{S}} \|\mathbf{P}_{\mathcal{D}}\mathbf{y}\|_2^2 / (s \log p)\right)$. The limiting behavior of the obtained maximal process turns out to be very challenging to analyze and hence we do not directly study this maximal process. Instead we focus on a related maximal process (will be defined shortly) derived from the earlier one and we use the results from Fan et al. (2018) to study its asymptotic behaviour. Now let us denote the set $\mathcal{S} \cap \mathcal{D}$ by $\mathcal{I}_0$ and $\mathcal{D} \setminus \mathcal{S}$ by $\mathcal{I}_1$, i.e., $\mathcal{D} = \mathcal{I}_0 \cup \mathcal{I}_1$. Next define the class $\mathcal{J}_{\mathcal{I}_0} = \{\mathcal{I}_1 \subseteq [p] : \mathcal{I}_1 \cap S = \varnothing, |\mathcal{I}_1 \cup \mathcal{I}_0| = s\}$ for each $\mathcal{I}_0 \subset S$. Note that $0 \leq |\mathcal{I}_0| \leq s-1$ from the construction (if $|\mathcal{I}_0| = s$ then $\mathcal{D} = \mathcal{S}$). The random variable of interest can be rewritten as follows:

$$\max_{\mathcal{D}\neq\mathcal{S}} \frac{\|\mathbf{P}_{\mathcal{D}}\mathbf{y}\|_2^2}{s \log p} = \max_{\mathcal{I}_0 : \mathcal{I}_0 \subset \mathcal{S}} \max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\mathbf{y}\|_2^2}{s \log p}.$$

Using union bound we get,

$$\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) \leq \sum_{\mathcal{I}_0 \subset \mathcal{S}} \mathbb{P}\left(\max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\mathbf{y}\|_2^2}{s \log p} > \frac{\|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2}{s \log p} \right). \tag{A.2}$$

Now fix a subset $\mathcal{I}_0$ of the true support $\mathcal{S}$. Similar to previous section define $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}_{\mathcal{I}_0}\boldsymbol{\beta}_{\mathcal{I}_0}$ and this independent of the features in $\mathcal{I}_0 \cup \mathcal{I}_1$. Also we have

$$\begin{aligned}\|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\mathbf{y}\|_2^2 &= \|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\tilde{\mathbf{y}}\|_2^2 + \left\|\mathbf{X}_{\mathcal{I}_0}\boldsymbol{\beta}_{\mathcal{I}_0}\right\|_2^2 + 2\boldsymbol{\beta}_{\mathcal{I}_0}^\top \mathbf{X}_{\mathcal{I}_0}^\top \tilde{\mathbf{y}}, \\ \|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2 &= \|\mathbf{P}_{\mathcal{S}}\tilde{\mathbf{y}}\|_2^2 + \left\|\mathbf{X}_{\mathcal{I}_0}\boldsymbol{\beta}_{\mathcal{I}_0}\right\|_2^2 + 2\boldsymbol{\beta}_{\mathcal{I}_0}^\top \mathbf{X}_{\mathcal{I}_0}^\top \tilde{\mathbf{y}}.\end{aligned}$$

Thus the summands in the right hand side of (A.2) can be written as the probability of the event $\{\max_{\mathcal{I}_1:\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}} \|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\mathbf{g}\|_2^2/(s\log p) > \|\mathbf{P}_{\mathcal{S}}\mathbf{g}\|_2^2/(s\log p)\}$, where $\mathbf{g} := (1+\left\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\right\|_2^2)^{-1/2}\tilde{\mathbf{y}}$. Note that $\mathbf{g}\sim\mathsf{N}_n(0,\mathbb{I}_n)$ and is independent of the features in $\mathcal{D}$. Now fix a specific $\mathcal{I}_0$. In the analysis we encounter the maximal process

$$\max_{\mathcal{I}_1:\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\mathbf{g}\|_2^2}{s\log p},$$

Now consider the set of indices $F_{\mathcal{I}_0} = (\{1,\cdots,p\}\setminus\mathcal{S})\cup\mathcal{I}_0$. Hence it is easy to see that $\tilde{p} := |F_{\mathcal{I}_0}| = p - s + |\mathcal{I}_0|$. Without loss of generality, let $F_{\mathcal{I}_0} = \{1,\ldots,\tilde{p}\}$. Also define $\tilde{s} := s - |\mathcal{I}_0|$. Let the set $\mathcal{V}_{\mathcal{I}_0} = \{\alpha\in\mathbb{R}^p : \|\boldsymbol{\alpha}\|_0 = s, \|\boldsymbol{\alpha}\|_2 = 1, \mathcal{I}_0\subseteq\mathcal{S}(\boldsymbol{\alpha}), \boldsymbol{\alpha}_{F_{\mathcal{I}_0}^c} = 0\}$. Here $\boldsymbol{\alpha}_J$ denotes the sub-vector of $\boldsymbol{\alpha}$ corresponding to the indices in $J\subseteq[p]$. Next we will focus on the random variable,

$$\widehat{L}_n := \widehat{L}_n(\tilde{s},\tilde{p}) = \sup_{\boldsymbol{\alpha}\in\mathcal{V}_{\mathcal{I}_0}} \frac{1}{n^{1/2}}\sum_{i=1}^n \frac{\boldsymbol{\alpha}^\top(g_i\mathbf{x}_i)}{(\boldsymbol{\alpha}^\top\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\alpha})^{1/2}}, \tag{A.3}$$

here $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$. Now recall that $\mathcal{D} = \mathcal{I}_0\cup\mathcal{I}_1$ for all $\mathcal{D}\neq\mathcal{S}$ with $|\mathcal{D}| = s$. To see the connection, first note that the above optimization problem can be viewed as the following:

$$\begin{aligned}
\widehat{L}_n &= \max_{\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}} \max_{\boldsymbol{\alpha}\in\mathcal{V}_{\mathcal{I}_0\cup\mathcal{I}_1}} \frac{\boldsymbol{\alpha}_{\mathcal{D}}^\top(\sum_{i=1}^n g_i\mathbf{x}_{i,\mathcal{D}}/n^{1/2})}{\{\boldsymbol{\alpha}_{\mathcal{D}}^\top(\widehat{\boldsymbol{\Sigma}}_{n,\mathcal{D}\mathcal{D}})\alpha_{\mathcal{D}}\}^{1/2}} \\
&= \max_{\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}} \left\{(\sum_{i=1}^n g_i\mathbf{x}_{i,\mathcal{D}}/n^{1/2})^\top\widehat{\boldsymbol{\Sigma}}_{n,\mathcal{D}\mathcal{D}}^{-1}(\sum_{i=1}^n g_i\mathbf{x}_{i,\mathcal{D}}/n^{1/2})\right\}^{1/2} \\
&= \max_{\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}} \{\mathbf{g}^\top\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^\top\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^\top\mathbf{g}\}^{1/2} \\
&= \max_{\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}} \|\mathbf{P}_{\mathcal{I}_0\cup\mathcal{I}_1}\mathbf{g}\|_2 .
\end{aligned}$$

Thus it is essential to study the asymptotic property of $\widehat{L}_n$. Now we define the standardized version of $\widehat{L}_n$ as follows

$$L_n := L_n(\tilde{s},\tilde{p}) = \sup_{\boldsymbol{\alpha}\in\mathcal{V}_{\mathcal{I}_0}} \frac{1}{n^{1/2}}\sum_{i=1}^n \boldsymbol{\alpha}^\top(g_i\mathbf{x}_i).$$

Let $\boldsymbol{Z} = (Z_1,\cdots,Z_{\tilde{p}})$ be $\tilde{p}-$variate Gaussian random variable with covariance matrix $\mathbb{I}_{\tilde{p}}$ and define the random variable $T^* := T^*(\tilde{s},\tilde{p}) = \sup_{\boldsymbol{\alpha}\in\mathcal{V}_{\mathcal{I}_0}} \boldsymbol{\alpha}_{F_{\mathcal{I}_0}}^\top\boldsymbol{Z}$.

Lemma A.1.1. *There exists universal constants K_0, K_1 such that for any $\delta_1 \in (0, K_0K_1]$,*

$$|L_n - T^*| \lesssim n^{-1}c_n^{1/2}(\tilde{s}, \tilde{p}) + K_0K_1n^{-3/2}c_n^2(\tilde{s}, \tilde{p}) + \delta_1 \tag{A.4}$$

holds with probability at least $1 - C\Delta_n(s, \tilde{p}; \delta_1)$ where $c_n(s, \tilde{p}) = s\log(e\tilde{p}/s) \vee \log n$ and

$$\Delta_n(\tilde{s}, \tilde{p}; \delta_1) = (K_0K_1)^3\frac{\{\tilde{s}b_n(\tilde{s}, \tilde{p})\}^2}{\delta_1^3 n^{1/2}} + (K_0K_1)^4\frac{\{\tilde{s}b_n(\tilde{s}, \tilde{p})\}^5}{\delta_1^4 n}$$

with $b_n(\tilde{s}, \tilde{p}) = \log(\tilde{p}/\tilde{s}) \vee \log n$.

Lemma A.1.2. *Assume that the sample size satisfies $n \geq C_1(K_0 \vee K_1)^4 c_n(\tilde{s}, \tilde{p})$. then with probability at least $1 - C_2 n^{-1/2}c_n^{1/2}(\tilde{s}, \tilde{p})$,*

$$\left|\widehat{L}_n - L_n\right| \lesssim (K_0 \vee K_1)^2 K_0K_1 n^{-1/2}c_n(\tilde{s}, \tilde{p}), \tag{A.5}$$

where $c_n(\tilde{s}, \tilde{p}) = \tilde{s}\log(e\tilde{p}/\tilde{s}) \vee \log n$.

Proof of the above two lemmas are omitted as it is in the same line of the proofs of Fan et al. (2018). Now applying Lemma A.1.1 and A.1.2 with

$$\delta_1 = \delta_1(s, \tilde{p}) = (K_0K_1)^{3/4}\min[1, n^{-1/8}\{\tilde{s}b_n(\tilde{s}, \tilde{p})^{3/8}\}]$$

yields that with probability at least $1 - C(K_0K_1)^{3/4}n^{-1/8}\{sb_n(s, \tilde{p})\}^{7/8}$,

$$\left|\widehat{L}_n - T^*\right| \lesssim (K_0K_1)^{3/4}n^{-1/8}\{\tilde{s}b_n(\tilde{s}, \tilde{p})\}^{3/8}.$$

Together with Lemma 2.3 from Chernozhukov et al. (2014) we can conclude that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(\widehat{L}_n \leq t) - \mathbb{P}(T^* \leq t)\right| \lesssim C(K_0K_1)^{3/4}n^{-1/8}\{\tilde{s}b_n(\tilde{s}, \tilde{p})\}^{7/8}. \tag{A.6}$$

Next by the definition of T^* it follows that

$$T^{*2} = \max_{\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}}\|\boldsymbol{Z}_{\mathcal{I}_0\cup\mathcal{I}_1}\|_2^2 = \sum_{j\in\mathcal{I}_0}Z_j^2 + \max_{\mathcal{I}_1\in\mathcal{J}_{\mathcal{I}_0}}\sum_{k\in\mathcal{I}_1}Z_k^2.$$

Let $\boldsymbol{W} \sim \mathsf{N}_{(p-s)}(0, \mathbb{I}_{(p-s)})$ be a Gaussian vector independent of $\boldsymbol{Z}$. Thus it follows that

$$T^{*2} \overset{d}{=} \sum_{j\in\mathcal{I}_0}Z_j^2 + \sum_{k=p-2s+|\mathcal{I}_0|+1}^{p-s}W_{(k:p-s)}^2 \leq \sum_{j\in\mathcal{I}_0}Z_j^2 + (s - |\mathcal{I}_0|)W_{(p-s:p-s)}^2.$$

From Equation (A.6) it also follows that

$$\sup_{t\geq 0}\left|\mathbb{P}(\widehat{L}_n^2 \leq t) - \mathbb{P}(T^{*2} \leq t)\right| \lesssim C(K_0K_1)^{3/4}n^{-1/8}\{\tilde{s}b_n(\tilde{s},\tilde{p})\}^{7/8}. \tag{A.7}$$

Now from the assumption, we have $r = 1 + 8\delta_0$. Assume that

$$s \leq 0.5\min\{\delta_0, \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2} + (1+4\delta_0)^{1/2}\}^2}\}\log p. \tag{A.8}$$

Hence $|\mathcal{I}_0| \leq s - 1 \leq 0.5\delta_0 \log p < \delta_0 \log p$. Thus we have,

$$\begin{aligned}
&\mathbb{P}(T^{*2} > 2(1+4\delta_0)(s-|\mathcal{I}_0|)\log p)\\
&= \mathbb{P}(\sum_{j\in\mathcal{I}_0} Z_j^2 + (s-|\mathcal{I}_0|)W^2_{(p-s:p-s)} > 2(1+4\delta_0)(s-|\mathcal{I}_0|)\log p)\\
&\leq \mathbb{P}\left(\sum_{j\in\mathcal{I}_0} Z_j^2 > 4\delta_0(s-|\mathcal{I}_0|)\log p\right) + \mathbb{P}\left((s-|\mathcal{I}_0|)W^2_{(p-s:p-s)} > 2(1+2\delta_0)(s-|\mathcal{I}_0|)\log p\right)\\
&\overset{(a)}{\leq} \mathbb{P}\left(\frac{\sum_{j\in\mathcal{I}_0} Z_j^2 - |\mathcal{I}_0|}{|\mathcal{I}_0|} > (4\delta_0\log p - |\mathcal{I}_0|)/|\mathcal{I}_0|\right) + \mathbb{P}\left(W^2_{(p-s:p-s)} > 2(1+2\delta_0)\log p\right)\\
&\overset{(b)}{\leq} \mathbb{P}\left(\frac{\sum_{j\in\mathcal{I}_0} Z_j^2 - |\mathcal{I}_0|}{|\mathcal{I}_0|} > (3\delta_0\log p)/|\mathcal{I}_0|\right) + \mathbb{P}\left(W^2_{(p-s:p-s)} > 2(1+2\delta_0)\log p\right)\\
&\lesssim \exp(-0.75\delta_0\log p) + (p-s)\mathbb{P}\left(W_1^2 > 2(1+2\delta_0)\log p\right)\\
&\lesssim p^{-0.75\delta_0} + C\frac{p^{-2\delta_0}}{\sqrt{\log p}}.
\end{aligned}$$

Inequality (a) uses $s - |\mathcal{I}_0| \geq 1$ and inequality (b) uses $|\mathcal{I}_0| < s < \delta_0 \log p$ (Condition (A.8)). Also, the first probability bound in (b) follows from Equation (56) in Wainwright (2009a) and the fact that $(3\delta_0\log p)/|\mathcal{I}_0| \geq 6 > 4$. The last inequality in (b) follows from tail bound of standard Gaussian distribution. Now define the event $\mathcal{E}_{\mathcal{I}_0} := \{\|\mathbf{P}_{\mathcal{S}}\mathbf{g}\|_2^2/(s\log p) > 2(1+4\delta_0)R_{\mathcal{I}_0}\}$ where $R_{\mathcal{I}_0} := (s-|\mathcal{I}_0|)/s$. Recall that

$$\begin{aligned}
\frac{\|\mathbf{P}_{\mathcal{S}}\mathbf{g}\|_2^2}{s\log p} &\geq \frac{\left\|\frac{\sum_{j\in\mathcal{S}\setminus\mathcal{I}_0}\mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{P}_{\mathcal{S}}\mathbf{w}}{(1+\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\|_2^2)^{1/2}}\right\|_2^2}{s\log p}\\
&\geq \frac{\left\{\frac{\left\|\sum_{j\in\mathcal{S}\setminus\mathcal{I}_0}\mathbf{X}_j\boldsymbol{\beta}_j\right\|_2}{(1+\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\|_2^2)^{1/2}} - \frac{\|\mathbf{P}_{\mathcal{S}}\mathbf{w}\|_2}{(1+\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\|_2^2)^{1/2}}\right\}^2}{s\log p} = (T_1^{1/2} - T_2^{1/2})^2.
\end{aligned} \tag{A.9}$$

where

$$T_1 := \frac{\left\|\sum_{j\in\mathcal{S}\setminus\mathcal{I}_0}\mathbf{X}_j\boldsymbol{\beta}_j\right\|_2^2}{(1+\left\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\right\|_2^2)s\log p} \geq \frac{2rR_{\mathcal{I}_0}}{1+2r(s-|\mathcal{I}_0|)(\log p)/n}\frac{V_n}{n},$$

and $V_n := \frac{\|\sum_{j\in\mathcal{S}\setminus\mathcal{I}_0}\mathbf{X}_j\boldsymbol{\beta}_j\|_2^2}{\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\|_2^2}$ is an χ_n^2 random variable. Also we have

$$T_2 := \frac{\|\mathbf{P}_{\mathcal{S}}\mathbf{w}\|_2^2}{(1+\left\|\boldsymbol{\beta}_{\mathcal{S}\setminus\mathcal{I}_0}\right\|_2^2)s\log p} \leq V_s/(s\log p)$$

where $V_s := \|\mathbf{P}_{\mathcal{S}}\mathbf{w}\|_2^2$ is an χ_s^2 random variable independent of $\mathbf{X}_S$. Next, we state the following simple algebraic relationship:

$$(1+6\delta_0)^{1/2} - \frac{\delta_0}{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}} \geq (1+4\delta_0)^{1/2}.$$

In light of Equation (A.9) and using the above algebraic inequality we have the following:

$$\mathcal{E}_{\mathcal{I}_0}^c \subseteq \{T_1 \leq 2(1+6\delta_0)R_{\mathcal{I}_0}\}\bigcup\{T_2 \geq 2\delta_0^2\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^{-2}R_{\mathcal{I}_0}\}$$

Next, we have

$$\mathbb{P}(T_1 \leq 2(1+6\delta_0)R_{\mathcal{I}_0}) \leq \mathbb{P}\left(\frac{V_n}{n} \leq \frac{1+6\delta_0}{1+8\delta_0}(1+2rs\log p/n)\right).$$

Now choose large n such that $(1+6\delta_0)(1+2rs\log p/n)) < (1+7\delta_0)$. Then for large n we have,

$$\mathbb{P}(T_1 \leq 2(1+6\delta_0)R_{\mathcal{I}_0}) \leq \mathbb{P}\left(|V_n/n-1| \geq \frac{\delta_0}{1+8\delta_0}\right) \lesssim \exp\left\{-C^*\frac{\delta_0^2}{(1+8\delta_0)^2}n\right\},$$

where C^* is a universal constant. Now we analyze the quantity T_2. We have the following

inequalities:

$$
\begin{aligned}
&\mathbb{P}(T_2 \geq \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2} R_{\mathcal{I}_0}) \\
&\leq \mathbb{P}(V_s/s \geq \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2} R_{\mathcal{I}_0} \log p) \\
&\leq \mathbb{P}(V_s \geq \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2} \log p) \\
&\leq \mathbb{P}(|V_s/s - 1| \geq 0.5 \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2} (\log p)/s) \\
&\leq \exp(-C' \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2} \log p) \quad \text{(Using Condition (A.8))} \\
&= p^{-C' \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2}} \quad (C' > 0 \text{ is universal constant}).
\end{aligned}
$$

Ultimately it shows that $\mathbb{P}(\mathcal{E}_{\mathcal{I}_0}^c) \lesssim \exp\left\{-C^* \frac{\delta_0^2}{(1+8\delta_0)^2} n\right\} + p^{-C' \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2}}$. Now we

are ready to show that the error probability goes to 0.

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\beta}}(\widehat{\mathcal{S}} \neq \mathcal{S}) &\leq \sum_{\mathcal{I}_0 \subset S} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0 \cup \mathcal{I}_1} \mathbf{y}\|_2^2}{s \log p} > \frac{\|\mathbf{P}_{\mathcal{S}} \mathbf{y}\|_2^2}{s \log p} \right) \\
&\leq \sum_{k=0}^{s-1} \sum_{\mathcal{I}_0 : |\mathcal{I}_0| = k} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0 \cup \mathcal{I}_1} \mathbf{y}\|_2^2}{s \log p} > \frac{\|\mathbf{P}_{\mathcal{S}} \mathbf{y}\|_2^2}{s \log p} \right) \\
&\leq \sum_{k=0}^{s-1} \sum_{\mathcal{I}_0 : |\mathcal{I}_0| = k} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0 \cup \mathcal{I}_1} \mathbf{y}\|_2^2}{s \log p} > \frac{\|\mathbf{P}_{\mathcal{S}} \mathbf{y}\|_2^2}{s \log p}, \mathcal{E}_{\mathcal{I}_0} \right) + \mathbb{P}(\mathcal{E}_{\mathcal{I}_0}^c) \\
&\leq \sum_{k=0}^{s-1} \sum_{\mathcal{I}_0 : |\mathcal{I}_0| = k} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0 \cup \mathcal{I}_1} \mathbf{g}\|_2^2}{s \log p} > \frac{\|\mathbf{P}_{\mathcal{S}} \mathbf{g}\|_2^2}{s \log p}, \mathcal{E}_{\mathcal{I}_0} \right) + \mathbb{P}(\mathcal{E}_{\mathcal{I}_0}^c) \\
&\leq \sum_{k=0}^{s-1} \sum_{\mathcal{I}_0 : |\mathcal{I}_0| = k} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{\mathcal{I}_1 : \mathcal{I}_1 \in \mathcal{J}_{\mathcal{I}_0}} \frac{\|\mathbf{P}_{\mathcal{I}_0 \cup \mathcal{I}_1} \mathbf{g}\|_2^2}{s \log p} > 2(1 + 4\delta_0) R_{\mathcal{I}_0} \right) + \mathbb{P}(\mathcal{E}_{\mathcal{I}_0}^c) \\
&\overset{(a)}{\lesssim} \sum_{k=0}^{s-1} \sum_{\mathcal{I}_0 : |\mathcal{I}_0| = k} \left[\mathbb{P}\left(T^{*2} > 2(1 + 4\delta_0) s R_{\mathcal{I}_0} \log p \right) + \mathbb{P}(\mathcal{E}_{\mathcal{I}_0}^c) + C(K_0 K_1)^{3/4} n^{-1/8} \{ s b_n(s, p) \}^{7/8} \right] \\
&\lesssim \sum_{k=0}^{s-1} \sum_{\mathcal{I}_0 : |\mathcal{I}_0| = k} p^{-0.75\delta_0} + C \frac{p^{-2\delta_0}}{\sqrt{\log p}} + \exp\left\{ -C^* \frac{\delta_0^2}{(1 + 8\delta_0)^2} n \right\} + p^{-C' \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2} + (1+4\delta_0)^{1/2}\}^2}} \\
&\quad + n^{-1/8} \{ s b_n(s, p) \}^{7/8} \\
&\lesssim \sum_{k=0}^{s-1} \binom{s}{k} \left[p^{-0.75\delta_0} + \exp\left\{ -C^* \frac{\delta_0^2}{(1 + 8\delta_0)^2} n \right\} + p^{-C' \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2} + (1+4\delta_0)^{1/2}\}^2}} + n^{-1/8} \{ s b_n(s, p) \}^{7/8} \right] \\
&\lesssim 2^s \left[p^{-0.75\delta_0} + \exp\left\{ -C^* \frac{\delta_0^2}{(1 + 8\delta_0)^2} n \right\} + p^{-C' \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2} + (1+4\delta_0)^{1/2}\}^2}} + n^{-1/8} \{ s b_n(s, p) \}^{7/8} \right].
\end{aligned}
$$

Inequality (a) uses $\tilde{s} b_n(\tilde{s}, \tilde{p}) \leq s b_n(s, p)$ for large p. Thus if

$$
\begin{aligned}
s &\lesssim \left(\delta_0 \wedge \frac{2\delta_0^2}{\{(1 + 6\delta_0)^{1/2} + (1 + 4\delta_0)^{1/2}\}^2} \wedge \frac{k}{16} \right) \log p \\
&= \left(\frac{2\delta_0^2}{\{(1 + 6\delta_0)^{1/2} + (1 + 4\delta_0)^{1/2}\}^2} \wedge \frac{k}{16} \right) \log p,
\end{aligned}
$$

then error probability goes to 0 uniformly over $\boldsymbol{\beta} \in \mathcal{M}_s^a$.

A.1.2 Model consistency of BSS for sub-Gaussian model

In this section, we will show that Theorem 4.1 also holds beyond the Gaussian model. We assume that the entries of the design matrix $\mathbf{X}$ are i.i.d. *mean-zero* and *sub-Gaussian* with

unit variance. We also assume that the entries of $\mathbf{w}$ are also i.i.d. *mean-zero* and *sub-Gaussian* with *unit variance* and independent of $\mathbf{X}$.

In this setup, the results of Theorem 3.1 in Fan et al. (2018) are also valid and the proof steps follow exactly the same steps as the proof of Theorem 4.1 until the introduction of the random variables V_n and V_s.

We note that Gaussianity was only used to characterize the distributions of V_n and V_s. In particular, due to Gaussianity, we have $V_n \sim \chi^2_n$ and $V_s \sim \chi^2_s$. However, under the sub-Gaussian case, V_n is the sum of n independent sub-Exponential random variables with *unit-mean*. Hence, the probability bound for T_1 shown in the original proof is also valid.

Next, for V_s, we can use Theorem 1.1 of Rudelson and Vershynin (2013). Note that, there exists a constant $K_{\psi_2} > 1$ such that $\|\varepsilon\|_{\psi_2} \leq K_{\psi_2}$, where $\|\varepsilon\|_{\psi_2} := \inf_{t>0}\{t : \mathbb{E}\exp(\varepsilon^2/t^2) \leq 2\}$. By Theorem 1.1 of Rudelson and Vershynin (2013), we can obtain the same probability bound for T_2 if $s \leq (0.5/K^2_{\psi_2}) \min\{\delta_0, \frac{2\delta_0^2}{\{(1+6\delta_0)^{1/2}+(1+4\delta_0)^{1/2}\}^2}\} \log p$. Hence, the rest of the proof is verbatim to the proof in the Gaussian case.

A.1.3 Proof of Theorem 2.4.3

In this section, we will show that BSS fails to recover the exact support when $r = 1$. We highlight three main steps of the proof:

1. Convert the BSS problem in the problem of selecting the model with maximum spurious correlation.

2. Use results from Fan et al. (2018) to find the asymptotic distribution of the maximum spurious correlation statistics.

3. Use the exact form of the asymptotic distribution along with scaling and centering parameters to approximate the recovery probability.

Recall the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}$ with $\mathcal{S} := \mathcal{S}_{\boldsymbol{\beta}}$ as the set of active features and $s = |\mathcal{S}| = O(\log p)$. As $r = 1$, there exists $j_0 \in \mathcal{S}$ such that $\beta_{j_0} = \{(2\log p)/n\}^{1/2}$. WLOG , let us assume that $j_0 = 1$ and define $\mathcal{I}_0 = \mathcal{S} \setminus \{1\}$. In order for BSS to recover the exact support $\mathcal{S}$, it is necessary that

$$
\begin{aligned}
&\max_{j \notin \mathcal{S}} \left\|\mathbf{P}_{\mathcal{I}_0 \cup \{j\}}\mathbf{y}\right\|_2^2 < \|\mathbf{P}_{\mathcal{S}}\mathbf{y}\|_2^2 \\
&\Leftrightarrow \max_{j \notin \mathcal{S}} \left\|\mathbf{P}_{\mathcal{I}_0 \cup \{j\}}\tilde{\mathbf{y}}\right\|_2^2 < \|\mathbf{P}_{\mathcal{S}}\tilde{\mathbf{y}}\|_2^2 \quad (\text{where } \tilde{\mathbf{y}} = \mathbf{X}_1\beta_1 + \mathbf{w}) \\
&\Leftrightarrow \max_{j \notin \mathcal{S}} \left\|\mathbf{P}_j^{\dagger}\tilde{\mathbf{y}}\right\|_2^2 < \left\|\mathbf{P}_1^{\dagger}\tilde{\mathbf{y}}\right\|_2^2,
\end{aligned}
$$

where $\mathbf{P}_j^\dagger$ is the orthogonal projection operator onto the sub-space span$\{\tilde{\mathbf{X}}_j\}$ for all $j \in \mathcal{S}^c \cup \{1\}$, where $\tilde{\mathbf{X}}_j = (\mathbb{I}_n - \mathbf{P}_{\mathcal{I}_0})\mathbf{X}_j$. Due to Gaussianity, it follows that $\left\|\tilde{\mathbf{X}}_j\right\|_2^2 \sim \chi^2_{n-s+1}$.

Now, note that

$$\begin{aligned}\left\|\mathbf{P}_1^\dagger \tilde{\mathbf{Y}}\right\|_2^2 &= \beta_1^2 \left\|\tilde{\mathbf{X}}_1\right\|_2^2 + 2\beta_1 \tilde{\mathbf{X}}_1^\top \mathbf{w} + \left\|\mathbf{P}_1^\dagger \mathbf{w}\right\|_2^2 \\ &= (2\log p)\frac{\left\|\tilde{\mathbf{X}}_1\right\|_2^2}{n} + 2(2\log p)^{1/2}\frac{\tilde{\mathbf{X}}_1^\top \mathbf{w}}{\sqrt{n}} + \left\|\mathbf{P}_1^\dagger \mathbf{w}\right\|_2^2.\end{aligned}$$

Next, define the events

$$\mathcal{E}_1 := \left\{ \left| \frac{\left\|\tilde{\mathbf{X}}_1\right\|_2^2}{n-s+1} - 1 \right| \le 1/\sqrt{\log p} \right\}, \quad \mathcal{E}_2 := \left\{ \frac{\tilde{\mathbf{X}}_1^\top \mathbf{w}}{\left\|\tilde{\mathbf{X}}_1\right\|_2} \le -1 \right\}, \quad \mathcal{E}_3 := \left\{ \left\|\mathbf{P}_1^\dagger \mathbf{w}\right\|_2^2 \le 16 \right\}.$$

When $p > 4$, by Bernstein's inequality, we have $\mathbb{P}(\mathcal{E}_1^c) \le e^{-c_1 \frac{n}{\log p}}$, where $c_1 > 0$ is a universal constant. Recall that $n \asymp p^k$, which tells that $\lim_{p\to\infty} \mathbb{P}(\mathcal{E}_1^c) = 0$. Next, note that $\frac{\tilde{\mathbf{X}}_1^\top E}{\left\|\tilde{\mathbf{X}}_1\right\|_2} \sim \mathsf{N}(0,1)$. Next, we introduce a useful lemma.

Lemma A.1.3 (Gordon (1941))**.** *Let $\Phi(\cdot)$ denote the cumulative distribution function of standard Gaussian distribution. Then for all $x \ge 0$, the following inequalities are true:*

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

By the above lemma we can conclude $\mathbb{P}(\mathcal{E}_2^c) \le 1 - \frac{e^{-1/2}}{2\sqrt{2\pi}}$. Finally, as $\left\|\mathbf{P}_1^\dagger \mathbf{w}\right\|_2^2 \sim \chi_1^2$, we have $\mathbb{P}(\mathcal{E}_3^c) \le 2e^{-8}$. Since $n + 4 > 4s$ for large p, we have the following under $\mathcal{E}_1 \cap \mathcal{E}_2$:

$$\frac{\tilde{\mathbf{X}}_1^\top \mathbf{w}}{\sqrt{n}} = \frac{\tilde{\mathbf{X}}_1^\top \mathbf{w}}{\left\|\tilde{\mathbf{X}}_1\right\|_2} \times \frac{\left\|\tilde{\mathbf{X}}_1\right\|_2}{\sqrt{n-s+1}} \times \sqrt{\frac{n-s+1}{n}} \le -\frac{\sqrt{3}\{1-(\log p)^{-1/2}\}}{2}.$$

Here we used the fact that $\sqrt{1-(\log p)^{-1/2}} > 1 - (\log p)^{-1/2}$. We define the event $\mathcal{E} :=$

$\cap_{i=1}^{3}\mathcal{E}_i$. Then, we have

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{S}}_{\text{best}} = \mathcal{S}) &\leq \mathbb{P}\left(\max_{j\notin\mathcal{S}} \left\|\mathbf{P}_j^{\dagger}\tilde{\mathbf{y}}\right\|_2^2 < \left\|\mathbf{P}_1^{\dagger}\tilde{\mathbf{y}}\right\|_2^2\right) \\
&\leq \mathbb{P}\left(\max_{j\notin\mathcal{S}} \left\|\mathbf{P}_j^{\dagger}\tilde{\mathbf{y}}\right\|_2^2 < \left\|\mathbf{P}_1^{\dagger}\tilde{\mathbf{y}}\right\|_2^2, \mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq \mathbb{P}\left\{\max_{j\notin\mathcal{S}} \left\|\mathbf{P}_j^{\dagger}\tilde{\mathbf{y}}\right\|_2^2 < 2\log p\left(1+\frac{1}{\sqrt{\log p}}\right) - (6\log p)^{1/2}\left(1-\frac{1}{\sqrt{\log p}}\right) + 16\right\} + \mathbb{P}(\mathcal{E}^c).
\end{aligned}$$

We further note that $\mathbf{g} := \tilde{\mathbf{y}}/(1+\beta_1^2)^{1/2}$ follows a standard isotropic Gaussian distribution. Using Theorem 3.1 of Fan et al. (2018) and the fact that $1+\beta_1^2 > 1$, we get

$$\begin{aligned}
&\mathbb{P}(\widehat{\mathcal{S}}_{\text{best}} = \mathcal{S}) \\
&\leq \mathbb{P}\left\{Z^2_{(p-s:p-s)} \leq 2\log p\left(1+\frac{1}{\sqrt{\log p}}\right) - (6\log p)^{1/2}\left(1-\frac{1}{\sqrt{\log p}}\right) + 16\right\} + \mathbb{P}(\mathcal{E}^c) + o(1) \\
&\leq \mathbb{P}\left\{Z^2_{(p-s:p-s)} - 2\log(p-s) + \log\log(p-s) \leq \underbrace{-\left(\sqrt{6}-2\right)(\log p)^{1/2} + \log\log p + O(1)}_{:=t_p}\right\} \\
&\quad + \mathbb{P}(\mathcal{E}^c) + o(1).
\end{aligned}$$

where $Z^2_{(p-s:p-s)}$ is the maximum order statistics of $\{Z_j^2\}_{j\in[p-s]}$ with $\{Z_j\}_{j\in[p-s]}$ being i.i.d. standard Gaussian. Finally from Remark 3.3 of Fan et al. (2018), we know

$$Z^2_{(p-s:p-s)} - 2\log(p-s) + \log\log(p-s) \xrightarrow{\text{d}} \Lambda,$$

where $\mathbb{P}(\Lambda \leq t) = \exp(-\pi^{-1/2}\exp(-t/2))$. As $t_p \to -\infty$, we have

$$\lim_{p\to\infty} \mathbb{P}(\widehat{\mathcal{S}}_{\text{best}} = \mathcal{S}) \leq 1 - \frac{e^{-1/2}}{2\sqrt{2\pi}} + 2e^{-8} + \lim_{p\to\infty}\mathbb{P}(\mathcal{E}_1^c) < 0.9.$$

In other words, when $r = 1$, i.e., $a = \{2\times 1\times(\log p)/n\}^{1/2}$, we have

$$\lim_{p\to\infty}\sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}(\widehat{\mathcal{S}}_{\text{best}} \neq \mathcal{S}) > \frac{1}{10}.$$

A.1.4 Minimax 0-1 loss under AURWM regime

We first present a result form Wang et al. (2010) which gives us the necessary condition for asymptotic exact recovery.

Theorem A.1.4 (Wang et al. (2010))**.** *Consider the model 2.1 with the design matrix* $\mathbf{X} \in$

$\mathbb{R}^{n\times p}$ *be drawn with i.i.d elements from any distribution with zero mean and unit variance. Let* $a := \min_{j\in\mathcal{S}_{\boldsymbol{\beta}}} |\beta_j|$*, i.e., it denote the minimum signal strength of* $\boldsymbol{\beta}$*. Define the function*

$$f_m(p,s,a) := \frac{\log\binom{p-s+m}{m} - 1}{\frac{1}{2}\log\Big(1 + ma^2(1-\frac{m}{p-s+m})\Big)}, \quad 1 \le m \le s.$$

Then $n \ge \max\{f_1(p,s,a),\ldots,f_s(p,s,a), s\}$ *is necessary for asymptotic exact recovery.*

In the light of Theorem A.1.4 the proof of Proposition 4.4 follows immediately. To see this note that if $r < 1$ then there exists $\alpha \in (0,1)$ such that $r = 1-\alpha$. Also recall that $a = \{2r(\log p)/n\}^{1/2}, s = O(\log p)$ and $n = \lfloor p^k \rfloor$. Note that $f_1(p,s,a)/n \sim \frac{1}{1-\alpha}$. This shows that asymptotically the necessary condition in the above theorem is violated and hence $r \ge 1$ is necessary.

A.2 Results related to ETS

A.2.1 Proof of Theorem 2.5.2

We first briefly describe the main steps of the proof:

1. We first establish the ℓ_2-error bound of $\widehat{\boldsymbol{\beta}}$, i.e., we show that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le \epsilon^{1/2}$.
2. Next, we upper bound the 0-1 loss $\mathbb{P}_{\boldsymbol{\beta}}(\hat{\eta} \ne \eta)$ by decomposing it across the coordinates.
3. We analyze each of the terms separately and use ℓ_2-error bound along with Gaussian tail inequalities to establish model consistency.

Now we are ready for the main proof. Due to Assumption 5.1, there exists a sequence $\{\alpha_p\}_{p\ge1} \subseteq [0,\infty)$ converging to 0 such that with probability $1-\alpha_p$ the following is true: $\mathcal{A}$ requires no more than $T(\epsilon,p,\boldsymbol{\beta})$ iterations to output $\widehat{\boldsymbol{\beta}}$ that satisfies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le \epsilon^{1/2}$.

For notational brevity, we write η instead of $\eta_{\boldsymbol{\beta}}$. Define the event $\mathcal{H} = \Big\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le \epsilon^{1/2}\Big\}$ and we have $\mathbb{P}(\mathcal{H}^c) \le \alpha_p$. Note that $\widehat{\boldsymbol{\beta}}$ is based on the subsample $\mathcal{D}_1$.

Next, for algebraic convenience we again reparametrize δ as $8\delta_0$ and set $\epsilon = 6\delta_0, \varsigma = (1+\epsilon)^{1/2}$. Now note that for any $\boldsymbol{\beta} \in \mathcal{M}_a^s$, we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\beta}}(\hat{\eta} \ne \eta |\, \mathcal{D}_1) &\le \sum_{j:\beta_j=0} \mathbb{P}(\hat{\eta}_j = 1, \mathcal{H} \,|\, \mathcal{D}_1) + \sum_{j:\beta_j\ne0} \mathbb{P}_{\boldsymbol{\beta}}(\hat{\eta}_j \ne 1, \mathcal{H} \,|\, \mathcal{D}_1) + \mathbb{P}(\mathcal{H}^c \,|\, \mathcal{D}_1) \\
&= \sum_{j:\beta_j=0} \mathbb{P}_{\boldsymbol{\beta}}(|\Delta_j| > \kappa_\varsigma(\mathbf{X}_j^{(2)}), \mathcal{H} \,|\, \mathcal{D}_1) + \sum_{j:\beta_j\ne0} \mathbb{P}_{\boldsymbol{\beta}}(|\Delta_j| \le \kappa_\varsigma(\mathbf{X}_j^{(2)}), \mathcal{H} \,|\, \mathcal{D}_1) + \mathbb{P}(\mathcal{H}^c \,|\, \mathcal{D}_1),
\end{aligned}$$

Using the fact that conditionally on $\widehat{\boldsymbol{\beta}}$ and X_j, the random variable Δ_j has the same distribution as the random variable in (11) in the main paper, we conclude that for all $j \notin \mathcal{S}(\boldsymbol{\beta})$,

$$\begin{aligned}\mathbb{P}(\eta_j = 1, \mathcal{H} \,|\, \mathcal{D}_1) &\leq \mathbb{P}\left((1+\epsilon)^{1/2} |g_j| > \frac{a \left\| \mathbf{X}_j^{(2)} \right\|_2}{2} + \frac{(1+\epsilon)\log p}{a \left\| \mathbf{X}_j^{(2)} \right\|_2}, \mathcal{H} \,\Big|\, \mathcal{D}_1 \right) \\ &= 2\,\mathbb{E}\left\{ \overline{\Phi}\left(\frac{a \left\| \mathbf{X}_j^{(2)} \right\|_2}{2(1+\epsilon)^{1/2}} + \frac{(1+\epsilon)^{1/2}\log p}{a \left\| \mathbf{X}_j^{(2)} \right\|_2} \right) \right\}.\end{aligned}$$

Here $\overline{\Phi}(\cdot)$ denotes the survival function of the standard Gaussian random variable. Now note that for each j we have $\left\| \mathbf{X}_j^{(2)} \right\|_2^2 \stackrel{\mathrm{d}}{=} V_{n_2}$, where V_{n_2} is a chi-squared random variable with n_2 degrees of freedom. Thus we have

$$\mathbb{P}(\eta_j = 1, \mathcal{H} \,|\, \mathcal{D}_1) \leq 2\,\mathbb{E}\left\{ \overline{\Phi}\left(\frac{a V_{n_2}^{1/2}}{2(1+\epsilon)^{1/2}} + \frac{(1+\epsilon)^{1/2}\log p}{a V_{n_2}^{1/2}} \right) \right\}.$$

Analogous argument and the fact that $|\beta_j| \geq a$ for all $\beta_j \neq 0$, leads to the fact that for all $j \in \mathcal{S}_{\boldsymbol{\beta}}$,

$$\mathbb{P}(\eta_j \neq 1, \mathcal{H} \,|\, \mathcal{D}_1) \leq 2\mathbb{E}\left\{ \overline{\Phi}\left(\max\left\{ \frac{a V_{n_2}^{1/2}}{2(1+\epsilon)^{1/2}} - \frac{(1+\epsilon)^{1/2}\log p}{a V_{n_2}^{1/2}}, 0 \right\} \right) \right\}.$$

Now recall that $\epsilon = 6\delta_0$ and $\gamma \in (0, \frac{\delta_0}{1+8\delta_0})$. With this choice of tuning parameters, it is easy to see that $r(1-\gamma)/(1+\epsilon) \geq \frac{1+7\delta_0}{1+6\delta_0} > 1$ and hence as $p \to \infty$ we have

$$\begin{aligned}W_{n_2} &:= \frac{1}{(\log p)^{1/2}} \left(\frac{a V_{n_2}^{1/2}}{2(1+\epsilon)^{1/2}} - \frac{(1+\epsilon)^{1/2}\log p}{a V_{n_2}^{1/2}} \right) \\ &\xrightarrow{\mathrm{p}} \frac{1}{(2r)^{1/2}} \left\{ r\left(\frac{1-\gamma}{1+\epsilon} \right)^{1/2} - \left(\frac{1+\epsilon}{1-\gamma} \right)^{1/2} \right\} > 0.\end{aligned}$$

The above display uses the fact that $n_2/n \to 1-\gamma$ and $V_{n_2}/n_2 \xrightarrow{\mathrm{p}} 1$ as $p \to \infty$. Next let is define the following quantity q:

$$q := q(\epsilon, \delta_0, \gamma) = \frac{1}{\{2(1+8\delta_0)\}^{1/2}} \left\{ (1+8\delta_0)\left(\frac{1-\gamma}{1+\epsilon} \right)^{1/2} - \left(\frac{1+\epsilon}{1-\gamma} \right)^{1/2} \right\}.$$

Due to the choice of ϵ and γ it is easy to show $q > 0$. Now define the event $G_{n_2} := \{W_{n_2} > q/2\}$. Before we proceed it is useful to note the following:

$$W_{n_2} = \frac{1}{(2r)^{1/2}} \left(\frac{r\{V_{n_2}/(n(1-\gamma))\}^{1/2}(1-\gamma)^{1/2}}{(1+\epsilon)^{1/2}} - \frac{(1+\epsilon)^{1/2}}{\{V_{n_2}/(n(1-\gamma))\}^{1/2}(1-\gamma)^{1/2}} \right).$$

Next, define the function

$$H(u) := \frac{1}{(2+16\delta_0)^{1/2}} \left\{ u(1+8\delta_0) \left(\frac{1-\gamma}{1+\epsilon} \right)^{1/2} - \frac{1}{u} \left(\frac{1+\epsilon}{1-\gamma} \right)^{1/2} \right\}, \quad u > 0.$$

As $r = 1 + 8\delta_0$ we have $W_{n_2} = H(\{V_{n_2}/(n(1-\gamma))\}^{1/2})$. It is also easy to see that $H(\cdot)$ is strictly increasing function on $(0, \infty)$ and $H(1) = q$. Hence $\lambda_{\delta_0} := H^{-1}(q/2) \in (0, 1)$. Now $G_{n_2}^c = \{W_{n_2} \leq q/2\} \subseteq \{H(\{V_{n_2}/(n(1-\gamma))\}^{1/2}) \leq q/2\}$. Thus a straightforward calculation shows that

$$\mathbb{P}(G_{n_2}^c) \leq \mathbb{P}\left(\frac{V_{n_2}}{n_2} \leq \frac{n(1-\gamma)}{n_2} \lambda_{\delta_0}^2 \right).$$

Choose p large enough such that $n_2/n > \lambda_{\delta_0}(1-\gamma)$ and hence we have,

$$\mathbb{P}(G_{n_2}^c) \leq \mathbb{P}\left(\frac{V_{n_2}}{n_2} \leq \lambda_{\delta_0} \right) \lesssim \exp(-K_{\delta_0} n_2),$$

where $K_{\delta_0} = (1 - \lambda_{\delta_0})^2/8$. Note that $\overline{\Phi}(t) \leq e^{-t^2/2}$ for all $t > 0$. Using this fact we have the following:

$$\mathbb{P}(\eta_j = 1, \mathcal{H} \,|\, \mathcal{D}_1) \leq \mathbb{E}\left[\exp\left\{ -\left(1 + \frac{W_{n_2}^2}{2} \right) \log p \right\} \mathbb{1}_{G_{n_2}} \right] + \mathbb{P}(G_{n_2}^c) \lesssim p^{-(1+q^2/8)} + \exp(-K_{\delta_0} n_2),$$

for all $j \notin \mathcal{S}_{\boldsymbol{\beta}}$. Similarly,

$$\mathbb{P}(\eta_j \neq 1, \mathcal{H} \,|\, \mathcal{D}_1) \lesssim p^{-q^2/8} + \exp(-K_{\delta_0} n_2), \quad \forall j \in \mathcal{S}(\boldsymbol{\beta}).$$

Remark A.2.1. *Note that $q^2 = \Omega(\frac{\delta_0^2}{1+\delta_0^2})$ and it shows that the upper bound in the above display deteriorates as $\delta_0 \to 0$. Also, as δ_0 approaches 0, the term K_{δ_0} also approaches 0. Hence, the rate of decay worsens as $\delta_0 \to 0$, and ETS continues to lose statistical power.*

A.2.2 Proof of Corollary 2.5.3

Similar to previous proofs, we reparametrize δ by $8\delta_0$ and set $\epsilon = 6\delta_0, \varsigma = (1+\epsilon)^{1/2}$. Now note that it is enough to prove the following:

$$\lim_{p\to\infty} \inf_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{j\notin\mathcal{S}_{\boldsymbol{\beta}}} |\Delta_j| < \min_{j\in\mathcal{S}_{\boldsymbol{\beta}}} |\Delta_j| \right) \to 1$$

as $p \to \infty$. To this end first define the following quantity:

$$t_p := \frac{\left(\frac{2rn_2 \log p}{n}\right)^{1/2}}{2} + \frac{\varsigma^2 \log p}{\left(\frac{2rn_2 \log p}{n}\right)^{1/2}}.$$

We will show that $\lim_{p\to\infty} \inf_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}} \left(\min_{j\in\mathcal{S}_{\boldsymbol{\beta}}} |\Delta_j| > t_p, \max_{j\notin\mathcal{S}_{\boldsymbol{\beta}}} |\Delta_j| \le t_p \right) \to 1$ as $p \to \infty$. For convenience let us define the events $G_{\min} := \{\min_{j\in\mathcal{S}_{\boldsymbol{\beta}}} |\Delta_j| > t_p\}$ and $G_{\max} := \{\max_{j\notin\mathcal{S}_{\boldsymbol{\beta}}} |\Delta_j| \le t_p\}$. Let $\mathcal{H}$ be the event as defined in Section A.2.1. First we will analyze $\mathbb{P}_{\boldsymbol{\beta}}(G_{\min}^c)$. Note that $\mathbb{P}_{\boldsymbol{\beta}}(G_{\min}^c) \le \mathbb{P}_{\boldsymbol{\beta}}(G_{\min}^c \cap \mathcal{H}) + \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{H}^c)$. Now the second term goes to 0 uniformly over $\boldsymbol{\beta} \in \mathcal{M}_s^a$. Also using Equation (11) under the event $\mathcal{H}$ we get

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(G_{\min}^c \cap \mathcal{H}) \\
&\le \sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}} \left(\min_{j\in\mathcal{S}_{\boldsymbol{\beta}}} \left| \boldsymbol{\beta}_j \|X_j^{(2)}\|_2 + (1+\epsilon)^{1/2} g_j \right| \le t_p \right) \\
&\le \sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{j\in\mathcal{S}_{\boldsymbol{\beta}}} \frac{|g_j|}{(\log p)^{1/2}} \ge \frac{1}{(1+\epsilon)^{1/2}(\log p)^{1/2}} \left\{ a \min_{j\in\mathcal{S}_{\boldsymbol{\beta}}} \|\mathbf{X}_j^{(2)}\|_2 - t_p \right\} \right) \\
&\le \sup_{\boldsymbol{\beta}\in\mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{j\in\mathcal{S}_{\boldsymbol{\beta}}} \frac{|g_j|}{(\log p)^{1/2}} \ge \frac{1}{(1+\epsilon)^{1/2}(\log p)^{1/2}} \left\{ a \min_{j\in[p]} \|\mathbf{X}_j^{(2)}\|_2 - t_p \right\} \right)
\end{aligned}$$

where $\{g_j\}_{j\in\mathcal{S}_{\boldsymbol{\beta}}}$ are non i.i.d. standard Gaussian. Note that $|\mathcal{S}_{\boldsymbol{\beta}}| = O(\log p)$. Hence

$$\max_{j\in\mathcal{S}_{\boldsymbol{\beta}}} |g_j| = O_{\mathbb{P}}(\log\log p),$$

which tells that

$$\max_{j\in\mathcal{S}_{\boldsymbol{\beta}}} \frac{|g_j|}{(\log p)^{1/2}} \xrightarrow{\text{p}} 0.$$

Also using lemma 3 from Fletcher et al. (2009) we have

$$\frac{1}{(1+\epsilon)^{1/2}(\log p)^{1/2}} \left(a \min_{j\in[p]} \|\mathbf{X}_j^{(2)}\|_2 - t_p \right) \xrightarrow{\text{p}} \frac{1}{(2r)^{1/2}} \left\{ r \left(\frac{1-\gamma}{1+\epsilon} \right)^{1/2} - \left(\frac{1+\epsilon}{1-\gamma} \right)^{1/2} \right\}.$$

The right-hand side of the above display is at least $q(\epsilon, \delta_0, \gamma)$ (defined in Section A.2.1) which is strictly positive. Again for compactness we use q instead of $q(\epsilon, \delta_0, \gamma)$. The above display motivates us to define the following event:

$$\mathcal{E}_p = \left\{ \frac{1}{(1+\epsilon)^{1/2}(\log p)^{1/2}} \left(a \min_{j \in [p]} \|\mathbf{X}_j^{(2)}\|_2 - t_p \right) \geq q/2 \right\},$$

and it follows that $\mathbb{P}(\mathcal{E}_p^c) \to 0$ as $p \to \infty$. This leads to the following inequality:

$$\begin{aligned} \sup_{\boldsymbol{\beta} \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(G_{\min}^c \cap \mathcal{H}) &\leq \sup_{\boldsymbol{\beta} \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}} \left(\max_{j \in \mathcal{S}_{\boldsymbol{\beta}}} \frac{|g_j|}{(\log p)^{1/2}} \geq q/2 \right) + \mathbb{P}(\mathcal{E}_p^c) \\ &\lesssim p^{-q^2/8} \log p + \mathbb{P}(\mathcal{E}_p^c) \to 0. \end{aligned}$$

Thus we have $\sup_{\boldsymbol{\beta} \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(G_{\min}^c) \to 0$. Similarly it can be shown that $\sup_{\boldsymbol{\beta} \in \mathcal{M}_s^a} \mathbb{P}_{\boldsymbol{\beta}}(G_{\max}^c) \to 0$ as $p \to \infty$. These two claims together complete the proof.

A.2.3 Discussion on Remark 2.5.4

As $r > 1 + \delta_*$, by reparameterizing δ_* by $8\tilde{\delta}$, we have $r > 1 + 8\tilde{\delta}$. Now we are basically back to the setting of the proof of Theorem 5.1 and all of the proof steps are exactly the same as that of Theorem 5.1 with $\tilde{\delta}$ in place of δ_0. This allows us to choose the threshold using the knowledge of δ_*, and we do not need the knowledge of a. In particular, one can construct the threshold $\kappa_\varsigma(\mathbf{X}_i^{(2)})$ with $\varsigma = (1 + A_2\delta_*)^{1/2}$, where A_2 is the same universal constant as described in Theorem 5.2

A.2.4 Discussion on examples of ETS

Solving ℓ_0-regularized problem:

Proofs for ETS-IHT:

We first introduce some standard assumptions for analyzing ETS-IHT.

Definition A.2.2 (RSC property)**.** *A differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is said to satisfy restricted strong convexity (RSC) at sparsity level $s = s_1 + s_2$ with strong convexity constraint ℓ_s if the following holds for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ s.t. $\|\boldsymbol{\theta}_1\|_0 \leq s_1$ and $\|\boldsymbol{\theta}_2\|_0 \leq s_2$:*

$$F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_2) \geq \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_2) \rangle + \frac{\ell_s}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 .$$

Definition A.2.3 (RSS property)**.** *A differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ is said to satisfy restricted strong smoothness (RSS) at sparsity level $s = s_1 + s_2$ with strong smoothness constraint L_s if the following holds for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ s.t. $\|\boldsymbol{\theta}_1\|_0 \leq s_1$ and $\|\boldsymbol{\theta}_2\|_0 \leq s_2$:*

$$F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_2) \leq \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_2) \rangle + \frac{L_s}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 .$$

Now we quote an important theorem from Jain et al. (2014) which quantifies the sub-optimality gap of Algorithm 2.

Theorem A.2.4 (Jain et al. (2014))**.** *Let F has RSC and RSS parameters given by $\ell_{2\hat{s}+s}(F) = \alpha$ and $L_{2\hat{s}+\hat{\pi}}(F) = L$ respectively. Call Algorithm 2 with $\widehat{s} \geq 32L^2\ell^{-2}s$ and $h = 2/(3L)$. Also let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\theta}, \|\boldsymbol{\theta}\|_0 \leq s} F(\boldsymbol{\theta})$. Then tth iterate of Algorithm 2 for $t = O(L\ell^{-1} \log(F(\boldsymbol{\beta}^{(0)})/\epsilon))$ satisfies:*

$$F(\boldsymbol{\beta}^{(t)}) - F(\widehat{\boldsymbol{\beta}}) \leq \epsilon.$$

In our setup, the observations $\{\mathbf{x}_i\}_{i=1}^n$ are coming from i.i.d. mean zero isotropic Gaussian distribution. Thus, lemma 6 from Agarwal et al. (2012) immediately tells that RSC and RSS at any sparsity level m hold for $f_{n_1}(\cdot)$ with probability at least $1 - \exp(-c_0 n_1)$ with $\ell_m = \frac{1}{2} - c_1(m \log p)/n_1$ and $L_m = 2 + c_1(m \log p)/n_1$, where c_0, c_1 are universal constants. Now set $m = 2\hat{s} + s$ and recall that $n_1 \sim \gamma p^k$. If $n_1 > 4c_1(2\hat{s} + s) \log p$ then we have $\ell_m \geq 1/4$ and $L_m \leq 9/4$, which means that $L_m/(9\ell_m) \leq 1$. Thus to apply Theorem A.2.4 it is enough to choose $\hat{s} = 2592s$. Also by the assumption on n for large p we have $n_1 > 4c_1(2\hat{s} + s) \log p$. Let $f_{n_1}(\boldsymbol{\theta}) := n_1^{-1} \|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\theta}\|_2^2$ for $\boldsymbol{\theta} \in \mathbb{R}^p$. Note that $f_{n_1}(0) = n_1^{-1}\|\mathbf{y}^{(1)}\|_2^2 \stackrel{\mathrm{d}}{=} (1 + \|\boldsymbol{\beta}\|_2^2)V_{n_1}/n_1$, where V_{n_1} is chi-square random variable with n_1 degrees of freedom. Also by Bernstein's type inequality it follows that $|(V_{n_1}/n_1) - 1| \leq 1/2$ with probability at least $1 - \exp(-c_4 n_1)$, where c_4 is a universal positive constant. Thus if $t = O(L_m \ell_m^{-1} \log((1 + \|\boldsymbol{\beta}\|_2^2)/\epsilon_0)) = O(\log p + \log((1 + \|\boldsymbol{\beta}\|_\infty)/\epsilon_0))$, then we have $f_{n_1}(\boldsymbol{\beta}^{(t)}) - f_{n_1}(\widehat{\boldsymbol{\beta}}) \leq \epsilon_0$. Thus by Theorem 3 of Jain et al. (2014) it follows that with probability at lest $1 - \exp(-c_0 n_1) - \exp(-c_4 n_1) - c_2 p^{-c_3}$ (c_2, c_3 are universal constants) we have

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}\|_2 \leq C \left(\frac{s \log p}{n_1} \right)^{1/2} + (8\epsilon_0)^{1/2} \leq (9\epsilon_0)^{1/2}, \quad \text{for large } p.$$

C is a positive universal constant in the above inequality. If we set $\epsilon = 9\epsilon_0$, then Assumption 5.1 holds with α_p equal to $\exp(-c_0 n_1) + \exp(-c_4 n_1) + c_2 p^{-c_3}$ and $T(\epsilon, p, \boldsymbol{\beta}) = O(\log p +$

$\log((1 + \|\boldsymbol{\beta}\|_\infty)/\epsilon)))$.

Solving ℓ_1-regularized problem:

We start by recalling the definition of the loss function $f_{n_1,\lambda}(\theta) = n_1^{-1}\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1$, and define the minimizer of the loss function $\widehat{\boldsymbol{\beta}}_L := \arg\min_{\boldsymbol{\theta}} f_{n_1,\lambda}(\boldsymbol{\theta})$. First, we will prove Proposition 5.5.

Proof of Proposition 2.5.5:

Let us assume

$$f_{n_1,\lambda}(\widehat{\boldsymbol{\beta}}) - f_{n_1,\lambda}(\widehat{\boldsymbol{\beta}}_L) \leq \epsilon_0 < 1.$$

Now, we will establish the ℓ_2-error rate between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. We write $\widehat{\mathbf{b}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Since, $\widehat{\boldsymbol{\beta}}_L$ is optimal, we have

$$f_{n_1,\lambda}(\widehat{\boldsymbol{\beta}}) \leq f_{n_1,\lambda}(\widehat{\boldsymbol{\beta}}_L) + \epsilon_0 \leq f_{n_1,\lambda}(\boldsymbol{\beta}) + \epsilon_0.$$

Rearrangement of the above inequality yields

$$0 \leq \frac{1}{n_1}\|\mathbf{X}^{(1)}\widehat{\mathbf{b}}\|_2^2 \leq \frac{2\mathbf{w}^\top \mathbf{X}^{(1)}\widehat{\mathbf{b}}}{n_1} + \lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) + \epsilon_0. \tag{A.10}$$

Since, $\|\mathbf{X}_j^{(1)}\|_2^2 \sim \chi^2_{n_1}$, we have

$$\mathbb{P}\left(\underbrace{\max_{j\in[p]} n_1^{-1/2}\left\|\mathbf{X}_j^{(1)}\right\|_2 < \sqrt{4/3}}_{:=\mathcal{E}}\right) \geq 1 - 2p^{-2}, \quad \text{for large } p.$$

Using Gaussianity of $\mathbf{X}_j^{(1)\top}\mathbf{w}/\|\mathbf{X}_j^{(1)}\|_2$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n_1}\left\|\mathbf{X}^{(1)\top}\mathbf{w}\right\|_\infty > 2\sqrt{\frac{\log p}{n_1}}\right) &\leq \mathbb{P}\left(\max_{j\in[p]}\left|\frac{\mathbf{X}_j^{(1)\top}\mathbf{w}}{\left\|\mathbf{X}_j^{(1)}\right\|_2}\right| > \sqrt{3\log p}\right) + \mathbb{P}(\mathcal{E}^c) \\ &\leq 2p^{-0.5} + 2p^{-2}.\end{aligned}$$

Setting $\lambda = 8\{(\log p)/n_1\}^{1/2}$, we have $\lambda \geq 2\left\|\mathbf{X}^{(1)\top}\mathbf{w}\right\|_\infty/n_1$ with probability at least $1 - 2p^{-0.5} - 2p^{-2}$. Now, since $\boldsymbol{\beta}$ is s-sparse with support on $\mathcal{S}$, we have

$$\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{\beta}_\mathcal{S}\|_1 - \left\|\boldsymbol{\beta}_\mathcal{S} + \widehat{\mathbf{b}}_\mathcal{S}\right\|_1 - \left\|\widehat{\mathbf{b}}_{\mathcal{S}^c}\right\| \leq \left\|\widehat{\mathbf{b}}_\mathcal{S}\right\|_1 - \left\|\widehat{\mathbf{b}}_{\mathcal{S}^c}\right\|_1.$$

Substituting this in the basic inequality (A.10) and using $\lambda \geq 2 \left\|\mathbf{X}^{(1)\top}\mathbf{w}\right\|_\infty / n_1$, we get

$$\begin{aligned} 0 \leq \frac{1}{n_1}\left\|\mathbf{X}^{(1)}\widehat{\mathbf{b}}\right\|_2^2 &\leq 2\left\|\frac{E^\top \mathbf{X}^{(1)}}{n_1}\right\|_\infty \left\|\widehat{\mathbf{b}}\right\|_1 + \lambda(\left\|\widehat{\mathbf{b}}_{\mathcal{S}}\right\|_1 - \left\|\widehat{\mathbf{b}}_{\mathcal{S}^c}\right\|_1) + \epsilon_0 \\ &\leq (\lambda/2)\left\|\widehat{\mathbf{b}}\right\|_1 + \lambda(\left\|\widehat{\mathbf{b}}_{\mathcal{S}}\right\|_1 - \left\|\widehat{\mathbf{b}}_{\mathcal{S}^c}\right\|_1) + \epsilon_0 \\ &\leq (\lambda/2)\{3\left\|\widehat{\mathbf{b}}_{\mathcal{S}}\right\|_1 - \left\|\widehat{\mathbf{b}}_{\mathcal{S}^c}\right\|_1\} + \epsilon_0. \end{aligned} \tag{A.11}$$

Hence, we have

$$\left\|\widehat{\mathbf{b}}\right\|_1^2 = (\left\|\widehat{\mathbf{b}}_{\mathcal{S}}\right\|_1 + \left\|\widehat{\mathbf{b}}_{\mathcal{S}^c}\right\|_1)^2 \leq (4\left\|\widehat{\mathbf{b}}_{\mathcal{S}}\right\|_1 + (2\epsilon_0/\lambda))^2 \leq 32\left\|\widehat{\mathbf{b}}_{\mathcal{S}}\right\|_1^2 + \frac{8\epsilon_0^2}{\lambda^2}.$$

Now by Theorem 7.16 of Wainwright (2019), we have the following with probability at least $1 - 2\exp(-n_1/32)$:

$$\frac{\left\|\mathbf{X}^{(1)}\boldsymbol{\theta}\right\|_2^2}{n_1} \geq c_1 \left\|\boldsymbol{\theta}\right\|_2^2 - c_2\frac{\log p}{n_1}\left\|\boldsymbol{\theta}\right\|_1^2 \quad \text{for every } \theta \in \mathbb{R}^p,$$

where $c_1, c_2 > 0$ are universal constants. Using the above fact we have

$$\frac{\left\|\mathbf{X}^{(1)}\widehat{\mathbf{b}}\right\|_2^2}{n_1} \geq c_1\left\|\widehat{\mathbf{b}}\right\|_2^2 - 32c_2\frac{s\log p}{n_1}\left\|\widehat{\mathbf{b}}\right\|_2^2 - 8c_2\frac{\log p}{n_1\lambda^2}\epsilon_0^2 \geq \frac{c_1}{2}\left\|\widehat{\mathbf{b}}\right\|_2^2 - \epsilon_0^2,$$

when $32c_2 s\log p/(n_1) < c_1/2$ and $8c_2\log p/(n_1\lambda^2) < 1$. This is possible for large enough values of p and λ.

Case 1: If $(c_1/4)\left\|\widehat{\mathbf{b}}\right\|_2^2 > \epsilon_0^2$, then using (A.11), we get

$$\frac{c_1}{4}\left\|\widehat{\mathbf{b}}\right\|_2^2 \leq \frac{3\lambda\sqrt{s}}{2}\left\|\widehat{\mathbf{b}}\right\|_2 + \epsilon_0.$$

This bound involves a quadratic form of $\left\|\widehat{\mathbf{b}}\right\|_2$; computing the roots of the quadratic form we get the following bound:

$$\|b\|_2 \leq \underbrace{\frac{6\lambda\sqrt{s}}{c_1}}_{=O(\sqrt{(s\log p)/n_1})} + \frac{2\sqrt{\epsilon_0}}{\sqrt{c_1}}.$$

Case 2: If $(c_1/4)\left\|\widehat{\mathbf{b}}\right\|_2^2 \leq \epsilon_0^2$, then $\left\|\widehat{\mathbf{b}}\right\|_2 \leq 2\epsilon_0/\sqrt{c_1} < 2\sqrt{\epsilon_0/c_1}$. The last inequality uses the fact that $\epsilon_0 < 1$.

Combining the bounds obtained in Case 1 and Case 2 and with probability at least $1 - 2p^{-0.5} - 2p^{-2} - e^{-n_1/32}$(which is $\geq 1 - 3p^{-0.5}$ for large p), we finally have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq (C_3\epsilon_0)^{1/2}$ for large enough p and C_3 being an absolute constant. Now, using the reparameterization $\epsilon = C_3\epsilon_0$, we have $\epsilon < C_3$, and the initial sub-optimality gap turns out to be ϵ/C_3. This finishes the proof.

Proof for ETS-PICASSO:

First, we will show that the estimator generated by the PICASSO algorithm is a good estimate of $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}^{\text{picasso}}$ be the estimator obtained by applying PICASSO for minimizing $f_{n_1}(\theta; \lambda, D_1)$. Now define the largest and smallest s_0-sparse eigenvalues of $\mathbf{G} := \mathbf{X}^{(1)\top}\mathbf{X}^{(1)}/n_1$ as:

$$\rho_+(s_0) := \max_{\mathbf{v}:\|\mathbf{v}\|_0 \leq s} \frac{\mathbf{v}^\top \mathbf{G}\mathbf{v}}{\|\mathbf{v}\|_2^2}; \quad \text{and} \quad \rho_-(s_0) := \min_{\mathbf{v}:\|\mathbf{v}\|_0 \leq s} \frac{\mathbf{v}^\top \mathbf{G}\mathbf{v}}{\|\mathbf{v}\|_2^2}.$$

Let $\tilde{s}_\psi = (484\psi^2 + 100\psi)s$, where $\psi > 0$ is constant. An application of union bound and Equation 4.22 in Vershynin (2018a) yields that

$$\mathbb{P}\left\{ \max_{\mathcal{D}:|\mathcal{D}| \leq s + 2\tilde{s}_2} \left\| \frac{\mathbf{X}_{\mathcal{D}}^{(1)\top}\mathbf{X}_{\mathcal{D}}^{(1)}}{n_1} - \mathsf{I}_p \right\|_{\text{op}} \lesssim \frac{\log p}{p^{k/2}} \right\} \geq 1 - 2p^{-2}, \tag{A.12}$$

for large values of p. This ensures that for large values of p, we have

$$0.99 \leq \rho_-(s + 2\tilde{s}_2) \leq \rho_+(s + 2\tilde{s}_2) \leq 1.1. \tag{A.13}$$

Defining $\kappa := \rho_+(s + 2\tilde{s}_2)/\rho_-(s + 2\tilde{s}_2)$, we have $\kappa < 2$. Thus, Assumption 3.5 of Zhao et al. (2018) holds with $\tilde{s} = \tilde{s}_2 > \tilde{s}_\kappa$. Also, (A.12) shows that for large p

$$\mathbb{P}\left(\underbrace{\max_{j \in [p]} n_1^{-1/2} \left\| \mathbf{X}_j^{(1)} \right\|_2 < \sqrt{4/3}}_{:=\mathcal{E}} \right) \geq 1 - 2p^{-2}.$$

Hence, we have

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n_1} \left\| \mathbf{X}^{(1)\top}\mathbf{w} \right\|_\infty > 2\sqrt{\frac{\log p}{n_1}} \right) &\leq \mathbb{P}\left(\max_{j \in [p]} \left| \frac{\mathbf{X}_j^{(1)\top}\mathbf{w}}{\left\| \mathbf{X}_j^{(1)} \right\|_2} \right| > \sqrt{3\log p} \right) + \mathbb{P}(\mathcal{E}^c) \\ &\leq 2p^{-0.5} + 2p^{-2}. \end{aligned}$$

Hence, Assumption 3.1 of Zhao et al. (2018) holds with high probability when $\lambda_N = \lambda \geq 8\{\log p/n_1\}^{1/2}$, where N denotes the final iteration count of the outermost loop of PICASSO and λ_K denotes the regularization parameter at the Kth iteration of the outer loop. Also, in this case, $N = O(\log(\|\mathbf{X}^{(1)\top}\mathbf{y}^{(1)}/n_1\|_\infty \sqrt{n_1/\log p}))$ which follows from the description of PICASSO (see Algorithm 3 in Zhao et al. (2018)). From triangle inequality, it follows that

$$\left\| \frac{\mathbf{X}^{(1)\top}\mathbf{y}^{(1)}}{n_1} \right\|_\infty \leq \|\mathbf{G}\boldsymbol{\beta}\|_\infty + \left\|\mathbf{X}^{(1)\top}\mathbf{w}/n_1\right\|_\infty,$$

and $\|\mathbf{G}\boldsymbol{\beta}\|_\infty \leq \|\mathbf{G}\|_{\infty,\infty} \|\boldsymbol{\beta}\|_\infty \leq \sqrt{p}\, \|\mathbf{G}\|_{\text{op}} \|\boldsymbol{\beta}\|_\infty$. Note that

$$\|\mathbf{G}\|_{\text{op}} = \frac{p}{n_1} \left\| \frac{\mathbf{X}^{(1)}\mathbf{X}^{(1)\top}}{p} \right\|_{\text{op}}.$$

Thus, applying Equation (4.22) in Vershynin (2018a), we get $\|\mathbf{G}\boldsymbol{\beta}\|_\infty \lesssim (p^{1.5}/n)\, \|\boldsymbol{\beta}\|_\infty$ with probability at least $2\exp(-n_1)$. Hence, we have $N = O(\log p)$ with probability at least $1 - 2p^{-0.5} - 2p^{-2} - \exp(n_1)$.

Now we will make sure that Assumption 3.7 of Zhao et al. (2018) also holds. Before, going any further let us clarify some notations. At Kth outer iteration of PICASSO, Zhao et al. (2018) denotes the inner and middle loop precision parameters as τ_K and δ_K, and the active set initialization parameter is φ. To be consistent with our notation we set $\epsilon_0 = \delta_K$ for every K. Also, we choose the parameters in such a way so that

$$\epsilon_0 \leq \min\{1/8, C_3\}, \quad \tau_K \leq \frac{\epsilon_0}{\rho_+(s+2\tilde{s})} \sqrt{\frac{\rho_-(1)}{\rho_+(1)(s+2\tilde{s})}}, \quad \varphi \leq 1/8, \tag{A.14}$$

C_3 is the same constant ad in Proposition 5.5. Using (A.13), one can set $\tau_K = O(\epsilon_0/\sqrt{\log p})$ which will be less than 1 for large p. So, under the above conditions, Assumption 3.7 in Zhao et al. (2018) holds. Hence, part (iii) of Theorem 3.12 in Zhao et al. (2018) tells that

$$f_{n_1}(\widehat{\boldsymbol{\beta}}^{\text{picasso}}; \lambda, D_1) - f_{n_1}(\widehat{\boldsymbol{\beta}}_L; \lambda, D_1) \leq \epsilon_0 \frac{500\lambda^2 s}{11}.$$

If $\lambda = 8\sqrt{(\log p)/n_1}$, then for large p

$$f_{n_1}(\widehat{\boldsymbol{\beta}}^{\text{picasso}}; \lambda, D_1) - f_{n_1}(\widehat{\boldsymbol{\beta}}_L; \lambda, D_1) \leq \epsilon_0/C_3,$$

Hence, due to Proposition 5.5, we have $\|\widehat{\boldsymbol{\beta}}^{\text{picasso}} - \boldsymbol{\beta}\|_2^2 \leq \epsilon_0$. Also, using Theorem 3.12 and

Lemma 3.13 of Zhao et al. (2018), we get that PICASSO needs no more than

$$T(\epsilon_0, \boldsymbol{\beta}, p) = O\bigg((\log p + \log \|\boldsymbol{\beta}\|_\infty)(\log p)^3\{\log\log p + \log(\epsilon_0^{-1})\}\bigg).$$

But the above facts are true when $\epsilon_0 \leq \min\{1/8, C_3\}$. In order to extend the above results to a bigger range of ϵ_0 we first define $C_\epsilon := \min\{1/8, C_3\}/2$. If $\epsilon \leq C_\epsilon$, then one can get the same result by setting δ_K and τ_K appropriately as prescribed in (A.14). If $\epsilon > C_\epsilon$, then setting $\epsilon_0 = C_\epsilon$ and using (A.14), we again get $\|\widehat{\boldsymbol{\beta}}^{\mathsf{picasso}} - \boldsymbol{\beta}\|_2^2 \leq \epsilon_0 < \epsilon$ within $O((\log p + \log \|\boldsymbol{\beta}\|_\infty)(\log p)^3\{\log\log p + \log(C_\epsilon^{-1}))$ iterations. Thus, conditions in Assumption 5.1 is met with

$$T(\epsilon, p, \boldsymbol{\beta}) = O\bigg((\log p + \log \|\boldsymbol{\beta}\|_\infty)(\log p)^3\{\log\log p + \log(\epsilon^{-1} \vee C_\epsilon^{-1})\bigg),$$

and with probability at least $1 - O(p^{-0.5})$.

Proof for ETS-PGH:

Let $\widehat{\boldsymbol{\beta}}^{\mathsf{pgh}}$ be the solution obtained by minimizing $f_{n_1}(\boldsymbol{\theta}; \lambda, D_1)$ via PGH method. For the proof of this part we will use Theorem 3.2 of Xiao and Zhang (2013). To apply that theorem we need to make sure that Assumption 3.2 of Xiao and Zhang (2013) is satisfied. To avoid notational confusion, we use $\tilde{\gamma}$ and $\tilde{\delta}$ to denote the parameters γ, δ' considered in Xiao and Zhang (2013) respectively. Let $\tilde{\delta} = 0.1$ and $\tilde{\gamma} = 2$. By a similar argument as before, it can be shown that with probability at least $1 - 2p^2$

$$\kappa(G, s_0) := \frac{\rho_+(s_0)}{\rho_-(s_0)} \leq \frac{1+\nu}{1-\nu},$$

where $s_0 = \lfloor 46(1+\tilde{\gamma})s \rfloor$, $\nu = 0.1$ and p is sufficiently large. Also, we choose

$$\lambda = 8\max\{2, \frac{\tilde{\gamma}+1}{\tilde{\gamma}(1-\tilde{\delta}) - (1+\tilde{\delta})}\}\{\log p/n_1\}^{1/2} \geq 4\max\{2, \frac{\tilde{\gamma}+1}{\tilde{\gamma}(1-\tilde{\delta} - (1+\tilde{\delta}))}\}\left\|\mathbf{X}^{(1)^\top}\mathbf{w}/n_1\right\|_\infty. \tag{A.15}$$

The last inequality of the above display shows that $\lambda \geq 8\|\mathbf{w}^\top\mathbf{X}^{(1)}/n_1\|_\infty$ and it happens with the probability at least $1 - O(p^{-0.5})$. Hence, following the arguments of the second bullet point on page 10 of Xiao and Zhang (2013), we can conclude that the Assumption 3.2 of Xiao and Zhang (2013) holds with $\tilde{s} = \lfloor 22(1+\tilde{\gamma})s \rfloor$, $\gamma_{\text{inc}} = 1.2$ (see Xiao and Zhang (2013)), $L_{\min} = 1.32$ and λ. Using part 3 of Theorem 3.2 in Xiao and Zhang (2013), for a

given precision level ϵ_0, we have

$$f_{n_1}(\widehat{\boldsymbol{\beta}}^{\mathsf{pgh}}; \lambda, D_1) - f_{n_1}(\widehat{\boldsymbol{\beta}}_L; \lambda, D_1) \leq O(\epsilon_0 s\sqrt{(\log p)/n_1}) < \epsilon_0/C_3, \quad \text{for large } p.$$

C_3 is the same universal constant as in Proposition 5.5. If $\epsilon_0 = \min\{1, C_3\}/2 =: C_4$, then $\|\widehat{\boldsymbol{\beta}}^{\mathsf{pgh}} - \boldsymbol{\beta}\|_2^2 \leq \epsilon_0$, and the total iteration complexity is

$$O\left(\log p \log\log p + \log\max\{1, (\lambda^2/\epsilon_0^2)\log p\}\right),$$

which for large value of of p, the the form $O((\log p + \log\|\boldsymbol{\beta}\|_\infty)\log\log p + \log\max\{1, (1/\epsilon_0^2)\log p\})$ (as $\lambda < 1$). If $\epsilon_0 \leq C_4$, then the order becomes

$$O((\log p + \log\|\boldsymbol{\beta}\|_\infty)\log\log p + \log(1/\epsilon_0)).$$

Otherwise, i.e., if $\epsilon_0 > C_4$ one can set the tolerance level at $\epsilon = C_4$ and the overall order in that case is $O((\log p + \log\|\boldsymbol{\beta}\|_\infty)\log\log p + \log(1/C_4))$. Thus, for a given tolerance level ϵ, the total iteration complexity is $O((\log p + \log\|\boldsymbol{\beta}\|_\infty)\log\log p + \log(\epsilon^{-1} \vee C_4^{-1}))$. Thus, Assumption 5.1 holds with probability at least $1 - O(p^{-0.5})$ and $T(\epsilon, p, \boldsymbol{\beta}) = O((\log p + \log\|\boldsymbol{\beta}\|_\infty)\log\log p + \log(\epsilon^{-1} \vee C_4^{-1}))$.

APPENDIX B

Appendix for Chapter 3

B.1 Proof of main results under linear model

B.1.1 Proof of Lemma 3.2.1

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.

B.1.2 Proof of Proposition 3.4.2

In this section, we will show that the SRC condition (4.7) is strictly stronger than the condition in Assumption 1. Recall that the features are normalized, i.e., $\|\mathbf{X}_j\|_2 = \sqrt{n}$ for all $j \in [p]$. Now, we will prove the proposition.

Proof. **SRC implies Assumption 3.4.1:**

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 $\boldsymbol{\alpha}_1 \perp \text{col}(\mathbf{X}_{\mathcal{M}})$, 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

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

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

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

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

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 3.4.1 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

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

Now assume that

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

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

B.1.3 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

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

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

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

Also, let $\widetilde{\mathbf{w}} := (-\mathbf{w})$. 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:

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

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

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

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

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 $f_{\mathcal{D}} := \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\mathbf{w}}$ 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 Wainwright (2019)tells that there exists a constant $A_1 > 0$ such that

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

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

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

Writing A_1 as c_1, we get

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

Quadratic term: Here we study the quadratic supremum process in Equation (B.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,

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

Also, note that $\mathbb{E}(\mathbf{w}^{\top}\mathbf{P}_{\mathcal{D}|\mathcal{I}}\mathbf{w}) \leq k\sigma^2$ 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 $d_{\text{op}}(\mathcal{G}_{\mathcal{I}}) = 1$ and $d_F(\mathcal{G}_{\mathcal{I}}) = \sqrt{k}$ (defined in Section B.3). Also, it follows that the quantities M, V and U (defined in Theorem B.3.2) 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 B.3.2, we get that there exists a universal positive constants A_2, A_3, such that for $t = A_3 c_{\mathcal{G}} k\log(ep)$, we get

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

Due to Theorem 1.1 of Rudelson and Vershynin (2013), setting $u_0 = k\log(ep)/(2n)$ we can show that there exists an absolute constant $A_4 > 0$ such that

$$\begin{aligned}\mathbb{P}\left\{n^{-1}\left|\mathbf{w}^\top \mathbf{P}_{\mathcal{S}|\mathcal{I}}\mathbf{w} - k\sigma^2\right| > \frac{c_{\mathcal{G}}\sigma^2 k\log(ep)}{2n}\right\} &\leq 2\exp\left\{-A_4 c_{\mathcal{G}} k\log(ep)\right\} \\ &= 2(ep)^{-A_4 c_{\mathcal{G}} k},\end{aligned} \tag{B.14}$$

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

$$\mathbb{P}\left\{n^{-1}\max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}}\mathbf{w}^\top(\mathbf{P}_{\mathcal{D}} - \mathbf{P}_{\mathcal{S}})\mathbf{w} > c_2(\mathscr{E}^2_{\mathcal{G}_{\mathcal{I}}} + c_{\mathcal{G}})\sigma^2\frac{k\log(ep)}{n}\right\} \leq (ep)^{-2c_{\mathcal{G}}^2 k} + 2(ep)^{-A_4 c_{\mathcal{G}} k}, \tag{B.15}$$

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

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

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

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

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 (B.16):

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

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

B.1.4 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

$$\begin{aligned}
\mathbb{P}\left(R_\mathcal{D} - R_\mathcal{S} < 0\right) &= \mathbb{P}\left(\|\mathbf{y} - \mathbf{P}_\mathcal{D}\mathbf{X}_\mathcal{S}\boldsymbol{\beta}^*_\mathcal{S}\|_2^2 < \|\mathbf{y} - \mathbf{X}_\mathcal{S}\boldsymbol{\beta}^*_\mathcal{S}\|_2^2\right) \\
&= \mathbb{P}\left(\left\|\mathbf{w} + n^{1/2}\boldsymbol{\gamma}_\mathcal{D}\right\|_2^2 < \|\mathbf{w}\|_2^2\right) \\
&\geq \frac{1}{2}\frac{e^{-0.5}}{\sqrt{2\pi}} > 0.1 \quad \text{(By Lemma B.4.2)},
\end{aligned}$$

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

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

First decomposition:

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

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

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

Recall that

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

Next, by Sudakov's lower bound, we have

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

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

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

for any $c_{\mathcal{T}} > 0$. Choosing $c_{\mathcal{T}} = \mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}(2^{-1/2} - 2^{-1})$, and using the fact that $\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}} \geq 16\{\log(ep)\}^{-1}$, we get

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{C}_{j_0}} \widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^\top\widetilde{\mathbf{w}} \geq \sigma\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}\sqrt{\log(ep)}/2\right\} \geq 1 - e^{-1}. \tag{B.20}$$

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 $\widehat{\tau}_{j_0}^{1/2} \leq \sigma\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}\sqrt{\log(ep)}/(2n^{1/2})$, then we have

$$\min_{\mathcal{D}\in\mathscr{C}_{j_0}} \left[\|\boldsymbol{\gamma}_{\mathcal{D}}\|_2 \left\{\|\boldsymbol{\gamma}_{\mathcal{D}}\|_2 - 2n^{-1/2}\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^\top\widetilde{\mathbf{w}}\right\}\right] \leq -\frac{\sigma^2\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}\sqrt{\log(ep)}}{2n}$$

Thus, using (B.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 $\mathscr{E}^* \tau_{\mathcal{I}_0} \sqrt{\log(ep)} > 4$ (Assumption 2). Thus the necessary condition turns out to be

$$\widehat{\tau}_{j_0} \geq \frac{\mathscr{E}^{*2} \tau_{\mathcal{I}_0}}{4} \frac{\sigma^2 \log(ep)}{n}. \tag{B.21}$$

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 $\mathcal{S}$:

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

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

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

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

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

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

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

By Borell-TIS inequality (Adler et al., 2007, 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

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

$$\mathbb{E}(\|f\|) \leq 4\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}}\sigma\sqrt{\log(ep)},$$

which ultimately yields

$$\mathbb{P}\left\{\|f\| \geq (4\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} + c_{\mathcal{T}})\sigma\sqrt{\log(ep)}\right\} \leq (ep)^{-c_{\mathcal{T}}^2/2}.$$

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

$$\mathbb{P}\left\{\max_{\mathcal{D}\in\mathscr{C}_{j_0}}\left|\widehat{\boldsymbol{\gamma}}_{\mathcal{D}}^{\top}\widetilde{\mathbf{w}}\right| \geq (4\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} + c_{\mathcal{T}})\sigma\sqrt{\log(ep)}\right\} \leq 2(ep)^{-c_{\mathcal{T}}^2/2}. \tag{B.26}$$

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

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

For the maximal process we will use Theorem 2.10 of Adamczak (2015). We begin with the definition of concentration property.

Definition B.1.1 (Adamczak (2015))**.** *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 $\mathbf{w}/\sigma$ enjoys concentration property with $K = \sqrt{2}$ (Boucheron et al., 2013, Theorem 5.6). Let $Q_1 := \max_{j\notin\mathcal{S}} \mathbf{w}^{\top}(\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^{\top} - \widehat{\mathbf{u}}_{j_0}\widehat{\mathbf{u}}_{j_0}^{\top})\mathbf{w}$. By Theorem 2.10 of Adamczak (2015) 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

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

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

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

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

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

By Sudakov's lower bound, we have

$$\begin{aligned}\mathbb{E}\max_{j\in\mathcal{S}^c} (\widehat{\mathbf{u}}_j^\top\mathbf{w}) \vee (-\widehat{\mathbf{u}}_j^\top\mathbf{w}) &\geq \sup_{\delta>0} \frac{\sigma\delta}{2}\sqrt{\log \mathcal{M}(\delta, \mathcal{U}_{\text{sym}}, \|\cdot\|_2)} \\ &\geq \sup_{\delta>0} \frac{\sigma\delta}{2}\sqrt{\log \mathcal{M}(\delta, \{\widehat{\mathbf{u}}_j\widehat{\mathbf{u}}_j^\top\}_{j\in\mathcal{S}^c}, \|\cdot\|_{\text{op}})} \\ &\geq \sigma\frac{\mathscr{E}^*_{\mathcal{G}_{\mathcal{I}_0}}}{\sqrt{4/3}}\sqrt{\log(ep)}.\end{aligned}$$

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

$$\mathbb{E}(Q_1) \geq \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}\log(ep)}{4/3} - \sigma^2.$$

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

$$\mathbb{P}\left[Q_1 \geq (3/4)\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}\log(ep) - \sigma^2 - 2\delta\sigma^2\log(ep)\right] \geq 1 - 2(ep)^{-\delta/2}.$$

Setting $\delta = \frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{8}$, we get

$$\mathbb{P}\left[Q_1 \geq \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{2}\log(ep) - \sigma^2\right] \geq 1 - 2(ep)^{-\frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{16}}.$$

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

$$\begin{aligned}&\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{S}})\\ &\leq \widehat{\tau}_{j_0} + 2\widehat{\tau}_{j_0}^{1/2}(4\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} + c_{\mathcal{T}})\sigma\sqrt{\frac{\log(ep)}{n}} - n^{-1}\left(\frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{2}\log(ep) - \sigma^2\right)\end{aligned} \tag{B.29}$$

with probability at least $1 - 2(ep)^{-c_{\mathcal{T}}^2/2} - 2(ep)^{-\frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{16}}$. Thus, for large p we have

$$\begin{aligned}&\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{D}})\\ &\leq \widehat{\tau}_{j_0} + 2\widehat{\tau}_{j_0}^{1/2}(4\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} + c_{\mathcal{T}})\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{4}\frac{\log(ep)}{n}\end{aligned}$$

with probability at least $1 - 2(ep)^{c_{\mathcal{T}}^2/2} - 2(ep)^{-\frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{16}}$. Now choose $c_{\mathcal{T}} = 4/\sqrt{\log(ep)}$ and use Assumption 2 to get

$$\begin{aligned}&\min_{\mathcal{D}\in\mathscr{C}_{j_0}} n^{-1}(R_{\mathcal{D}} - R_{\mathcal{D}})\\ &\leq \widehat{\tau}_{j_0} + 2\widehat{\tau}_{j_0}^{1/2}\left(4\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} + \frac{4}{\sqrt{\log(ep)}}\right)\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{4}\frac{\log(ep)}{n}\\ &\leq \widehat{\tau}_{j_0} + 8\sqrt{2}\widehat{\tau}_{j_0}^{1/2}\left(\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} + \frac{1}{\sqrt{2\log(ep)}}\right)\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{4}\frac{\log(ep)}{n}\\ &\leq \widehat{\tau}_{j_0} + 10\sqrt{2}\widehat{\tau}_{j_0}^{1/2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}}\sigma\sqrt{\frac{\log(ep)}{n}} - \frac{\sigma^2\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{4}\frac{\log(ep)}{n}\end{aligned}$$

with probability at least 1/5. Thus, in light of (B.23), the following is necessary:

$$\widehat{\tau}_{j_0} \geq \left\{\frac{\sqrt{200\mathscr{E}^2_{\mathcal{T}_{\mathcal{I}_0}} + \mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}} - 10\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}}}{2}\right\}^2 \frac{\sigma^2\log(ep)}{n}. \tag{B.30}$$

Case 1: If $\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} \leq \mathscr{E}^*{}_{\mathcal{G}_{\mathcal{I}_0}}$, then the right hand side of (B.30) is lower bounded by

$$\frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{(\sqrt{201} + 10\sqrt{2})^2}\frac{\sigma^2\log(ep)}{n}.$$

Thus (B.30) yields the necessary condition

$$\widehat{\tau}_{j_0} \geq \frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{(\sqrt{201}+10\sqrt{2})^2}\frac{\sigma^2\log(ep)}{n}.$$

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

$$\widehat{\tau}_{j_0} \geq \tilde{C}_1 \max\{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}\}\frac{\sigma^2\log(ep)}{n},$$

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

Case 2: If $\mathscr{E}^*{}_{\mathcal{G}_{\mathcal{I}_0}} \leq \mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}$, then using the inequality $\sqrt{1+t}-\sqrt{t}<1$ for all $t>0$, we can conclude that the right hand side of (B.30) is always smaller than

$$\frac{\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}}{4}\frac{\sigma^2\log(ep)}{n},$$

which is further smaller than

$$\frac{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}}{4}\frac{\sigma^2\log(ep)}{n}.$$

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

$$\widehat{\tau}_{j_0} \geq \frac{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}}{4}\frac{\sigma^2\log(ep)}{n} = \max\{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}\}\frac{\sigma^2\log(ep)}{4n}.$$

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:

$$\widehat{\tau}_{j_0} \geq C_1 \max\{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}\}\frac{\sigma^2\log(ep)}{n}, \quad \text{if } \mathscr{E}^*{}_{\mathcal{G}_{\mathcal{I}_0}} \notin (\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}}), \tag{B.31}$$

or,

$$\widehat{\tau}_{j_0} \geq C_2 \max\left\{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}, \left(\sqrt{200\mathscr{E}^2_{\mathcal{T}_{\mathcal{I}_0}} + \mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}} - 10\sqrt{2}\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}}\right)^2\right\}\frac{\sigma^2\log(ep)}{n},$$

if $\mathscr{E}^*{}_{\mathcal{G}_{\mathcal{I}_0}} \in (\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}})$.

If there exists $\mathcal{I}_0 \in \mathcal{J}$ such that $\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}/\mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}} \in (\alpha, 1)$, then in the preceding case it follows that the last display can be simplified in the following form:

$$\widehat{\tau}_{j_0} \geq C_2 \max\left\{\mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}, A_\alpha\mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}}\right\}\frac{\sigma^2\log(ep)}{n}, \quad \text{if } \mathscr{E}^*{}_{\mathcal{G}_{\mathcal{I}_0}} \in (\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}}),$$

where

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

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

$$\widehat{\tau}_{j_0} \geq C_2 A_\alpha \max \left\{ \mathscr{E}^{*2}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}^{*2}_{\mathcal{G}_{\mathcal{I}_0}} \right\} \frac{\sigma^2 \log(ep)}{n}. \tag{B.32}$$

Since, (B.31) and (B.32) hold for all choices of j_0 and $\mathcal{I}_0$ depending on whether the $\mathscr{E}^*{}_{\mathcal{G}_{\mathcal{I}_0}} \in (\mathscr{E}^*{}_{\mathcal{T}_{\mathcal{I}_0}}, \mathscr{E}_{\mathcal{T}_{\mathcal{I}_0}})$ is satisfied or not, the claim follows. □

B.1.5 Correlated random feature model example

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

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

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

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

We start with providing an upper bound on the margin quantity $\widehat{\tau}$. Using Equation (B.53) and (B.54), for any $\varepsilon_{n,p} \in (0,1)$ we get

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

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

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

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

Let $\mathbf{X}_j=(x_{1,j},\ldots,x_{n,j})^\top$ for all $j\in[p]$. Now note 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

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

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

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

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

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

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

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

Note that

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

and

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

Next, we consider the event

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

Due to (B.33), (B.34), (B.36) and (B.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

$$\Phi(\mathbf{u}, \mathbf{t}) := \frac{u_1 - (t_1^2/u_2) - (t_2^2/u_3) + (t_1 t_2 t_3)/(u_2 u_3)}{\sqrt{u_1 - t_1^2/u_2}\sqrt{u_1 - t_2^2/u_3}},$$

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

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

Noting the fact that

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

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

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

Applying Lemma B.4.4 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

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

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

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

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,

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

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

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

B.2 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

$$\begin{aligned}
\delta\mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D} + \boldsymbol{\Delta}; \bar{\boldsymbol{\beta}}_\mathcal{D}) &:= \mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D} + \boldsymbol{\Delta}) - \mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D}) - \nabla\mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D})^\top \boldsymbol{\Delta} \\
&= \frac{1}{2}\boldsymbol{\Delta}^\top \nabla^2 \mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D} + t\boldsymbol{\Delta})\boldsymbol{\Delta} \quad \text{(for some } t \in (0,1)) \\
&= \frac{1}{n}\sum_{i=1}^n b''(\mathbf{x}_{i,\mathcal{D}}^\top(\bar{\boldsymbol{\beta}}_\mathcal{D} + t\boldsymbol{\Delta}))(\mathbf{x}_{i,\mathcal{D}}^\top \boldsymbol{\Delta})^2 \\
&\geq \psi(x_0 R + x_0 r)\kappa_0 \|\boldsymbol{\Delta}\|_2^2 \quad \text{(Using Assumption 3.5.1(a), 3.5.1(b), 3.5.1(d))} \\
&\geq \psi(x_0 R + x_0 R_0)\kappa_0 \|\boldsymbol{\Delta}\|_2^2
\end{aligned}$$

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

$$\psi(x_0 R + x_0 R_0)\|\boldsymbol{\Delta}_\alpha\|_2^2 \leq \delta\mathcal{L}_\mathcal{D}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},\alpha}, \bar{\boldsymbol{\beta}}_\mathcal{D}) \leq -\nabla\mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D})^\top \boldsymbol{\Delta}_\alpha \leq \left\|\nabla\mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D})\right\|_2 \|\boldsymbol{\Delta}_\alpha\|_2.$$

Hence we have

$$\|\boldsymbol{\Delta}_\alpha\|_2 \leq \frac{\left\|\nabla\mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D})\right\|_2}{\psi(x_0 R + x_0 R_0)} \leq \frac{\sqrt{s}\left\|\nabla\mathcal{L}_\mathcal{D}(\bar{\boldsymbol{\beta}}_\mathcal{D})\right\|_\infty}{\psi(x_0 R + x_0 R_0)}. \tag{B.41}$$

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 $\mathbb{E}\{\exp(\lambda\epsilon_i[\mathbf{x}_{i,\mathcal{D}}]_j) \le \exp(\lambda^2 x_0^2/2)\}$, 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

$$\mathbb{P}\left(\left\|\nabla \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_\infty \ge 2tx_0(\phi B)^{1/2}\right) \le 2s\exp\left(-\frac{nt^2}{2}\right). \tag{B.42}$$

Setting $t = 4(\log n/n)^{1/2}$ in (B.42) we get

$$\mathbb{P}\left(\left\|\nabla \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})\right\|_\infty \ge 8x_0(\phi B)^{1/2}\sqrt{\frac{\log n}{n}}\right) \le \frac{2}{n^7}. \tag{B.43}$$

Using the above fact and (B.41) 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

$$\mathbb{P}\left(\max_{\mathcal{D}\in\mathscr{A}_s}\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}}\right\|_2 > \frac{9x_0(\phi B)^{1/2}}{\psi(x_0R + x_0R_0)}\sqrt{\frac{s\log n}{n}}\right) \le \frac{4s\log p}{n^7}. \tag{B.44}$$

By a similar argument, it can be shown that

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\right\|_2 > \frac{9x_0(\phi B)^{1/2}}{\psi(x_0R + x_0R_0)}\sqrt{\frac{s\log n}{n}}\right) \le \frac{2}{n^7}. \tag{B.45}$$

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

$$\widehat{\boldsymbol{\beta}}_{\mathcal{D}} = \bar{\boldsymbol{\beta}}_{\mathcal{D}} - [\nabla^2 \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}})]^{-1} \left(\nabla \mathcal{L}_{\mathcal{D}}(\bar{\boldsymbol{\beta}}_{\mathcal{D}}) + \mathbf{R}_{\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}})^{\otimes 2} \right) \tag{B.46}$$

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 (B.46) 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) \\
&\quad + (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) \\
&\quad + (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}}) - \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}) \\
&\quad + (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 $\widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}}$ be the orthogonal projector onto the $\mathrm{col}([\widetilde{\mathbf{X}}_{\mathcal{D}}]_{\mathcal{I}})$. 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}}_{\mathcal{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}}_{\mathcal{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}}_{\mathcal{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}}_{\mathcal{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{B.47}
$$

Now recall that

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

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

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 Wainwright (2019). First, note that $\mathsf{diam}(\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)}) \le \sqrt{2}$ 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

$$\begin{aligned} &\mathbb{P}\left\{ \max_{\mathcal{D}\in\mathscr{A}_{\mathcal{I}}} \widehat{\boldsymbol{r}}_{\mathcal{D}}^\top \boldsymbol{\epsilon} \ge A_1(\mathscr{E}_{\widetilde{\mathcal{T}}_{\mathcal{I}}^{(s)}} \sqrt{k\log(ep)} + \sqrt{2k\{\log(es) \vee \log\log(ep)\}}) \right\} \\ &\le 3\{(es) \vee \log(ep)\}^{-2k}, \end{aligned} \tag{B.48}$$

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 B.3.2. By the properties

of projection matrices, we have $d_{\text{op}}(\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}) = 1$ and $d_F(\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}) = \sqrt{k}$. Hence, equipped with Assumption 3.5.1(e), the quantities M, V and U (defined in Theorem B.3.2) has the following properties:

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

Due to Theorem B.3.2, 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

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

for a universal positive constant A_2. As $\max_{\mathcal{D} \in \mathscr{A}_{\mathcal{I}}} \mathbb{E}\{\boldsymbol{\xi}^\top (\widetilde{\mathbf{P}}_{\mathcal{D}} - \widetilde{\mathbf{P}}_{\mathcal{I}|\mathcal{D}})\boldsymbol{\xi}\} \leq k\sigma_0^2 \leq k\sigma_0^2 \mathscr{E}^2_{\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}} \log(ep)$, we finally have

$$\begin{aligned}
&\mathbb{P}\left(Q_{\mathscr{A}_s} \leq A_4 \mathscr{E}^2_{\widetilde{\mathcal{G}}_{\mathcal{I}}^{(s)}} k \log(ep) + A_3 k \sqrt{\log(ep)\{\log(es) \vee \log\log(ep)\}}\right) \\
&\leq \{(es) \vee \log(ep)\}^{-2k},
\end{aligned} \tag{B.49}$$

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,

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

By Theorem 1 of Rudelson and Vershynin (2013), 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

$$\mathbb{P}\left[\boldsymbol{\epsilon}^\top \mathbf{P}_{\mathcal{I}}\boldsymbol{\epsilon} \geq A_6 s \max\{\log(es), \log\log(ep)\}\right] \leq \{\log(ep)\}^{-2s} \quad \text{for all } \mathcal{D} \in \mathscr{A}_s \cup \{\mathcal{S}\}. \quad \text{(B.50)}$$

Final 0-1 error bound

Define the event

$$\Omega = \left\{ \max_{\mathcal{D} \in \mathscr{A}_s \cup \{\mathcal{S}\}} \left\| \widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \bar{\boldsymbol{\beta}}_{\mathcal{D}} \right\|_2 \leq \frac{9x_0(\phi B)^{1/2}}{\psi(x_0 R + x_0 R_0)} \sqrt{\frac{s \log n}{n}} \right\}$$

.

By (B.44) and (B.45) 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 following

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

where

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

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

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

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

where $\psi_{**} = \psi(x_0 R + x_0 R_0)$. Now, recall the property (3.15) of $\Delta_{\mathrm{kl}}(\mathcal{D})$. Thus, for the

aforementioned condition to hold for $\widetilde{\tau}_{\text{glm}}(s)$, it is sufficient to have

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

Under the above inequality and due to (B.48), (B.49), and (B.50), we can finally conclude

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

B.3 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 Krahmer et al. (2014) and Banerjee et al. (2019). 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 B.3.1. *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 $\mathcal{N}(\epsilon,T,d)$ by the well-known Dudley's integral (See Talagrand (2005)). 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

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

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

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

Theorem B.3.2 (Banerjee et al. (2019)). *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*

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

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.

B.4 Technical lemmas

Lemma B.4.1 (Gordon (1941)). *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 B.4.2. *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 B.4.1 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 B.4.3 (Laurent and Massart (2000))**.** *Let W be chi-squared random variable with degrees of freedom m. Then, we have the following large-deviation inequalities for all $x > 0$*

$$\mathbb{P}(W - m > 2\sqrt{mx} + 2x) \leq \exp(-x), \quad \textit{and} \tag{B.51}$$

$$\mathbb{P}(W - m < -2\sqrt{mx}) \leq \exp(-x). \tag{B.52}$$

If we set $x = mu$ in Equation (B.51) 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

$$\mathbb{P}\left(\frac{W}{m} - 1 \geq \delta\right) \leq \exp(-m\delta^2/16). \tag{B.53}$$

Similarly, setting $x = mu$ and $u = \delta^2/4$ in Equation (B.52), we get

$$\mathbb{P}\left(\frac{W}{m} - 1 \leq -\delta\right) \leq \exp(-m\delta^2/4). \tag{B.54}$$

Lemma B.4.4. *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 $f_\delta(x) \leq f_\delta(\delta) = \sqrt{2\delta}$. □

APPENDIX C

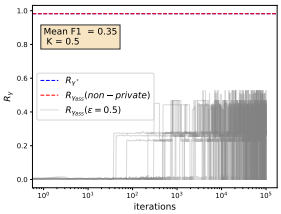
Appendix for Chapter 4

C.1 More simulation details

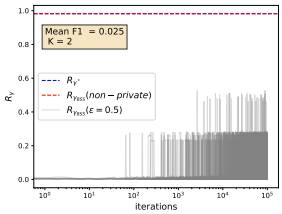
Under the setup of Section 4.5, we consider the privacy parameter $\varepsilon \in \{0.5, 1, , 3, 5, 10\}$. For the Metropolis-Hastings random walk, we vary $K \in \{0.5, 2, 3, 3.5\}$ and initialize 10 independent Markov chains from random initializations and record the `F-score` of the last iteration. We also track the qualities of the model through its explanatory power. In particular, we calculate the scale factor $R_\gamma := \mathbf{y}^\top \mathbf{\Phi}_\gamma \mathbf{y} / \|\mathbf{y}\|_2^2$ for each model $\gamma \in \{\gamma_t\}_{t \geq 1}$ along the random walks. Typically, a high value of R_γ will indicate the superior quality of the model γ. Note that $-\|\mathbf{y}\|_2^2 (1 - R_\gamma)$ is proportional to the log of the probability mass function function of γ. Thus, tracking R_γ is equivalent to tracking the log-likelihood of γ along the random walks.

Strong signal: Under this setup, note that the model estimate of ABESS exactly matches the true model. For $\varepsilon \geq 3$ and $K \geq 2$, Figure C.1 shows that all the chains have identified a reasonably good estimate of the true model γ^* within $50p$ iterations. This empirical phenomenon validates theoretical findings in Theorem 4.4.3. However, for larger values of K the performance is worse as the noise level is also large. On the other hand, for the case of $K = 0.5$, the performance is also worse due to too much shrinkage that results in a bad estimate of $\boldsymbol{\beta}$. The mean `F-score`'s also suggest the same phenomenon. For smaller values of ε, the performance is generally bad due to increased noise level. This is expected as higher privacy usually entails a worse performance in terms of utility.

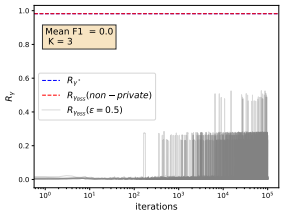
Weak signal: We perform the same experiments under a weak signal regime. As expected, both Figure C.2 and Table 4.1 show that the performance of the proposed algorithm is generally inferior to that in the strong signal regime for $K \geq 2$. However, note that our algorithm enjoys a better utility for $K = 0.5$. In fact, performance is as good as the non-private BSS for $\varepsilon \geq 3$. This is not surprising as $K = 0.5$ closer to $\|\boldsymbol{\beta}\|_1 \approx 0.7$ in the weak



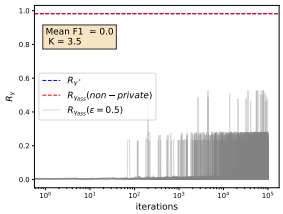
(a) $\varepsilon = 0.5, K = 0.5$



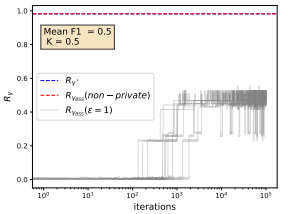
(b) $\varepsilon = 0.5, K = 2$



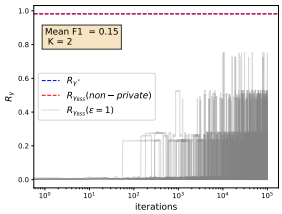
(c) $\varepsilon = 0.5, K = 3$



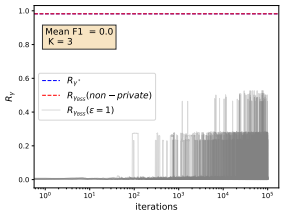
(d) $\varepsilon = 0.5, K = 3.5$



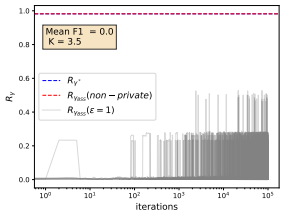
(e) $\varepsilon = 1, K = 0.5$



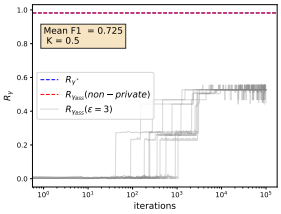
(f) $\varepsilon = 1, K = 2$



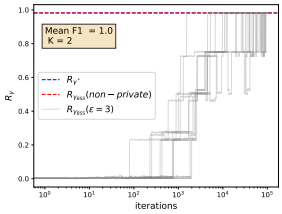
(g) $\varepsilon = 1, K = 3$



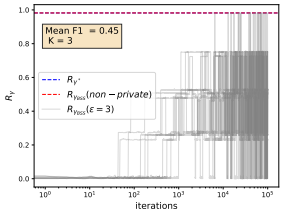
(h) $\varepsilon = 1, K = 3.5$



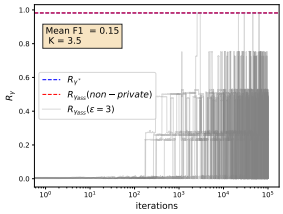
(i) $\varepsilon = 3, K = 0.5$



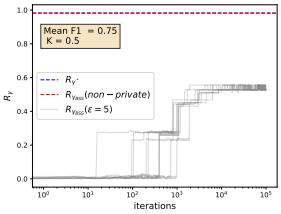
(j) $\varepsilon = 3, K = 2$



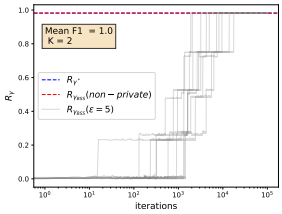
(k) $\varepsilon = 3, K = 3$



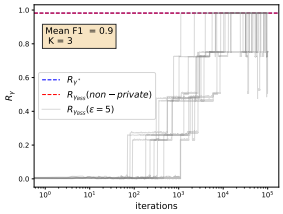
(l) $\varepsilon = 3, K = 3.5$



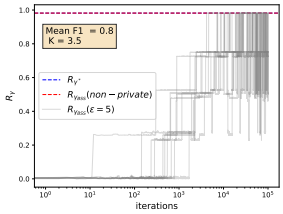
(m) $\varepsilon = 5, K = 0.5$



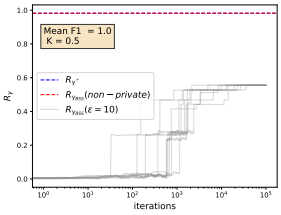
(n) $\varepsilon = 5, K = 2$



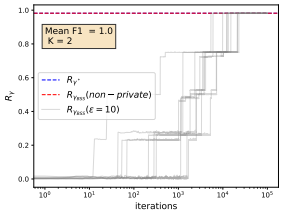
(o) $\varepsilon = 5, K = 3$



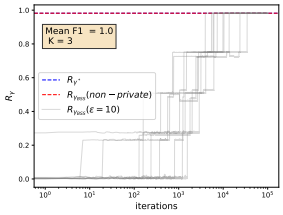
(p) $\varepsilon = 5, K = 3.5$



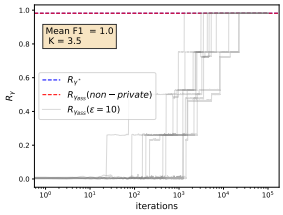
(q) $\varepsilon = 10, K = 0.5$



(r) $\varepsilon = 10, K = 2$



(s) $\varepsilon = 10, K = 3$



(t) $\varepsilon = 10, K = 3.5$

Figure C.1: Metropolis-Hastings random walk under different privacy budgets and ℓ_1 regularization. (Strong signal)

signal case and results in better estimation for $\boldsymbol{\beta}$. On the contrary, larger values of K inject more noise into the algorithm and the utility deteriorates.

C.2 Proof of Main results

C.2.1 Proof of Lemma 4.3.1

Recall that $u_K(\gamma; \mathbf{X}, \mathbf{y}) = -L_{\gamma,K}(\mathbf{X}, \mathbf{y})$ where $L_{\gamma,K}(\mathbf{X}, \mathbf{y}) := \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta})^2$. Therefore, it suffices to show that $L_{\gamma,K}(\cdot,\cdot)$ is data monotone. Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and $D' = D \cup \{\mathbf{x}_{n+1}, y_{n+1}\}$. We define

$$\widehat{\boldsymbol{\beta}}_{n,\gamma} := \arg\min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_1 \le K} \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta})^2,$$

$$\widehat{\boldsymbol{\beta}}_{n+1,\gamma} := \arg\min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_1 \le K} \sum_{i=1}^{n+1}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta})^2.$$

Therefore, we have the following inequalities:

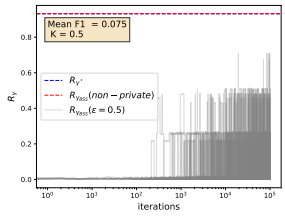
$$L_{\gamma,K}(D') = \sum_{i=1}^{n+1}(y_i - \mathbf{x}_{i,\gamma}^\top \widehat{\boldsymbol{\beta}}_{n+1,\gamma})^2 \ge \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \widehat{\boldsymbol{\beta}}_{n+1,\gamma})^2 \ge \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \widehat{\boldsymbol{\beta}}_{n,\gamma})^2 = L_{\gamma,K}(D).$$

The above inequalities conclude the proof.

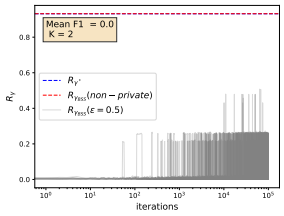
C.2.2 Proof of Utility Guarantee (Theorem 4.3.5)

Consider the notation in Section C.3.2 and recall the event $\mathcal{E}_K := \cap_{\gamma:\gamma\in\mathscr{A}_s}\{\boldsymbol{\beta}_{\gamma,K} = \boldsymbol{\beta}_{\gamma,ols}\}$. We will For notational brevity, we use L_γ to denote $L_\gamma(\mathbf{X}, \mathbf{y})$. Now, we restrict ourselves to the event $\mathcal{E}_K$. therefore we have $L_{\gamma,K} = L_\gamma$ for all γ. To establish a lower bound $\pi(\gamma^*)$, we make use of its specific form, thereby obtaining the following inequality:

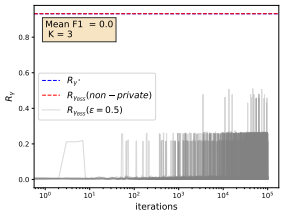
$$\pi(\gamma^*) = \frac{1}{1 + \sum_{\gamma'\in\mathscr{A}_s} \exp\left\{-\frac{\varepsilon(L_{\gamma'} - L_{\gamma^*})}{\Delta u}\right\}}.$$



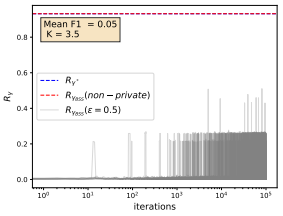
(a) $\varepsilon = 0.5, K = 0.5$



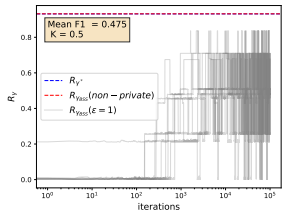
(b) $\varepsilon = 0.5, K = 2$



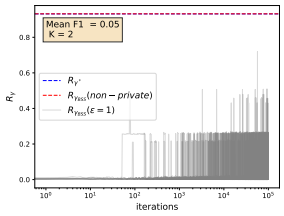
(c) $\varepsilon = 0.5, K = 3$



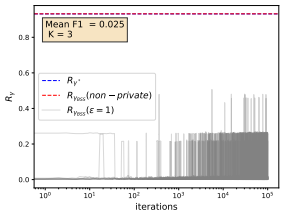
(d) $\varepsilon = 0.5, K = 3.5$



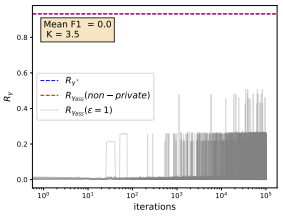
(e) $\varepsilon = 1, K = 0.5$



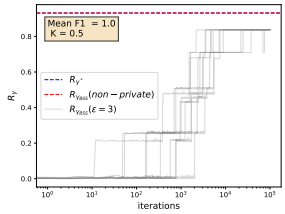
(f) $\varepsilon = 1, K = 2$



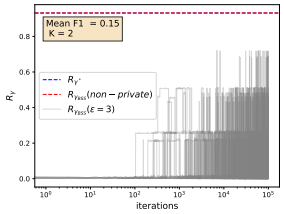
(g) $\varepsilon = 1, K = 3$



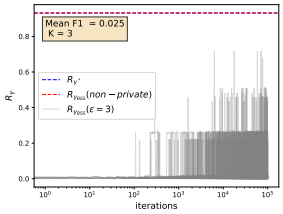
(h) $\varepsilon = 1, K = 3.5$



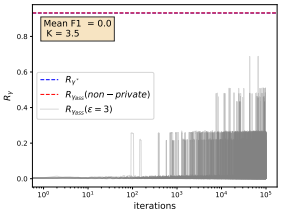
(i) $\varepsilon = 3, K = 0.5$



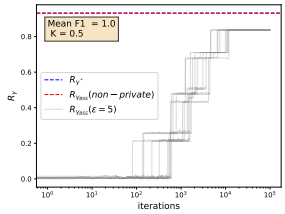
(j) $\varepsilon = 3, K = 2$



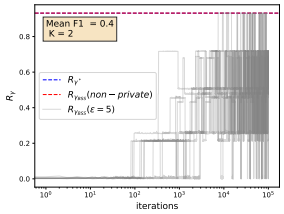
(k) $\varepsilon = 3, K = 3$



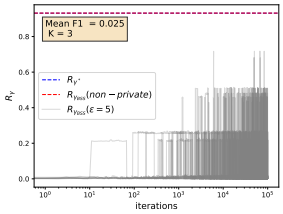
(l) $\varepsilon = 3, K = 3.5$



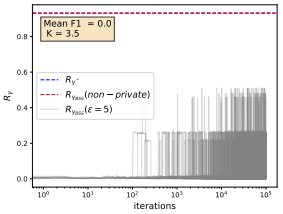
(m) $\varepsilon = 5, K = 0.5$



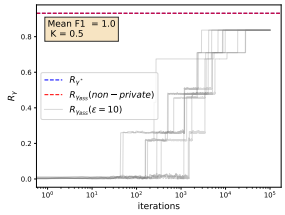
(n) $\varepsilon = 5, K = 2$



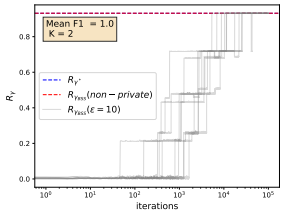
(o) $\varepsilon = 5, K = 3$



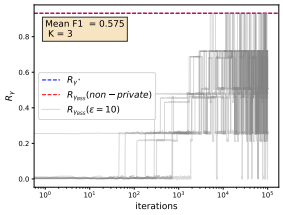
(p) $\varepsilon = 5, K = 3.5$



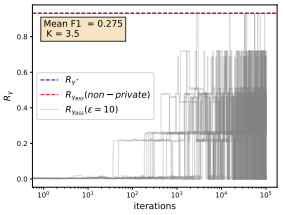
(q) $\varepsilon = 10, K = 0.5$



(r) $\varepsilon = 10, K = 2$



(s) $\varepsilon = 10, K = 3$



(t) $\varepsilon = 10, K = 3.5$

Figure C.2: Metropolis-Hastings random walk under different privacy budgets and ℓ_1 regularization. (Weak signal)

Now we fix $k \in [s]$, and and consider any $\gamma \in \mathscr{A}_{s,k}$. For any $\eta \in [0,1]$, note that

$$\begin{aligned}
&n^{-1}(L_\gamma - L_{\gamma^*}) = n^{-1}\{\mathbf{y}^\top(\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{y} - \mathbf{y}^\top(\mathbb{I}_n - \mathbf{\Phi}_{\gamma^*})\mathbf{y}\} \\
&= n^{-1}\left\{(\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma} + \mathbf{w})^\top(\mathbb{I}_n - \mathbf{\Phi}_\gamma)(\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma} + \mathbf{w}) - \mathbf{w}^\top(\mathbb{I}_n - \mathbf{\Phi}_{\gamma^*})\mathbf{w}\right\} \\
&= \eta\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} + 2^{-1}(1-\eta)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} - 2\left\{n^{-1}(\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}\right\}^\top(-\mathbf{w}) \\
&\quad + 2^{-1}(1-\eta)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} - n^{-1}\mathbf{w}^\top(\mathbf{\Phi}_\gamma - \mathbf{\Phi}_{\gamma^*})\mathbf{w}.
\end{aligned}$$

Consider the random variable Following the analysis of Guo et al. (2020), we have

$$\mathbb{P}\left[\max_{\gamma\in\mathscr{A}_{s,k}}\left|2n^{-1}\{(\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}\}^\top\mathbf{w}\right| \geq 2^{-1}(1-\eta)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}\right] \leq 2e^{-6k\log p},$$

and,

$$\mathbb{P}\left[\max_{\gamma\in\mathscr{A}_{s,k}} n^{-1}\left|\mathbf{w}^\top(\mathbf{\Phi}_\gamma - \mathbf{\Phi}_{\gamma^*})\mathbf{w}\right| \geq 2^{-1}(1-\eta)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}\right] \leq 4e^{-2k\log p},$$

whenever

$$\frac{\min_{\gamma\in\mathscr{A}_{s,k}}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}}{k} \geq C\sigma^2\left\{\frac{\log p}{n(1-\eta)}\right\}$$

for large enough universal constant $C > 0$. Setting $\eta = 1/2$, we note that whenever $\mathfrak{m}_*(s) \geq 2C\sigma^2\{(\log p)/n\}$, we get

$$n^{-1}(L_\gamma - L_{\gamma^*}) \geq \frac{1}{2}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} \geq \frac{k\mathfrak{m}_*(s)}{2} \quad \text{for all } \gamma \in \mathscr{A}_s,$$

with probability at least $1 - 2p^{-6} - 4p^{-2}$. Also, note that $\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} \leq \kappa_+ s b_{\mathsf{max}}^2$. Hence, if we have

$$\mathfrak{m}_*(s) \geq \max\left\{2C, \frac{16\Delta u}{\varepsilon\sigma^2}\right\}\frac{\sigma^2\log p}{n},$$

the following are true:

$$\begin{aligned}
&\sum_{\gamma'\in\mathscr{A}_s}\exp\left\{-\frac{\varepsilon(L_{\gamma'} - L_{\gamma^*})}{\Delta u}\right\} \\
&\leq \sum_{\gamma'\in\mathscr{A}_s}\exp\left\{-\frac{nk\varepsilon\mathfrak{m}_*(s)}{2\Delta u}\right\} \\
&\leq \sum_{k=1}^{s}\binom{p-s}{k}\binom{s}{k}\exp\left\{-\frac{nk\varepsilon\mathfrak{m}_*(s)}{2\Delta u}\right\} \\
&\leq \sum_{k=1}^{s} p^{2k}.p^{-4k} \leq p^{-2}.
\end{aligned}$$

Therefore, we have

$$\min_{\gamma \in \mathscr{A}_s \cup \{\gamma^*\}} \pi(\gamma) \geq \frac{1}{1+p^{-2}} \geq 1 - p^{-2}$$

with probability $1 - 2p^{-6} - 4p^{-2}$. Now by the discussion in Section C.3.2, we have $\mathbb{P}(\mathcal{E}_K) \geq 1 - 2p^{-2}$ for $K \geq \sqrt{s}\left\{(\frac{\kappa_+}{\kappa_-})b_{\mathsf{max}} + (\frac{8}{\kappa_-})\sigma x_{\mathsf{max}}\right\}$. This finishes the proof.

C.2.3 Proof of Lemma 4.4.1

For clarity, we first specify some notations. Let γ_t^D denote the model update of MH chain run over dataset D. Let τ_η^D be the corresponding η-mixing time. Let D and D' be two neighboring datasets, and π^D and $\pi^{D'}$ be the corresponding probability mass functions for the exponential mechanism. Then, we have the following:

$$\begin{aligned}
\mathbb{P}(\gamma^D_{\tau^D_\eta} = \gamma) &\leq \pi^D(\gamma) + \eta \\
&\leq e^\varepsilon \pi^{D'}(\gamma) + \eta \\
&\leq e^\varepsilon \mathbb{P}(\gamma^{D'}_{\tau^{D'}_\eta} = \gamma) + \eta(1 + e^\varepsilon).
\end{aligned}$$

This finishes the proof.

C.2.4 Proof of Mixing Time (Theorem 4.4.3)

We again restrict ourselves to the event $\mathcal{E}_K$ with $K \geq \sqrt{s}\left\{(\frac{\kappa_+}{\kappa_-})b_{\mathsf{max}} + (\frac{8}{\kappa_-})\sigma x_{\mathsf{max}}\right\}$. For the proof, let $\widetilde{\mathbf{P}}$ denote the transition matrix of the original Metropolis-Hastings sampler (4.10). In this case, the state space is $\mathscr{S} = \mathscr{A}_s \cup \{\gamma^*\}$. Now consider the transition matrix $\mathbf{P} = \widetilde{\mathbf{P}}/2 + \mathbb{I}_n/2$, corresponding to the lazy version of the random walk that stays in its current position with a probability of at least $1/2$. Due to the construction, the smallest eigenvalue of $\mathbf{P}$ is always non-negative, and the mixing time of the chain is completely determined by the second largest eigenvalue λ_2 of $\mathbf{P}$. To this end, we define the spectral gap $\mathsf{Gap}(\mathbf{P}) = 1 - \lambda_2$, and for any lazy Markov chain, we have the following sandwich relation (Sinclair, 1992; Woodard and Rosenthal, 2013)

$$\frac{1}{2}\frac{(1 - \mathsf{Gap}(\mathbf{P}))}{\mathsf{Gap}(\mathbf{P})}\log(1/(2\eta)) \leq \tau_\eta \leq \frac{\log[1/\min_{\gamma \in \mathscr{S}} \pi(\gamma)] + \log(1/\eta)}{\mathsf{Gap}(\mathbf{P})}. \tag{C.1}$$

Lower Bound on $\pi(\cdot)$:

To establish a lower bound on the target distribution in (4.7), we make use of its specific form, thereby obtaining the following inequality:

$$\begin{aligned}\pi(\gamma) &= \pi(\gamma^*).\frac{\pi(\gamma)}{\pi(\gamma^*)} \\ &= \frac{1}{1+\sum_{\gamma'\in\mathscr{A}_s}\exp\left\{-\frac{\varepsilon(L_{\gamma'}-L_{\gamma^*})}{\Delta u}\right\}}.\exp\left\{-\frac{\varepsilon(L_\gamma - L_{\gamma^*})}{\Delta u}\right\}.\end{aligned}$$

Now we fix $k \in [s]$, and and consider any $\gamma \in \mathscr{A}_{s,k}$.

Similar to the proof of Section C.2.2, we note that whenever $\mathfrak{m}_*(s) \geq 2C\sigma^2\{(\log p)/n\}$ for a large enough universal constant $C > 0$, we get

$$\frac{3}{2}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} \geq n^{-1}(L_\gamma - L_{\gamma^*}) \geq \frac{1}{2}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} \geq \frac{k\mathfrak{m}_*(s)}{2} \quad \text{for all } \gamma \in \mathscr{A}_s,$$

with probability at least $1-2p^{-6}-4p^{-2}$. Also, note that $\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma} \leq \kappa_+ s b_{\mathsf{max}}^2$. Hence, if we have

$$\mathfrak{m}_*(s) \geq \max\left\{2C, \frac{16\Delta u}{\varepsilon\sigma^2}\right\}\frac{\sigma^2\log p}{n},$$

the following are true:

$$\begin{aligned}&\sum_{\gamma'\in\mathscr{A}_s}\exp\left\{-\frac{\varepsilon(L_{\gamma'}-L_{\gamma^*})}{\Delta u}\right\} \\ &\leq \sum_{\gamma'\in\mathscr{A}_s}\exp\left\{-\frac{nk\varepsilon\mathfrak{m}_*(s)}{2\Delta u}\right\} \\ &\leq \sum_{k=1}^{s}\binom{p-s}{k}\binom{s}{k}\exp\left\{-\frac{nk\varepsilon\mathfrak{m}_*(s)}{2\Delta u}\right\} \\ &\leq \sum_{k=1}^{s}p^{2k}.p^{-4k} \leq p^{-2},\end{aligned}$$

and,

$$\begin{aligned}\exp\left\{-\frac{\varepsilon(L_\gamma - L_{\gamma^*})}{\Delta u}\right\} &\geq \exp\left\{-\frac{3n\varepsilon\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top\Gamma(\gamma)\boldsymbol{\beta}_{\gamma^*\setminus\gamma}}{2\Delta u}\right\} \\ &\geq \exp\left\{-\frac{3ns\varepsilon\kappa_+ b_{\mathsf{max}}^2}{2\Delta u}\right\}\end{aligned}$$

Combining these two facts we have

$$\min_{\gamma\in\mathscr{A}_s\cup\{\gamma^*\}}\pi(\gamma) \geq \frac{1}{1+p^{-2}}\exp\left\{-\frac{3ns\varepsilon\kappa_+ b_{\mathsf{max}}^2}{2\Delta u}\right\} \geq \frac{1}{2}\exp\left\{-\frac{3ns\varepsilon\kappa_+ b_{\mathsf{max}}^2}{2\Delta u}\right\} \tag{C.2}$$

with probability $1-2p^{-6}-4p^{-2}$.

Lower Bound on Spectral Gap:

Now it remains to prove a lower bound on the spectral gap $\mathsf{Gap}(\mathbf{P})$, and we do so via the canonical path argument (Sinclair, 1992). We begin by describing the idea of a canonical path ensemble associated with a Markov chain. Given a Markov chain $\mathcal{C}$ with state space $\mathscr{S}$, consider the weighted directed graph $G(\mathcal{C}) = (V, E)$ with vertex set $V = \mathscr{S}$ and the edge set E in which a ordered pair $e = (\gamma, \gamma')$ is included as an edge with weight $\mathbf{Q}(e) = \mathbf{Q}(\gamma, \gamma') = \pi(\gamma)\mathbf{P}(\gamma, \gamma')$ iff $\mathbf{P}(\gamma, \gamma') > 0$. A *canonical path ensemble* $\mathcal{T}$ corresponding to $\mathcal{C}$ is a collection of paths that contains, for each ordered pair (γ, γ') of distinct vertices, a unique simple path $T_{\gamma,\gamma'}$ connecting γ and γ'. We refer to any path in the ensemble $\mathcal{T}$ as a canonical path.

Sinclair (1992) shows that for any reversible Markov chain and nay choice of a canonical path ensemble $\mathcal{T}$, the spectral gap of $\mathbf{P}$ is lower bounded as

$$\mathsf{Gap}(\mathbf{P}) \geq \frac{1}{\rho(\mathcal{T})\ell(\mathcal{T})}, \tag{C.3}$$

where $\ell(\mathcal{T})$ corresponds to the length of the longest path in the ensemble $\mathcal{T}$, and the quantity $\rho(\mathcal{T}) := \max_{e \in E} \frac{1}{Q(e)} \sum_{(\gamma,\gamma'): e \in T_{\gamma,\gamma'}} \pi(\gamma)\pi(\gamma')$ is known as the *path congestion parameter*.

Thus, it boils down to the construction of a suitable canonical path ensemble $\mathcal{T}$. Before going into further details, we introduce some working notations. For any two given paths T_1 and T_2:

- Their intersection $T_1 \cap T_2$ denotes the collection of overlapping edges.
- If $T_2 \subset T_1$, then $T_1 \setminus T_2$ denotes the path obtained by removing all the edges of T_2 from T_1.
- We use $\bar{T}_1$ to denote the reverse of T_1.
- If the endpoint of T_1 is same as the starting point of T_2, then $T_1 \cup T_2$ denotes the path obtained by joining T_1 and T_2 at that point.

We will now shift focus toward the construction of the canonical path ensemble. At a high level, our construction follows the same scheme as in Yang et al. (2016).

Canonical path ensemble construction:

First, we need to construct the canonical path T_{γ,γ^*} from any $\gamma \in \mathscr{S}$ to the true model γ^*. To this end, we introduce the concept of *memoryless* paths. We call a set $\mathcal{T}_M$ of canonical paths memoryless with respect to the central state γ^* if

1. for any state $\gamma \in \mathscr{S}$ satisfying $\gamma \neq \gamma^*$, there exists a unique simple path T_{γ,γ^*} in $\mathcal{T}_M$ connecting γ and γ^*;

2. for any intermediate state $\tilde{\gamma} \in \mathscr{S}$ on any path $T_{\gamma,\gamma^*} \in \mathcal{T}_M$, the unique path connecting $\tilde{\gamma}$ and γ^* is the sub-path of T_{γ,γ^*} starting from $\tilde{\gamma}$ and ending at γ^*.

Intuitively, this memoryless property tells that for any intermediate step in any canonical path, the next step towards the central state does not depend on history. Specifically, the memoryless canonical path ensemble has the property that in order to specify the canonical path connecting any state $\gamma \in \mathscr{S}$ and the central state γ^*, we only need to specify the next state from $\gamma \in \mathscr{S} \setminus \{\gamma^*\}$, i.e., we need a transition function $\mathcal{G} : \mathscr{S} \setminus \{\gamma^*\} \to \mathscr{S}$ that maps the current state γ to the next state. For simplicity, we define $\mathcal{G}(\gamma^*) = \gamma^*$ to make $\mathscr{S}$ as the domain of $\mathcal{G}$. For a more detailed discussion, we point the readers to Section 4 of Yang et al. (2016). We now state a useful lemma that is pivotal to the construction of the canonical path ensemble.

Lemma C.2.1 (Yang et al. (2016)). *If a function $\mathcal{G} : \mathscr{S} \setminus \{\gamma^*\} \to \mathscr{S}$ satisfies the condition $d_H(\mathcal{G}(\gamma), \gamma^*) < d_H(\gamma, \gamma^*)$ for any state $\gamma \in \mathscr{S} \setminus \{\gamma^*\}$, then $\mathcal{G}$ is a valid transition map.*

Using the above lemma, we will now construct the memoryless set of canonical paths from any state $\gamma \in \mathscr{S}$ to γ^* by explicitly specifying a transition map $\mathcal{G}$. In particular, we consider the following transition function:

- If $\gamma \neq \gamma^*$, we define $\mathcal{G}(\gamma)$ to be γ', whch is formed by replacing the least influential covariate in γ with most influential covariate in $\gamma^* \setminus \gamma$. In notations, we have $\gamma'_j = \gamma_j$ for all $j \notin \{j_\gamma, k_\gamma\}$, $\gamma'_{j_\gamma} = 1$ and $\gamma'_{k_\gamma} = 0$, where $j_\gamma := \arg\max_{j \in \gamma^* \setminus \gamma} \left\| \mathbf{\Phi}_{\gamma \cup \{j\}} \mathbf{X}_{\gamma^*} \boldsymbol{\beta}_{\gamma^*} \right\|_2^2$ and $k_\gamma := \arg\min_{k \in \gamma \setminus \gamma^*} \left\| \mathbf{\Phi}_{\gamma \cup \{j\}} \mathbf{X}_{\gamma^*} \boldsymbol{\beta}_{\gamma^*} \right\|_2^2 - \left\| \mathbf{\Phi}_{\gamma \cup \{j\} \setminus \{k\}} \mathbf{X}_{\gamma^*} \boldsymbol{\beta}_{\gamma^*} \right\|_2^2$. Thus, the transition step involves a double flip which entails that $d_H(\mathcal{G}(\gamma), \gamma^*) = d_H(\gamma, \gamma^*) - 2$.

Due to Lemma C.2.1, it follows that the above transition map $\mathcal{G}$ is valid and gives rise to a unique memoryless set $\mathcal{T}_M$ of canonical paths connecting any $\gamma \in \mathscr{S}$ and γ^*.

Based on this, we are now ready to construct the canonical path ensemble $\mathcal{T}$. Specifically, due to memoryless property, two simple paths T_{γ,γ^*} and T_{γ',γ^*} share an identical subpath to γ^* starting from their first common intermediate state. Let $T_{\gamma \cap \gamma'}$ denote the common sub-path $T_{\gamma \cap \gamma^*} \cap T_{\gamma' \cap \gamma *}$, and $T_{\gamma \setminus \gamma'} := T_{\gamma,\gamma^*} \setminus T_{\gamma \cap \gamma'}$ denotes the remaining path of T_{γ,γ^*} after removing the segment $T_{\gamma \cap \gamma'}$. The path $T_{\gamma' \setminus \gamma}$ is defined in a similar way. Then it follows that $T_{\gamma \setminus \gamma'}$ and $T_{\gamma' \setminus \gamma}$ have the same endpoint. Therefore, it is allowed to consider the path $T_{\gamma \setminus \gamma'} \cup \bar{T}_{\gamma' \setminus \gamma}$.

We call γ a *precedent* of γ' if γ' is on the canonical path $T_{\gamma,\gamma^*} \in \mathcal{T}$, and a pair of states γ, γ' are *adjacent* if the canonical path $T_{\gamma,\gamma'}$ is $e_{\gamma,\gamma'}$, the edge connecting γ and γ'. Next, for $\gamma \in \mathscr{S}$, define

$$\Lambda(\gamma) := \{\tilde{\gamma} \mid \gamma \in T_{\tilde{\gamma},\gamma^*}\} \tag{C.4}$$

denote the set of all precedents. We denote by $|T|$ the length of the path T. The following lemma provides some important properties of the previously constructed canonical path ensemble.

Lemma C.2.2. *For any distinct pair $(\gamma, \gamma') \in \mathscr{S} \times \mathscr{S}$:*

(a) We have

$$|T_{\gamma,\gamma^*}| \leq d_H(\gamma, \gamma^*)/2 \leq s, \quad \textit{and}$$
$$|T_{\gamma,\gamma'}| \leq \frac{1}{2}\{d_H(\gamma, \gamma^*) + d_H(\gamma', \gamma^*)\} \leq 2s.$$

(b) If γ and γ' are adjacent and γ is precedent of of γ', then

$$\{(\bar{\gamma}, \bar{\gamma}') \mid e_{\gamma,\gamma'} \in T_{\bar{\gamma},\bar{\gamma}'}\} \subset \Lambda(\gamma) \times \mathscr{S}.$$

Proof. For the first claim, let us first assume that $|T_{\gamma,\gamma^*}| = k$, i.e., $\mathcal{G}^k(\gamma) = \gamma^*$ for the appropriate transition map $\mathcal{G}$. Also, recall that $|\gamma| = |\gamma^*| = s$. Hence, due to an elementary iterative argument, it follows that

$$\begin{aligned} 2s \geq d_H(\gamma, \gamma^*) &= d_H(\mathcal{G}(\gamma), \gamma^*) + 2 \\ &= d_H(\mathcal{G}^2(\gamma), \gamma^*) + 4 \\ &\vdots \\ &= 2k. \end{aligned}$$

Also, note that $|T_{\gamma,\gamma'}| \leq |T_{\gamma,\gamma^*}| + |T_{\gamma',\gamma^*}|$. Hence, the claim follows using the previous inequality.

For the second claim, note that for any pair $(\bar{\gamma}, \bar{\gamma}')$ such that $T_{\bar{\gamma},\bar{\gamma}'} \ni e_{\gamma,\gamma'}$, we have two possible options : (i) $e_{\gamma,\gamma'} \in T_{\bar{\gamma}\setminus\bar{\gamma}'}$, or (ii) $e_{\gamma,\gamma'} \in T_{\bar{\gamma}'\setminus\bar{\gamma}}$. As γ is precedent of γ', the only possibility that we have is $e_{\gamma,\gamma'} \in T_{\gamma\setminus\gamma'}$. This shows that γ belongs to the path $T_{\bar{\gamma},\gamma^*}$ and $\bar{\gamma} \in \Lambda(\gamma)$. □

According to Lemma C.2.2(b), the path congestion parameter $\rho(T)$ satisfies

$$\rho(T) \leq \max_{(\gamma,\gamma')\in\Gamma_*} \frac{1}{\mathbf{Q}(\gamma,\gamma')} \sum_{\bar{\gamma}\in\Lambda(\gamma),\bar{\gamma}'\in\mathscr{S}} \pi(\bar{\gamma})\pi(\bar{\gamma}') = \max_{(\gamma,\gamma')\in\Gamma_*} \frac{\pi[\Lambda(\gamma)]}{\mathbf{Q}(\gamma,\gamma')}, \tag{C.5}$$

where the set $\Gamma_* := \{(\gamma,\gamma') \in \mathscr{S}\times\mathscr{S} \mid T_{\gamma,\gamma'} = e_{\gamma,\gamma'}, \gamma\in\Lambda(\gamma')\}$. Here we used the fact that the weight function $\mathbf{Q}$ satisfies the reversibility condition $\mathbf{Q}(\gamma,\gamma') = \mathbf{Q}(\gamma',\gamma)$ in order to restrict the range of the maximum to pairs (γ,γ') where $\gamma\in\Lambda(\gamma')$.

For the lazy form of the Metropolis-Hastings walk (4.10), we have

$$\begin{aligned}\mathbf{Q}(\gamma,\gamma') &= \pi(\gamma)\mathbf{P}(\gamma,\gamma') \\ &\geq \frac{1}{ps}\pi(\gamma)\min\left\{1, \frac{\pi(\gamma')}{\pi(\gamma)}\right\} \geq \frac{1}{ps}\min\left\{\pi(\gamma),\pi(\gamma')\right\}.\end{aligned}$$

Substituting this bound in (C.5), we get

$$\begin{aligned}\rho(T) &\leq ps \max_{(\gamma,\gamma')\in\Gamma_*} \frac{\pi(\Lambda(\gamma))}{\min\{\pi(\gamma),\pi(\gamma')\}} \\ &= ps \max_{(\gamma,\gamma')\in\Gamma_*} \left\{\max\left\{1,\frac{\pi(\gamma)}{\pi(\gamma')}\right\}\cdot\frac{\pi(\Lambda(\gamma))}{\pi(\gamma)}\right\}.\end{aligned} \tag{C.6}$$

In order to prove that $\rho(T) = O(ps)$ with high probability, it is sufficient to prove that the two terms inside the maximum are $O(1)$. To this end, we introduce two useful lemmas.

Lemma C.2.3. *Consider the event*

$$\mathcal{A}_n = \left\{\max_{\gamma\in\mathscr{S},\ell\notin\gamma} \mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{\ell\}} - \mathbf{\Phi}_\gamma)\mathbf{w} \leq 12\sigma^2 s\log p\right\}$$

Then we have $\mathbb{P}(\mathcal{A}_n) \geq 1 - p^{-2}$.

Proof. First note that $\mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{\ell\}} - \mathbf{\Phi}_\gamma)\mathbf{w} = (\mathbf{h}_{\gamma,\ell}^\top\mathbf{w})^2$ for an appropriate unit vector $\mathbf{h}_{\gamma,\ell}$ depending only upon $\mathbf{X}_\gamma$ and $\mathbf{X}_\ell$. By Sub-gaussian tail inequality, we have

$$\mathbb{P}\left\{(\mathbf{h}_{\gamma,\ell}^\top\mathbf{w})^2 \geq t\right\} \leq 2e^{-\frac{t}{2\sigma^2}}.$$

Setting $t = 12\sigma^2 s \log p$ and applying an union bound we get

$$\begin{aligned}\mathbb{P}\left\{\max_{\gamma\in\mathscr{S},\ell\notin\gamma}(\mathbf{h}_{\gamma,\ell}^\top\mathbf{w})^2 \geq 12\sigma^2 s\log p\right\} &\leq 2\binom{p}{s}(p-s)p^{-6s}\\ &\leq 2p^{-3s}\\ &\leq p^{-2}.\end{aligned}$$

□

Lemma C.2.4. *Suppose that, in addition to the conditions in Theorem 4.4.3, the event $\mathcal{A}_n$ holds. Then for all $\gamma \neq \gamma^*$, we have*

$$\frac{\pi(\gamma)}{\pi(\mathcal{G}(\gamma))} \leq p^{-3}.$$

Moreover, for all γ,

$$\frac{\pi[\Lambda(\gamma)]}{\pi(\gamma)} \leq 2.$$

Therefore, both Lemma C.2.3 and Lemma C.2.4 give $\rho(T) \leq 2ps$ with probability $1-p^{-2}$. Lemma C.2.2(a) suggests that $\ell(T) \leq 2s$. Therefore, Equation (C.3) shows that $\mathsf{Gap}(\mathbf{P}) \geq \frac{1}{4ps^2}$ with probability $1-p^{-2}$. Finally, combining (C.2) and (C.1), we get the following with $1-8p^{-2}$

$$\tau_\eta \leq C_2 ps^2\left(\frac{n\varepsilon\kappa_+ b_{\mathsf{max}}^2}{\left\{r + (\frac{\kappa_+}{\kappa_-})b_{\mathsf{max}}x_{\mathsf{max}} + (\frac{\sigma}{\kappa_-})x_{\mathsf{max}}^2\right\}^2} + \log(1/\eta)\right),$$

where $C_2 > 0$ is a universal constant. Finally, the proof is concluded by arguing that $\mathbb{P}(\mathcal{E}_K^c) \leq 2p^{-2}$.

C.3 Proof of Auxiliary Results

C.3.1 Proof of Sensitivity Bound (Lemma 4.3.3)

Let $(\mathbf{X}, \mathbf{y})$ and $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$ be two neighboring datasets with n and $n+1$ observation respectively. For a subset $\gamma \in \mathscr{A}_s \cup \{\gamma^*\}$, consider the OLS estimators as follows:

$$\boldsymbol{\beta}_{\gamma,K} := \arg\min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_1\leq K}\|\mathbf{y}-\mathbf{X}_\gamma\boldsymbol{\theta}\|_2^2, \quad \text{and} \quad \tilde{\boldsymbol{\beta}}_{\gamma,K} := \arg\min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_1\leq K}\left\|\tilde{\mathbf{y}}-\tilde{\mathbf{X}}_\gamma\boldsymbol{\theta}\right\|_2^2.$$

From the definition of the score function $u(\gamma; \mathbf{X}, \mathbf{y})$, we have

$$u(\gamma; \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta}_{\gamma,K})^2$$
$$u(\gamma; \tilde{\mathbf{X}}, \tilde{\mathbf{y}}) = -\sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \tilde{\boldsymbol{\beta}}_{\gamma,K})^2 - (\tilde{\mathbf{y}}_{n+1} - \tilde{\mathbf{x}}_{n+1,\gamma}^\top \tilde{\boldsymbol{\beta}}_{\gamma,K})^2.$$

By the property of the OLS estimators, we have

$$\begin{aligned}
&u(\gamma; \mathbf{X}, \mathbf{y}) - u(\gamma; \tilde{\mathbf{X}}, \tilde{\mathbf{y}}) \\
&= \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \tilde{\boldsymbol{\beta}}_{\gamma,K})^2 + (\tilde{y}_{n+1} - \tilde{\mathbf{x}}_{n+1,\gamma}^\top \tilde{\boldsymbol{\beta}}_{\gamma,K})^2 - \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta}_{\gamma,K})^2 \\
&\le \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta}_{\gamma,K})^2 + (\tilde{y}_{n+1} - \tilde{\mathbf{x}}_{n+1,\gamma}^\top \boldsymbol{\beta}_{\gamma,K})^2 - \sum_{i=1}^{n}(y_i - \mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta}_{\gamma,K})^2 \\
&= (\tilde{y}_{n+1} - \tilde{\mathbf{x}}_{n+1,\gamma}^\top \boldsymbol{\beta}_{\gamma,K})^2 \\
&\le (r + x_{\mathsf{max}} K)^2.
\end{aligned}$$

Similarly, we have $u(\gamma; \tilde{\mathbf{X}}, \tilde{\mathbf{y}}) - u(\gamma; \mathbf{X}, \mathbf{y}) \le (r + x_{\mathsf{max}} K)^2$. Next, the $(\varepsilon, 0)$-DP follows from Lemma 4.2.4. This finishes the proof.

C.3.2 Constrained problem to unconstrained OLS problem

Now we are ready to bound $\left\|\boldsymbol{\beta}_{\gamma,K}\right\|_1$. Define the OLS estimator corresponding to the model γ as

$$\boldsymbol{\beta}_{\gamma,ols} = \underbrace{(\frac{\mathbf{X}_\gamma^\top \mathbf{X}_\gamma}{n})^{-1}\frac{\mathbf{X}_\gamma^\top \mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}}{n}}_{:=\mathbf{u}_1} + \underbrace{(\frac{\mathbf{X}_\gamma^\top \mathbf{X}_\gamma}{n})^{-1}\frac{\mathbf{X}_\gamma^\top \mathbf{w}}{n}}_{:=\mathbf{u}_2}.$$

In this section, we will show that there exists a choice for K such that the event $\mathcal{E}_K := \cap_{\gamma:|\gamma|=s}\{\boldsymbol{\beta}_{\gamma,ols} = \boldsymbol{\beta}_{\gamma,K}\}$ holds with high probability. By Holder's inequality we have $\|\mathbf{u}_1\|_2 \le \left\|(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma/n)^{-1}\right\|_{\mathrm{op}} \left\|\mathbf{X}_\gamma^\top \mathbf{X}_{\gamma^*}/n\right\|_{\mathrm{op}} \left\|\boldsymbol{\beta}_{\gamma^*}\right\|_2 \le (\frac{\kappa_+}{\kappa_-})b_{\mathsf{max}}$. Hence, an application of Cauchy-Schwarz inequality directly yields that $\|\mathbf{u}_1\|_1 \le 2(\frac{\kappa_+}{\kappa_-})\sqrt{s}b_{\mathsf{max}}$. Next, note that

$$\|\mathbf{u}_2\|_2 \le \left\|(\frac{\mathbf{X}_\gamma^\top \mathbf{X}_\gamma}{n})^{-1}\right\|_2 \left\|\frac{\mathbf{X}_\gamma^\top \mathbf{w}}{n}\right\|_2 \le \sqrt{s}\left\|(\frac{\mathbf{X}_\gamma^\top \mathbf{X}_\gamma}{n})^{-1}\right\|_2 \left\|\frac{\mathbf{X}_\gamma^\top \mathbf{w}}{n}\right\|_\infty \le \sqrt{s}\left\|(\frac{\mathbf{X}_\gamma^\top \mathbf{X}_\gamma}{n})^{-1}\right\|_2 \left\|\frac{\mathbf{X}^\top \mathbf{w}}{n}\right\|_\infty.$$

Therefore, we get $\|\mathbf{u}_2\|_1 \le \frac{s}{\kappa_-}\left\|\mathbf{X}^\top \mathbf{w}/n\right\|_\infty$. In order to upper bound the last term in the previous inequality, we define $D_{i,j} = \mathbf{X}[i,j]w_j$ for all $(i,j) \in [s]\times[n]$. Using the sub-Gaussian

property of w_j, we have $\mathbb{E}(e^{\lambda w_j}) \leq e^{\lambda^2 x_{\mathsf{max}}^2 \sigma^2/2}$. Therefore, due to Hoeffding's inequality, we have

$$\mathbb{P}\left(\frac{1}{n}\Big|\sum_{j\in[n]} D_{i,j}\Big| \geq 8\sigma x_{\mathsf{max}}\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-4}.$$

Note that $\left\|\mathbf{X}^\top \mathbf{w}/n\right\|_\infty = \max_{i\in[s]} n^{-1}\big|\sum_{j\in[n]} D_{i,j}\big|$. Hence, by simple union-bound argument, it follows that

$$\mathbb{P}\left(\max_{\gamma:|\gamma|=s}\left\|\frac{\mathbf{X}_\gamma^\top \mathbf{w}}{n}\right\|_\infty \geq 8\sigma x_{\mathsf{max}}\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-4} \leq 2p^{-2}.$$

Thus, Assumption 4.3.4(c) yields that $\left\|\boldsymbol{\beta}_{\gamma,K}\right\|_1^2 \leq s\left\{(\frac{\kappa_+}{\kappa_-})b_{\mathsf{max}} + (\frac{8}{\kappa_-})\sigma x_{\mathsf{max}}\right\}^2$. Therefore, if $K \geq \sqrt{s}\left\{(\frac{\kappa_+}{\kappa_-})b_{\mathsf{max}} + (\frac{8}{\kappa_-})\sigma x_{\mathsf{max}}\right\}$ then $\mathbb{P}(\mathcal{E}_K) \geq 1 - 2p^{-2}$.

C.3.3 Proof of Corollary 4.4.4

Based on Theorem 4.4.3, we have $\|\pi_t - \pi\|_{\mathsf{TV}} \leq eta$ with probability at least $1 - c_2 p^{-2}$ whenever t is sufficiently large. Also, by Theorem 4.3.5, we know $\pi(\gamma^*) \geq 1 - p^{-2}$ with probability $1 - c_1 p^{-2}$. Therefore, we have $\pi_t(\gamma^*) \geq 1 - \eta - p^{-2}$ with probability at least $1 - (c_! + c_2)p^{-2}$. This finishes the proof.

C.4 Proof of Lemma C.2.4

part (a):

Let j_γ, k_γ be the indices defined in the construction of $\mathcal{G}(\gamma)$. The we have $\gamma' = \gamma \cup \{j_\gamma\} \setminus \{k_\gamma\}$. Let $\mathbf{v}_1 = (\boldsymbol{\Phi}_{\gamma\cup\{j_\gamma\}} - \boldsymbol{\Phi}_\gamma)\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}$ and $\mathbf{v}_2 = (\boldsymbol{\Phi}_{\gamma\cup\{j_\gamma\}} - \boldsymbol{\Phi}_{\gamma'})\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}$. Then Lemma C.4.1 guarantees that

$$\|\mathbf{v}_1\|_2^2 \geq n\kappa_- \mathfrak{m}_*(s), \quad \text{and} \quad \|\mathbf{v}_2\|_2^2 \leq n\kappa_- \mathfrak{m}_*(s)/2. \tag{C.7}$$

By the form in (4.7), we have

$$\frac{\pi(\gamma)}{\pi(\gamma')} = \exp\left\{-\frac{\mathbf{y}^\top(\boldsymbol{\Phi}_{\gamma'} - \boldsymbol{\Phi}_\gamma)\mathbf{y}}{(\Delta u/\varepsilon)}\right\}.$$

To show that the above ration is $O(1)$, it suffices to show that $\mathbf{y}^\top(\boldsymbol{\Phi}_{\gamma'} - \boldsymbol{\Phi}_\gamma)\mathbf{y}$ is large. By

simple algebra, it follows that

$$\begin{aligned}
&\mathbf{y}^\top(\mathbf{\Phi}_{\gamma'} - \mathbf{\Phi}_\gamma)\mathbf{y} = \mathbf{y}^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_\gamma)\mathbf{y} - \mathbf{y}^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_{\gamma'})\mathbf{y} \\
&= \|\mathbf{v}_1\|_2^2 + 2\mathbf{v}_1^\top\mathbf{w} + \mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_\gamma)\mathbf{w} - \left\{\|\mathbf{v}_2\|_2^2 + 2\mathbf{v}_2^\top\mathbf{w} + \mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_{\gamma'})\mathbf{w}\right\} \\
&= \|\mathbf{v}_1\|_2^2 + 2\mathbf{v}_1^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_\gamma)\mathbf{w} + \mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_\gamma)\mathbf{w} \\
&\quad - \left\{\|\mathbf{v}_2\|_2^2 + 2\mathbf{v}_2^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_{\gamma'})\mathbf{w} + \mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_{\gamma'})\mathbf{w}\right\} \\
&\geq \|\mathbf{v}_1\|_2\left(\|\mathbf{v}_1\|_2 - 2\left\|(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_\gamma)\mathbf{w}\right\|_2\right) - \|\mathbf{v}_2\|_2\left(\|\mathbf{v}_2\|_2 + 2\left\|(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_{\gamma'})\mathbf{w}\right\|_2\right) \\
&\quad - \left\|(\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}} - \mathbf{\Phi}_{\gamma'})\mathbf{w}\right\|_2^2.
\end{aligned} \tag{C.8}$$

Now, we recall the event

$$\mathcal{A}_n = \left\{\max_{\gamma\in\mathscr{S},\ell\notin\gamma} \mathbf{w}^\top(\mathbf{\Phi}_{\gamma\cup\{\ell\}} - \mathbf{\Phi}_\gamma)\mathbf{w} \leq 12\sigma^2 s\log p\right\}.$$

Let $A^2 := n\kappa_-\mathfrak{m}_*(s) \geq \kappa_- C_0\sigma^2\log p$. Then for C_0 large enough so that $\kappa_- C_0 \geq (128\times 12)s$, Equation (C.8) leads to the following inequality under event $\mathcal{A}_n$:

$$\mathbf{y}^\top(\mathbf{\Phi}_{\gamma'} - \mathbf{\Phi}_\gamma)\mathbf{y} \geq A(A - A/4) - (A/\sqrt{2})(A/\sqrt{2} + A/4) - A^2/16 \geq A/8.$$

This readily yields that

$$\frac{\pi(\gamma)}{\pi(\gamma')} \leq \exp\left\{-\frac{n\kappa_-\mathfrak{m}_*(s)}{(16\Delta u/\varepsilon)}\right\} \leq p^{-3} \tag{C.9}$$

under the margin condition of Theorem 4.4.3.

Part (b):

From the previous part, the bound (C.9) implies that $\pi(\gamma)/\pi(\mathcal{G}(\gamma)) \leq p^{-3}$. For each $\bar{\gamma} \in \Lambda(\gamma)$, we have that $\gamma \in T_{\bar{\gamma},\gamma} \subset T_{\bar{\gamma},\gamma^*}$. Let the path $T_{\bar{\gamma},\gamma}$ be $\gamma_0 \to \gamma_1 \to \ldots \to \gamma_k$, where $k = |T_{\bar{\gamma},\gamma}|$ is the length of the path, and $\gamma_0 = \bar{\gamma}$ and $\gamma_k = \gamma$ are the two endpoints. Now note that $\{\gamma_\ell\}_{\ell\leq k-1} \subset \mathscr{S}$, and (C.9) ensures that

$$\frac{\pi(\bar{\gamma})}{\pi(\gamma)} = \prod_{\ell=1}^{k}\frac{\pi(\gamma_{\ell-1})}{\pi(\gamma_\ell)} \leq p^{-3k}.$$

Also, by Lemma C.2.2(a) we have $k \in [s]$. Now, we count the total number of sets in $\Lambda(\gamma)$ for each $k \in [s]$. Recall that by the construction of the canonical path, we update the current state by adding a new influential covariate and deleting one unimportant one. Hence any state in $\mathscr{S}$ has at most sp adjacent precedents, implying that there could be at most $s^k p^k$

distinct paths of length k. This entails that

$$\frac{\pi(\Lambda(\gamma))}{\pi(\gamma)} \leq \sum_{\bar{\gamma}\in\Lambda(\gamma)} \frac{\pi(\bar{\gamma})}{\pi(\gamma)} \leq \sum_{k=1}^{s} (ps)^k p^{-3k} \leq \sum_{k=1}^{s} p^{-k} \leq \frac{1}{1-1/p} \leq 2.$$

C.4.1 Supporting lemmas

Recall the definition of j_γ and k_γ. The first result in the following lemma shows that the gain in adding j_γ to the current model γ is at least $n\kappa_- \mathfrak{m}_*(s)$. The second result shows that the loss incurred by removing k_γ from the model $\gamma \cup \{j_\gamma\}$ is at most $n\kappa_- \mathfrak{m}_*(s)/2$. As a result, it follows that it is favorable to replace $\mathbf{X}_{k_\gamma}$ with the more influential feature $\mathbf{X}_{j_\gamma}$ in the current model γ.

Lemma C.4.1. *Under Assumption 4.3.4(b) and Assumption 4.4.2, the following hold for all $\gamma \in \mathscr{A}_s$:*

(a) $\left\|\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 \geq n\kappa_- \mathfrak{m}_*(s)$, *and*

(b) $\left\|\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}\setminus\{k\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 \leq n\kappa_- \mathfrak{m}_*(s)/2.$

Proof. For each $\ell \in \gamma^* \setminus \gamma$, we have

$$\begin{aligned}
\left\|\mathbf{\Phi}_{\gamma\cup\{\ell\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 &= \boldsymbol{\beta}_{\gamma^*}^\top \mathbf{X}_{\gamma^*}^\top (\mathbf{\Phi}_{\gamma\cup\{\ell\}} - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*} \\
&= \frac{\boldsymbol{\beta}_{\gamma^*}^\top \mathbf{X}_{\gamma^*}^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_\ell \mathbf{X}_\ell^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}}{\mathbf{X}_\ell^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_\ell} \\
&\geq \frac{\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top \mathbf{X}_{\gamma^*\setminus\gamma}^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_\ell \mathbf{X}_\ell^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}}{n},
\end{aligned}$$

where the second equality simply follows from Gram-Schmidt orthogonal decomposition. By summing the preceding inequality over $\ell \in \gamma^* \setminus \gamma$, we get

$$\begin{aligned}
\sum_{\ell\in\gamma^*\setminus\gamma} \left\|\mathbf{\Phi}_{\gamma\cup\{\ell\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 &\geq \frac{\boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top \mathbf{X}_{\gamma^*\setminus\gamma}^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma} \mathbf{X}_{\gamma^*\setminus\gamma}^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma}}{n} \\
&\geq \kappa_- \boldsymbol{\beta}_{\gamma^*\setminus\gamma}^\top \mathbf{X}_{\gamma^*\setminus\gamma}^\top (\mathbb{I}_n - \mathbf{\Phi}_\gamma)\mathbf{X}_{\gamma^*\setminus\gamma}\boldsymbol{\beta}_{\gamma^*\setminus\gamma} \\
&\geq n\kappa_- \left|\gamma \setminus \gamma^*\right| \mathfrak{m}_*(s) \\
&= n\kappa_- \left|\gamma^* \setminus \gamma\right| \mathfrak{m}_*(s).
\end{aligned}$$

The last inequality follows from the fact that $|\gamma| = |\gamma^*| = s$. Since j_γ maximizes

$\left\|\mathbf{\Phi}_{\gamma\cup\{\ell\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2$ over all $\ell \in \gamma^* \setminus \gamma$, the preceding inequality implies that

$$\left\|\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 \geq n\kappa_{-}\mathfrak{m}_*(s).$$

Similarly, to prove the second claim, first note that for any $k \in \gamma \setminus \gamma^*$, we have

$$\begin{aligned}
\left\|\mathbf{\Phi}_{\gamma'\cup\{k\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma'}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 &= \boldsymbol{\beta}_{\gamma^*\setminus\gamma'}^\top \mathbf{X}_{\gamma^*\setminus\gamma'}^\top (\mathbf{\Phi}_{\gamma'\cup\{k\}} - \mathbf{\Phi}_{\gamma'})\mathbf{X}_{\gamma^*\setminus\gamma'}\boldsymbol{\beta}_{\gamma^*\setminus\gamma'} \\
&= \frac{\boldsymbol{\beta}_{\gamma^*\setminus\gamma'}^\top \mathbf{X}_{\gamma^*\setminus\gamma'}^\top (\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k\mathbf{X}_k^\top(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_{\gamma^*\setminus\gamma'}\boldsymbol{\beta}_{\gamma^*\setminus\gamma'}}{\mathbf{X}_k^\top(\mathbb{I}_n - \mathbf{\Phi}_{\gamma})\mathbf{X}_k} \\
&= \left\langle (\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_{\gamma^*\setminus\gamma'}\boldsymbol{\beta}_{\gamma^*\setminus\gamma'}, \frac{(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k}{\|(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k\|_2} \right\rangle^2 \\
&\leq \left\|\boldsymbol{\beta}_{\gamma^*\setminus\gamma}\right\|_1^2 \left\| \frac{\mathbf{X}_{\gamma^*\setminus\gamma'}^\top(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k}{\|(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k\|_2} \right\|_\infty^2 \\
&\leq b_{\mathsf{max}}^2 \left\| \frac{\mathbf{X}_{\gamma^*\setminus\gamma'}^\top(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k}{\|(\mathbb{I}_n - \mathbf{\Phi}_{\gamma'})\mathbf{X}_k\|_2} \right\|_\infty^2 .
\end{aligned}$$

□

Since k_γ minimizes $\left\|\mathbf{\Phi}_{\gamma'\cup\{k\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma'}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2$ over all possible $k \in \gamma \setminus \gamma^*$, by Assumption 4.4.2 we have

$$\left\|\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 - \left\|\mathbf{\Phi}_{\gamma\cup\{j_\gamma\}\setminus\{k\}}\mathbf{X}_{\gamma^*}\boldsymbol{\beta}_{\gamma^*}\right\|_2^2 \leq n\kappa_{-}\mathfrak{m}_*(s)/2.$$

APPENDIX D

Appendix for Chapter 5

D.1 More simulations

We set the number of arms $K = 10$ and we generate the context vectors $\{x_i(t)\}_{i=1}^K$ from multivariate d-dimensional Gaussian distribution $\mathsf{N}_d(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|\wedge 1}$ and $\rho = 0.3$. We consider $d = 1000$ and the sparsity $s^* = 5$. We choose the set of active indices S^* uniformly over all the subsets of $[d]$ of size s^*. Next, for each choice of d, we consider two types generating scheme for β:

- **Setup 1:** $\{U_i\}_{i\in S^*} \overset{i.i.d.}{\sim} \text{Uniform}(0.3, 1)$ and set $\beta_j = U_j(\sum_{\ell\in S^*} U_\ell^2)^{-1/2}\mathbb{1}(j \in S^*)$.
- **Setup 2:** $\{Z_i\}_{i\in S^*} \overset{i.i.d.}{\sim} \mathsf{N}(0, 1)$ and set $\beta_j = Z_j(\sum_{\ell\in S^*} Z_\ell^2)^{-1/2}\mathbb{1}(j \in S^*)$.

We run 40 independent simulations and plot the mean cumulative regret with 95% confidence band in Figure D.1. In all the setups, we see that VBTS outperforms its competitors by a wide margin. VBTS also enjoys superior empirical performance under the autoregressive (AR) model (see Figure D.2) with an auto-correlation coefficient 0.3.

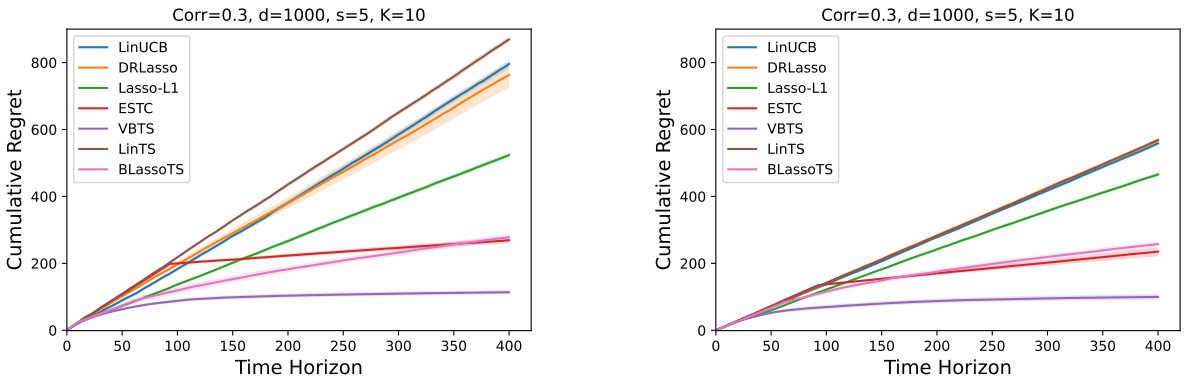


Figure D.1: Regret bound for equi-correlated design: (Left) Setup 1, (Right) Setup 2

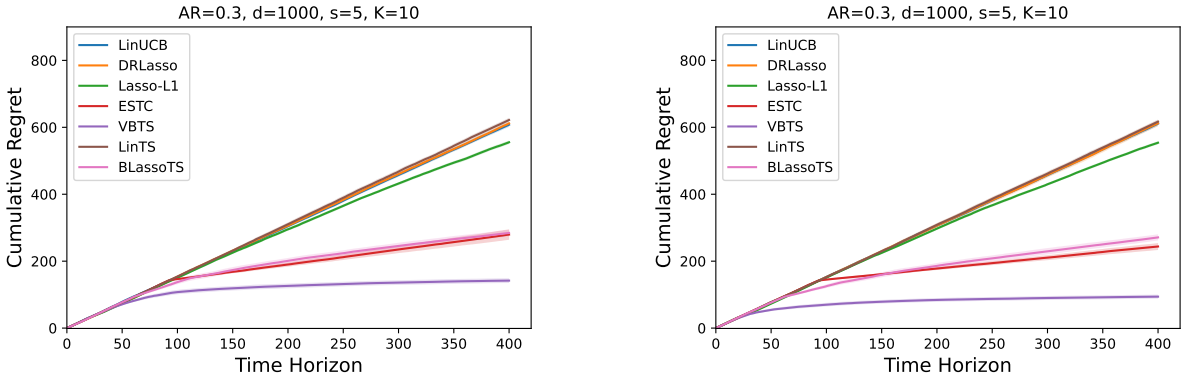


Figure D.2: Regret bound for AR(1) design: (Left) Setup 1, (Right) Setup 2

D.2 Proof of Theorem 5.3.2

This section presents the detailed proof of Theorem 5.3.2. First, for clarity of presentation, we introduce some notations. We use X_t to denote the matrix $(x_{a_1}(1), \ldots, x_{a_t}(t))^\top \in \mathbb{R}^{t\times d}$. Given this, we denote the covariance matrix $\widehat{\Sigma}_t = X_t^\top X_t/t$. Next, we define the set

$$\mathbb{S}_0^{d-1}(s) \triangleq \mathbb{S}^{d-1} \cap \{v \ : \ \|v\|_0 \leq s\}.$$

We also define the following:

Definition D.2.1. *For a index set $I \subseteq [d]$ and $\alpha \in \mathbb{R}^+$, we define the restricted cone as*

$$\mathbb{C}_\alpha(I) := \{v \in \mathbb{R}^d \ : \ \|v_{I^c}\|_1 \leq \alpha \|v_I\|_1, v_I \neq 0\}$$

.

In high-dimensional literature one typically assumes compatibility condition on the design matrix X, i.e.,

$$\phi_{\mathsf{comp}}(S^*; X) := \inf_{\delta \in \mathbb{C}_7(S^*)} \frac{\|X\delta\|_2 |S^*|^{1/2}}{t^{1/2}\|\delta\|_1} > 0, \tag{D.1}$$

where $S^* = \{j \ : \ \beta_j^* \neq 0\}$. This is mainly to guarantee the estimation accuracy of high-dimensional estimators like LASSO (Bickel et al., 2009) or to show the posterior consistency in Bayesian high dimensional literature (Castillo et al., 2015).

As discussed in the main paper, we prove the theorem in three parts, the subsequent sections deal with each part separately.

D.2.1 Proof of part (i)

In this section we will show that the matrix $\widehat{\Sigma}_t$ enjoys the SRC condition with high probability. As a warm up, we recall the definition of Orlicz norms:

Definition D.2.2 (Orlicz norms). *For random variable Z we have the followings:*

(a) The sub-Gaussian *norm of a random variable Z, denoted $\|Z\|_{\psi_2}$, is defined as*

$$\|Z\|_{\psi_2} := \inf\{\lambda > 0 : \mathbb{E}\{\exp(Z^2/\lambda^2)\} \leq 2\}.$$

(b) The sub-Exponential *norm of a random variable Z, denoted $\|Z\|_{\psi_1}$, is defined as*

$$\|Z\|_{\psi_1} := \inf\{\lambda > 0 : \mathbb{E}\{\exp(|Z|/\lambda)\} \leq 2\}.$$

The details and related properties can be found in Section 2.5.2 and Section 2.7 in Vershynin (2018b). The following is a relationship between sub-gaussian and sub-exponential random variables.

Lemma D.2.3 (Sub-Exponential and sub-Gaussian squared). *A random variable Z is sub-Gaussian iff Z^2 is sub-Exponential. Moreover,*

$$\left\|Z^2\right\|_{\psi_1} = \|Z\|_{\psi_2}^2 .$$

Proof. The proof can be found in Lemma 2.7.4 of Vershynin (2018b) □

Lemma D.2.4 (Bernstein's inequality). *Let $Z_1, \ldots, Z_N$ be independent mean-zero sub-exponential random variable. Then for every $\delta \geq 0$ we have*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i\right| \geq \delta\right) \leq 2\exp\left\{-c_2 \min\left(\frac{\delta^2}{K_0^2}, \frac{\delta}{K_0}\right) N\right\},$$

where $K_0 = \max_{i\in[N]} \|Z_i\|_{\psi_1}$.

Proof. The proof can be found in Corollary 2.8.3 of Vershynin (2018b). □

Proposition D.2.5 (Empirical SRC). *Let $\varepsilon = \min\{1/4, 1/(\tilde{c}\phi_u\vartheta^2\xi K\log K + 3)\}$ for some universal constant $\tilde{c} > 0$ and also define the quantity $\kappa(\xi, \vartheta, K) \triangleq \min\{(4c_3 K\xi\vartheta^2)^{-1}, 1/2\}$ for some universal positive constants c_3. Then the followings are true for any constant $C > 0$:*

$$\mathbb{P}\left(\phi_{\max}(Cs^*; \widehat{\Sigma}_t) \geq \frac{c_9\vartheta^2\phi_u \log K}{1-3\varepsilon}\right) \leq \exp\left\{-c_8 t\log K + Cs^*\log K + Cs^*\log(3d/\varepsilon)\right\},$$

$$\mathbb{P}\left(\phi_{\min}(Cs^*;\widehat{\Sigma}_t) \le \frac{1}{8K\xi}\right) \le 2\exp\{-c_2\kappa^2(\xi,\vartheta,K)t + Cs^*\log K + Cs^*\log(3d/\varepsilon)\} + \exp\left\{-c_8\log(K)t + Cs^*\log K + Cs^*\log(3d/\varepsilon)\right\},$$

where all c_j's in the above display are universal positive constants.

Proof. We will first show the SRC condition for a fixed vector $v \in \mathbb{S}_0^{d-1}(Cs^*)$. Then, the whole argument will be extended via a ε-net argument.

Analysis for a fixed vector v: Let $v \in \mathbb{S}_0^{d-1}(Cs^*)$ be a fixed vector. Now note that following fact:

$$v^\top\widehat{\Sigma}_t v = \frac{1}{t}\sum_{\tau=1}^{t}\{v^\top x_{a_\tau}(\tau)\}^2 \ge \frac{1}{t}\sum_{\tau=1}^{t}\min_{i\in[K]}\{v^\top x_i(\tau)\}^2.$$

We define $Z_{\tau,v} \triangleq \min_{i\in[K]}\{v^\top x_i(\tau)\}^2$ and note that for a fixed $v \in \mathbb{S}_0^{d-1}(Cs^*)$, the random variables $\{Z_{\tau,v}\}_{\tau=1}^t$ are i.i.d. across the time points. Moreover, due to Assumption 5.2.2(b) and Lemma D.2.3, we have

$$\|Z_{\tau,v}\|_{\psi_1} = \left\|\min_{i\in[K]}\left|v^\top x_i(\tau)\right|\right\|_{\psi_2}^2 \le c_3\vartheta^2.$$

Thus, $\{Z_{\tau,v}\}_{\tau=1}^t$ are i.i.d sub-exponential random variables. First we will show that $\mathbb{E}(Z_{\tau,v})$ is uniformly lower bounded. Due to Assumption 5.2.2(c) we have

$$\mathbb{P}(Z_{\tau,v} \le h) \le \sum_{i=1}^{K}\mathbb{P}\{(v^\top x_i(\tau))^2 \le h\} \le K\xi h.$$

Thus we have the following:

$$\begin{aligned}\mathbb{E}(Z_{\tau,v}) &= \int_0^\infty \mathbb{P}(Z_{\tau,v} \ge u)\,du \\ &\ge \int_0^h \mathbb{P}(Z_{\tau,v} \ge u)\,du \\ &\ge \int_0^h (1 - K\xi u)\,du \\ &= h(1 - K\xi h/2).\end{aligned} \tag{D.2}$$

Setting $h = 1/(K\xi)$ in Equation (D.2) yields $\mathbb{E}(Z_{\tau,v}) \ge \frac{1}{2K\xi}$. Now, using Lemma D.2.4, we have the following for a $\mu \in (0, c_1/\|Z_{1,v}\|_{\psi_1})$ and $\delta > 0$:

$$\mathbb{P}\left(\frac{1}{t}\sum_{\tau=1}^{t}\{Z_{\tau,v}-\mathbb{E}(Z_{\tau,v})\}\geq\delta\right)\leq 2\exp\left\{-c_2\min\left(\frac{\delta^2}{\|Z_{1,v}\|_{\psi_1}^2},\frac{\delta}{\|Z_{1,v}\|_{\psi_1}}\right)t\right\}$$
$$\leq 2\exp\left\{-c_2\min\left(\frac{\delta^2}{c_3^2\vartheta^4},\frac{\delta}{c_3\vartheta^2}\right)t\right\}.$$

Now choose $\delta=\min\{(4K\xi)^{-1},c_3\vartheta^2/2\}$ to finally get

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{\tau=1}^{t}\{Z_{\tau,v}-\mathbb{E}(Z_{\tau,v})\}\right|\geq\frac{1}{4K\xi}\right)\leq 2\exp\left\{-c_2\kappa^2(\xi,\vartheta,K)t\right\},\tag{D.3}$$

where $\kappa(\xi,\vartheta,K)=\min\{(4c_3K\xi\vartheta^2)^{-1},1/2\}$. Now recall that $\mathbb{E}(Z_{\tau,v})\geq 1/(2K\xi)$, which shows that

$$\mathbb{P}\left\{\frac{1}{t}\sum_{\tau=1}^{t}Z_{\tau,v}\geq\frac{1}{4K\xi}\right\}\leq 2\exp\left\{-c_2\kappa^2(\xi,\vartheta,K)t\right\},\quad\forall\, v\in\mathbb{S}_0^{d-1}(Cs^*).\tag{D.4}$$

ε-net argument: We consider a ε-net of the space $\mathbb{S}_0^{d-1}(Cs^*)$ constructed in a specific way which will be described shortly. We denote it by $\mathcal{N}_\varepsilon$. Let $J\subseteq[d]$ such that $|J|=Cs^*$ and consider the set $E_J=\mathbb{S}^{d-1}\cap\mathsf{span}\{e_j\ :\ j\in J\}$. Here e_j denotes the jth canonical basis of $\mathbb{R}^d$. Thus we have

$$\mathbb{S}_0^{d-1}(Cs^*)=\bigcup_{J:|J|=Cs^*}E_J.$$

Now we describe the procedure of constructing a net for $\mathbb{S}_0^{d-1}(Cs^*)$ which is essential for controlling the parse eigenvalues.

Greedy construction of net:

- Construct a ε-net of E_J for each J of size Cs^*. We denote this net by $\mathcal{N}_{\varepsilon,J}$. Note that $|\mathcal{N}_{\varepsilon,J}|\leq(3/\varepsilon)^{Cs^*}$ (Vershynin, 2018b, Corollary 4.2.13) for $\varepsilon\in(0,1)$ as E_J can be viewed as an unit ball embedded in $\mathbb{R}^{Cs^*}$.
- Then the net $\mathcal{N}_\varepsilon$ is constructed by taking union over all the $\mathcal{N}_{\varepsilon,J}$, i.e.,

$$\mathcal{N}_\varepsilon=\bigcup_{J:|J|=Cs^*}\mathcal{N}_{\varepsilon,J}.$$

Thus, from from the construction we have

$$|\mathcal{N}_\varepsilon|\leq\binom{d}{Cs^*}\left(\frac{3}{\varepsilon}\right)^{Cs^*}\leq\exp\{Cs^*\log(3d/\varepsilon)\},\tag{D.5}$$

whenever $\varepsilon \in (0,1)$. Now, we state an useful lemma on evaluating minimum eigenvalue on ε-net.

Lemma D.2.6. *Let A be a $m \times m$ symmetric positive-definite matrix and $\varepsilon \in (0,1)$. Then, for ε-net $\mathcal{N}_\varepsilon$ of $\mathbb{S}_0^{d-1}(s)$ constructed in greedy way, we have*

$$\phi_{\min}(s; A) \geq \min_{u \in \mathcal{N}_\varepsilon} u^\top A u - 3\varepsilon \phi_{\max}(s; A).$$

The proof of the lemma is deferred to Appendix D.4.1. Note that from Equation (D.4) and an union bound argument we get

$$\mathbb{P}\left(\min_{u \in \mathcal{N}_\varepsilon} v^\top \widehat{\Sigma}_t v \geq \frac{1}{4K\xi}\right) \geq 1 - 2\exp\{-c_2\kappa^2(\xi, \vartheta, K)t + Cs^* \log K + Cs^* \log(3d/\varepsilon)\}. \quad \text{(D.6)}$$

If $\phi_{\max}(Cs^*, \widehat{\Sigma}_t)$ is bounded with high probability, then for small ε, then along with Lemma D.2.6 we will readily have an uniform lower bound on $\phi_{\min}(Cs^*, \widehat{\Sigma}_t)$.

Bounding $\phi_{\max}(Cs^*, \widehat{\Sigma}_t)$: Here we gain start with $v \in \mathcal{N}_\varepsilon$. Similar, to previous discussion we have

$$v^\top \widehat{\Sigma}_t v = \frac{1}{t}\sum_{\tau=1}^{t} \{v^\top x_{a_\tau}(\tau)\}^2 \leq \frac{1}{t}\sum_{\tau=1}^{t} \max_{i \in [K]} \{v^\top x_i(\tau)\}^2.$$

We define $W_{\tau,v} \triangleq \max_{i \in [K]} \{v^\top x_i(\tau)\}^2$ and note that for a fixed $v \in \mathbb{S}_0^{d-1}(Cs^*)$, the random variables $\{W_{\tau,v}\}_{\tau=1}^t$ are i.i.d. across the time points. Moreover, due to Assumption 5.2.2(b) and Lemma D.2.3, we have

$$\|W_{\tau,v}\|_{\psi_1} = \left\| \max_{i \in [K]} \{v^\top x_i(\tau)\}^2 \right\|_{\psi_1} \leq c_4 K \vartheta^2.$$

Thus, $\{W_{\tau,v}\}_{\tau=1}^t$ are i.i.d sub-exponential random variables. Recall, that $\phi_{\max}(Cs^*, \Sigma_i) \leq \phi_u$ for all $i \in [K]$. The next lemma provides an upper bound on the moment generating function (MGF) of sub-Exponential random variables.

Lemma D.2.7. *(Vershynin, 2018b, Lemma 2.8.1) Let X be a mean-zero, sub-Exponential random variable. Then there exists positive constants c_5, c_6, such that for any λ with $|\lambda| \leq c_5 / \|X\|_{\psi_1}$, the following is true:*

$$\mathbb{E}\{\exp(\lambda X)\} \leq \exp(c_6 \lambda^2 \|X\|_{\psi_1}^2).$$

Equipped with the above lemma we have the following;

$$\begin{aligned}\mathbb{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} W_{\tau,v} - \phi_u \geq \delta\right) &= \mathbb{P}\left(\sum_{\tau=1}^{t}\{W_{\tau,v} - \phi_u\} \geq \delta t\right) \\ &= \mathbb{P}\left(\exp\left\{\mu\sum_{\tau=1}^{t}(W_{\tau,v} - \phi_u)\right\} \geq e^{\mu\delta t}\right) \\ &\leq e^{-\mu\delta t}\prod_{\tau=1}^{t}\mathbb{E}\{e^{\mu(W_{\tau,v}-\phi_u)}\}.\end{aligned} \tag{D.7}$$

For a fixed $\tau \in [t]$ we note the following;

$$\begin{aligned}\mathbb{E}\{e^{\mu(W_{\tau,v}-\phi_u)}\} &\leq \sum_{i=1}^{K}\mathbb{E}\{e^{\mu[(v^\top x_i(\tau))^2-\phi_u]}\} \\ &\leq \sum_{i=1}^{K}\mathbb{E}\{e^{\mu[(v^\top x_i(\tau))^2-v^\top\Sigma_i v]}\}.\end{aligned}$$

For brevity let $\kappa_i \triangleq \left\|\{v^\top x_i(\tau)\}^2\right\|_{\psi_1}$. If we choose $\mu \leq c_5/\max_{i\in[K]}\kappa_i$, then by Lemma D.2.7 we have

$$\mathbb{E}\{e^{\mu(W_{\tau,v}-\phi_u)}\} \leq \exp\left(c_6\mu^2\max_{i\in[K]}\kappa_i^2 + \log K\right).$$

Using the above inequality in Equation (D.7), it follows that

$$\mathbb{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} W_{\tau,v} - \phi_u \geq \delta\right) \leq \exp\left(-\mu\delta t + c_6 t\mu^2\max_{i\in[K]}\kappa_i^2 + t\log K\right). \tag{D.8}$$

The right hand side of Equation (D.8) is minimized at $\mu = \delta/(2c_2\max_{i\in[K]}\kappa_i^2)$ with the minimum value of

$$\exp\left(-\frac{\delta^2 t}{4c_6\max_{i\in[K]}\kappa_i^2} + t\log K\right).$$

If $\delta/(2c_6\max_{i\in[K]}\kappa_i^2) > c_5/\max_{i\in[K]}\kappa_i^2$, then the right hand side of Equation (D.8) is minimized at $\mu = c_5/\max_{i\in[K]}\kappa_i^2$ and we get

$$\mathbb{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} W_{\tau,v} - \phi_u \geq \delta\right) \leq \exp\left(-\frac{c_5\delta t}{\max_i\kappa_i} + c_6 c_5^2 t + t\log K\right).$$

Using the fact that $\delta/(2c_6\max_{i\in[K]}\kappa_i^2) > c_5/\max_{i\in[K]}\kappa_i^2$, the right hand side of the above

display can be upper bounded by

$$\exp\left(-\frac{c_5\delta t}{2\max_i \kappa_i} + t\log K\right).$$

Thus we have for all $\delta > 0$

$$\mathbb{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} W_{\tau,v} - \phi_u \geq \delta\right) \leq \exp\left(-\min\left\{\frac{\delta^2 t}{4c_6 \max_{i\in[K]} \kappa_i^2}, \frac{c_5\delta t}{2\max_i \kappa_i}\right\} + t\log K\right). \quad \text{(D.9)}$$

Next we set $\delta = c_7\vartheta^2\phi_u \log K$ for sufficiently large $c_7 > 0$. Then Equation (D.9) yields

$$\mathbb{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} W_{\tau,v} - \phi_u \geq \delta\right) \leq \exp\left(-c_8 t\log K\right).$$

Finally taking union bond over all vectors in $\mathcal{N}_\varepsilon$ we get

$$\mathbb{P}\left(\forall v \in \mathcal{N}_\varepsilon : v^\top \widehat{\Sigma}_t v \geq c_9\vartheta^2\phi_u \log K\right) \leq \exp\left\{-c_8 t\log K + Cs^*\log K + Cs^*\log(3d/\varepsilon)\right\}. \quad \text{(D.10)}$$

Next, to prove the same for all $v \in \mathbb{S}_0^{d-1}(Cs^*)$ we need the following lemma.

Lemma D.2.8 (maximum sparse eigenvalue on net)**.** *Let A be a $m \times m$ symmetric positive-definite matrix and $\varepsilon \in (0, 1/3)$. Then, for ε-net $\mathcal{N}_\varepsilon$ of $\mathbb{S}_0^{d-1}(s)$ constructed in greedy way, we have*

$$\max_{v\in\mathcal{N}_\varepsilon} v^\top A v \leq \max_{v\in\mathbb{S}_0^{d-1}(Cs^*)} v^\top A v \leq \frac{1}{1-3\varepsilon}\max_{v\in\mathcal{N}_\varepsilon} v^\top A v.$$

The proof of the above lemma is deferred to Appendix D.4.2. Now we set some $\varepsilon \in (0, 1/3)$. In light of the above lemma we immediately have that

$$\mathbb{P}\left(\phi_{\max}(Cs^*; \widehat{\Sigma}_t) \geq \frac{c_9\vartheta^2\phi_u\log K}{1-3\varepsilon}\right) \leq \exp\left\{-c_8 t\log K + Cs^*\log K + Cs^*\log(3d/\varepsilon)\right\}. \quad \text{(D.11)}$$

Finally using Equation (D.6), (D.11) and Lemma D.2.6 we have

$$\begin{aligned}
&\mathbb{P}\left(\phi_{\min}(Cs^*; \widehat{\Sigma}_t) \geq \frac{1}{4K\xi} - \frac{3\varepsilon c_9\vartheta^2\log K}{1-3\varepsilon}\phi_u\right) \\
&\geq 1 - 2\exp\{-c_2\kappa^2(\xi,\vartheta,K)t + Cs^*\log K + Cs^*\log(3d/\varepsilon)\} \\
&\quad - \exp\left\{-c_8 t\log K + Cs^*\log K + Cs^*\log(3d/\varepsilon)\right\}.
\end{aligned} \quad \text{(D.12)}$$

Now the result follows from taking $\varepsilon = \min\{1/4, 1/(24\phi_u\vartheta^2\xi K\log K + 3)\}$. □

D.2.2 Proof of part (ii)

In this section we will show that the matrix $\widehat{\Sigma}_t$ enjoys the compatibility condition (D.1) with high probability. This is equivalent to showing that the quantity

$$\Psi(S^*; \widehat{\Sigma}_t) \triangleq \inf_{\delta \in \mathbb{C}_7(S^*)} \left(\frac{\delta^\top \widehat{\Sigma}_t \delta}{\|\delta\|_1^2} \right) s^* = \phi_{\mathsf{comp}}(S^*; X_t)^2.$$

is bounded away from 0 with high probability. First we present the Transfer lemma (Oliveira, 2013, Lemma 5) below.

Lemma D.2.9 (Transfer lemma)**.** *Suppose $\widehat{\Sigma}_t$ and Σ are matrix with non-negative diagonal entries, and assume $\eta \in (0,1)$, $m \in [d]$ are such that*

$$\forall v \in \mathbb{R}^d \text{ with } \|v\|_0 \leq m, v^\top \widehat{\Sigma}_t v \geq (1-\eta) v^\top \Sigma v. \tag{D.13}$$

Assume D is a diagonal matrix whose diagonal entries $D_{j,j}$ are non-negative and satisfy $D_{j,j} \geq (\widehat{\Sigma}_t)_{j,j} - (1-\eta)\Sigma_{j,j}$. Then

$$\forall x \in \mathbb{R}^d, x^\top \widehat{\Sigma}_t x \geq (1-\eta) x^\top \Sigma x - \frac{\left\|D^{1/2} x\right\|_1^2}{m-1}. \tag{D.14}$$

Condition (D.13) basically demands that $\widehat{\Sigma}_t$ enjoys SRC condition with the sparsity parameter m. Then under the proper choice of diagonal matrix D with sufficiently large diagonal elements $\{D_{j,j}\}_{j=1}^d$, Equation (D.14) will yield the desired compatibility condition for $\widehat{\Sigma}_t$. We formally state the result in the following lemma:

Proposition D.2.10 (Empirical compatibility condition)**.** *Assume the conditions of Proposition D.2.5 hold. Also assume that Assumption 5.2.2(d) holds with $C = C_0 \phi_u \vartheta^2 \xi K \log K$ for some sufficiently large universal constant $C_0 > 0$. Then there exists a positive constant C_1 such that the following is true:*

$$\begin{aligned}
&\mathbb{P}\left(\Psi(S^*; \widehat{\Sigma}_t) \geq \frac{1}{C_1 \xi K} \right) \\
&= 1 - 2\exp\{-c_2 \kappa^2(\xi, \vartheta, K) t + C s^* \log K + C s^* \log(3d/\varepsilon)\} \\
&\quad - 2\exp\left\{-c_8 t \log K + C s^* \log K + C s^* \log(3d/\varepsilon)\right\}
\end{aligned}$$

with $\varepsilon = \min\{1/4, 1/(\tilde{c}\phi_u \vartheta^2 \xi K \log K + 3)\}$ for the same universal constant $\tilde{c} > 0$ in Proposition D.2.5 and $\kappa(\xi, \vartheta, K) = \min\{(4c_3 K \xi \vartheta^2)^{-1}, 1/2\}$.

Proof. As suggested before we will make use of Lemma D.2.9. Towards this, we set $\Sigma = \frac{1}{4K\xi}\mathbb{I}_d$ and $D = \mathsf{diag}(\widehat{\Sigma}_t)$. Next, we define the following two events:

$$\mathcal{G}_{t,1} := \left\{ \phi_{\max}(Cs^*; \widehat{\Sigma}_t) \leq \frac{c_9 \vartheta^2 \phi_u \log K}{1 - 3\varepsilon} \right\},$$

$$\mathcal{G}_{t,2} := \left\{ \phi_{\min}(Cs^*; \widehat{\Sigma}_t) \geq \frac{1}{8K\xi} \right\},$$

where the constants ε and c_9 are same as in Proposition D.2.5. Under $\mathcal{G}_{t,1}$ and $\mathcal{G}_{t,2}$, the inequality in Equation (D.13) holds with $\eta = 1/2$ and $m = Cs^*$. Also, due construction of D, we trivially have

$$D_{j,j} \geq (\widehat{\Sigma}_t)_{j,j} - (1 - \eta)\Sigma_{j,j}.$$

Lastly, note that

$$\max_{j \in [d]} D_{j,j} = \max_{j \in [d]} (\widehat{\Sigma}_t)_{j,j} = \max_{j \in [d]} e_j^\top \widehat{\Sigma}_t e_j \leq \frac{c_9 \vartheta^2 \phi_u \log K}{1 - 3\varepsilon} \tag{D.15}$$

under $\mathcal{G}_{t,1}$. Equipped with Lemma D.2.9, under $\mathcal{G}_{t,1}$ and $\mathcal{G}_{t,2}$, for all $x \in \mathbb{C}_7(S^*) \cap \mathbb{S}^{d-1}$ we have the following:

$$\begin{aligned} x^\top \widehat{\Sigma}_t x &\geq \frac{1}{8K\xi} - \frac{\left\| D^{1/2} x \right\|_1^2}{Cs^* - 1} \\ &\geq \frac{1}{8K\xi} - \frac{\left(\frac{c_9 \vartheta^2 \phi_u \log K}{1 - 3\varepsilon} \right) \|x\|_1^2}{Cs^* - 1} \\ &\geq \frac{1}{8K\xi} - \frac{\left(\frac{c_9 \vartheta^2 \phi_u \log K}{1 - 3\varepsilon} \right) 64 s^*}{Cs^* - 1}. \end{aligned}$$

The last inequality follows form the fact that

$$\|x\|_1 = \|x_{S^*}\|_1 + \left\| x_{(S^*)^c} \right\|_1 \leq 8 \|x_{S^*}\|_1 \leq 8\sqrt{s^*} \|x_{S^*}\|_2 \leq 8\sqrt{s^*}. \tag{D.16}$$

Thus, if $C \gtrsim \frac{\phi_u \vartheta^2 \xi K \log K}{1 - 3\varepsilon} + \frac{1}{s^*}$ then $x^\top \widehat{\Sigma}_t x \geq 1/(16K\xi)$. Also, note that from the choice of ε in Proposition D.2.5, we have $\varepsilon < 1/4$. This further tells that if $C = C_0 \phi_u \vartheta^2 \xi K \log K$ for large enough $C_0 > 0$, then

$$\inf_{\delta \in \mathbb{C}_7(S^*)} \frac{\delta^\top \widehat{\Sigma}_t \delta}{\|\delta\|_2^2} \geq \frac{1}{16K\xi}.$$

Then, the result follows from Proposition D.2.5. Using this and Equation (D.16) we also have

$$\Psi(S^*; \widehat{\Sigma}_t) \geq \frac{1}{64} \inf_{\delta \in \mathbb{C}_7(S^*)} \frac{\delta^\top \widehat{\Sigma}_t \delta}{\|\delta\|_2^2} \geq \frac{1}{C_1 \xi K}$$

where $C_1 = 1024$. Finally, the result follows from conditioning over the events $\mathcal{G}_{t,1}$ and $\mathcal{G}_{t,2}$ and using Proposition D.2.5.

□

D.2.3 Proof of part (iii)

In this section we will establish the desired regret bound in Theorem 5.3.2. The main tools that has been used to prove the regret bound is the Bayesian contraction in high-dimensional linear regression problem. In particular, we will use Theorem D.3.1 to control the ℓ_1-distance between $\widetilde{\beta}_t$ and β^* at each time point $t \in [T]$.

We recall that $X_t = (x_{a_1}(1), \ldots, x_{a_t}(t))^\top$. Also, note that the sequence $\{x_{a_\tau}(\tau)\}_{\tau=1}^t$ forms an adapted sequence of observations, i.e., $x_{a_\tau}(\tau)$ may depend on the history $\{x_{a_u}(u), r(u)\}_{u=1}^{\tau-1}$. Also, recall that $\{\epsilon(\tau)\}_{\tau=1}^t$ are mean-zero σ-sub-Gaussian errors.

Lemma D.2.11 (Bernstein Concentration)**.** *Let $\{D_k, \mathcal{F}_k\}_{k=1}^\infty$ be a martingale difference sequence, and suppose that D_k is a σ-sub-Gaussian in adapted sense, i.e., for all $\alpha \in \mathbb{R}, \mathbb{E}[e^{\alpha D_k} \mid \mathcal{F}_{k-1}] \leq e^{\alpha^2\sigma^2/2}$ almost surely. Then, for all $t \geq 0$, $\mathbb{P}\left(\left|\sum_{k=1}^t D_k\right| \geq \delta\right) \leq 2\exp\{-\delta^2/(2t\sigma^2)\}$.*

Proof. Proof of Lemma D.2.11 follows from Theorem 2.19 of Wainwright (2019) by setting $\alpha_k = 0$ and $\nu_k = \sigma$ for all k. □

Lemma D.2.11 is the main tool that is used to control the correlation between $\epsilon_t := (\epsilon(1), \ldots, \epsilon(t))^\top$ and the chosen contexts X_t which is important to control the Bayesian contraction of the posterior distribution in each round. To elaborate, let $X_t^{(j)}$ be the jth column for $j \in [d]$ and define $D_{t,j} := \epsilon(t) x_{a_t,j}(t)$. Note that for a fixed $j \in [d]$, $\{D_{\tau,j}\}_{\tau=1}^t$ forms a martingale difference sequence with respect to the filtration $\{\mathcal{F}_\tau\}_{\tau=1}^{t-1}$ with $\mathcal{F}_\tau := \sigma(\mathcal{H}_\tau)$ is the σ-algebra generated by $\mathcal{H}_\tau$ and $\mathcal{F}_1 = \varnothing$. Also note that

$$\mathbb{E}(e^{\alpha D_{t,j}} \mid \mathcal{F}_{t-1}) \leq \mathbb{E}\{e^{\alpha^2\sigma^2 x_{a_t,j}^2(t)/2}\} \leq \mathbb{E}\{e^{\alpha^2\sigma^2 x_{\max}^2/2}\}.$$

Thus, using Lemma D.2.11, we have the following proposition:

Proposition D.2.12 (Lemma EC.2, Bastani and Bayati (2020))**.** *Define the event*

$$\mathcal{T}_t(\lambda_0(\gamma)) := \left\{ \max_{j \in [d]} \frac{\left|\epsilon_t^\top X_t^{(j)}\right|}{t} \leq \lambda_0(\gamma) \right\},$$

where $\lambda_0(\gamma) = x_{\mathsf{max}}\sigma\sqrt{(\gamma^2 + 2\log d)/t}$. *Then we have* $\mathbb{P}\left\{\mathcal{T}_t(\lambda_0(\gamma))\right\} \geq 1 - 2\exp(-\gamma^2/2)$.

The proof of the above proposition mainly relies on the martingale difference structure and Lemma D.2.4. It is important to mention that the proof does not depend on any particular algorithm.

For notational brevity we define $\|X_t\| := \max_{j\in[d]}\sqrt{(X_t^\top X_t)_{j,j}}$. Next, we will set $\gamma = \gamma_t := \sqrt{2\log t}$. Hence by Proposition D.2.12 we have $\mathbb{P}\left\{\mathcal{T}_t(\lambda_0(\gamma_t))\right\} \geq 1 - 2t^{-1}$. Also recall that, under $\mathcal{G}_{t,1}$ and $\mathcal{G}_{t,2}$, we have

$$\frac{1}{\sqrt{8K\xi}} \leq \|X_t\|/\sqrt{t} \leq \sqrt{4c_9\phi_u\vartheta^2\log K}. \tag{D.17}$$

Also, under $\mathcal{G}_{t,2}\cap\mathcal{T}_t(\lambda_0(\gamma_t))$ it follows that

$$\max_{j\in[d]}\frac{\left|\epsilon_t^\top X_t^{(j)}\right|}{\sigma} \leq x_{\mathsf{max}}\sqrt{2t(\log d + \log t)} = \overline{\lambda}_t \tag{D.18}$$

Now, we are ready to present the proof of the main regret bound.

Main regret bound

Recall the definition of regret is $R(T) = \sum_{t=1}^T \Delta_{a_t}(t)$, where $\Delta_{a_t}(t) = x_{a_t^*}(t)^\top\beta^* - x_{a_t}(t)^\top\beta^*$. Next, we partition the whole time horizon $[T]$ in to two parts, namely $\{t : 1 \leq t \leq T_0\}$ and $\{t : T_0 \leq t \leq T\}$, where T_0 will be chosen later. Thus, the regret can be written as

$$R(T) = \underbrace{\sum_{t=1}^{T_0}\Delta_{a_t}(t)}_{R(T_0)} + \underbrace{\sum_{t=T_0+1}^{T}\Delta_{a_t}(t)}_{\tilde{R}(T)}.$$

All notations for expectation operators and probability measures are given in Appendix D.3.

Now by Assumption 5.2.2(a) and 5.2.3(a) we have the following inequality:

$$\mathbb{E}\{R(T_0)\} \leq 2x_{\mathsf{max}}b_{\mathsf{max}}T_0. \tag{D.19}$$

Next, we focus on the term $\tilde{R}(T)$. First, we define a few quantities below:

$$\begin{aligned}
\overline{\phi}_t(s) &:= \inf_\delta\left\{\frac{\|X_t\delta\|_2\,|S_\delta|^{1/2}}{t^{1/2}\,\|\delta\|_1} : 0 \neq |S_\delta| \leq s\right\},\\
\widetilde{\phi}_t(s) &:= \inf_\delta\left\{\frac{\|X_t\delta\|_2}{t^{1/2}\,\|\delta\|_2} : 0 \neq |S_\delta| \leq s\right\}.
\end{aligned} \tag{D.20}$$

Now set

$$\overline{\psi}_t(S) = \bar{\phi}_t\left(\left(2 + \frac{40}{A_4} + \frac{128A_4^{-1}x_{\mathsf{max}}^2}{\Psi(S,\widehat{\Sigma}_t)}\right)|S|\right),$$

$$\widetilde{\psi}_t(S) = \widetilde{\phi}_t\left(\left(2 + \frac{40}{A_4} + \frac{128A_4^{-1}x_{\mathsf{max}}^2}{\Psi(S,\widehat{\Sigma}_t)}\right)|S|\right).$$

Note that $\bar{\phi}_t(s) \geq \widetilde{\phi}_t(s)$, hence $\bar{\psi}_t(S) \geq \widetilde{\psi}_t(S)$.

Recall that

$$5\overline{\lambda}_t/3 \leq \lambda_t \leq 2\overline{\lambda}_t, \tag{D.21}$$

under $\mathcal{G}_{t,1}$ and $\mathcal{G}_{t,2}$. Also define the following events:

$$\mathcal{E}_t := \left\{\left\|\tilde{\beta}_{t+1} - \beta^*\right\|_1 \leq Q_4\sigma x_{\mathsf{max}}K\xi(D_* + s^*)\sqrt{\frac{\log d + \log t}{t}}\right\}.$$

where $D_* = \{1 + (40/A_4) + 128A_4^{-1}x_{\mathsf{max}}^2/\Psi(S^*,\widehat{\Sigma}_t)\}s^*$ and Q_4 is large enough universal constant as specified in Theorem D.3.1. Also we have

$$\bar{\psi}_t(S) \leq \bar{\phi}_t\left(\left(2 + \frac{40}{A_4} + \frac{64A_4^{-1}x_{\mathsf{max}}^2}{\Psi(S,\widehat{\Sigma}_t)}\frac{\lambda}{\overline{\lambda}}\right)|S|\right), \text{and}$$

$$\widetilde{\psi}_t(S) \leq \widetilde{\phi}_t\left(\left(2 + \frac{40}{A_4} + \frac{64A_4^{-1}x_{\mathsf{max}}^2}{\Psi(S,\widehat{\Sigma}_t)}\frac{\lambda}{\overline{\lambda}}\right)|S|\right).$$

Next, by Proposition D.2.10, the event

$$\mathcal{G}_{t,3} := \left\{\Psi(S^*;\widehat{\Sigma}_t) \geq \frac{1}{C_1\xi K}\right\}$$

holds with probability of at least $1 - 2\exp\{-c_2\kappa^2(\xi,\vartheta,K)t + Cs^*\log K + Cs^*\log(3d/\varepsilon)\} - 2\exp\{-c_8 t\log K + Cs^*\log K + Cs^*\log(3d/\varepsilon)\}$. For, notational brevity, we define

$$\tilde{C} := C_1\xi K.$$

Also, define the event

$$\mathcal{G}_{t,4} := \left\{\widetilde{\psi}_t^2(S^*) \geq \frac{1}{8K\xi}\right\}.$$

Noting that $\widetilde{\psi}_t^2(S^*) = \phi_{\min}(\tilde{C}_1 s^*;\widehat{\Sigma}_t)$ with

$$\tilde{C}_1 = 2 + \frac{40}{A_4} + 128A_4^{-1}x_{\mathsf{max}}^2\tilde{C},$$

an argument similar to the proof of Proposition D.2.5 yields

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}_{t,4}\right) \geq 1 - 2\exp\{-c_2\kappa^2(\xi,\vartheta,K)t + \tilde{C}_1 s^* \log K + \tilde{C}_1 s^* \log(3d/\varepsilon)\} \\ - \exp\left\{-c_8 \log(K)t + \tilde{C}_1 s^* \log K + \tilde{C}_1 s^* \log(3d/\varepsilon)\right\}.\end{aligned} \tag{D.22}$$

Finally, Using Proposition D.2.12 with $\gamma = \gamma_d$ and the result of Theorem D.3.1, we have the following:

$$\begin{aligned}\mathbb{P}(\mathcal{E}_t^c) &= \mathbb{E}_{X_t}\mathbb{E}_t^X\left(\mathbb{1}_{\mathcal{E}_t^c}\right) \\ &= \mathbb{E}_{X_t}\mathbb{E}_{t,\mathbf{r}_t}^X\left\{\Pi_t^X\left(\mathcal{E}_t^c \mid \mathbf{r}_t\right)\right\} \\ &= \mathbb{E}_{X_t}\mathbb{E}_{t,\mathbf{r}_t}^X\left\{\Pi_t^X\left(\mathcal{E}_t^c \mid \mathbf{r}_t\right)\mathbb{1}_{\mathcal{T}_t(\lambda_0(\gamma_t))\cap\mathcal{G}_{t,2}}\right\} + \mathbb{E}_{X_t}\mathbb{E}_{t,\mathbf{r}_t}^X\left\{\Pi_t^X\left(\mathcal{E}_t^c \mid \mathbf{r}_t\right)\mathbb{1}_{\mathcal{T}_t^c(\lambda_0(\gamma_t))\cup\mathcal{G}_{t,2}^c}\right\} \\ &\leq \frac{M_1}{d^{s^*}} + \frac{2}{t} + 2\exp\{-c_2\kappa^2(\xi,\vartheta,K)t + Cs^*\log K + Cs^*\log(3d/\varepsilon)\} \\ &\quad + \exp\{-c_8 t\log K + Cs^*\log K + Cs^*\log(3d/\varepsilon)\}\end{aligned} \tag{D.23}$$

for some large universal constant $M_1 > 0$. Next, let $\mathcal{G}_t = \cap_{i=1}^4 \mathcal{G}_{t,i}$. Under the event $\mathcal{E}_t \cap \mathcal{G}_t$, we have

$$D_* + s^* \leq \underbrace{\left(2 + \frac{40}{A_4} + \frac{128C_1K\xi x_{\mathsf{max}}^2}{A_4}\right)}_{:=\rho} s^*,$$

and,

$$\left\|\tilde{\beta}_{t+1} - \beta^*\right\|_1 \leq M_2\rho\sigma x_{\mathsf{max}}\xi K\left\{\frac{s^{*2}(\log d + \log t)}{t}\right\}^{1/2},$$

where M_2 is an universal constant depending on A_4. Now we set

$$\delta_t = M_2\rho\sigma x_{\mathsf{max}}^2\xi K\left\{\frac{s^{*2}(\log d + \log t)}{t}\right\}^{1/2}.$$

It follows that under $\mathcal{E}_{t-1}\cap\mathcal{G}_{t-1}$, we have the following almost sure inequality:

$$\begin{aligned}\Delta_{a_t}(t) &= x_{a_t^*}^\top(t)\beta^* - x_{a_t}^\top(t)\beta^* \\ &= x_{a_t^*}^\top(t)\beta^* - x_{a_t^*}^\top(t)\tilde{\beta}_t + \underbrace{(x_{a_t^*}^\top(t)\tilde{\beta}_t - x_{a_t}^\top(t)\tilde{\beta}_t)}_{\leq 0} + x_{a_t^*}^\top(t)\tilde{\beta}_t - x_{a_t}^\top(t)\beta^* \\ &\leq \left\|x_{a_t^*}(t)\right\|_\infty\|\tilde{\beta}_t - \beta^*\|_1 + \|x_{a_t}(t)\|_\infty\|\tilde{\beta}_t - \beta^*\|_1 \\ &\leq 2\delta_{t-1}.\end{aligned}$$

Finally define the event

$$\mathcal{M}_t := \left\{ x_{a_t^*}^\top \beta^* > \max_{i \neq a_t^*} x_{a_t}^\top \beta^* + h_{t-1} \right\}.$$

Under $\mathcal{M}_t \cap \mathcal{E}_{t-1} \cap \mathcal{G}_{t-1}$, we have the following for any $i \neq a_t^*$:

$$\begin{aligned} x_{a_t^*}^\top(t)\tilde{\beta}_t - x_i^\top(t)\tilde{\beta}_t &= \left\langle x_{a_t^*}(t), \tilde{\beta}_t - \beta^* \right\rangle + \langle x_{a_t}(t) - x_i(t), \beta^* \rangle + \left\langle x_i(t), \beta^* - \tilde{\beta}_t \right\rangle \\ &\geq -\delta_{t-1} + h_{t-1} - \delta_{t-1}. \end{aligned}$$

Thus, if we set $h_{t-1} = 3\delta_{t-1}$ then $x_{a_t^*}^\top(t)\tilde{\beta}_t - \max_{i \neq a_t^*} x_i^\top(t)\tilde{\beta}_t \geq \delta_{t-1}$. As a result, in tth round the regret is 0 almost surely as the optimal arm will be chosen with probability 1. Thus, finally using Assumption 5.2.3(b), we have

$$\begin{aligned} \mathbb{E}(\Delta_{a_t}(t)) &= \mathbb{E}\{\Delta_{a_t}(t)\mathbb{1}_{\mathcal{M}_t^c}\} \\ &= \mathbb{E}\{\Delta_{a_t}(t)\mathbb{1}_{\mathcal{M}_t^c \cap \mathcal{E}_{t-1} \cap \mathcal{G}_{t-1}}\} + \mathbb{E}\{\Delta_{a_t}(t)\mathbb{1}_{\mathcal{M}_t^c \cap (\mathcal{E}_{t-1} \cap \mathcal{G}_{t-1})^c}\} \\ &\leq 2\delta_{t-1}\mathbb{P}(\mathcal{M}_t^c) + 2x_{\mathsf{max}} b_{\mathsf{max}} \mathbb{P}(\mathcal{E}_t^c \cup \mathcal{G}_t^c) \\ &\leq 2\delta_{t-1}\mathbb{P}(\mathcal{M}_t^c) + \frac{2M_1 x_{\mathsf{max}} b_{\mathsf{max}}}{d^{s^*}} + \frac{2x_{\mathsf{max}} b_{\mathsf{max}}}{d} \\ &\quad + M_3 x_{\mathsf{max}} b_{\mathsf{max}} \exp\{-c_2\kappa^2(\xi, \vartheta, K)t + Ds^* \log K + Ds^* \log(3d/\varepsilon)\} \\ &\quad + M_4 x_{\mathsf{max}} b_{\mathsf{max}} \exp\left\{-c_8 \log(K)t + Ds^* \log K + Ds^* \log(3d/\varepsilon)\right\}, \end{aligned} \tag{D.24}$$

where M_3, M_4 are large enough universal positive constants and $D = \max\{C, \tilde{C}_1\} = \Theta(\phi_u \vartheta^2 \xi K \log K)$. Thus, if we set

$$T_0 = M_5 \max\left\{ \frac{1}{\kappa^2(\xi, \vartheta, K)}, \frac{1}{\log K} \right\} (Ds^* \log K + Ds^* \log(3d/\varepsilon)), \tag{D.25}$$

for some large universal constant $M_5 > 0$. Thus, we have

$$\mathbb{E}\{\tilde{R}(T)\} \leq \underbrace{2 \sum_{t=T_0+1}^{T} \delta_{t-1}\mathbb{P}(\mathcal{M}_t^c)}_{I_\omega} + M_6 x_{\mathsf{max}} b_{\mathsf{max}} \exp\left\{-M_7(Ds^* \log K + Ds^* \log(3d/\varepsilon))\right\} + O(\log T).$$

Recall that

$$\delta_t = M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K \left\{ \frac{s^{*2}(\log d + \log t)}{t} \right\}^{1/2}.$$

For $\omega \in [0,1]$ we have

$$\begin{aligned} I_\omega &\leq 2 \sum_{t=T_0+1}^{T} \delta_{t-1} \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^\omega \\ &\asymp \frac{\{3M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K\}^{1+\omega} \, s^{*1+\omega}}{\Delta_*^\omega} \int_{T_0}^{T} (\log d + \log u)^{\frac{1+\omega}{2}} u^{-\frac{1+\omega}{2}} \, du \\ &\lesssim \begin{cases} \frac{\left[3M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K\right]^{1+\omega} s^{*1+\omega} (\log d)^{\frac{1+\omega}{2}} T^{\frac{1-\omega}{2}}}{\Delta_*^\omega}, & \text{for } \omega \in [0,1), \\ \frac{\left[3M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K\right]^2 s^{*2} (\log d + \log T) \log T}{\Delta_*}, & \text{for } \omega = 1. \end{cases} \end{aligned} \tag{D.26}$$

For $\omega \in (1, \infty)$ we have

$$I_\omega \leq 2 \sum_{t=T_0+1}^{T} \delta_{t-1} \min \left\{ 1, \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^\omega \right\}. \tag{D.27}$$

Note that

$$\frac{3\delta_{t-1}}{\Delta_*} \leq 1 \Rightarrow t \geq T_1 := \left[3M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K\right]^2 \frac{s^{*2} \log d}{\Delta_*^2} + 1.$$

Thus. from Equation (D.27) we have

$$\begin{aligned} I_\omega &\leq 2 \sum_{t=T_0+1}^{T_1} \delta_{t-1} + 2 \sum_{t=T_1+1}^{T} \delta_{t-1} \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^\omega \\ &\leq 2 \int_{T_0}^{T_1} M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K \left\{ \frac{s^{*2} (\log d + \log u)}{u} \right\}^{1/2} du + \quad 2 \sum_{t=T_1+1}^{T} \delta_{t-1} \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^\omega \\ &\lesssim 4 \left[M_2 \rho \sigma x_{\mathsf{max}}^2 \xi K\right]^2 \left\{ \frac{s^{*2} (\log d + \log T)}{\Delta_*} \right\} + 2 J_\omega, \end{aligned}$$

where $J_\omega := \sum_{t=T_1+1}^{T} \delta_{t-1} \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^\omega$.

Finally, we give bound on the term J_ω:

$$
\begin{aligned}
J_\omega &= \sum_{t=T_1+1}^{T} \delta_{t-1} \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^{\omega} \\
&= \sum_{t=2}^{T} \delta_{t-1} \left(\frac{3\delta_{t-1}}{\Delta_*} \right)^{\omega} \mathbb{1}\{3\delta_{t-1}/\Delta_* \leq 1\} \\
&\leq \left(\frac{3^{\frac{\omega}{1+\omega}} M_2 \rho\sigma x_{\mathsf{max}}^2 \xi K}{\Delta_*^{\frac{\omega}{1+\omega}}} \right)^{1+\omega} \int_1^T \{s^{*2}(\log d + \log u)\}^{\frac{1+\omega}{2}} u^{-\frac{1+\omega}{2}} \mathbb{1}\{u \geq T_1\}\, du \\
&\leq \left(\frac{3^{\frac{\omega}{1+\omega}} M_2 \rho\sigma x_{\mathsf{max}}^2 \xi K}{\Delta_*^{\frac{\omega}{1+\omega}}} \right)^{1+\omega} \{s^{*2}(\log d + \log T)\}^{\frac{1+\omega}{2}} \int_{T_1}^{\infty} u^{-\frac{1+\omega}{2}}\, du \\
&= 2 \left(\frac{3^{\frac{\omega}{1+\omega}} M_2 \rho\sigma x_{\mathsf{max}}^2 \xi K}{\Delta_*^{\frac{\omega}{1+\omega}}} \right)^{1+\omega} \{s^{*2}(\log d + \log T)\}^{\frac{1+\omega}{2}} \frac{T_1^{-\frac{\omega-1}{2}}}{\omega - 1} \\
&\asymp \left\{ \frac{6\left[M_2 \rho\sigma x_{\mathsf{max}}^2 \xi K\right]^2}{(\omega - 1)} \right\} \left(\frac{s^{*2} \log d}{\Delta_*} \right).
\end{aligned}
\tag{D.28}
$$

Finally, for $\omega = \infty$ it is easy to see that $J_\omega = 0$. Hence, the result follows from combing Equation (D.19), (D.26), (D.27) and (D.28).

D.3 Posterior contraction result

We briefly describe the probability space under which we are working. Given β, the bandit environment (along with the specific policy π) gives rise to the chosen contexts X_t and rewards $\mathbf{r}_t$. Here we note that the chosen contexts depend not only on the arm-specific distributions, but also on the sequence of actions taken under π till time t. Let $\mathcal{Q}_t$ denote the joint distribution of $(\beta, X_t, \mathbf{r}_t)$ under $\beta \sim \Pi$ (prior) and $(X_t, \mathbf{r}_t) \mid \beta \sim \mathrm{SLCB}_t(\beta, \pi, \mathcal{P}_\epsilon)$ where the latter indicates the joint distribution of the observed contexts and rewards (till time t) under the SLCB environment with policy π, true parameter β and $\mathcal{P}_\epsilon$ denotes the noise distribution. We work under a likelihood misspecified regime, which we now discuss.

We assume that the true parameter is β^* and the observations $(X_t, \mathbf{r}_t)$ is generated from $\mathcal{Q}_t^* := \mathrm{SLCB}_t(\beta^*, \pi, \mathcal{P}_\epsilon^*)$, where π is the policy given by the TS and $\mathcal{P}_\epsilon^*$ is an arbitrary sub-Gaussian distribution. We denote by $\mathcal{Q}_{t,\mathbf{r}_t}^{*X}$ the conditional distribution of $\mathbf{r}_t$ given X_t arising from the joint $\mathcal{Q}_t^*$ and $\mathbb{E}_{t,\mathbf{r}_t}^X$ to be the expectation under $\mathcal{Q}_{t,\mathbf{r}_t}^{*X}$. Furthermore, we denote by $\mathcal{Q}_{X_t}^*$ the marginal distribution of X_t under the joint $\mathcal{Q}_t^*$ and $\mathbb{E}_{X_t}$ to be the corresponding expectation.

For modelling purpose, we place prior Π on β and model the likelihood as $(X_t, \mathbf{r}_t) \mid \beta \sim$

$\mathrm{SLCB}_t(\beta, \pi, \mathcal{P}_\epsilon)$, where $\mathcal{P}_\epsilon$ is taken to be $\mathsf{N}(0, \sigma^2)$. This gives rise to a joint distribution $\mathcal{Q}_t$, as discussed above. Now, let $\Pi_t^X(\cdot \mid \mathbf{r}_t)$ denote the posterior distribution of β given all others, i.e. it is the conditional measure of β given $X_t, \mathbf{r}_t$ arising from the joint $\mathcal{Q}_t$.

Thus, given a measurable set B, $\Pi_t^X(B|\mathbf{r}_t)$ is a random measure, whose randomness is due to (X_t, r_t). In the following result, we consider $\mathbb{E}_{t,\mathbf{r}_t}^X \Pi_t^X(B|\mathbf{r}_t)$, which is the expectation of the above under $\mathcal{Q}_{t,\mathbf{r}_t}^{*X}$. Thus, this quantity itself is a random variable, whose randomness is due to X_t. The following result shows that, for B taken as the complement of an appropriate ball around the true β^*, this random variable is small, almost surely $\mathcal{Q}_{X_t}^*$.

Theorem D.3.1. *Consider the bandit problem in* (5.1) *and let Assumption 5.2.2-5.2.4 hold. Also, assume that the prior on parameter β is modeled as* (5.2) *with*

$$(5/3)\overline{\lambda}_t \leq \lambda \leq 2\overline{\lambda}_t, \quad \overline{\lambda}_t = x_{\mathsf{max}}\sqrt{2t(\log d + \log t)}$$

Then the following is true:

$$\mathbb{E}_{t,\mathbf{r}_t}^X \left\{ \Pi_t^X \left(\|\beta - \beta^*\|_1 \geq Q_4 \sigma x_{\mathsf{max}} K \xi (D_* + s^*) \sqrt{\frac{\log d + \log t}{t}} \;\middle|\; \mathbf{r}_t \right) \mathbb{1}_{\mathcal{G}_t \cap \mathcal{T}_t(\lambda_0(\gamma_t))} \right\} \lesssim d^{-s^*},$$

almost sure X_t, where Q_4 is a universal constant and $D_* = D_1 s^* + \frac{D_2 x_{\mathsf{max}}^2 s^*}{\phi_{\mathsf{comp}}^2(S^*; X_t)}$ with $D_1 = 1 + (40/A_4)$ and $D_2 = 128 A_4^{-1}$.

Proof. Without loss of generality we assume that $\sigma = 1$ as the bandit reward model can be viewed as

$$(\mathbf{r}_t/\sigma) = X_t(\beta^*/\sigma) + (\epsilon_t/\sigma).$$

In this case $\overline{\lambda}_t = x_{\mathsf{max}}\sqrt{2t(\log d + \log t)} =: \overline{\lambda}$.

Next, define the event

$$\mathcal{T}_0 := \left\{ \max_{j \in [d]} \left| \epsilon_t^\top X_t^{(j)} \right| \leq \bar{\lambda} \right\}.$$

By Lemma D.2.12 and condition (D.18), it follows that fro any measurable set $\mathcal{B} \subset \mathbb{R}^d$,

$$\mathbb{E}_{t,\mathbf{r}_t} \Pi_t^X(\mathcal{B} \mid \mathbf{r}_t) \leq \left[\mathbb{E}_{t,\mathbf{r}_t} \left\{ \Pi_t^X(\mathcal{B} \mid \mathbf{r}_t) \mathbb{1}_{\mathcal{T}_0} \right\} \right]^{1/2} + \frac{2}{t}.$$

Recall that the errors ϵ_t is modeled as isotropic standard Gaussian independent of the features. Thus, conditioned on the matrix X_t, model likelihood ration takes the following form:

$$\mathcal{L}_{t,\beta,\beta^*}(\mathbf{r}_t) := \exp\left\{ -\frac{\|X_t\beta - X_t\beta^*\|_2^2}{2} + (\mathbf{r}_t - X_t\beta^*)^\top (X_t\beta - X_t\beta^*) \right\}.$$

Then by Lemma 2 of Castillo et al. (2015) it follows that

$$\int \mathcal{L}_{t,\beta,\beta^*}(\mathbf{r}_t)\, d\Pi(\beta) \geq \frac{\pi_d(s^*)}{p^{2s^*}} e^{-\lambda\|\beta^*\|_1} e^{-1},$$

where Π is given by (5.2). The only change that is needed in their proof to run the argument in our case is the following upper bound:

$$\|X\beta\|_2 \leq \|\beta\|_1 \|X\| \leq c_9 \vartheta^2 \phi_u \log K/(1-3\varepsilon).$$

The last inequality follows from the fact that we are on the event $\mathcal{G}_{t,1}$ by assumption. The rest of the proof follows from the fact that $\lambda \in (5\overline{\lambda}/3, 2\overline{\lambda})$.

Thus by Bayes's formula it follows that

$$\begin{aligned}\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t) &= \frac{\int_{\mathcal{B}} \mathcal{L}_{t,\beta,\beta^*}(\mathbf{r}_t)\, d\Pi(\beta)}{\int \mathcal{L}_{t,\beta,\beta^*}(\mathbf{r}_t)\, d\Pi(\beta)} \\ &\leq \frac{ed^{2s^*}}{\pi_d(s^*)} e^{\lambda\|\beta^*\|_1} \int_{\mathcal{B}} \exp\left\{ -\frac{\|X_t\beta - X_t\beta^*\|_2^2}{2} + (\mathbf{r}_t - X_t\beta^*)^\top (X_t\beta - X_t\beta^*) \right\} d\Pi(\beta). \end{aligned} \tag{D.29}$$

Using Holder's inequality, we see that on $\mathcal{T}_0$,

$$\begin{aligned}(\mathbf{r}_t - X_t\beta^*)^\top X_t(\beta - \beta^*) &= \epsilon_t^\top X_t(\beta - \beta^*) \\ &\leq \left\|\epsilon_t^\top X_t\right\|_\infty \|\beta - \beta^*\|_1 \\ &\leq \bar{\lambda} \|\beta - \beta^*\|_1 .\end{aligned} \tag{D.30}$$

Therefore, on the event $\mathcal{T}_0$, the expected value under $\mathbb{E}_{\beta^*}$ of the integrand on the right hand side of (D.29) is bounded above by

$$e^{-(1/2)\|X_t(\beta-\beta^*)\|_2^2 + \bar{\lambda}\|\beta-\beta^*\|_1}.$$

Thus, we have

$$\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0} \leq \frac{ep^{2s^*}}{\pi_d(s^*)} \int_{\mathcal{B}} e^{\lambda\|\beta\|_1 - (1/2)\|X_t(\beta-\beta^*)\|_2^2 + \bar{\lambda}\|\beta-\beta^*\|_1}\, d\Pi(\beta). \tag{D.31}$$

Now, by triangle inequality,

$$\begin{aligned}\lambda \left\|\beta^*\right\|_1 + \overline{\lambda} \left\|\beta - \beta^*\right\|_1 &= \lambda \left\|\beta^*_{S^*}\right\|_1 + \overline{\lambda} \left\|\beta - \beta^*\right\|_1 \\ &\leq \lambda \left\|\beta^*_{S^*} - \beta_{S^*}\right\|_1 + \lambda \left\|\beta_{S^*}\right\|_1 + \overline{\lambda} \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 + \overline{\lambda} \left\|\beta_{S^{*c}}\right\| \\ &= \lambda \left\|\beta_{S^*}\right\|_1 + \overline{\lambda} \left\|\beta_{S^{*c}}\right\|_1 + (\lambda + \overline{\lambda}) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &= \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - \lambda) \left\|\beta_{S^{*c}}\right\|_1 + (\lambda + \overline{\lambda}) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 .\end{aligned} \tag{D.32}$$

Case 1: Suppose $7 \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \leq \left\|\beta_{S^{*c}}\right\|_1$. Then the following holds:

$$\begin{aligned}(\lambda + \overline{\lambda}) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 &= (\overline{\lambda} - 3\lambda/4) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 + (7\lambda/4) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &\leq (\overline{\lambda} - 3\lambda/4) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 + (\lambda/4) \left\|\beta_{S^{*c}}\right\|_1 .\end{aligned}$$

Using the above inequality in (D.32) we get

$$\begin{aligned}\lambda \left\|\beta^*\right\|_1 + \overline{\lambda} \left\|\beta - \beta^*\right\|_1 &\leq \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - 3\lambda/4) \left\|\beta_{S^{*c}}\right\|_1 + (\overline{\lambda} - 3\lambda/4) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &= \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - 3\lambda/4) \left\|\beta - \beta^*\right\|_1\end{aligned} \tag{D.33}$$

Case 2: Now assume $7 \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 > \left\|\beta_{S^{*c}}\right\|_1$. We again focus on the inequality (D.32), i.e.,

$$\begin{aligned}\lambda \left\|\beta^*\right\|_1 + \overline{\lambda} \left\|\beta - \beta^*\right\|_1 &\leq \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - \lambda) \left\|\beta_{S^{*c}}\right\|_1 + (\lambda + \overline{\lambda}) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &= \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - \lambda) \left\|\beta_{S^{*c}}\right\|_1 + (\overline{\lambda} - \lambda) \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &\quad + 2\lambda \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &= \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - \lambda) \left\|\beta - \beta^*\right\|_1 + 2\lambda \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 \\ &\leq \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - 3\lambda/4) \left\|\beta - \beta^*\right\|_1 + 2\lambda \left\|\beta_{S^*} - \beta^*_{S^*}\right\|_1 .\end{aligned} \tag{D.34}$$

Finally, by compatibility condition and Young's inequality we get

$$2\lambda \left\|\beta_{S^*} - \beta^*_{S^*}\right\| \leq 2\lambda \frac{\left\|X_t(\beta - \beta^*)\right\|_2 s^{*1/2}}{t^{1/2} \phi_{\mathsf{comp}}(S^*; X_t)} \leq \frac{\left\|X_t(\beta - \beta^*)\right\|_2^2}{2} + \frac{2s^*\lambda^2}{t\phi^2_{\mathsf{comp}}(S^*; X_t)}.$$

Using the above inequality in (D.34) we get

$$\lambda \left\|\beta^*\right\|_1 + \overline{\lambda} \left\|\beta - \beta^*\right\|_1 \leq \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - 3\lambda/4) \left\|\beta - \beta^*\right\|_1 + \frac{\left\|X_t(\beta - \beta^*)\right\|_2^2}{2} + \frac{2s^*\lambda^2}{t\phi^2_{\mathsf{comp}}(S^*; X_t)}. \tag{D.35}$$

Thus combining (D.33) and (D.35) we can conclude

$$\lambda \left\|\beta^*\right\|_1 + \overline{\lambda} \left\|\beta - \beta^*\right\|_1 \leq \lambda \left\|\beta\right\|_1 + (\overline{\lambda} - 3\lambda/4) \left\|\beta - \beta^*\right\|_1 + \frac{\|X_t(\beta - \beta^*)\|_2^2}{2} + \frac{2s^*\lambda^2}{t\phi^2_{\mathsf{comp}}(S^*; X_t)}. \tag{D.36}$$

Using the above result and recalling that $5\overline{\lambda}/3 \leq \lambda \leq 2\overline{\lambda}$, we see that the right hand side of (D.31) is bounded by

$$\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0} \leq \frac{ed^{2s^*}}{\pi_d(s^*)} e^{\frac{2s^*\lambda^2}{t\phi^2_{\mathsf{comp}}(S^*;X_t)}} \int_{\mathcal{B}} e^{\lambda\|\beta\|_1 - (\lambda/4)\|\beta - \beta^*\|_1} \, d\Pi(\beta)$$

Controlling sparsity: For the set $\mathcal{B} = \{\beta : |S_\beta| > L\}$ and $L \geq s^*$, the above integral can be bounded by

$$\begin{aligned}
&\sum_{S: s=|S|>L} \frac{\pi_d(s)}{\binom{d}{s}} \left(\frac{\lambda}{2}\right)^s \int e^{-(\lambda/4)\left\|\beta_S - \beta^*_S\right\|} \, d\beta_S \\
&\leq \sum_{s=L+1}^{\infty} \pi_d(s) 4^s \\
&\leq \pi_d(s^*) 4^{s^*} \left(\frac{4A_2}{d^{A_4}}\right)^{L+1-s^*} \sum_{j=0}^{\infty} \left(\frac{4A_2}{d^{A_4}}\right)^j
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}^X_{t,\mathbf{r}_t}\left\{\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0}\right\} \\
&\lesssim \exp\left\{4s^*\log d + \frac{2s^*\lambda^2}{t\phi^2_{\mathsf{comp}}(S^*; X_t)} + s^*\log 4 - (A_4/4)(L + 1 - s^*)\log d\right\} \\
&\leq \exp\left\{5s^*\log d + \frac{4s^*\lambda\overline{\lambda}}{t\phi^2_{\mathsf{comp}}(S^*; X_t)} - (A_4/4)(L + 1 - s^*)\log d\right\}
\end{aligned}$$

Now recall that $\overline{\lambda}^2 = 2tx^2_{\mathsf{max}}(\log d + \log t) \leq 4tx^2_{\mathsf{max}} \log d$. Using this inequality in the above display we have

$$\mathbb{E}^X_{t,\mathbf{r}_t}\left\{\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0}\right\} \lesssim \exp\left\{5s^*\log d + \frac{16s^*(\lambda/\overline{\lambda})x^2_{\mathsf{max}}\log d}{\phi^2_{\mathsf{comp}}(S^*; X_t)} - (A_4/4)(L + 1 - s^*)\log d\right\}.$$

Thus, setting $L \geq 40s^*/A_4 + s^* + \frac{64A_4^{-1}s^*x^2_{\mathsf{max}}}{\phi^2_{\mathsf{comp}}(S^*;X_t)}(\lambda/\overline{\lambda})$ then there exists a universal constant $Q_1 > 0$ such that

$$\mathbb{E}^X_{t,\mathbf{r}_t}\left\{\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0}\right\} \leq Q_1 d^{-s^*}.$$

Control on prediction: Recall that $\lambda/\overline{\lambda} \leq 2$. Using this and the result in the previous part, we can conclude that the posterior distribution is asymptotically supported on the even $\mathcal{B}_1 = \{\beta : |S_\beta| \leq D_*\}$, where $D_* = D_1 s^* + \frac{D_2 x_{\mathsf{max}}^2 s^*}{\phi_{\mathsf{comp}}^2(S^*;X_t)}$ where $D_1 = 1 + (40/A_4)$ and $D_2 = 128A_4^{-1}$. By combining (D.29), (D.30) and the inequality $\lambda \|\beta^*\|_1 \leq 2\overline{\lambda}\|\beta - \beta^*\|_1 + \lambda\|\beta\|_1$ we can conclude that any Borel set $\mathcal{B}$,

$$\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0} \leq \frac{ed^{2s^*}}{\pi_d(s^*)} \int_{\mathcal{B}} \exp\left\{ -\frac{\|X_t\beta - X_t\beta^*\|_2^2}{2} + 3\overline{\lambda}\|\beta - \beta^*\|_1 + \lambda\|\beta\|_1 \right\} d\Pi(\beta).$$

By the definition in (D.20) we have,

$$\begin{aligned}(4-1)\overline{\lambda}\|\beta - \beta^*\|_1 &\leq \frac{4\overline{\lambda}\|X_t(\beta-\beta^*)\|_2 |S_{\beta-\beta^*}|^{1/2}}{t^{1/2}\overline{\phi}_t(|S_{\beta-\beta^*}|)} - \overline{\lambda}\|\beta - \beta^*\|_1 \\ &\leq \frac{1}{4}\|X_t(\beta-\beta^*)\|_2^2 + \frac{16\overline{\lambda}^2|S_{\beta-\beta^*}|}{t\overline{\phi}_t(|S_{\beta-\beta^*}|)^2} - \overline{\lambda}\|\beta-\beta^*\|_1 .\end{aligned} \tag{D.37}$$

Since $|S_{\beta-\beta^*}| \leq |S_\beta| + s^* \leq D_* + s^*$ on the event $\mathcal{B}_1$, it follows that

$$\begin{aligned}\Pi_t^X(\mathcal{B} \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0} \leq \frac{ed^{2s^*}}{\pi_d(s^*)} &e^{16\overline{\lambda}^2(D_*+s^*)/(t\overline{\psi}_t(S^*)^2)} \\ &\times \int_{\mathcal{B}} e^{-(1/4)\|X_t(\beta-\beta^*)\|_2^2 - \overline{\lambda}\|\beta-\beta^*\|_1 + \lambda\|\beta\|_1} \, d\Pi(\beta).\end{aligned} \tag{D.38}$$

Now we set $\mathcal{B} = \mathcal{B}_2 := \{\beta \in \mathcal{B}_1 : \|X_t(\beta-\beta^*)\|_2 > L\}$, where L will be chosen shortly. Recall that $\pi_d(s^*) \geq (A_1 p^{-A_3})^{s^*}\pi_p(0)$. It follows that for set $\mathcal{B}$, the right hand side of (D.38) is upper bounded by

$$\begin{aligned}&\frac{ed^{2s^*}}{\pi_d(s^*)} e^{16\overline{\lambda}^2(D_*+s^*)/(t\overline{\psi}_t(S^*)^2)} e^{-(1/4)L^2} \int e^{-\overline{\lambda}\|\beta-\beta^*\|_1 + \lambda\|\beta\|_1} \, d\Pi(\beta) \\ &\lesssim d^{(2+A_3)s^*} A_1^{-s^*} e^{16\overline{\lambda}^2(D_*+s^*)/(t\overline{\psi}_t(S^*)^2)} e^{-(1/4)L^2} \underbrace{\sum_{s=0}^{d} \pi_d(s^*) 2^s}_{O(1)} .\end{aligned}$$

Hence by a calculation similar to previous discussion yields that for

$$\frac{1}{4}L^2 = (3+A_3)s^*\log d + \frac{16\overline{\lambda}^2(D_*+s^*)}{t\overline{\psi}_t(S^*)^2} \leq Q_2 x_{\mathsf{max}}^2 (D_*+s^*)\frac{\log d + \log t}{\overline{\psi}_t(S^*)^2} =: L_*^2,$$

where $Q_2 > 0$ is sufficiently large universal constant, then we have

$$\mathbb{E}^X_{t,\mathbf{r}_t}\left\{\Pi^X_t(\mathcal{B}_2 \mid \mathbf{r}_t)\mathbb{1}_{\mathcal{T}_0}\right\} \leq \frac{1}{d^{s^*}}.$$

Control on estimation: Similar to (D.37) we have

$$\begin{aligned}\overline{\lambda}\left\|\beta-\beta^*\right\|_1 &\leq \frac{\overline{\lambda}\left\|X_t(\beta-\beta^*)\right\|_2 |S_{\beta-\beta^*}|^{1/2}}{t^{1/2}\overline{\phi}_t(|S_{\beta-\beta^*}|)} \\ &\leq \frac{1}{4}\left\|X_t(\beta-\beta^*)\right\|_2^2 + \frac{\overline{\lambda}^2\,|S_{\beta-\beta^*}|}{t\overline{\phi}_t(|S_{\beta-\beta^*}|)^2}.\end{aligned}$$

On the event $\mathcal{B}_2$, we thus have

$$\overline{\lambda}\left\|\beta-\beta^*\right\|_1 \leq Q_3 x^2_{\mathsf{max}}(D_* + s^*)\frac{\log d + \log t}{\overline{\psi}_t(S^*)^2}.$$

Finally on the event $\mathcal{B}_2 \cap \mathcal{G}_t$ we have $\overline{\lambda} = x_{\mathsf{max}}\sqrt{2t(\log d + \log t)}$ and $\overline{\psi}_t(S^*)^2 \gtrsim (K\xi)^{-1}$ and it follows that

$$\left\|\beta-\beta^*\right\|_1 \leq Q_4 K\xi x_{\mathsf{max}}(D_* + s^*)\sqrt{\frac{\log d + \log t}{t}}.$$

□

D.4 Technical lemmas

D.4.1 Proof of Lemma D.2.6

As A is symmetric positive definite matrix, by Cholesky decomposition there exists as lower triangular matrix L such that $A = LL^\top$. Let $v \in \mathbb{S}_0^{d-1}(s)$. Then there exists a index set J of size s, such that $\mathsf{supp}(v) \subseteq J$. Hence we have $v \in E_J$. Now consider the net $\mathcal{N}_{\varepsilon,J}$ and let u be the nearest point to v in $\mathcal{N}_{\varepsilon,J}$. Thus we have $\left\|v-u\right\|_2 \leq \varepsilon < 1$ and $\left\|v-u\right\|_0 \leq s$. Then we have the following:

$$\begin{aligned}v^\top Av &= (v-u)^\top A(v-u) + 2(v-u)^\top Au + u^\top Au \\ &\geq u^\top Au - 3\varepsilon\phi_{\max}(s;A).\end{aligned} \tag{D.39}$$

The second inequality follows from the following facts:

$$\left|(v-u)^\top A(v-u)\right| \leq \left\|v-u\right\|_2^2 \phi_{\max}(s;A) \leq \varepsilon\phi_{\max}(s;A),$$

$$\begin{aligned}\left|(v-u)^\top Au\right| &= \left|(v-u)^\top LL^\top u\right| \\ &\le \sqrt{(v-u)^\top LL^\top (v-u)}\sqrt{u^\top LL^\top u} \\ &= \sqrt{(v-u)^\top A(v-u)}\sqrt{u^\top Au} \\ &\le \varepsilon\phi_{\max}(s;A)\end{aligned}$$

Then the result follows from taking infimum over u and v in both sides.

D.4.2 Proof of Lemma D.2.8

The lower bound result is trivial. Hence, we focus on the upper bound part. As A is symmetric positive definite matrix, by Cholesky decomposition there exists as lower triangular matrix L such that $A = LL^\top$. Let $v \in \mathbb{S}_0^{d-1}(s)$. Then there exists a index set J of size s, such that $\mathsf{supp}(v) \subseteq J$. Hence we have $v \in E_J$. Now consider the net $\mathcal{N}_{\varepsilon,J}$ and let u be the nearest point to v in $\mathcal{N}_{\varepsilon,J}$. Thus we have $\|v-u\|_2 \le \varepsilon < 1/3$ and $\|v-u\|_0 \le s$. By a similar argument as in the proof of Lemma D.2.6, we can conclude that

$$v^\top Av \le 3\varepsilon\phi_{\max}(s;A) + \max_{u\in\mathcal{N}_\varepsilon} u^\top Au.$$

Thus by taking supremum over v on the left-hand side of the above display, we get

$$(1-3\varepsilon)\phi_{\max}(s;A) \le \max_{u\in\mathcal{N}_\varepsilon} u^\top Au \iff \phi_{\max}(s;A) \le \frac{1}{1-3\varepsilon}\max_{u\in\mathcal{N}_\varepsilon} u^\top Au.$$

D.5 Pseudo code of VBTS and other tables

Algorithm 3: Variational Bayes Thompson Sampling

```
Set H_0 = ∅;
for t = 1, ··· , T do
    if t ≤ 1 then
        Choose action a_t uniformly over [K];
    end
    else
        Compute VB posterior Π̃_{t−1} from Π(· | H_{t−1}) using CAVI ;
        Generate sample β̃_t ∼ Π̃_{t−1};
        Play arm: a_t = arg max_{i∈[K]} x_i(t)^⊤ β̃_t;
    end
    Observe reward r_{a_t}(t);
    Update H_t ← H_{t−1} ∪ {(a_t, r_{a_t}(t), x_{a_t}(t))}.
end
```

Table D.1: Time comparison among the competing algorithms.

Type	Algorithm	Mean time of execution (seconds)	
		Equicorrelated	Auto-regressive
Non-TS	LinUCB	15.41	16.30
	DR Lasso	3.12	3.13
	Lasso-L1	3.57	3.59
	ESTC	1.01	1.05
TS	LinTS	1344.39	1346.46
	BLasso TS	1511.68	1455.53
	VBTS	**29.33**	**27.65**

Bibliography

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24.

Abe, N., Biermann, A. W., and Long, P. M. (2003). Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293.

Adamczak, R. (2015). A note on the hanson-wright inequality for random vectors with dependencies. *Electron. Commun. Probab.*, 20:1–13.

Adler, R. J., Taylor, J. E., et al. (2007). *Random fields and geometry*, volume 80. Springer.

Aeron, S., Saligrama, V., and Zhao, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory*, 56(10):5111–5130.

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control.*, 19(6):716–723.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Akçakaya, M. and Tarokh, V. (2009). Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inf. Theory*, 56(1):492–504.

Alotaiby, T., El-Samie, F. E. A., Alshebeili, S. A., and Ahmad, I. (2015). A review of channel selection algorithms for eeg signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015:1–21.

Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):971–989.

Apple (2017). Learning with privacy at scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.*, 39:2533–2556.

Ariu, K., Abe, K., and Proutière, A. (2022). Thresholded lasso bandit. In *International Conference on Machine Learning*, pages 878–928. PMLR.

Atmaca, U. I., Maple, C., Epiphaniou, G., and Dianati, M. (2021). A privacy-preserving route planning scheme for the internet of vehicles. *Ad Hoc Networks*, 123:102680.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

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.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413.

Bastani, H. and Bayati, M. (2015). Online decision-making with high-dimensional covariates, 2015. *Available at SSRN 2661896*.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294.

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., and Liss, J. (2021). Digital medicine and the curse of dimensionality. *NPJ digital medicine*, 4(1):153.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.*, 44(2):813–852.

Bertsimas, D. and Parys, B. V. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *Ann. Statist.*, 48(1):300–323.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Cai, T. T., Wang, Y., and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509.

Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010a). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010b). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Chakraborty, S., Roy, S., and Tewari, A. (2023). Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24.

Chen, C., Luo, L., Zhang, W., Yu, Y., and Lian, Y. (2021). Efficient and robust high-dimensional linear contextual bandits. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4259–4265.

Chen, Y., Wang, Y., Fang, E. X., Wang, Z., and Li, R. (2022). Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *Journal of the American Statistical Association*, (just-accepted):1–31.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Clara, G., Szabo, B., and Ray, K. (2021). sparsevb: Spike-and-slab variational bayes for linear and logistic regression: R package version 0.1. 0.

Consortium, W. T. C. C. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306.

Durfee, D. and Rogers, R. M. (2019). Practical differentially private top-k selection with pay-what-you-get composition. *Advances in Neural Information Processing Systems*, 32.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015a). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. (2015b). Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pages 486–503. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Dwork, C., Su, W., and Zhang, L. (2015c). Private false discovery rate control. *arXiv preprint arXiv:1511.03803*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, 70(5):849–911.

Fan, J., Shao, Q.-M., and Zhou, W.-X. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *Ann. Statist.*, 46(3):989–1017.

Ferdowsi, A., Ali, S., Saad, W., and Mandayam, N. B. (2019). Cyber-physical security and safety of autonomous connected vehicles: Optimal control meets multi-armed bandit learning. *IEEE Transactions on Communications*, 67(10):7228–7244.

Fletcher, A. K., Rangan, S., and Goyal, V. K. (2009). Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory*, 55(12):5758–5772.

Foster, D., Karloff, H., and Thaler, J. (2015). Variable selection is hard. In *Conference on Learning Theory*, pages 696–709. PMLR.

Garfinkel, S. (2022). Differential privacy and the 2020 us census. https://mit-serc.pubpub.org/pub/differential-privacy-2020-us-census/release/1.

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.*, 13:2107–2143.

George, E. I. and McCulloch, R. E. (1993a). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1993b). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gilton, D. and Willett, R. (2017). Sparse linear contextual bandits via relevance vector machines. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 518–522. IEEE.

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems*, 3(1):230–261.

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. Stat.*, 12(3):364–366.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., De Rycke, Y., Pierga, J.-Y., Lucchesi, C., Reyal, F., et al. (2010). A prognostic dna signature for t1t2 node-negative breast cancer patients. *Genes, chromosomes and cancer*, 49(12):1125–1134.

Guo, H., Li, T., and Wang, Z. (2023). Pleiotropic genetic association analysis with multiple phenotypes using multivariate response best-subset selection. *BMC genomics*, 24(1):759.

Guo, X., Zhang, H., Wang, Y., and Wu, J.-L. (2015). Model selection and estimation in high dimensional regression models with group SCAD. *Stat. Probab. Lett.*, 103:86–92.

Guo, Y., Zhu, Z., and Fan, J. (2020). Best subset selection is robust against design dependence. *arXiv preprint arXiv:2007.01478*.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516.

Hao, B., Lattimore, T., and Deng, W. (2021). Information directed sampling for sparse linear bandits. *Advances in Neural Information Processing Systems*, 34:16738–16750.

Hao, B., Lattimore, T., and Qin, C. (2022). Contextual information-directed sampling. In *International Conference on Machine Learning*, pages 8446–8464. PMLR.

Hao, B., Lattimore, T., and Wang, M. (2020). High-dimensional sparse linear bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10753–10763. Curran Associates, Inc.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Stat. Sci.*, 35(4):579–592.

Hazimeh, H., Mazumder, R., and Saab, A. (2022). Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Math. Program.*, 196(1-2):347–388.

He, Q. and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics*, 27(1):1–8.

Hong, H. G., Kang, J., and Li, Y. (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Anal.*, 24(1):45–71.

Huang, D., Li, R., and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *J. Bus. Econ. Stat.*, 32(2):237–244.

Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018). A constructive approach to l0 penalized regression. *The Journal of Machine Learning Research*, 19(1):403–439.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Stat. Sin.*, pages 1603–1618.

Huang, T.-J., McKeague, I. W., and Qian, M. (2019). Marginal screening for high-dimensional predictors of survival outcomes. *Stat. Sin.*, 29(4):2105–2139.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730 – 773.

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*, volume 27.

Jain, P. and Thakurta, A. G. (2014). (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484. PMLR.

Jang, K., Zhang, C., and Jun, K.-S. (2022). Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits. In *Advances in Neural Information Processing Systems*.

Ji, P., Jin, J., et al. (2012). Ups delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.*, 40(1):73–103.

Jin, J., Zhang, C.-H., and Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.*, 15(1):2723–2772.

Jin, Y., Kim, Y.-H., and Rao, B. D. (2011). Limits on support recovery of sparse signals via multiple-access communication techniques. *IEEE Trans. Inf. Theory*, 57(12):7877–7892.

Kasiviswanathan, S. P. and Jin, H. (2016). Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497. PMLR.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer.

Kifer, D., Smith, A., and Thakurta, A. (2012). Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings.

Kim, G.-S. and Paik, M. C. (2019). Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32:5877–5887.

Kim, H., Ben-Othman, J., and Mokdad, L. (2019). Udipp: A framework for differential privacy preserving movements of unmanned aerial vehicles in smart cities. *IEEE Transactions on Vehicular Technology*, 68(4):3933–3943.

Kong, W., Zhu, J., Bi, S., Huang, L., Wu, P., and Zhu, S.-J. (2023). Adaptive best subset selection algorithm and genetic algorithm aided ensemble learning method identified a robust severity score of covid-19 patients. *iMeta*, 2(3):e126.

Kowshik, S. S. and Polyanskiy, Y. (2021). Fundamental limits of many-user mac with finite payloads and fading. *IEEE Trans. Inf. Theory*, 67(9):5853–5884.

Krahmer, F., Mendelson, S., and Rauhut, H. (2014). Suprema of chaos processes and the restricted isometry property. *Commun. Pure. Appl. Math.*, 67(11):1877–1904.

Kuzborskij, I., Cella, L., and Cesa-Bianchi, N. (2019). Efficient linear bandits through matrix sketching. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 177–185. PMLR.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2013). Adaptive monte carlo for bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics*, 22(3):729–748.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, pages 1302–1338.

Lei, J., Charest, A.-S., Slavkovic, A., Smith, A., and Fienberg, S. (2018). Differentially private model selection with penalized and constrained likelihood. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):609–633.

Leisenring, W., Pepe, M. S., and Longton, G. (1997). A marginal regression modelling framework for evaluating medical diagnostic tests. *Stat. Med.*, 16(11):1263–1281.

Li, J., Zheng, Q., Peng, L., and Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72(4):1145–1154.

Li, K., Yang, Y., and Narisetty, N. N. (2021). Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. *Electronic Journal of Statistics*, 15(2):5652–5695.

Li, W., Barik, A., and Honorio, J. (2022). A simple unified framework for high dimensional bandit problems. In *International Conference on Machine Learning*, pages 12619–12655. PMLR.

Li, X., Jiang, H., Haupt, J., Arora, R., Liu, H., Hong, M., and Zhao, T. (2020). On fast convergence of proximal algorithms for sqrt-lasso optimization: Don't worry about its nonsmooth loss function. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 49–59. PMLR.

Li, Y., Hong, H. G., Ahmed, S. E., and Li, Y. (2019). Weak signals in high-dimensional regression: Detection, estimation and prediction. *Appl. Stoch. Model. Bus. Ind.*, 35(2):283–298.

Lifshits, M. A. (1995). *Gaussian random functions*, volume 322. Springer Science & Business Media.

Liu, H. and Foygel Barber, R. (2020). Between hard and soft thresholding: optimal iterative thresholding algorithms. *Inf. Inference: J. IMA*, 9(4):899–933.

Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X., and Dehmer, M. (2018). Feature selection of gene expression data for cancer classification using double rbf-kernels. *BMC bioinformatics*, 19(1):1–14.

Lu, L., Yan, J., and de Silva, C. W. (2016). Feature selection for ecg signal processing using improved genetic algorithm and empirical mode decomposition. *Measurement*, 94:372–381.

Lu, W. (2005). Marginal regression of multivariate event times based on linear transformation models. *Lifetime Data Anal.*, 11(3):389–404.

Malandraki, C. and Daskin, M. S. (1992). Time dependent vehicle routing problems: Formulations, properties and heuristic algorithms. *Transportation science*, 26(3):185–200.

Mallows, C. L. (2000). Some comments on cp. *Technometrics*, 42(1):87–94.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660.

Marttinen, P., Gillberg, J., Havulinna, A., Corander, J., and Kaski, S. (2013). Genome-wide association studies with high-dimensional phenotypes. *Stat. Appl. Genet. Mol. Biol.*, 12(4):413–431.

Matsumoto, Y. K., Himoto, Y., Nishio, M., Kikkawa, N., Otani, S., Ito, K., Yamanoi, K., Kato, T., Fujimoto, K., Kurata, Y., et al. (2023). Nodal infiltration in endometrial cancer: a prediction model using best subset regression. *European Radiology*, pages 1–10.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246 – 270.

Miller, A. (2002). *Subset selection in regression*. chapman and hall/CRC.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12:229–244.

Narisetty, N. N., Shen, J., and He, X. (2018). Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*.

Ndaoud, M. and Tsybakov, A. B. (2020). Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory*, 66(4):2517–2532.

Near, J. and Darais, D. (2022). Differential privacy: Future work & open challenges. https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84.

Oh, M.-h., Iyengar, G., and Zeevi, A. (2021). Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR.

Oliveira, R. I. (2013). The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903*.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors.

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(9).

Pochet, Y. and Wolsey, L. A. (2006). *Production planning by mixed integer programming*, volume 149. Springer.

Rad, K. R. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Inf. Theory*, 57(7):4672–4679.

Ray, K. and Szabó, B. (2021). Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12.

Ren, Z. and Zhou, Z. (2020). Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.

Rogers, R., Cardoso, A. R., Mancuhan, K., Kaura, A., Gahlawat, N., Jain, N., Ko, P., and Ahammad, P. (2020). A members first approach to enabling linkedin's labor market insights at scale. *arXiv preprint arXiv:2010.13981*.

Roy, S. and Tewari, A. (2023). On the computational complexity of private high-dimensional model selection via the exponential mechanism. *arXiv preprint arXiv:2310.07852*.

Roy, S., Tewari, A., and Zhu, Z. (2022). High-dimensional variable selection with heterogeneous signals: A precise asymptotic perspective. *arXiv preprint arXiv:2201.01508*.

Roy, S., Tewari, A., and Zhu, Z. (2023). Tale of two c (omplex) ities. *arXiv preprint arXiv:2301.06259*.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:1–9.

Scarlett, J. and Cevher, V. (2016). Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Trans. Inf. Theory*, 63(1):593–620.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, pages 461–464.

Sinclair, A. (1992). Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, probability and Computing*, 1(4):351–370.

Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822.

Steil, J., Hagestedt, I., Huang, M. X., and Bulling, A. (2019). Privacy-aware eye tracking using differential privacy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–9.

Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the lasso path. *Ann. Statist.*, pages 2133–2150.

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

Talwar, K., Guha Thakurta, A., and Zhang, L. (2015). Nearly optimal private lasso. *Advances in Neural Information Processing Systems*, 28.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pages 495–517.

Thakurta, A. G. and Smith, A. (2013). Differentially private feature selection via stability arguments, and the robustness of the lasso. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 819–850, Princeton, NJ, USA. PMLR.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288.

Tikhonov, A. (1943). On the stability of inverse problems. *Proc. USSR Acad. Sci.*, 39:195–198.

Tong, W., Hua, J., and Zhong, S. (2017). A jointly differentially private scheduling protocol for ridesharing services. *IEEE Transactions on Information Forensics and Security*, 12(10):2444–2456.

Van De Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, 3:1360–1392.

Vershynin, R. (2018a). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vershynin, R. (2018b). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Vicente-Saez, R. and Martinez-Fuentes, C. (2018). Open science now: A systematic literature review for an integrated definition. *Journal of business research*, 88:428–436.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199.

Wainwright, M. J. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory*, 55(12):5728–5741.

Wainwright, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, C. D., Chen, Z., Lian, Y., and Chen, M. (2022a). Asset selection based on high frequency sharpe ratio. *J. Econom.*, 227(1):168–188. Annals Issue: Time Series Analysis of Higher Moments and Distributions of Financial Data.

Wang, C.-H. and Cheng, G. (2020). Online batch decision-making with high-dimensional covariates. In *International Conference on Artificial Intelligence and Statistics*, pages 3848–3857. PMLR.

Wang, D. and Xu, J. (2019). On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, pages 6628–6637. PMLR.

Wang, D., Ye, M., and Xu, J. (2017). Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30.

Wang, H., Yang, Y., and Su, W. J. (2022b). The price of competition: Effect size heterogeneity matters in high dimensions. *IEEE Trans. Inf. Theory*, 68(8):5268–5294.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory*, 56(6):2967–2979.

Wang, X., Wei, M., and Yao, T. (2018). Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208. PMLR.

Wang, Y.-X. (2018). Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Ann. Statist.*, 37(5A):2178–2201.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.

Wei, J., Lin, Y., Yao, X., Zhang, J., and Liu, X. (2020). Differential privacy-based genetic matching in personalized medicine. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1109–1125.

White, T., Blok, E., and Calhoun, V. D. (2022). Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. *Human Brain Mapping*, 43(1):278–291.

Woodard, D. B. and Rosenthal, J. S. (2013). Convergence rate of Markov chain methods for genomic motif discovery. *The Annals of Statistics*, 41(1):91 – 124.

Xiao, L. and Zhang, T. (2013). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM J. Optim.*, 23(2):1062–1091.

Xie, W. and Deng, X. (2020). Scalable algorithms for the sparse ridge regression. *SIAM J. Optim.*, 30(4):3359–3386.

Yang, S., Wen, J., Eckert, S. T., Wang, Y., Liu, D. J., Wu, R., Li, R., and Zhan, X. (2020). Prioritizing genetic variants in gwas with lasso using permutation-assisted tuning. *Bioinformatics*, 36(12):3811–3817.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497 – 2532.

Ye, K. and Lim, L.-H. (2016). Schubert varieties and distances between subspaces of different dimensions. *SIAM J. Matrix Anal. Appl.*, 37(3):1176–1197.

Young, M., Paré, M.-A., Bergmann, H., et al. (2020). Differential privacy for expanding access to building energy data. In *ACEEE Summer Study on Energy Efficiency in Buildings Proceedings*.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Stat. Sci.*, 27(4):576–593.

Zhang, F., Li, W., Zhang, Y., and Feng, Z. (2018). Data driven feature selection for machine learning algorithms in computer vision. *IEEE Internet of Things Journal*, 5(6):4262–4272.

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*, pages 921–948. PMLR.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The J. Mach. Learn. Res.*, 7:2541–2563.

Zhao, T., Liu, H., and Zhang, T. (2018). Pathwise coordinate optimization for sparse learning: Algorithm and theory. *Ann. Statist.*, 46(1):180–218.

Zheng, Q., Hong, H. G., and Li, Y. (2020). Building generalized linear models with ultrahigh dimensional features: A sequentially conditional approach. *Biometrics*, 76(1):47–60.

Zheng, Z., Fan, Y., and Lv, J. (2014). High dimensional thresholded regression and shrinkage effect. *J. R. Stat. Soc. Ser. B*, pages 627–649.

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(202):1–7.

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(52):33117–33123.

Zhu, Z. and Wu, S. (2021). On the early solution path of best subset selection. *arXiv preprint arXiv:2107.06939*.