Signal-to-noise ratio aware minimaxity and its asymptotic expansion

Yilin Guo

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Signal-to-noise ratio aware minimaxity and its asymptotic expansion

Yilin Guo

Since its development, the minimax framework has been one of the corner stones of theoretical statistics, and has contributed to the popularity of many well-known estimators, such as the regularized M-estimators for high-dimensional problems. In this thesis, we will first show through the example of sparse Gaussian sequence model, that the theoretical results under the classical minimax framework are insufficient for explaining empirical observations. In particular, both hard and soft thresholding estimators are (asymptotically) minimax, however, in practice they often exhibit sub-optimal performances at various signal-to-noise ratio (SNR) levels. To alleviate the descrepancy, we first demonstrate that this issue can be resolved if the signal-to-noise ratio is taken into account in the construction of the parameter space. We call the resulting minimax framework the signal-to-noise ratio aware minimaxity. Then, we showcase how one can use higher-order asymptotics to obtain accurate approximations of the SNR-aware minimax risk and discover minimax estimators. Theoretical findings obtained from this refined minimax framework provide new insights and practical guidance for the estimation of sparse signals.

In a broader context, we investigated the same problem for sparse linear regression. We assume the random design and allow the feature matrix to be high dimensional as $X \in \mathbb{R}^{n \times p}$ and $p \gg n$. This adds an extra layer of challenge to the estimation of coefficients. Previous studies have largely relied on results expressed in rate-minimaxity, where estimators are compared based on minimax risk with order-wise accuracy, without specifying the precise constant in the

approximation. This lack of precision contributes to the notable gap between theoretical conclusions of the asymptotic minimax estimators and empirical findings of the sub-optimality. This thesis addresses this gap by initially refining the classical minimax result, providing a characterization of the constant in the first-order approximation. Subsequently, by following the framework of SNR-aware minimaxity we introduced before, we derived improved approximations of minimax risks under different SNR levels. Notably, these refined results demonstrated better alignment with empirical findings compared to classical minimax outcomes. As showcased in the thesis, our enhanced SNR-aware minimax framework not only offers a more accurate depiction of sparse estimation but also unveils the crucial role of SNR in the problem. This insight emerges as a pivotal factor in assessing the optimality of estimators.

Table of Contents

List of Figures

Acknowledgements

First and foremost, I extend my heartfelt gratitude to my PhD advisors, Prof. Arian Maleki and Prof. Haolei Weng. Throughout my PhD journey and research endeavors, their unwavering guidance and encouragement played a pivotal role in shaping my current achievements. I am profoundly thankful for their patience and confidence, which were instrumental in completing not only this thesis but also various other projects that significantly influenced my personal and academic growth. I would like to express special thanks to Prof. Arian Maleki for introducing me to the field of theoretical statistics, his consistent patience in addressing every query, his guidance in charting the research path of my PhD, and his encouragement in navigating uncertainties related to my future career. He has been a supreme mentor and a benevolent advisor. Deserving of unique appreciation is Prof. Haolei Weng, whose continuous support has been instrumental in completing this research journey. His conscientious and responsible guidance has provided crucial insights, and his unwavering support has been present at every significant milestone. Without the mentorship of Prof. Arian Maleki and Prof. Haolei Weng, I would not have reached the point I stand at in my PhD path.

I wish to express my gratitude to my dissertation committee: Prof. Cynthia Rush, Prof. Sumit Mukherjee and Prof. Daniel Hsu. Their valuable time and insightful suggestions have played a crucial role in enhancing the completeness and interest of this work. Special thanks go to Prof. Yang Feng, who introduced me to the world of research, instilling in me the initial confidence to embark on a Ph.D. journey and stand where I am today. I extend my appreciation to all the faculty and staff in the Department of Statistics at Columbia. Their unwavering support has

smoothed my Ph.D. path, guiding me in various aspects towards the completion of this dissertation and paving the way for my future career. My sincere thanks also go to my fellow Ph.D. colleagues in the Department of Statistics. Whether I encountered them here or have known them from before, I am grateful for the kindness and support.

Chapter 1: Introduction

1.1 Objective and organization

The minimax framework is one of the most popular approaches for comparing the performance of estimators and obtaining the optimal ones. Since its development, the minimax framework has been used in a broad range of areas including, among others, classical statistical decision theory [1, 2], non-parametric statistics [3, 4], high-dimensional statistics [5], and mathematical data science [6]. Despite its popularity, when the parameter space is set too general, since the minimax framework focuses on particular areas of the parameter space, its conclusions can be misleading if translated and used in practice. Take the high-dimensional sparse linear regression for example. It has been proved that the best subset selection is minimax rate-optimal over the class of k-sparse parameters [7]. Nevertheless, recent empirical and theoretical works demonstrate the inferior performance of the best subset selection in low signal-to-noise ratio (SNR) [8, 9, 10]. The key issue in this problem is that the parameter space in the minimax analysis only incorporates sparsity structure and does not control the signal strength for non-zero components of the sparse vector. In this thesis, we aim to answer the following question:

(*) How can we enhance the minimax framework to improve the accuracy of responses concerning the optimality of estimators?

We address this question through two canonical examples: (1) sparse signal denoising, and (2) sparse linear regression. We clarify these two problems as well as the thesis's contributions in the following sections.

1.2 Sparse signal denoising

Let $y_i = \theta_i + \sigma_n z_i,\ i = 1, 2, \ldots, n.$ where $y = (y_1, \ldots, y_n)$ denote our observations of the unknown parameters $\theta = (\theta_1, \ldots, \theta_n)$ corrupted by i.i.d. standard Gaussian noise $z = (z_1, \ldots, z_n)$. Let $\sigma_n > 0$ denote the noise level which may vary with n. The goal is to estimate θ assuming that $\theta \in \Theta(k_n) = \left\{\theta \in \mathbb{R}^n : \ \|\theta\|_0 \leq k_n\right\}$. Under the classical minimax framework, the following minimax risk is often studied:

$$R(\Theta(k_n), \sigma_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(k_n)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2. \tag{1.1}$$

This estimation problem has been well studied in statistical decision theory since 1990s. To clarify some of the existing results and the challenges they face, we first introduce soft and hard thresholding estimators. Define the soft thresholding estimator $\hat{\eta}_S(y, \lambda) \in \mathbb{R}^n$ and hard thresholding estimator $\hat{\eta}_H(y, \lambda) \in \mathbb{R}^n$ with coordinates:

$$[\hat{\eta}_S(y, \lambda)]_i = \arg\min_{\mu \in \mathbb{R}} \ (y_i - \mu)^2 + 2\lambda|\mu| = \text{sign}(y_i)(|y_i| - \lambda)_+, \tag{1.2}$$

$$[\hat{\eta}_H(y, \lambda)]_i = \arg\min_{\mu \in \mathbb{R}} \ (y_i - \mu)^2 + \lambda^2 I(\mu \neq 0) = y_i I(|y_i| > \lambda), \tag{1.3}$$

where $\text{sign}(u), u_+$ represent the sign and positive part of u respectively, $I(\cdot)$ denotes the indicator function, and $\lambda \geq 0$ is a tuning parameter. Also, the subscript i denotes the coordinate of a vector. The following theorem states a classical asymptotic minimax result.

Theorem 1 ([11, 12, 3])**.** *Assume the Gaussian sequence model and parameter space $\Theta(k_n)$ with $k_n/n \to 0$ as $n \to \infty$. Then the minimax risk, defined in* (1.1)*, satisfies*

$$R(\Theta(k_n), \sigma_n) = (2 + o(1)) \cdot \sigma_n^2 k_n \log(n/k_n).$$

Moreover, both the soft and hard thresholding estimators with tuning $\lambda_n = \sigma_n \sqrt{2 \log(n/k)}$ are

asymptotically minimax, i.e., for $\hat{\theta} = \hat{\eta}_S(y, \lambda_n)$ *or* $\hat{\eta}_H(y, \lambda_n)$*, it holds that*

$$\sup_{\theta \in \Theta(k_n)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = (2 + o(1)) \cdot \sigma_n^2 k_n \log(n/k_n).$$

This theorem indicates that both soft and hard thresholding estimators can achieve the asymptotic minimax risk when the sparsity satisfies $k_n/n \to 0$. However, we will show in Chapter 2 that, empirically, soft and hard thresholding estimators have divergent average mean squared errors in different noise levels. Specifically, the experiment in Chapter 2 (Figure 2.2) shows that in low noise level, hard thresholding performs better than both linear estimator and soft thresholding; as the noise level increases, soft thresholding starts to outperform hard thresholding, and eventually both hard and soft thresohlding are outperformed by the linear estimator. This implies that SNR has significant impact on the sparse estimation. Such phenomenon is not clearly reflected under classical minimax framework. This leads us to think about Question (*). Particularly, in Chapter 2, we introduce the SNR-aware minimax framework, where we control the SNR level in addition to the sparsity in the current minimax parameter space. As will be described in Chapter 2, this more constrained minimax framework is capable of discovering SNR regimes under which estimators show different behaviors.

In addition, one of the main challenges in minimax analysis is to estimate the minimax risk and find the corresponding optimal estimators. As can be guessed, it is even more challenging to solve this new constrained minimax framework than the original minimax problem. In response to the difficulty of evaluating the minimax risk, [3] suggested finding an approximation of the minimax estimator. This asymptotic approximation is also useful in our SNR-aware minimax analysis. However, we will show in Chapter 2 that, the approximation proposed by [3] is not sufficiently accurate to solve the SNR-aware minimaxity.

Hence, in Chapter 2, we introduce the higher-order asymptotic analysis to obtain more accurate approximations of the minimax risk. We show that the combination of the SNR aware minimax framework and higher order approximation provide much more accurate analysis of estimators.

More specifically, in Chapter 2, we will show that when the SNR level approaches zero, the linear estimator achieves up to the second-order minimax optimality whereas soft thresholding is proved to be suboptimal. Furthermore, when the SNR level can be arbitrary large, hard thresholding is the only optimal estimator in our second-order asymptotic analysis of the minimax risk. Finally, when the SNR level is large but below a certain threshold, we prove that an optimally tuned combination of linear and soft thresholding estimators (resembles elastic net in linear regression) is much closer to the optimal estimator than the soft or hard thresholding estimators. More interestingly, the threshold dividing the SNR level that leads to different minimax conclusions turns out to be $\sqrt{2\log(n/k_n)}$, the threshold at which the signals in n/k_n density can be detected from i.i.d. standard normal noises. Therefore, our analysis of the new SNR aware minimax framework brings new insights into the impact of SNR in sparse estimation.

1.3 Sparse linear regression

As one of the most recognized extensions of the sequence model discussed in previous section, the linear regression model is considered:

$$y_i = x_i^T \beta + \sigma z_i, \qquad i = 1, \ldots, n, \tag{1.4}$$

where $y_i \in \mathbb{R}$ denotes the response, $x_i \in \mathbb{R}^p$ represents the feature or covariate vector, $\beta \in \mathbb{R}^p$ is the unknown signal vector to be estimated, and finally $z_1, \ldots, z_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ are standard normal errors. The goal is to estimate $\beta \in \mathbb{R}^p$ given $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$, assuming the sparsity structure $\beta \in \Theta(k) := \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le k\}$. The classical minimax framework defines the minimax risk as

$$R(\Theta(k), \sigma) := \inf_{\hat{\beta}} \sup_{\beta\in\Theta(k)} \mathbb{E}_\beta \|\hat{\beta} - \beta\|^2. \tag{1.5}$$

Since considered, obtaining the exact minimax risk has remained mathematically challenging. As alternatives, there arose a line of research finding the approximation of the minimax risk. To review the prevailing results and develop our framework, in this thesis, we assume that $\{x_i\}_{i=1}^n \overset{i.i.d}{\sim}$

$\mathcal{N}(0, \frac{1}{n} I_p)$ and are independent with the noise errors $\{z_i\}_{i=1}^n$. In approximation of the minimax risk, previous literatures [13, 7, 14, 15] have studied the rate performance of the minimax risk when $n, p \to \infty$. Following the setting of this thesis, it was shown by [14] that

$$R(\Theta(k), \sigma) \sim \sigma^2 k \log(p/k), \tag{1.6}$$

where the notation "$\sim$" means that as $n, p \to \infty$ and $(k \log(p/k))/n \to 0$, the ratio $R(\Theta(k), \sigma)/(k \log(p/k))$ remains bounded. Furthermore, it has been studied in the literatures [13, 7, 14, 15] that many estimators, such as best subset selection [16, 17], Dantzig selector [18] and LASSO [19] achieve this rate-optimal minimax criteria, meaning that their risks (under optimal tuning) divided by $k \log(p/k)$ remain bounded.[1]

However, extensive simulation results reported in [8, 20] have confirmed that when the signal-to-noise ratio is low, all these estimators exhibit suboptimal performance and adding an ℓ_2-squared regularizer can improve the performance of the estimators. Hence, the rate-optimal minimax results could become misleading guidelines for practitioners. To figure out the mismatching between the rate-optimal and the simulation results, we propose the following conjectures:

- Conjecture 1: As is clear, the rate optimal minimax result does not evaluate the minimax risk exactly. It ignores the constant in the minimax risk approximation and only captures the rate behavior in view of k and p for mathematical simplicity. It is possible that if we calculate the exact maximum risk for estimators, the differences between constants can explain the discrepancies between the simulation studies and the rate-optimal minimax results.

- Conjecture 2: It could be that since the minimax framework only focuses on the spots of the parameter space that are hard for the estimation problem, its theoretical implications will be different from the simulation studies. Hence, the framework needs to be amended to provide more informative results.

To settle Conjecture 1, [21] has contributed to characterizing the constant of the minimax risk

[1] In some of these results, the risk is stated with high probability and the rate is $k \log p$ instead of $k \log(p/k)$.

to obtain a better approximation of the minimax risk. As the abovementioned literatures, [21] proved the result in a probabilistic statement, meaning that, fixing $\forall \beta \in \Theta(k)$ and considering a certain estimator $\hat{\beta}$, $\sigma^{-2}\|\hat{\beta} - \beta\|^2/(k \log(p/k)) \leq 2(1 + o(1))$ holds only with high probability tending to one. This expose the result of [21] to the doubt that there might exist some rare but possible event, under which the "optimal" estimator has unbounded risk. In this sense, the overall mean-squared error of the estimator might not achieve the exact constant characterized by [21]. As a complement, we deliver a result that is proved in Chapter 3:

Theorem 2. *Assume model* (1.4) *and parameter space* $\Theta(k)$. *Suppose* $n, p \to \infty$. *If* $k/p \to 0$ *and* $\big(k \log p\big)/n \to 0$, *then the minimax risk defined in* (1.5) *satisfies*

$$R(\Theta(k), \sigma) = 2\sigma^2 \cdot k \log(p/k)\Big(1 + o(1)\Big).$$

First, compared to the rate minimax result in (1.6), Theorem 2 characterizes the constant in the rate $\sim k \log(p/k)$, attaining more accurate approximation for the minimax risk. However, from this point, we have shown that under the current minimax framework, the same estimator remains optimal irrespective of different SNR settings in practice. This leads to a conclusion for Conjecture 1: By characterizing the exact constant on top of the current rate minimax results cannot explain the discrepancy between simulation and theoretical findings.

Therefore, we turn to Conjecture 2 for potential interpretations. As we discussed in Section 1.2, the disalignment between simulations and theories could result from these concerns of the classical minimaxity: (1) Since we do not impose any constraint on the signal strength, the minimax framework only focuses on a particular signal-to-noise ratio that makes the estimation problem the hardest. Hence, the factor of SNR affecting practical results is masked by the minimax framework. (2) The approximations we obtain for the minimax risk in rate-optimal minimax framework, and even in Theorem 2 are not accurate enough for distinguishing performances of different estimators and hence more accurate approximations are required for this purpose. This leads us to think about Question (*) under linear regression setting.

To address those concerns and answer Question (*), as we introduced in previous section, we add control of the SNR in the minimax framework by inserting a SNR constraint on the parameter space such that $\Theta(k, \tau) := \left\{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq k,\ \|\beta\|_2^2 \leq k\tau^2\right\}$. On top of this, the SNR-aware minimaxity is defined as the minimax risk over $\Theta(k, \tau)$. With the new constrained framework, we should expect that it is more challenging to mathematically solve the problem. In approximation, we will present in Theorem 14, Chapter 3 a first-order asymptotic result for the SNR-aware minimax risk. As will be clarified in the theorem, the first-order accuracy is still insufficient to identify the impact of SNR in the minimax risk. Finally, we will show that when we analyze the higher-order asymptotics, the minimax risk is decreased by different quantities in different SNR settings, which provides a clearer answer to the above question and becomes a more practical guideline for empirical applications.

Chapter 2: SNR-aware minimaxity in sparse signal denoising

In Chapter 2, we focus on the popular example of the sparse Gaussian sequence model – a special case of the sparse linear regression model with an orthogonal design. We first discuss in detail the limitations of classical minimaxity in Section 2.1. This is devoted to the development of a much more informative minimax framework that alleviates major drawbacks of the classical one. Then in Section 2.2, we introduces the SNR aware minimax framework by controlling and monitoring the signal-to-noise ratio and sparsity level through the parameter space. As will be discussed later, solving this new constrained minimax problem is much more challenging than the original minimax analysis. Hence, we resort to higher-order asymptotic analysis to obtain approximate minimax results. The conclusions of this signal-to-noise ratio aware minimax framework turn out to provide new insights into the estimation of sparse signals.

2.1 Classical minimaxity and its limitations in sparse Gaussian sequence model

We consider the Gaussian sequence model:

$$y_i = \theta_i + \sigma_n z_i, \quad i = 1, 2, \ldots, n. \tag{2.1}$$

Here, $y = (y_1, \ldots, y_n)$ is the vector of observations, $\theta = (\theta_1, \ldots, \theta_n)$ is the unknown signal consisting of n unknown parameters, z_i's are i.i.d. standard Gaussian error variables, and $\sigma_n > 0$ is the noise level that may vary with sample size n. The goal is to estimate θ from the sparse parameter space

$$\Theta(k_n) = \left\{\theta \in \mathbb{R}^n : \ \|\theta\|_0 \leq k_n\right\}, \tag{2.2}$$

where $\|\theta\|_0$ denotes the number of non-zero components of θ, and the sparsity k_n is allowed to change with n. The most popular approach for studying this estimation problem and obtaining the optimal estimators is the *minimax* framework. Considering the squared loss, the minimax framework aims to find the estimator that achieves the minimax risk given by

$$R(\Theta(k_n), \sigma_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(k_n)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2, \tag{2.3}$$

where $\mathbb{E}_\theta(\cdot)$ is the expectation taken under (2.1) with true parameter value θ.

Gaussian sequence model plays a fundamental role in non-parametric and high-dimensional statistics. There exists extensive literature on the minimax estimation of θ or its functionals over various structured parameter spaces such as Sobolev ellipsoids, hyperrectangles and Besov bodies. These parameter spaces usually characterize the smoothness properties of functions in terms of their Fourier or wavelet coefficients. We refer to [22, 3, 4] and references therein for a systematic treatment of this topic. The estimation problem over $\Theta(k_n)$ has been also well studied in statistical decision theory (e.g., with application to wavelet signal processing) since 1990s. Define the soft thresholding estimator $\hat{\eta}_S(y, \lambda) \in \mathbb{R}^n$ and hard thresholding estimator $\hat{\eta}_H(y, \lambda) \in \mathbb{R}^n$ with coordinates: for $1 \le i \le n$,

$$[\hat{\eta}_S(y, \lambda)]_i = \arg\min_{\mu \in \mathbb{R}} \ (y_i - \mu)^2 + 2\lambda|\mu| = \mathrm{sign}(y_i)(|y_i| - \lambda)_+, \tag{2.4}$$

$$[\hat{\eta}_H(y, \lambda)]_i = \arg\min_{\mu \in \mathbb{R}} \ (y_i - \mu)^2 + \lambda^2 I(\mu \neq 0) = y_i I(|y_i| > \lambda), \tag{2.5}$$

where $\mathrm{sign}(u), u_+$ represent the sign and positive part of u respectively, $I(\cdot)$ denotes the indicator function, and $\lambda \ge 0$ is a tuning parameter. We summarize a classical asymptotic minimax result in the following theorem.

Theorem 3 ([11, 12, 3])**.** *Assume model* (2.1) *and parameter space* (2.2) *with* $k_n/n \to 0$ *as* $n \to \infty$.

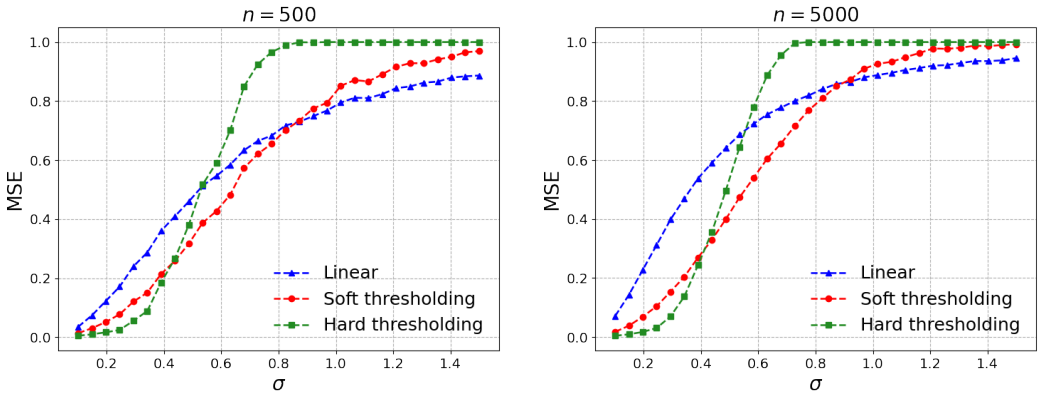

Figure 2.1: Mean squared error comparison at different noise levels. Data is generated according to (2.1) with $k_n = \lfloor n^{2/3} \rfloor$ and θ having k_n components equal to 1.5. "linear" denotes the simple linear estimator $\frac{1}{1+\lambda} y$. All the three estimators are optimally tuned. MSE is averaged over 20 repetitions along with standard error. Other details of the simulation can be found in Section 2.3.

Then the minimax risk, defined in (2.3)*, satisfies*

$$R(\Theta(k_n), \sigma_n) = (2 + o(1)) \cdot \sigma_n^2 k_n \log(n/k_n).$$

Moreover, both the soft and hard thresholding estimators with tuning $\lambda_n = \sigma_n \sqrt{2 \log(n/k)}$ *are asymptotically minimax, i.e., for* $\hat{\theta} = \hat{\eta}_S(y, \lambda_n)$ *or* $\hat{\eta}_H(y, \lambda_n)$*, it holds that*

$$\sup_{\theta \in \Theta(k_n)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = (2 + o(1)) \cdot \sigma_n^2 k_n \log(n/k_n).$$

Theorem 3 shows that both soft and hard thresholding estimators are minimax optimal for estimating sparse signals (with small values of k_n/n). Despite the mathematical beauty of the above results, its practical implications seem not clear. We demonstrate this point by a simulation in Figure 2.1. As is clear from the left panel, when the noise level is low, hard thresholding performs the best among the three estimators; as the noise level increases, hard thresholding starts to be outperformed by soft thresholding, and eventually both hard and soft thresohlding are outperformed by the linear estimator. The same comparison holds in the right panel as the sample size increases from 500 to 5000. This phenomenon can be widely observed for different types of sparse signals. We provide more simulations in Section 2.3.

In light of Theorem 3 and Figure 2.1, we would like to raise a few critical comments:

1. Despite their minimax optimality, both hard and soft thresholding estimators selected by the classical minimaxity do not perform well compared to a simple linear estimator when the noise is large.

2. The hard and soft thresholding estimators have distinct performances at different noise levels, despite they are both asymptotically minimax.

3. Figure 2.1 implies that the signal-to-noise ratio (SNR) has a significant impact on the estimation. However, the effect of SNR is not well captured in the classical minimax results (Theorem 3).

These observations lead us to the following question: is it possible to develop a refined minimax framework which addresses differences between hard and soft thresholding estimators and characterizes the role of SNR in the recovery of sparse signals? Such a framework will provide more proper insights and sound guidance for practical purpose.

2.2 SNR-aware minimaxity

To overcome the limitations of the classical minimaxity discussed in Section 2.1, in this chapter, we aim to develop a signal-to-noise-ratio-aware minimax framework. This framework imposes direct constraints on the signal strength over the parameter space and performs the corresponding minimax analysis that accounts for the impact of signal-to-noise ratio (SNR). To obtain accurate minimax results in the SNR-aware setting, we will derive higher-order asymptotics which provides asymptotic approximations precise up to the second order. As will be discussed in detail in Section 2.2, our proposed framework reveals three regimes in which distinct estimators achieve minimax optimality. In particular, hard-thresholding estimator outperforms soft-thresholding estimator and remains (asymptotically) minimax optimal in the high SNR regime; as SNR decreases, new optimal estimators will emerge. These new theoretical findings offer much better explanations for

what is happening in Figure 2.1, and are much more informative towards understanding the sparse estimation problem in practice.

We collect the notations used throughout this chapter here for convenience. For a scalar $x \in \mathbb{R}$, x_+ and $\text{sign}(x)$ denote the positive part of x and its sign respectively; $\lfloor x \rfloor$ is the largest integer less than or equal to x. For an integer n, $[n] = \{1, 2, \ldots, n\}$. We use I_A and $I(A)$ to represent the indicator function of the set A interchangeably. For a given vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, $\|v\|_0 = \#\{i : v_i \neq 0\}$, $\|v\|_\infty = \max_i |v_i|$, and $\|v\|_q = \left(\sum_{i=1}^p |v_i|^q\right)^{1/q}$ for $q \in (0, \infty)$. We use the notation δ_μ as the point mass at $\mu \in \mathbb{R}$. We also use $\{e_j\}_{j=1}^p$ to denote the natural basis in $\mathbb{R}^p$. For two non-zero real sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we use $a_n = o(b_n)$ to represent $|a_n/b_n| \to 0$ as $n \to \infty$, and $a_n = \omega(b_n)$ if and only if $b_n = o(a_n)$; $a_n = O(b_n)$ means $\sup_n |a_n/b_n| < \infty$, and $a_n = \Omega(b_n)$ if and only if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ denotes $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. For a distribution π, $\text{supp}(\pi)$ denotes its support. Finally, we reserve the notations $\phi(y)$ and $\Phi(y) = \int_{-\infty}^y \phi(s)ds$ for the standard normal density and its cumulative distribution function respectively.

2.2.1 SNR-aware minimax framework

We focus on the above-mentioned Gaussian sequence model (2.1). To develop the SNR-aware minimax framework, we start by inserting a notion of signal-to-noise ratio in the minimax setting. To this end, we consider the following SNR-aware parameter space:

$$\Theta(k_n, \tau_n) = \left\{\theta \in \mathbb{R}^n : \ \|\theta\|_0 \leq k_n, \ \|\theta\|_2^2 \leq k_n \tau_n^2\right\}. \tag{2.6}$$

Here, as before, k_n is the parameter that controls the number of nonzero components of the signal $\theta \in \mathbb{R}^n$. The new parameter τ_n can be considered as a measure of signal strength (on average) for each non-zero coordinate of θ. Unlike $\Theta(k_n)$, the new parameter space $\Theta(k_n, \tau_n)$ is responsive to changing signal strength. Minimax analysis based on it may thus provide a viable path for revealing the impact of SNR on the estimation of sparse signals. Define the corresponding minimax risk (for

squared loss):

$$R(\Theta(k_n, \tau_n), \sigma_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2. \tag{2.7}$$

We aim to investigate the following problems:

1. Characterizing the minimax risk, $R(\Theta(k_n, \tau_n), \sigma_n)$, for different choices of sparsity level and signal-to-noise ratio. This will help us understand the intertwined roles of SNR and sparsity on signal recovery.

2. Obtaining minimax optimal estimators in the aforementioned settings, along with evaluating the performance of some common estimators (e.g., soft thresholding).

The solutions to the above problems will help resolve the issues we raised before about the classical minimax results. First, we introduce two critical quantities associated with the target parameter space $\Theta(k_n, \tau_n)$ introduced in (2.6) under the model (2.1). Denote

$$\epsilon_n = \frac{k_n}{n}, \quad \mu_n = \frac{\tau_n}{\sigma_n}. \tag{2.8}$$

It is clear that ϵ_n represents the sparsity level and μ_n is a form of signal-to-noise ratio over the parameter space. We aim to study $R(\Theta(k_n, \tau_n), \sigma_n)$ for different values of (ϵ_n, μ_n). Since an explicit solution to exact minimaxity is very challenging to derive (it is not even available for $\Theta(k_n)$), we focus on obtaining asymptotic minimaxity, and consider the following regimes: as $n \to \infty$,

Regime (I) Low signal-to-noise ratio: $\mu_n \to 0$, $\epsilon_n \to 0$;

Regime (II) Moderate signal-to-noise ratio: $\mu_n \to \infty$, $\epsilon_n \to 0$, $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$;

Regime (III) High signal-to-noise ratio: $\epsilon_n \to 0$, $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$.

The condition $\epsilon_n \to 0$ is standard to model sparse signals. The above three regimes are classified according to the order of signal-to-noise ratio μ_n. As will be shown in Section 2.2.3 via

higher-order asymptotics, each regime exhibits unique minimaxity, and distinct minimax estimators emerge in different regimes. But before that, we first derive similar first-order asymptotic result as the classical one and reveal its limitations in the SNR-aware minimax setting.

2.2.2 First order analysis of SNR-aware minimaxity and its drawbacks

Our first theorem generalizes Theorem 3, to our SNR-aware minimax framework.

Theorem 4. *Assume model* (2.1) *and parameter space* (2.6)*. The following hold:*

- Regime (I). *When* $\mu_n \to 0, \epsilon_n \to 0$*,*

$$R(\Theta(k_n, \tau_n), \sigma_n) = (1 + o(1)) \cdot n\sigma_n^2 \epsilon_n \mu_n^2,$$

  and the zero estimator is asymptotically minimax optimal (up to the first order).

- Regime (II). *When* $\mu_n \to \infty$*,* $\epsilon_n \to 0$*,* $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$*,*

$$R(\Theta(k_n, \tau_n), \sigma_n) = (1 + o(1)) \cdot n\sigma_n^2 \epsilon_n \mu_n^2,$$

  and the zero estimator is asymptotically minimax optimal (up to the first order).

- Regime (III). *When* $\epsilon_n \to 0$*,* $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$*,*

$$R(\Theta(k_n, \tau_n), \sigma_n) = (2 + o(1)) \cdot n\sigma_n^2 \epsilon_n \log(\epsilon_n^{-1}).$$

  Furthermore, both soft and hard thresholding estimators (2.4)-(2.5) *with the tuning parameter* $\lambda_n = \sigma_n \sqrt{2 \log \epsilon_n^{-1}}$ *are asymptotically minimax optimal (up to the first order).*

This theorem is covered as a special case of Theorems 5, 6, and 8 we present in Section 2.2.3. Hence, the proof is skipped.

There are a few aspects of the above results that we would like to emphasize here:

1. As is clear, first-order analysis under the SNR-aware minimax framework already provides more information than in the previous framework. For instance, it implies that below a certain signal-to-noise-ratio, i.e. when $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, sparsity promoting estimators such as hard or soft thresholding do not seem to have any advantage over the zero estimator. In fact, the zero estimator is optimal up to the first order. Later in Section 2.2.3 we will argue that even these theorems should be interpreted carefully, and that the current interpretation is not fully accurate.

2. If we consider the rate of ϵ_n fixed and evaluate the minimax risk as a function of μ_n, we will see a phase transition happening in the first order term of the minimax risk. As long as the first order is concerned, the trivial zero estimator is minimax optimal for any $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$. Hence, it seems that unless $\mu_n = \Omega(\sqrt{\log \epsilon_n^{-1}})$, even the optimal minimax estimators will miss the signal. Once $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$, the first order result implies the optimality of non-trivial estimators, such as soft-thresholding. While it is challenging to provide an intuitive argument for the phase transition occurring at $\sqrt{\log \epsilon_n^{-1}} = \sqrt{\log(n/k_n)}$, the following explanation may offer some insight: Consider a k_n-sparse signal (with k_n non-zero components) in $\mathbb{R}^n$ with Gaussian noises. *On average*, there exists one non-zero signal component among n/k_n locations. The maximum absolute value of the noises at the n/k_n locations is on the order of $\sqrt{\log(n/k_n)}$. Consequently, from an intuitive perspective, it becomes easier to detect signals when their magnitudes exceed this threshold, but significantly more challenging when they fall below this threshold. It's important to note that heuristic arguments like the one above have their limitations and should not be solely relied upon for drawing conclusive results. This aspect will be further clarified in the next section, where we will demonstrate that minimax estimators can outperform zero estimators even when $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$.

One of the main issues in the above theorem is that the first-order asymptotic approximation of minimax risk does not seem to always offer accurate information. For example, as the signal-to-noise ratio significantly increases from Regime (I) to Regime (II), the first-order analysis falls short

of capturing any difference and continues to generate the naive zero estimator as the optimal one. Moreover, in Regime (III), the analysis is inadequate to explain the difference between hard and soft thresholding estimators. In the next section, we push the analysis one step further to develop second-order asymptotics. This refined version of the SNR-aware minimax analysis will provide a much more accurate approximation of the minimax risk, and can provide more useful information and resolve the confusing aspects of the first-order results presented above.

2.2.3 Second order analysis of SNR-aware minimaxity

In this section, we discuss how the analysis provided in Section 2.2.2 can be refined to resolve the issues we raised in Section 2.1.

Results in Regime (I)

We start with Regime (I). As discussed in Theorem 4, as far as the first order of minimax risk is concerned, the zero estimator is asymptotically optimal in this regime, and no other estimators can outperform the zero estimator. The reason this peculiar feature arises is that since the exact expression for $R(\Theta(k_n, \tau_n), \sigma_n)$ is very complicated, Theorem 4 resorts to an approximation that is asymptotically accurate. However, this approximation is coarse when n is not too large and/or ϵ_n is not too small. The conclusions that are based on such first order analysis are hence not reliable. Therefore, we pursue a second-order asymptotic analysis of minimax risk to achieve better approximations. This more delicate analysis turns out to be instructive for understanding the three regimes of varying SNRs. We first present the result in Regime (I). Define the simple linear estimator $\hat{\eta}_L(y, \lambda) \in \mathbb{R}^n$ with coordinates:

$$[\hat{\eta}_L(y, \lambda)]_i = \frac{y_i}{1+\lambda} = \arg\min_{\mu \in \mathbb{R}} \ (y_i - \mu)^2 + \lambda \mu^2, \quad 1 \le i \le n. \tag{2.9}$$

Theorem 5. *Consider model* (2.1) *and parameter space* (2.6)*. For Regime (I) in which* $\epsilon_n \to$

$0, \mu_n \to 0$ *as* $n \to \infty$, *we have*

$$R(\Theta(k_n, \tau_n), \sigma_n) = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - \epsilon_n^2 \mu_n^4 \left(1 + o(1)\right)\right).$$

In addition, the linear estimator $\hat{\eta}_L(y, \lambda_n)$ *with tuning* $\lambda_n = \left(\epsilon_n \mu_n^2\right)^{-1}$ *is asymptotically minimax up to the second order term, i.e.*

$$\sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_L(y, \lambda_n) - \theta \right\|_2^2 = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - \epsilon_n^2 \mu_n^4 \left(1 + o(1)\right)\right).$$

The proof of this theorem can be found in Section 2.5.2. Compared with Theorem 4, Theorem 5 obtains the additional second dominating term in the minimax risk. This negative term quantifies the amount of improvement that can be possibly achieved over the trivial zero estimator (whose supremum risk exactly equals $n\sigma_n^2 \epsilon_n \mu_n^2$). Indeed, the non-trivial linear estimator $\hat{\eta}_L(y, \lambda_n)$ has supremum risk matching with the minimax risk up to the second order. Therefore, through the lens of second-order asymptotics, we discover a new minimax optimal estimator that outperforms the zero estimator recommended from the first-order analysis.

The second-order optimality of the linear estimator $\hat{\eta}_L(y, \lambda_n)$ in Regime (I) raises the following question: how do non-linear estimators compare with $\hat{\eta}_L(y, \lambda_n)$? For instance, the soft thresholding estimator $\hat{\eta}_S(y, \lambda)$ in (2.4) with $\lambda = \infty$ recovers the zero estimator and is hence first-order optimal. Can $\hat{\eta}_S(y, \lambda)$ with proper tuning become second-order asymptotically optimal in this regime? The following theorem shows that the answer is negative.

Proposition 1. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (I) where* $\epsilon_n \to 0, \mu_n \to 0$ *as* $n \to \infty$, *the optimally tuned soft thresholding estimator* $\hat{\eta}_S(y, \lambda)$ *has supremum risk:*

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_S(y, \lambda) - \theta \right\|_2^2 = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - \exp\left[-\frac{1}{2}\frac{1}{\mu_n^2}\left(\log\frac{1}{\epsilon_n}\right)^2 \left(1 + o(1)\right)\right]\right).$$

The proof of this proposition can be found in Section 2.5.3.

It is straightforward to confirm that $\exp\left[-\frac{1}{2}\frac{1}{\mu_n^2}\left(\log\frac{1}{\epsilon_n}\right)^2\left(1+o(1)\right)\right]/(\epsilon_n^2\mu_n^4) = o(1)$ under the scaling $\epsilon_n \to 0, \mu_n \to 0$. Hence, soft thresholding $\hat{\eta}_S(y,\lambda)$ is outperformed by the linear estimator $\hat{\eta}_L(y,\lambda_n)$ and is sub-optimal (up to second order). A similar result can be proved for the hard thresholding estimator as well.

Proposition 2. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (I) where* $\epsilon_n \to 0, \mu_n \to 0$ *as* $n \to \infty$, *the optimally tuned hard thresholding estimator* $\hat{\eta}_H(y,\lambda)$ *has supremum risk:*

$$\inf_{\lambda} \sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda)-\theta\right\|_2^2 = n\sigma_n^2\epsilon_n\mu_n^2.$$

The proof of this proposition is presented in Section 2.5.4.

The fact that $\hat{\eta}_L(y,\lambda_n)$ is optimal and $\hat{\eta}_S(y,\lambda)$ and $\hat{\eta}_H(y,\lambda)$ are sub-optimal in Regime (I) is intriguing. It says that the former non-sparse estimator is better than the latter sparse one for recovering sparse signals. In fact, the result further implies that any sparsity-promoting procedure cannot improve over a simple linear shrinkage for the recovery of sparse signals. A high-level explanation is that since Regime (I) has low signal-to-noise ratio in which variance is the dominating factor of mean squared error, linear shrinkage achieves a better balance between bias and variance than those more "aggressive" sparsity-inducing operations. These results demonstrate the practical relevance of SNR-aware minimaxity as opposed to the classical minimax approach.

Results in Regime (II)

We now move on to discuss Regime (II) where new minimaxity results arise as the signal-to-noise ratio increases. Introduce an estimator $\hat{\eta}_E(y,\lambda,\gamma) = \frac{\hat{\eta}_S(y,\lambda)}{1+\gamma} \in \mathbb{R}^n$ with coordinates:

$$[\hat{\eta}_E(y,\lambda,\gamma)]_i = \frac{[\hat{\eta}_S(y,\lambda)]_i}{1+\gamma} = \arg\min_{u\in\mathbb{R}} \ (y_i-u)^2 + 2\lambda|u| + \gamma u^2, \quad 1 \le i \le n. \tag{2.10}$$

The estimator $\hat{\eta}_E(y, \lambda, \gamma)$ is a composition of soft thresholding and linear shrinkage. It can be considered as an "interpolation" between soft thresholding estimator and linear estimator.

Theorem 6. *Consider model* (2.1) *and parameter space* (2.6)*. For Regime (II) in which* $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ *as* $n \to \infty$*, we have*

$$R(\Theta(k_n, \tau_n), \sigma_n) \geq n\sigma_n^2 \Big(\epsilon_n \mu_n^2 - \frac{1}{2} \epsilon_n^2 \mu_n^2 e^{\mu_n^2} \left(1 + o(1)\right) \Big).$$

In addition, based on the estimator $\hat{\eta}_E(y, \lambda_n, \gamma_n)$ *with tuning parameters* $\lambda_n = 2\tau_n$*, and* $\gamma_n = (2\epsilon_n \mu_n^2 e^{\frac{3}{2}\mu_n^2})^{-1} - 1$*, we have*

$$R(\Theta(k_n, \tau_n), \sigma_n) \leq \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \big\| \hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta \big\|_2^2 = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - (\sqrt{2/\pi} + o(1)) \epsilon_n^2 \mu_n e^{\mu_n^2} \right).$$

The proof of this theorem can be found in Section 2.5.5.

Remark 1. *Theorem 6 does not provide a tight upper or lower bound for the minimax risk approximation. However, the upper bound given by* $\hat{\eta}_E(y, \lambda_n, \gamma_n)$ *only differs from the lower bound up to an order of* μ_n *in the second order term. Note that this difference is very small in view of the occurrence of* $e^{\mu_n^2}$ *in the second order term. In this sense, the estimator* $\hat{\eta}_E(y, \lambda_n, \gamma_n)$ *is nearly optimal in Regime (II). In this theorem, we believe that the upper bound is not necessarily sharp. In fact, we anticipate that there may be other estimators capable of outperforming* $\hat{\eta}_E(y, \lambda_n, \gamma_n)$*. Our next theorem (Theorem 7) gives an accurate second order term for the minimax risk in Regime (II), under a uniform boundedness condition on parameter coordinates in the parameter space. However, as will be elaborated in the proof, the technique employed to establish the upper bound on the minimax risk is not constructive and does not identify the minimax estimator.*

Theorem 7. *Consider model* (2.1) *with the following parameter space:*

$$\Theta^A(k_n, \tau_n) := \Big\{ \theta \in \mathbb{R}^n : \ \|\theta\|_0 \leq k_n, \ \|\theta\|_2^2 \leq k_n \tau_n^2, \ \|\theta\|_\infty \leq A\tau_n \Big\}. \tag{2.11}$$

For Regime (II) in which $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ *as* $n \to \infty$*, we have that for any*

constant $A > 1$,

$$R(\Theta^A(k_n, \tau_n), \sigma_n) = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - \frac{1}{2} \epsilon_n^2 \mu_n^2 e^{\mu_n^2} \left(1 + o(1)\right) \right).$$

The theorem is proved in Section 2.5.6.

Now let us interpret the above results. First note that in Regime (II), compared to Regime (I), the magnitude of the second order term (relative to the first order term) is much larger, so that the possible improvement over the zero estimator is much more significant. This is expected as the SNR is higher compared to Regime (I). Furthermore, the (near) optimality of $\hat{\eta}_E(y, \lambda_n, \gamma_n)$ showed in Theorem 6 indicates that thresohlding and linear shrinkage together play an important role in estimating sparse signals in Regime (II). To shed more light on it, the following two propositions prove that neither soft thresohlding $\hat{\eta}_S(y, \lambda)$ nor linear estimator $\hat{\eta}_L(y, \lambda)$ alone is close to optimal. To shed more light on it, the following three propositions prove that neither thresohlding estimators $\hat{\eta}_S(y, \lambda), \hat{\eta}_H(y, \lambda)$ nor linear estimator $\hat{\eta}_L(y, \lambda)$ alone is close to optimal.

Proposition 3. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (II) where* $\epsilon_n \to 0, \mu_n \to \infty$, $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, *as* $n \to \infty$, *the optimally tuned soft thresholding estimator has supremum risk:*

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_S(y, \lambda) - \theta \right\|_2^2 = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - \exp\left[-\frac{1}{2} \frac{1}{\mu_n^2} \left(\log \frac{1}{\epsilon_n} \right)^2 \left(1 + o(1)\right) \right] \right).$$

The proof of this proposition can be found in Section 2.5.7.

Proposition 4. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (II) where* $\epsilon_n \to 0, \mu_n \to \infty$, $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ *as* $n \to \infty$, *the optimally tuned hard thresholding estimator* $\hat{\eta}_H(y, \lambda)$ *has supremum risk:*

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_H(y, \lambda) - \theta \right\|_2^2 = n\sigma_n^2 \epsilon_n \mu_n^2.$$

The proof of this proposition is presented in Section 2.5.8.

Proposition 5. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (II) where* $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, *as* $n \to \infty$, *the optimally tuned linear estimator has supremum risk:*

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_L(y, \lambda) - \theta \right\|_2^2 = n\sigma_n^2 \left(\epsilon_n \mu_n^2 - \frac{\epsilon_n^2 \mu_n^4}{1 + \epsilon_n \mu_n^2} \right).$$

The proof of this proposition can be easily followed by the discussion in Section 2.5.2.

Comparing the second order term in Theorem 6 and Propositions 3-5 under the scaling condition $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, it is straightforward to verify that the supremum risk of $\hat{\eta}_E(y, \lambda_n, \gamma_n)$ is much smaller than that of optimally tuned soft thresholding and linear estimator. In light of what we have discussed in Regime (I), the results in Regime (II) deliver an interesting message: when SNR increases from low to moderate level, sparsity promoting operation becomes effective in estimating sparse signals; on the other hand, since SNR is not sufficiently high yet, a component of linear shrinkage towards zero still boosts the performance.

Results in Regime (III)

Finally, let us consider the high-SNR regime, i.e., Regime (III). As shown in Theorem 4, the first-order approximation of minimax risk claims that both hard and soft thresholding estimators are optimal. However, the refined second-order analysis will reveal that hard thresholding remains optimal while soft thresholding is in fact sub-optimal, up to the second order term.

Theorem 8. *Consider model* (2.1) *and parameter space* (2.6). *For Regime (III) in which* $\epsilon_n \to 0, \mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ *as* $n \to \infty$, *we have*

$$R(\Theta(k_n, \tau_n), \sigma_n) = n\sigma_n^2 \left(2\epsilon_n \log \epsilon_n^{-1} - 2\epsilon_n \nu_n \sqrt{2 \log \nu_n} \left(1 + o(1)\right) \right),$$

where $\nu_n := \sqrt{2 \log \epsilon_n^{-1}}$. *In addition, the hard thresholding* $\hat{\eta}_H(y, \lambda_n)$ *with tuning* $\lambda_n = \sigma_n \sqrt{2 \log \epsilon_n^{-1}}$ *is asymptotically minimax up to the second order term, i.e.*

$$\sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_H(y, \lambda_n) - \theta \right\|_2^2 = n\sigma_n^2 \left(2\epsilon_n \log \epsilon_n^{-1} - 2\epsilon_n \nu_n \sqrt{2 \log \nu_n} \left(1 + o(1)\right) \right).$$

The proof of this theorem can be found in Section 2.5.9. Before we interpret this result, let us obtain the risk of the soft thresholding estimator as well.

Proposition 6. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (III) where* $\epsilon_n \to 0, \mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ *as* $n \to \infty$*, the optimally tuned soft thresholding achieves the supremum risk:*

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_S(y, \lambda) - \theta \right\|_2^2 = n\sigma_n^2 \left(2\epsilon_n \log \epsilon_n^{-1} - 6\epsilon_n \log \nu_n \left(1 + o(1)\right) \right),$$

where $\nu_n = \sqrt{2 \log \epsilon_n^{-1}}$.

The proof of the proposition can be found in Section 2.5.10.

Proposition 7. *Consider model* (2.1) *and parameter space* (2.6). *In Regime (III) where* $\epsilon_n \to 0, \mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ *as* $n \to \infty$*, the optimally tuned linear estimator achieves the supremum risk:*

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_L(y, \lambda) - \theta \right\|_2^2 = \frac{n\sigma_n^2 \epsilon_n \mu_n^2}{1 + \epsilon_n \mu_n^2} = \omega(n\sigma_n^2 \epsilon_n \log(\epsilon_n^{-1})).$$

The proof of this proposition is presented in Section 2.5.11.

Combining the above two results, we can conclude that overall in Regime (III) hard thresholding offers a better estimate than soft thresholding. The intuition is that Regime (III) has a high SNR where bias becomes the dominating factor of mean squared error, therefore hard thresholding has an edge on soft thresholding by not shrinking the above-threshold coordinates. Moreover, note that the difference between the first order and second order terms in the minimax risk is smaller than $\sqrt{\log \epsilon_n^{-1}}$. This implies that the second order term in our approximations can be relevant in a wide range of sparsity levels.

2.3 Numerical experiments

As discussed in Section 2.1 through one simulation example, classical minimax results are inadequate for characterizing the role of signal-to-noise ratio (SNR) in the estimation of sparse signals. Hence, we developed the SNR-aware minimax framework in Section 2.2 to overcome

the limitations of the classical minimaxity. In this section, we provide more empirical results to evaluate the points we discussed above.

We generate the signal θ in the following way: for a sample size n, $\theta = (\theta_1, \ldots, \theta_n)$ is generated by assigning τ_n to a random choice of k_n coordinates and setting the others to zero. Then $y = (y_1, \ldots, y_n)$ and $z = (z_1, \ldots, z_n)$ are generated according to Model (2.1) for a certain noise level σ_n.

Given the sample size n, we consider three sparsity levels $k_n = \lfloor n^{2/3} \rfloor$, $\lfloor n^{3/4} \rfloor$, $\lfloor n^{1/2} \rfloor$, so that $\epsilon_n = k_n/n \to 0$ as $n \to \infty$. In addition, since SNR is decided by $\mu_n = \tau_n/\sigma_n$, without the loss of generality, we fix the value of the signal strength $\tau_n = 10$. We demonstrate our findings in two ways:

1. Let μ_n change from small to large values, and plot the mean squared error (MSE) of different estimators as a function of μ_n.

2. Let σ_n change from small to large values, and plot the MSE as a function of σ_n.

In our experiments, we consider moderate sample size $n = 500$ and large sample size $n = 5000$. We consider the four estimators that have been extensively discussed in the previous sections: linear estimator $\hat{\eta}_L$ defined in (2.9), soft thresholding $\hat{\eta}_S$ defined in (2.4), hard thresholding $\hat{\eta}_H$ defined in (2.5), and the soft-linear "interpolation" estimator $\hat{\eta}_E$ defined in (2.10) (since $\hat{\eta}_E$ is the composition of soft thresholding and linear shrinkage, we refer to it as soft-linear "interpolation" for convenience). We evaluate the performance of estimators using the empirical MSE scaled by the total signal strength: $\|\theta\|_2^{-2} \cdot \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2$. The MSEs shown in Figures 2.2 and 2.3 are averaged over 20 repetitions, plotted with 95% confidence intervals from t-distribution. For each estimator, tuning parameters are chosen by grid search to obtain the minimum possible MSE.

From Figures 2.2, when σ_n changes from small to large values, we observed that: (1) When σ_n is near zero, hard thresholding achieves the minimum MSE among the four estimators discussed in previous sections. This corresponds to Regime (III) in our theory. (2) When σ_n is in moderate area, the soft-linear 'interpolation' estimator $\hat{\eta}_E$ has the minimum empirical MSE. This corresponds to

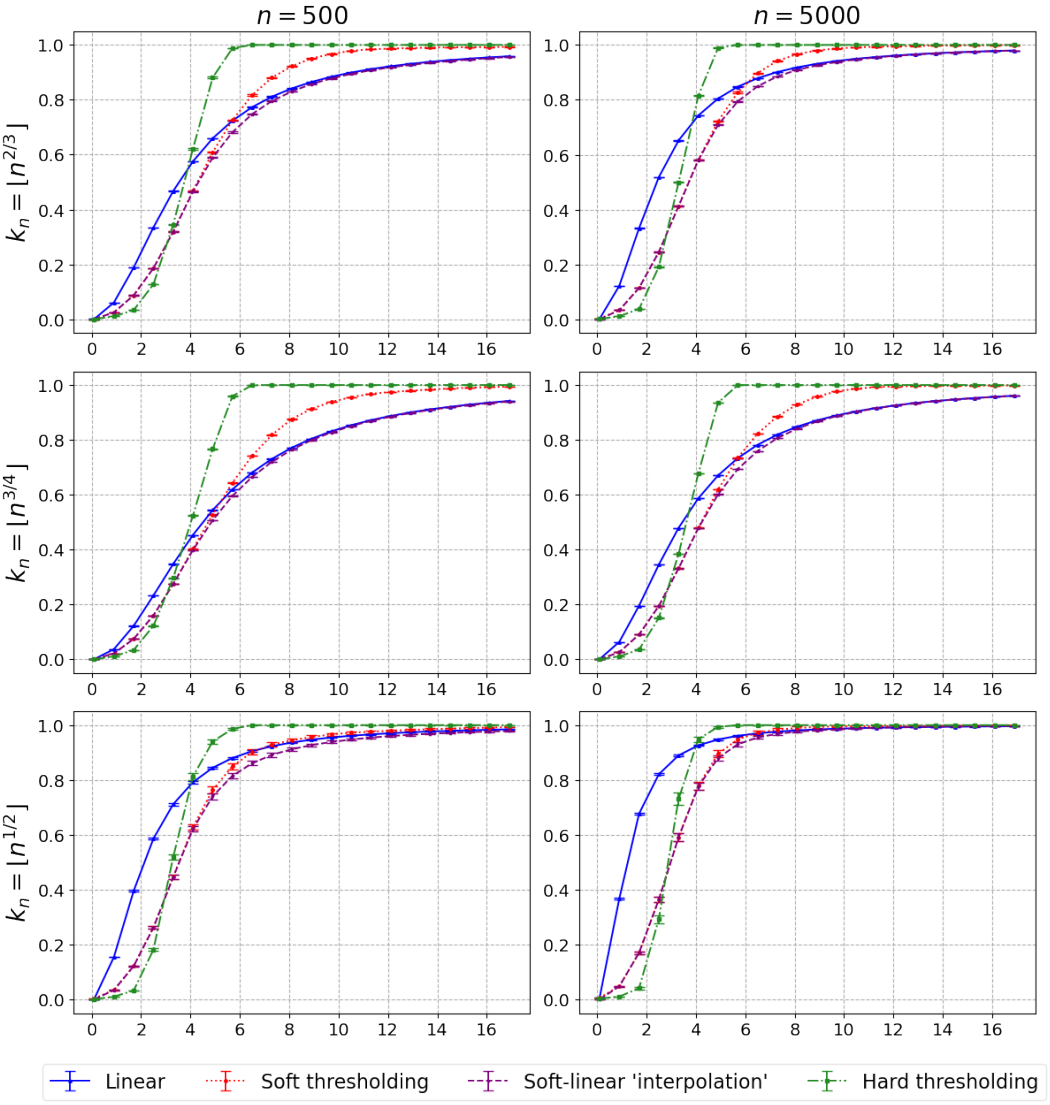

Figure 2.2: Mean squared error comparison at different noise levels. On each graph, the y-axis is the scaled MSE, and the x-axis is the noise standard deviation σ_n.

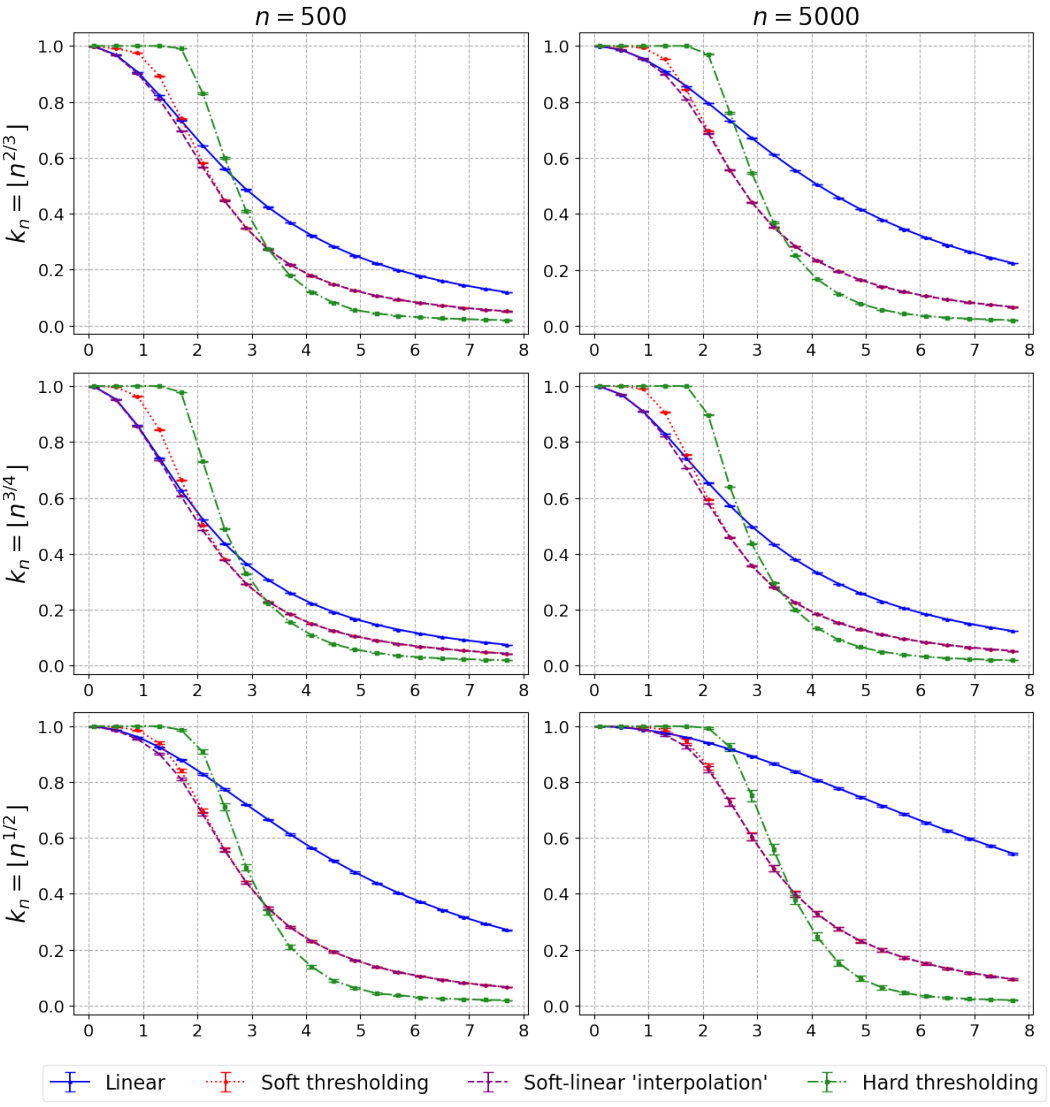

Figure 2.3: Mean squared error comparison at different SNR levels. On each graph, the y-axis is the scaled MSE, and the x-axis is the SNR μ_n.

Regime (II) in our theory. (3) When σ_n becomes large, the linear estimator $\hat{\eta}_L$ as well as the optimally tuned $\hat{\eta}_E$ (since $\hat{\eta}_E$ can achieve $\hat{\eta}_L$ when optimally tuned) have the minimum empirical MSE. Our theory in Regime (I) states that when SNR is small, $\hat{\eta}_L$ becomes asymptotically minimax optimal. The empirical studies align well with our current theory.

Figures 2.3 offer similar conclusions as the ones we mentioned above. The main difference is that instead of revealing MSE as a function of the noise level, we view it as a function of SNR. Due to this difference, the leftmost part of each graph corresponds to Regime (I). As μ_n increases, the curves will correspond to Regime (II) and Regime (III). In particular, when μ_n is large, it corresponds with the area of σ_n near zero in Figures 2.2. Here, it is shown more clearly that in the large SNR regime, hard thresholding has the minimum empirical MSE among all the estimators.

2.4 Discussions

2.4.1 Summary

We introduced two new notions that can make the minimax results more meaningful and appealing for practical purposes: (i) signal-to-noise-ratio aware minimaxity, (ii) second-order asymptotic approximation of minimax risk. We showed that these two notions can alleviate the major drawbacks of the classical minimax results. For instance, while the classical results prove that the hard and soft thresholding estimators are minimax optimal, the new results reveal that in a wide range of low signal-to-noise ratios the two estimators are in fact sub-optimal. Even when the signal-to-noise ratio is high, only hard thresholding is optimal and soft thresholding remains sub-optimal. Furthermore, our refined minimax analysis identified three optimal (or nearly optimal) estimators in three regimes with varying SNR: hard thresholding $\hat{\eta}_H(y, \lambda)$ of (2.5) in high SNR; $\hat{\eta}_E(y, \lambda, \gamma)$ of (2.10) in moderate SNR; linear estimator $\hat{\eta}_L(y, \lambda)$ of (2.9) in low SNR. As is clear from the definition of the three estimators, they are induced by ℓ_0-regularization, elastic net regularization [23] and ℓ_2-regularization, respectively. These regularization techniques have been widely used in statistics and machine learning [24]. In the next section, we discuss some related works.

The concepts of signal-to-noise ratio aware minimaxity and higher-order asymptotic approximations introduced in this thesis may open up new venues for investigating various estimation problems. As will be shown in next chapter, we have used the same framework to revisit the sparse estimation problem in high-dimensional linear regression and obtained new insights. That being said, it is important to acknowledge that the additional insights gained from this framework come with increased mathematical complexity when computing minimax estimators. Therefore, one direction we plan to explore in the future is the development of simpler and more general techniques for obtaining higher-order approximations of minimax risk or the supremum risk of well-established estimators.

2.4.2 Related works

There are some recent works on the significance of SNR for sparse learning. The extensive simulations conducted in the linear regression setting by [8] demonstrated that best subset selection (ℓ_0-regularization) performs better than the lasso (ℓ_1-regularization) in very high SNR, while the lasso outperforms best subset selection in low SNR regimes. [25, 9] developed new variants of subset selection that can perform consistently well in various levels of SNR. Some authors of the current paper (with their collaborators) established sharp theoretical characterizations of ℓ_q-regularization under varying SNR regimes in high-dimensional sparse regression and variable selection problems [20, 26, 10]. In particular, their results revealed that among the ℓ_q-regularization for $q \in [0, 2]$, as SNR decreases from high to low levels, the optimal value of q for parameter estimation and variable selection will move from 0 towards 2. All the aforementioned works studied the impact of SNR on several or a family of popular estimators. Hence their comparison conclusions are only applicable to a restricted set of estimators. In contrast, our work focused on minimax analysis that led to stronger optimality-type conclusions. For example, the preceding works showed that ℓ_2-regularization outperforms other ℓ_q-regularization when SNR is low. We obtained a stronger result that ℓ_2-regularization is in fact (minimax) optimal among all the estimators in low SNR.

In a separate work, the first order minimax optimality is also proved for other estimators, such as empirical Bayes estimators [27]. However, as we discussed before, first order minimax analysis is inherently incapable of evaluating the impact of the SNR on the performance of different estimators.

The second-order analysis of the minimax risk of the Gaussian sequence model under the sparsity constraint has been discussed in [28]. To compare this paper with our work, we have to mention the following points: (1) Such analysis still suffers from the fact that it disregards the effect of the signal-to-noise ratio. By restricting the signal-to-noise ratio, our SNR-aware minimax framework provides much more refined information about the minimax estimators. (2) In terms of the theoretical analysis, the SNR-aware minimax analysis requires much more delicate analysis compared to the classical settings where there is no constraint on the SNR. In particular, constructing and proving the least favorable distributions is more complicated in our settings compared to the classical setting. As a result, all the following steps of the proof become more complicated too.

We should also emphasize that minimax analysis over classes of ℓ_p balls (i.e., $\Theta = \{\theta : \|\theta\|_p \leq C_n\}$) for $p > 0$ under Gaussian sequence model has been performed in [11, 3, 29]. These works revealed that a notion of SNR involving C_n and σ_n plays a critical role in characterizing the asymptotic minimax risk and the optimality of linear or thresholding estimators. Finally, see [30, 31] for non-asymptotic minimax rate analysis of variable selection and functional estimation on sparse Gaussian sequence models.

2.4.3 Future research

Several important directions are left open for future research:

- The thesis considered estimating signals with sparsity $k_n/n \to 0$. The other denser regime where $k_n/n \to c > 0$ is also important to study. This will provide complementary asymptotic insights into the estimation of signals with varying sparsity. There exists classical minimax analysis along this line (see Chapter 8 in [3]). A generalization of SNR-aware minimaxity to this regime is an interesting future work.

- The obtained minimax optimal estimators involve tuning parameters that depend on unknown quantities such as sparsity k_n and signal strength τ_n from the parameter space. It is important to develop fully data-driven estimators that retain optimality for practical use. Hence, adaptive minimaxity is the next step, and classical adaptivity results (e.g., [3]) may be helpful for the development.

- In this thesis, we have focused on the parameter spaces that imposed the exact sparsity on θ. Sparsity promoting denoisers such as hard thresholding and soft thresholding have been also used over other structured parameter spaces such as Sobolev ellipsoids and Besov bodies. These parameter spaces usually characterize the smoothness properties of functions in terms of their Fourier or wavelet coefficients. We refer to [22, 3, 4] and references therein for a systematic treatment of this topic. An interesting future research would be to explore the implications of the SNR-aware minimaixty and higher-order approximation of the minimax risk for such spaces.

- The current work focused on the classical sparse Gaussian sequence model. It would be interesting to pursue a generalization to high-dimensional sparse linear regressions. Existing works (see [32, 5] and references there) established minimax rate optimality (with loose constants) which is not adequate to accurately capture the impact of SNR. Instead, the goal is to derive asymptotic approximations with sharp constants as we did for Gaussian sequence models. We believe that this is generally a very challenging problem without imposing specific constraint on the design matrix. A good starting point is to consider the "compressed sensing" model whose design rows follow independent isotropic Gaussian distribution. We have made some major progress along this line and look forward to further development.

2.5 Proofs of the main results

2.5.1 Preliminaries

Scale invariance

The minimax risk defined in (2.7) has the following scale invariance property

$$R(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R(\Theta(k_n, \mu_n), 1),$$

where we recall that $\mu_n = \tau_n/\sigma_n$. This can be easily verified by rescaling the Gaussian sequence model to have unit variance. Moreover, similar invariance holds for the four estimators considered in the chapter. We state it without proof in the following: $\forall \sigma > 0$,

$$\sigma \cdot \hat{\eta}_S(y, \lambda) = \hat{\eta}_S(\sigma y, \sigma\lambda), \quad \sigma \cdot \hat{\eta}_H(y, \lambda) = \hat{\eta}_S(\sigma y, \sigma\lambda),$$
$$\sigma \cdot \hat{\eta}_L(y, \lambda) = \hat{\eta}_L(\sigma y, \lambda), \quad \sigma \cdot \hat{\eta}_E(y, \lambda, \gamma) = \hat{\eta}_E(\sigma y, \sigma\lambda, \gamma).$$

These invariance properties will be frequently used in the proof to reduce a problem to a simpler one under unit variance.

Gaussian tail bound

Recall the notation that ϕ, Φ denote the probability density function and cumulative distribution function of a standard normal random variable, respectively. The following Gaussian tail bound will be extensively used in the proof.

Lemma 1 (Exercise 8.1 in [3])**.** *Define*

$$\tilde{\Phi}_l(\lambda) := \lambda^{-1}\phi(\lambda) \sum_{k=0}^{l} \frac{(-1)^k}{k!} \frac{\Gamma(2k+1)}{2^k \lambda^{2k}},$$

where $\Gamma(\cdot)$ *is the gamma function. Then, for each* $k \geq 0$ *and all* $\lambda > 0$*:*

$$\tilde{\Phi}_{2k+1}(\lambda) \leq 1 - \Phi(\lambda) \leq \tilde{\Phi}_{2k}(\lambda).$$

The minimax theorem

Consider the Gaussian sequence model:

$$y_i = \theta_i + \sigma z_i, \quad i = 1, 2, \ldots, n, \tag{2.12}$$

where $z_1, z_2, \ldots, z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. If π is a prior distribution of $\theta \in \mathbb{R}^n$, the integrated risk of an estimator $\hat{\theta}$ (with squared error loss) is $B(\hat{\theta}, \pi) = \mathbb{E}_\pi \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2$, and the Bayes risk of π is $B(\pi) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi)$. We state a version of minimax theorem suited to the Gaussian sequence model. The theorem allows to evaluate minimax risk by calculating the maximum Bayes risk over a class of prior distributions.

Theorem 9 (Theorem 4.12 in [3])**.** *Consider the Gaussian sequence model* (2.12)*. Let* $\mathcal{P}$ *be a convex set of probability measures on* $\mathbb{R}^n$*. Then*

$$\inf_{\hat{\theta}} \sup_{\pi \in \mathcal{P}} B(\hat{\theta}, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\hat{\theta}} B(\hat{\theta}, \pi) = \sup_{\pi \in \mathcal{P}} B(\pi).$$

A maximising π *is called a least favorable distribution (with respect to* $\mathcal{P}$*).*

Independence is less favorable

We present a useful result that can often help find the least favorable distributions. Let π be an arbitrary prior, so that the θ_j are not necessarily independent. Denote by π_j the marginal distribution of θ_j. Build a new prior $\bar{\pi}$ by making the θ_j independent: $\bar{\pi} = \prod_j \pi_j$. This product prior has a larger Bayes risk.

Theorem 10 (Lemma 4.15 in [3])**.** $B(\bar{\pi}) \geq B(\pi)$.

A machinery for obtaining lower bounds for the minimax risk

In our results, we are often interested in finding lower bounds for the minimax risk. The following elementary result taken from Chapter 4.3 of [3] will be useful in those cases.

Theorem 11. *Consider the minimax risk of a risk function $r(\cdot,\cdot)$ over a parameter set Θ:*

$$R(\Theta) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} r(\hat{\theta},\theta).$$

Recall that $B(\pi)$ is the Bayes risk of prior π: $B(\pi) = \inf_{\hat{\theta}} \int r(\hat{\theta},\theta)\pi(d\theta)$. Let $\mathcal{P}$ denote a collection of probability measure, and $\operatorname{supp}\mathcal{P}$ *denote the union of all* $\operatorname{supp}\pi$ *for π in $\mathcal{P}$. If*

$$B(\mathcal{P}) = \sup_{\pi\in\mathcal{P}} B(\pi),$$

then

$$\operatorname{supp}\mathcal{P} \subset \Theta \quad \Rightarrow \quad R(\Theta) \geq B(\mathcal{P}).$$

2.5.2 Proof of Theorem 5

To calculate the minimax risk $R(\Theta(k_n,\tau_n),\sigma_n)$, we first obtain an upper bound by computing the supremum risk of the linear estimator $\hat{\eta}_L(y,\lambda_n)$,

$$R(\Theta(k_n,\tau_n),\sigma_n) \leq \sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\|\hat{\eta}_L(y,\lambda_n)-\theta\|_2^2.$$

We then derive a matching lower bound based on Theorem 11. In particular, we construct a particular prior supported on $\Theta(k_n,\tau_n)$ (that is the least favorable prior at the level of approximation we require), and its corresponding Bayes risk leads to a sharp lower bound for the minimax risk. The detailed derivation of the upper and lower bounds is presented below.

Upper bound

Thanks to the simple form of the linear estimator $\hat{\eta}_L(y, \lambda_n)$, its supremum risk under tuning $\lambda_n = (\epsilon_n \mu_n^2)^{-1}$ can be computed in a straightforward way: for all $\theta \in \Theta(k_n, \tau_n)$,

$$\begin{aligned} & \mathbb{E}_\theta \|\hat{\eta}_L(y, \lambda_n) - \theta\|_2^2 = \mathbb{E}_\theta \sum_{i=1}^n \left(\frac{1}{1+\lambda_n} y_i - \theta_i \right)^2 \\ = & \sum_{i=1}^n \left[\left(\frac{\lambda_n}{1+\lambda_n} \right)^2 \theta_i^2 + \left(\frac{1}{1+\lambda_n} \right)^2 \sigma_n^2 \right] \leq \frac{\lambda_n^2 k_n \tau_n^2 + n\sigma_n^2}{(1+\lambda_n)^2} = \frac{n\sigma_n^2 \epsilon_n \mu_n^2}{1+\epsilon_n \mu_n^2} \\ = & n\sigma_n^2 \epsilon_n \mu_n^2 \cdot \left(1 - \epsilon_n \mu_n^2 (1+\epsilon_n \mu_n^2)^{-1} \right) = n\sigma_n^2 \epsilon_n \mu_n^2 \cdot \left(1 - \epsilon_n \mu_n^2 (1+o(1)) \right), \end{aligned}$$

where we have used the assumption $\epsilon_n = k_n/n \to 0, \mu_n = \tau_n/\sigma_n \to 0$, and the constraint $\|\theta\|_2^2 \leq k_n \tau_n^2, \forall \theta \in \Theta(k_n, \tau_n)$. As a result,

$$R(\Theta(k_n, \tau_n), \sigma_n) \leq \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_L(y, \lambda_n) - \theta\|_2^2 = n\sigma_n^2 \epsilon_n \mu_n^2 \cdot \left(1 - \epsilon_n \mu_n^2 (1+o(1)) \right).$$

Lower bound

First, due to the scale invariance property $R(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R(\Theta(k_n, \mu_n), 1)$ (see Section 3.4.1), it is sufficient to obtain lower bound for $R(\Theta(k_n, \mu_n), 1)$, i.e., the minimax risk under Gaussian sequence model: $y_i = \theta_i + z_i, 1 \leq i \leq n$, with $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. A general strategy for finding lower bounds of minimax risk in sparse Gaussian sequence model, is to employ i.i.d. univariate spike prior as the (asymptotically) least favorable prior. Although such product prior served as a suitable tool to establish a sharp lower bound for proving Theorem 3, we have since recognized its inadequacy in providing a sufficiently sharp lower bound for obtaining the second-order approximation of the minimax risk. Hence, in order to use Theorem 11, we utilize the family of *independent block priors* [33, 3]. The specific independent block prior $\pi^{IB}(\theta)$ on $\Theta(k_n, \mu_n)$ for our problem is constructed in the following steps:

1. Divide $\theta \in \mathbb{R}^n$ into k_n disjoint blocks of dimension $m = n/k_n$[1]:

$$\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k_n)}).$$

2. Sample each block $\theta^{(j)} \in \mathbb{R}^m$ from the symmetric spike prior $\pi_S^{\mu,m}$: for $1 \le i \le m$,

$$\pi_S^{\mu,m}\left(\theta^{(j)} = \mu e_i\right) = \pi_S^{\mu,m}\left(\theta^{(j)} = -\mu e_i\right) = \frac{1}{2m},$$

   where $\mu \in (0, \mu_n]$ is a location parameter.

3. Combine independent blocks:

$$\pi^{IB}(\theta) = \prod_{j=1}^{k_n} \pi_S^{\mu,m}(\theta^{(j)})$$

   .

In other words, the independent block prior π^{IB} picks a single spike (from $2m$ possible locations) in each of k_n non-overlapping blocks of θ, with the spike location within each block being independent and uniform. As is clear from the construction, $\operatorname{supp} \pi^{IB} \subseteq \Theta(k_n, \mu_n)$ so that

$$R(\Theta(k_n, \mu_n), 1) \ge B(\pi^{IB}) = k_n \cdot B(\pi_S^{\mu,m}). \tag{2.13}$$

Here, the last equation holds because when the prior has block independence and the loss function is additive, the Bayes risk can be decomposed into the sum of Bayes risk of prior for each block (see Chapter 4.5 in [3]).

As a result, the main goal of the rest of this section is to obtain a sharp lower bound (*up to the second order*) for the Bayes risk $B(\pi_S^{\mu,m})$, i.e., the risk of the posterior mean under the spike prior $\pi_S^{\mu,m}$. The following two lemmas are instrumental in obtaining such a sharp lower bound.

[1]For simplicity, here we assume n/k_n is an integer. In the case when it is not, we can slightly adjust the block size to obtain the same lower bound.

Lemma 2. *Consider the Gaussian sequence model:* $y_i = \theta_i + z_i, 1 \le i \le m$, *with* $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. *The Bayes risk of* $\pi_S^{\mu,m}$ *takes the form*

$$B(\pi_S^{\mu,m}) = \mathbb{E}_{\mu e_1}(\hat{\theta}_1 - \mu)^2 + (m-1)\mathbb{E}_{\mu e_2}\hat{\theta}_1^2,$$

where $\mathbb{E}_{\mu e_1}(\cdot)$ *is taken with respect to* $y \sim \mathcal{N}(\mu e_1, I)$ *and* $\mathbb{E}_{\mu e_2}(\cdot)$ *for* $y \sim \mathcal{N}(\mu e_2, I)$*;* $\hat{\theta}_1$ *is the posterior mean for the first coordinate having the expression*

$$\hat{\theta}_1 = \frac{\mu(e^{\mu y_1} - e^{-\mu y_1})}{\sum_{i=1}^m (e^{\mu y_i} + e^{-\mu y_i})}.$$

Proof. Let the posterior mean be $\hat{\theta} = \mathbb{E}[\theta|y]$. Using Bayes' Theorem we obtain

$$\begin{aligned}\hat{\theta}_1 &= \mu\mathbb{P}(\theta = \mu e_1 \mid y) - \mu\mathbb{P}(\theta = -\mu e_1 \mid y) \\ &= \frac{\mu[\mathbb{P}(y \mid \theta = \mu e_1) - \mathbb{P}(y \mid \theta = -\mu e_1)]}{\sum_{i=1}^m [\mathbb{P}(y \mid \theta = \mu e_i) + \mathbb{P}(y \mid \theta = -\mu e_i)]} = \frac{\mu(e^{\mu y_1} - e^{-\mu y_1})}{\sum_{i=1}^m (e^{\mu y_i} + e^{-\mu y_i})}.\end{aligned}$$

Moreover, since both θ_i's (under the prior) and z_i's are exchangeable, the pairs $\{(\hat{\theta}_i, \theta_i)\}_{i=1}^m$ are exchangeable as well. As a result,

$$\begin{aligned}&B(\pi_S^{\mu,m}) = \mathbb{E}\sum_{i=1}^m (\hat{\theta}_i - \theta_i)^2 = m\mathbb{E}(\hat{\theta}_1 - \theta_1)^2 \\ =&m\left[\frac{1}{2m}\mathbb{E}_{\mu e_1}(\hat{\theta}_1 - \mu)^2 + \frac{1}{2m}\mathbb{E}_{-\mu e_1}(\hat{\theta}_1 + \mu)^2 + \frac{1}{2m}\sum_{i=2}^m \left(\mathbb{E}_{\mu e_i}\hat{\theta}_1^2 + \mathbb{E}_{-\mu e_i}\hat{\theta}_1^2\right)\right] \\ =&\frac{1}{2}\left[\mathbb{E}_{\mu e_1}(\hat{\theta}_1 - \mu)^2 + \mathbb{E}_{-\mu e_1}(\hat{\theta}_1 + \mu)^2\right] + \frac{1}{2}\sum_{i=2}^m \left[\mathbb{E}_{\mu e_i}\hat{\theta}_1^2 + \mathbb{E}_{-\mu e_i}\hat{\theta}_1^2\right] \\ =&\mathbb{E}_{\mu e_1}(\hat{\theta}_1 - \mu)^2 + (m-1)\mathbb{E}_{\mu e_2}\hat{\theta}_1^2,\end{aligned}$$

where in the last equation we have used the facts that the distribution of $\hat{\theta}_1$ under $\theta = \mu e_1$ equals that of $-\hat{\theta}_1$ under $\theta = -\mu e_1$, and $\hat{\theta}_1$ has the same distribution when $\theta = \pm\mu e_i, i = 2, \ldots, m$. □

Lemma 3. *As $\mu \to 0, m \to \infty$, The Bayes risk of $\pi_S^{\mu,m}$ has the lower bound*

$$B(\pi_S^{\mu,m}) \geq \mu^2 - \frac{\mu^4}{m}(1 + o(1)).$$

Proof. Denote $p_m = \frac{e^{\mu y_1} - e^{-\mu y_1}}{\sum_{i=1}^m (e^{\mu y_i} + e^{-\mu y_i})}$. According to Lemma 2, the Bayes risk can be lower bounded in the following way:

$$B(\pi_S^{\mu,m}) \geq \mu^2 \cdot \left[1 - 2\mathbb{E}_{\mu e_1} p_m + (m-1)\mathbb{E}_{\mu e_2} p_m^2\right].$$

It is thus sufficient to prove that $\mathbb{E}_{\mu e_1} p_m \leq \frac{\mu^2}{m}(1 + o(1))$ and $(m-1)\mathbb{E}_{\mu e_2} p_m^2 \geq \frac{\mu^2}{m}(1 + o(1))$. We first prove the former one. We have

$$\begin{aligned}
\mathbb{E}_{\mu e_1} p_m &= \mathbb{E}\left[\frac{e^{\mu(\mu+z_1)} - e^{-\mu(\mu+z_1)}}{\sum_{j\neq 1}[e^{\mu z_j} + e^{-\mu z_j}] + e^{\mu(\mu+z_1)} + e^{-\mu(\mu+z_1)}}\right] \\
&= \mathbb{E}\left[\frac{(e^{\mu^2} - 1)e^{\mu z_1}}{\sum_{j\neq 1}[e^{\mu z_j} + e^{-\mu z_j}] + e^{\mu(\mu+z_1)} + e^{-\mu(\mu+z_1)}}\right] \\
&+ \mathbb{E}\left[\frac{(1 - e^{-\mu^2})e^{-\mu z_1}}{\sum_{j\neq 1}[e^{\mu z_j} + e^{-\mu z_j}] + e^{\mu(\mu+z_1)} + e^{-\mu(\mu+z_1)}}\right] \\
&+ \mathbb{E}\left[\frac{e^{\mu z_1} - e^{-\mu z_1}}{\sum_{j\neq 1}[e^{\mu z_j} + e^{-\mu z_j}] + e^{\mu(\mu+z_1)} + e^{-\mu(\mu+z_1)}}\right] \\
&=: E_1 + E_2 + E_3.
\end{aligned}$$

We study E_1, E_2 and E_3 separately. For E_1, given that the numerator inside the expectation is positive, we apply the basic inequality $a + b \geq 2\sqrt{ab}, \forall a, b \geq 0$ to the denominator to obtain

$$E_1 \leq \frac{e^{\mu^2} - 1}{2m}\mathbb{E}e^{\mu z_1} = \frac{\mu^2}{2m} \cdot \frac{(e^{\mu^2} - 1)e^{\mu^2/2}}{\mu^2} = \frac{\mu^2(1 + o(1))}{2m}.$$

Similarly, for E_2 we have

$$E_2 \leq \frac{1 - e^{-\mu^2}}{2m}\mathbb{E}e^{-\mu z_1} = \frac{\mu^2}{2m} \cdot \frac{(1 - e^{-\mu^2})e^{\mu^2/2}}{\mu^2} = \frac{\mu^2(1 + o(1))}{2m}.$$

To study E_3, define

$$A := \sum_{j\neq 1}[e^{\mu z_j} + e^{-\mu z_j}] + e^{\mu(\mu+z_1)} + e^{-\mu(\mu+z_1)},$$
$$B := \sum_{j\neq 1}[e^{\mu z_j} + e^{-\mu z_j}] + e^{\mu(\mu-z_1)} + e^{-\mu(\mu-z_1)}.$$

The basic inequality $a+b \geq 2\sqrt{ab}$ implies that $A \geq 2m, B \geq 2m$. This together with the symmetry of standard normal distribution yields

$$E_3 = \mathbb{E}\frac{e^{\mu z_1}}{A} - \mathbb{E}\frac{e^{-\mu z_1}}{A} = \mathbb{E}\frac{e^{\mu z_1}}{A} - \mathbb{E}\frac{e^{\mu z_1}}{B} = \mathbb{E}\left[\frac{(e^{\mu^2} - e^{-\mu^2})(e^{-\mu z_1} - e^{\mu z_1})e^{\mu z_1}}{AB}\right]$$
$$\leq \mathbb{E}\left[\frac{(e^{\mu^2} - e^{-\mu^2})(1 - e^{2\mu z_1})I_{(z_1\leq 0)}}{AB}\right] \leq \frac{e^{\mu^2} - e^{-\mu^2}}{4m^2}\mathbb{E}\left[(1 - e^{2\mu z_1})\mathbb{1}_{(z_1\leq 0)}\right] = O\Big(\frac{\mu^2}{m^2}\Big)$$

It remains to prove $(m-1)\mathbb{E}_{\mu e_2} p_m^2 \geq \frac{\mu^2}{m}(1+o(1))$. Denote

$$C := \left[e^{\mu b} + e^{-\mu b} + 2(m-2)e^{\frac{\mu^2}{2}} + e^{\frac{3}{2}\mu^2} + e^{-\frac{\mu^2}{2}}\right]^2,$$

where $b > 0$ is a scalar to be specified later. Then

$$\begin{aligned}
\mathbb{E}_{\mu e_2} p_m^2 &= \mathbb{E}\left[\frac{(e^{\mu z_1} - e^{-\mu z_1})^2}{\left[\sum_{j\neq 2}(e^{\mu z_j} + e^{-\mu z_j}) + e^{\mu(\mu+z_2)} + e^{-\mu(\mu+z_2)}\right]^2}\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[\frac{(e^{\mu z_1} - e^{-\mu z_1})^2}{[e^{\mu z_1} + e^{-\mu z_1} + 2(m-2)e^{\frac{\mu^2}{2}} + e^{\frac{3}{2}\mu^2} + e^{-\frac{\mu^2}{2}}]^2}\right] \\
&\geq \mathbb{E}\left[\frac{(e^{\mu z_1} - e^{-\mu z_1})^2 I_{(|z_1|\leq b)}}{[e^{\mu b} + e^{-\mu b} + 2(m-2)e^{\frac{\mu^2}{2}} + e^{\frac{3}{2}\mu^2} + e^{-\frac{\mu^2}{2}}]^2}\right] \\
&= \frac{2}{C}\left[\mathbb{E}e^{2\mu z_1} I_{(|z_1|\leq b)} - \mathbb{P}(|z_1| \leq b)\right] \\
&= \frac{2}{C}\left[e^{2\mu^2}\int_{-b-2\mu}^{b-2\mu}\phi(z)dz - \int_{-b}^{b}\phi(z)dz\right] \\
&= \frac{2}{C}\left[(e^{2\mu^2} - 1)\int_{-b-2\mu}^{b-2\mu}\phi(z)dz - \int_{b-2\mu}^{b}\phi(z)dz + \int_{-b-2\mu}^{-b}\phi(z)dz\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(b)}}{=} \quad \frac{2}{C}\left[2\mu^2(1+o(1)) + o(\mu^2) + o(\mu^2)\right] \\
&\overset{\text{(c)}}{\geq} \quad \frac{2}{4m^2e^{2\sqrt{\mu}}} \cdot 2\mu^2(1+o(1)) = \frac{\mu^2}{m^2}(1+o(1)).
\end{aligned}$$

Inequality (a) is obtained by conditioning on z_1 and applying Jensen's inequality on the convex function $1/(x+c)^2$ for $x > 0$. Equality (b) holds by setting $b = 1/\sqrt{\mu}$, for the purpose of matching the asymptotic order $\frac{\mu^2}{m}(1 + o(1))$. Finally, inequality (c) is because $C \leq 4m^2e^{2\sqrt{\mu}}$ when μ is sufficiently small. □

We are in the position to derive the matching lower bound for the minimax risk. Recall that in the block prior we have $m = n/k_n, \mu \in (0, \mu_n]$. Set $\mu = \mu_n$. The assumption $\epsilon_n = k_n/n \to 0, \mu_n \to 0$ guarantees that the condition $m \to \infty, \mu \to 0$ in Lemma 3 is satisfied. We therefore combine Lemma 3 and (2.13) to obtain

$$\begin{aligned}
R(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R(\Theta(k_n, \mu_n), 1) &\geq k_n\sigma_n^2 \cdot B(\pi_S^{\mu,m}) \\
&\geq k_n\sigma_n^2 \cdot \left[\mu_n^2 - \frac{\mu_n^4 k_n}{n}(1+o(1))\right] \\
&= n\sigma_n^2 \cdot \left(\epsilon_n\mu_n^2 - \epsilon_n^2\mu_n^4(1+o(1))\right).
\end{aligned}$$

2.5.3 Proof of Proposition 1

Define the supremum risk of optimally tuned soft thresholding estimator as

$$R_s(\Theta(k_n, \tau_n), \sigma_n) = \inf_{\lambda>0} \sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta \left\|\hat{\eta}_S(y, \lambda) - \theta\right\|_2^2,$$

where $y_i = \theta_i + \sigma_n z_i$, with $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. It is straightforward to verify that

$$R_s(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R_s(\Theta(k_n, \mu_n), 1). \tag{2.14}$$

Hence, without loss of generality, in the rest of the proof we will assume that $\sigma_n = 1$.

Since $\hat{\eta}_S(y,\lambda)$ is the special case of $\hat{\eta}_E(y,\lambda,\gamma)$ with $\gamma = 0$, the supremum risk result stated in Equation (2.43) for $\hat{\eta}_E(y,\lambda,\gamma)$ applies to $\hat{\eta}_S(y,\lambda)$ as well. It shows that the supremum risk of $\hat{\eta}_S(y,\lambda)$ is attained on a particular boundary of the parameter space:

$$\begin{aligned}\sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}\sum_{i=1}^{n} |\hat{\eta}_S(y_i,\lambda)-\theta_i|_2^2 &= (n-k_n)r_S(\lambda,0)+k_n r_S(\lambda,\mu_n)\\ &= n\left[(1-\epsilon_n)r_S(\lambda,0)+\epsilon_n r_S(\lambda,\mu_n)\right], \end{aligned} \tag{2.15}$$

with $\epsilon_n = k_n/n$ and $r_S(\lambda,\mu)$ defined as

$$r_S(\lambda,\mu) = \mathbb{E}(\hat{\eta}_S(\mu+z,\lambda)-\mu)^2, \;\; z\sim\mathcal{N}(0,1). \tag{2.16}$$

To prove Proposition 1, we need to find the optimal λ that minimizes the supremum risk in (2.15), or equivalently, the function

$$F(\lambda) := (1-\epsilon_n)r_S(\lambda,0)+\epsilon_n r_S(\lambda,\mu_n). \tag{2.17}$$

Lemma 4. *Denote the optimal tuning by* $\lambda_* = \arg\min_{\lambda\geq 0} F(\lambda)$. *It holds that*

$$\log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2} - 2\log\log\frac{2}{\epsilon_n} < \lambda_*\mu_n < \log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2}, \tag{2.18}$$

when n is sufficiently large.

Proof. Using integration by parts, we first obtain a more explicit expression for $F(\lambda)$:

$$F(\lambda) = (1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z,\lambda) + \epsilon_n\mu_n^2 - 2\epsilon_n\mu_n\mathbb{E}\hat{\eta}_S(\mu_n+z,\lambda) + \epsilon_n\mathbb{E}\hat{\eta}_S^2(\mu_n+z,\lambda), \tag{2.19}$$

where the three expectations take the form

$$\mathbb{E}\hat{\eta}_S^2(z,\lambda) = 2(1+\lambda^2)\int_\lambda^\infty \phi(z)dz - 2\lambda\phi(\lambda) \tag{2.20}$$

$$\mathbb{E}\hat{\eta}_S(\mu_n + z, \lambda) = \phi(\lambda - \mu_n) + (\mu_n - \lambda)\int_{\lambda-\mu_n}^{\infty} \phi(z)dz - \phi(\lambda + \mu_n) + (\mu_n + \lambda)\int_{\lambda+\mu_n}^{\infty} \phi(z)dz \quad (2.21)$$

$$\begin{aligned}\mathbb{E}\hat{\eta}_S^2(\mu_n + z, \lambda) =& \left[\left(1 + (\lambda - \mu_n)^2\right)\int_{\lambda-\mu_n}^{\infty} \phi(z)dz - (\lambda - \mu_n)\phi(\lambda - \mu_n)\right] \\ &+ \left[\left(1 + (\lambda + \mu_n)^2\right)\int_{\lambda+\mu_n}^{\infty} \phi(z)dz - (\lambda + \mu_n)\phi(\lambda + \mu_n)\right]. \quad (2.22)\end{aligned}$$

Therefore, $F(\lambda)$ is a differentiable function of λ, and as long as the infimum of $F(\lambda)$ is not achieved at 0 or $+\infty$, λ_* will satisfy $F'(\lambda_*) = 0$. From Equations (2.19)-(2.22), it is direct to compute $F(0) = 1 > F(+\infty) = \epsilon_n\mu_n^2$ for large n. Moreover, as we will show in the end of the proof, $F(\lambda)$ is increasing when λ is above a threshold. Hence, the optimal tuning $\lambda_* \in (0, \infty)$, and we can characterize it through the derivative equation:

$$\begin{aligned}0 = F'(\lambda_*) =& (1 - \epsilon_n)\left[4\lambda_*\int_{\lambda_*}^{\infty} \phi(z)dz - 4\phi(\lambda_*)\right] \\ &+ \epsilon_n\left[-2\phi(\lambda_* - \mu_n) - 2\phi(\lambda_* + \mu_n) + 2\lambda_*\left(\int_{\lambda_*-\mu_n}^{\infty} \phi(z)dz + \int_{\lambda^*+\mu_n}^{\infty} \phi(z)dz\right)\right]. \quad (2.23)\end{aligned}$$

First, we show that $\lambda_* \to \infty$. Suppose this is not true. Then $\lambda_* \le C$ for some constant $C > 0$ (take a subsequence if necessary). From (2.19), we have

$$F(\lambda_*) \ge (1 - \epsilon_n)r_S(C, 0) = 2(1 - \epsilon_n)\left[(1 + C^2)\int_C^{\infty} \phi(z)dz - C\phi(C)\right] > \epsilon_n\mu_n^2 = F(+\infty),$$

when n is large. This contradicts with the optimality of λ_*.

Second, we prove that $\lambda_*\mu_n \to \infty$. Otherwise, $\lambda_*\mu_n = O(1)$ (take a subsequence if necessary). We will show that it leads to a contradiction in (2.23). Using the Gaussian tail bound $\int_t^{\infty} \phi(z)dz = (\frac{1}{t} - \frac{1+o(1)}{t^3})\phi(t)$ as $t \to \infty$ from Section 2.5.1, since $\lambda_* \to \infty, \mu_n \to 0, \lambda_*\mu_n = O(1)$, we obtain

$$-\lambda_*\int_{\lambda_*}^{\infty} \phi(z)dz + \phi(\lambda_*) = (1 + o(1)) \cdot \lambda_*^{-2}\phi(\lambda_*), \quad (2.24)$$

$$-\phi(\lambda_* + \mu_n) + \lambda_*\int_{\lambda_*+\mu_n}^{\infty} \phi(z)dz = O(\lambda_*^{-2}\phi(\lambda_*)), \quad (2.25)$$

$$-\phi(\lambda_* - \mu_n) + \lambda_* \int_{\lambda_* - \mu_n}^{\infty} \phi(z)dz = O(\lambda_*^{-2}\phi(\lambda_*)). \tag{2.26}$$

Given that $\epsilon_n \to 0$, combining the above results with (2.23) implies that $0 = F'(\lambda_*) \cdot \lambda_*^2 \phi^{-1}(\lambda_*) = -4 + o(1)$, which is a contradiction.

Third, we show that $\lambda_* \mu_n < \log \frac{2}{\epsilon_n} + \frac{\mu_n^2}{2}$ for large n. Now that we have proved $\lambda_* \mu_n \to \infty$, results in (2.25)-(2.26) can be strengthened:

$$-\phi(\lambda_* + \mu_n) + \lambda_* \int_{\lambda_* + \mu_n}^{\infty} \phi(z)dz = o(\mu_n \lambda_*^{-1} \phi(\lambda_* - \mu_n)), \tag{2.27}$$

$$-\phi(\lambda_* - \mu_n) + \lambda_* \int_{\lambda_* - \mu_n}^{\infty} \phi(z)dz = (1 + o(1)) \cdot \mu_n \lambda_*^{-1} \phi(\lambda_* - \mu_n). \tag{2.28}$$

Plugging (2.24) and (2.27)-(2.28) into (2.23) gives $(4+o(1)) \cdot \lambda_*^{-2} \phi(\lambda_*) = (2+o(1)) \cdot \epsilon_n \mu_n \lambda_*^{-1} \phi(\lambda_* - \mu_n)$, which can be further simplified as

$$2 + o(1) = \epsilon_n \mu_n \lambda_* \exp(\lambda_* \mu_n - \mu_n^2/2). \tag{2.29}$$

The above equation implies that $\lambda_* \mu_n < \log \frac{2}{\epsilon_n} + \frac{\mu_n^2}{2}$ for large n. Otherwise, the right-hand side will be no smaller than $2\mu_n \lambda_n \to \infty$ contradicting with the left-hand side term.

Fourth, we prove that $\lambda_* \mu_n > \log \frac{2}{\epsilon_n} + \frac{\mu_n^2}{2} - 2 \log \log \frac{2}{\epsilon_n}$ when n is large. Otherwise, suppose $\lambda_* \mu_n \leq \log \frac{2}{\epsilon_n} + \frac{\mu_n^2}{2} - 2 \log \log \frac{2}{\epsilon_n}$ (take a subsequence if necessary). This leads to

$$0 \leq \epsilon_n \mu_n \lambda_* \exp(\lambda_* \mu_n - \mu_n^2/2) \leq \frac{2\mu_n \lambda_*}{(\log \frac{2}{\epsilon_n})^2} < \frac{2 \log \frac{2}{2\epsilon_n} + \mu_n^2}{(\log \frac{2}{\epsilon_n})^2} = o(1),$$

where we have used the upper bound $\lambda_* \mu_n < \log \frac{2}{\epsilon_n} + \frac{\mu_n^2}{2}$ derived earlier. The obtained result contradicts with (2.29).

Finally, as mentioned earlier in the proof, we need to show that $\lambda_* \neq +\infty$ for large n. It is sufficient to prove that $F'(\lambda) > 0, \forall \lambda \in [\frac{2}{\mu_n} \log \frac{1}{\epsilon_n}, \infty)$, when n is large. To this end, using the Gaussian tail bound $\int_t^{\infty} \phi(z)dz \geq (\frac{1}{t} - \frac{1}{t^3})\phi(t), \forall t > 0$ and the derivative expression (2.23), we

have

$$F'(\lambda) \geq \frac{\phi(\lambda)}{\lambda^2} \cdot \Big[-4 + 4\epsilon_n + \frac{\mu_n(\lambda-\mu_n)^2 - \lambda}{(\lambda-\mu_n)^3\lambda^{-2}} 2\epsilon_n e^{\lambda\mu_n - \mu_n^2/2} + \frac{-\mu_n(\lambda+\mu_n)^2 - \lambda}{(\lambda+\mu_n)^3\lambda^{-2}} 2\epsilon_n e^{-\lambda\mu_n - \mu_n^2/2} \Big]$$
$$\geq \frac{\phi(\lambda)}{\lambda^2} \cdot \Big[-4 + 4\epsilon_n + (2+o(1)) \cdot \epsilon_n e^{-\mu_n^2/2} \lambda\mu_n e^{\lambda\mu_n} \Big],$$

where we used that $\lambda \geq \frac{2}{\mu_n}\log\frac{1}{\epsilon_n}$ implies $\lambda\mu_n = \omega(1)$. Note that the above asymptotic notion $o(\cdot)$ is uniform for all $\lambda \geq \frac{2}{\mu_n}\log\frac{1}{\epsilon_n}$ when n is large. Since $\lambda\mu_n \geq 2\log\frac{1}{\epsilon_n}$, we can easily continue from the above inequality to obtain $F'(\lambda) > 0$ for sufficiently large n. □

The next lemma turns $F(\lambda_*)$ into a form that is more amenable to asymptotic analysis.

Lemma 5. *Define*

$$\begin{aligned}
\mathcal{A} &= -\mu_n(\lambda_*-\mu_n) + 1 + \frac{\mu_n(\lambda_*-\mu_n)^3 e^{-2\lambda_*\mu_n}}{(\lambda_*+\mu_n)^2} + \frac{(\lambda_*-\mu_n)^3 e^{-2\lambda_*\mu_n}}{(\lambda_*+\mu_n)^3} + O\Big(\frac{\mu_n}{\lambda_*}\Big), \\
\mathcal{B} &= \mu_n(\lambda_*-\mu_n)^2 - \lambda_* + (3+o(1))\lambda_*^{-1} + \frac{[-\mu_n\lambda_*^2 - \lambda_*(1+2\mu_n^2(1+o(1)))]}{(\lambda_*+\mu_n)^3} \cdot (\lambda_*-\mu_n)^3 e^{-2\lambda_*\mu_n}.
\end{aligned}$$

As $\epsilon_n \to 0, \mu_n \to 0$, it holds that

$$F(\lambda_*) = \epsilon_n\mu_n^2 + \frac{4(1-\epsilon_n)\phi(\lambda_*)}{\lambda_*^3} \cdot \Big[1 - 6\lambda_*^{-2} + O(\lambda_*^{-4}) + \Big(\lambda_* - \frac{3+o(1)}{\lambda_*}\Big)\frac{\mathcal{A}}{\mathcal{B}}\Big].$$

Proof. We use Gaussian tail bounds to evaluate the three expectations (2.20)-(2.22) in the expression of $F(\lambda_*)$ in (2.19). Note that as shown in Lemma 4, $\lambda_*\mu_n = \Theta(\log 2\epsilon_n^{-1})$. The first expectation is

$$\mathbb{E}\hat{\eta}_S^2(z,\lambda_*) = 2\phi(\lambda_*)\left[2\lambda_*^{-3} - 12\lambda_*^{-5} + O(\lambda_*^{-7})\right]. \tag{2.30}$$

Regarding the second one, we obtain

$$\phi(\lambda_*-\mu_n) - (\lambda_*-\mu_n)\int_{\lambda_*-\mu_n}^{\infty}\phi(z)dz = \left[(\lambda_*-\mu_n)^{-2} + O\left((\lambda_*-\mu_n)^{-4}\right)\right]\phi(\lambda_*-\mu_n),$$

and

$$\phi(\lambda_*+\mu_n)-(\lambda_*+\mu_n)\int_{\lambda_*+\mu_n}^{\infty}\phi(z)dz=\left[(\lambda_*+\mu_n)^{-2}e^{-2\lambda_*\mu_n}+O\left((\lambda_*+\mu_n)^{-4}e^{-2\lambda_*\mu_n}\right)\right]\cdot\phi(\lambda_*-\mu_n).$$

Therefore,

$$\mathbb{E}\hat{\eta}_S(\mu_n+z,\lambda_*)=\left[(\lambda_*-\mu_n)^{-2}-(\lambda_*+\mu_n)^{-2}e^{-2\lambda_*\mu_n}+O\left((\lambda_*-\mu_n)^{-4}\right)\right]\phi(\lambda_*-\mu_n). \quad (2.31)$$

For the third expectation, we first have

$$\left(1+(\lambda_*-\mu_n)^2\right)\int_{\lambda_*-\mu_n}^{\infty}\phi(z)dz-(\lambda_*-\mu_n)\phi(\lambda_*-\mu_n)=\left[2(\lambda_*-\mu_n)^{-3}+O\left((\lambda_*-\mu_n)^{-5}\right)\right]\phi(\lambda_*-\mu_n),$$
$$\left(1+(\lambda_*+\mu_n)^2\right)\int_{\lambda_*+\mu_n}^{\infty}\phi(z)dz-(\lambda_*+\mu_n)\phi(\lambda_*+\mu_n)=\left[2(\lambda_*+\mu_n)^{-3}+O\left((\lambda_*+\mu_n)^{-5}\right)\right]\phi(\lambda_*+\mu_n).$$

Thus,

$$\mathbb{E}\hat{\eta}_S^2(\mu_n+z,\lambda_*)=\left[2(\lambda_*-\mu_n)^{-3}+2(\lambda_*+\mu_n)^{-3}e^{-2\lambda_*\mu_n}+O\left((\lambda_*-\mu_n)^{-5}\right)\right]\phi(\lambda_*-\mu_n). \quad (2.32)$$

Plugging (2.30)-(2.32) into (2.19), we have

$$\begin{aligned}F(\lambda_*)&=2(1-\epsilon_n)\phi(\lambda_*)\left[2\lambda_*^{-3}-12\lambda_*^{-5}+O(\lambda_*^{-7})\right]+\epsilon_n\mu_n^2\\&\quad-2\epsilon_n\mu_n\left[(\lambda_*-\mu_n)^{-2}-(\lambda_*+\mu_n)^{-2}e^{-2\lambda_*\mu_n}+O\left((\lambda_*-\mu_n)^{-4}\right)\right]\phi(\lambda_*-\mu_n)\\&\quad+\epsilon_n\left[2(\lambda_*-\mu_n)^{-3}+2(\lambda_*+\mu_n)^{-3}e^{-2\lambda_*\mu_n}+O\left((\lambda_*-\mu_n)^{-5}\right)\right]\phi(\lambda_*-\mu_n)\\&=\epsilon_n\mu_n^2+2(1-\epsilon_n)\phi(\lambda_*)\left[2\lambda_*^{-3}-12\lambda_*^{-5}+O(\lambda_*^{-7})\right]+\frac{2\epsilon_n\mathcal{A}\phi(\lambda_*-\mu_n)}{(\lambda_*-\mu_n)^3}. \quad (2.33)\end{aligned}$$

Next, we utilize the derivative equation (2.23) to further simplify (2.33). We first list the asymptotic approximations needed:

$$-\lambda_*\int_{\lambda_*}^{\infty}\phi(z)dz+\phi(\lambda_*)=(1-(3+o(1))\lambda_*^{-2})\cdot\lambda_*^{-2}\phi(\lambda_*),$$

$$-\phi(\lambda_*+\mu_n)+\lambda_*\int_{\lambda_*+\mu_n}^{\infty}\phi(z)dz=\frac{[-\mu_n\lambda_*^2-\lambda_*(1+2\mu_n^2(1+o(1)))]e^{-2\lambda_*\mu_n}}{(\lambda_*+\mu_n)^3}\phi(\lambda_*-\mu_n),$$
$$-\phi(\lambda_*-\mu_n)+\lambda_*\int_{\lambda_*-\mu_n}^{\infty}\phi(z)dz=\frac{\mu_n(\lambda_*-\mu_n)^2-\lambda_*[1-(3+o(1))\lambda_*^{-2}]}{(\lambda_*-\mu_n)^3}\phi(\lambda_*-\mu_n).$$

Plugging them into (2.23) yields

$$4(1-\epsilon_n)\Big[\frac{1}{\lambda_*^2}-\frac{3+o(1)}{\lambda_*^4}\Big]\phi(\lambda_*)=2\epsilon_n\frac{\mathcal{B}\phi(\lambda_*-\mu_n)}{(\lambda_*-\mu_n)^3}.$$

Obtaining the expression for $\frac{\phi(\lambda_*-\mu_n)}{(\lambda_*-\mu_n)^3}$ from the above equation and plugging it into (2.33) completes the proof. □

We now apply Lemmas 4 and 5 to obtain the final form of $F(\lambda_*)$. Referring to the expression of $F(\lambda_*)$ in Lemma 5, the key term to compute is $1+\Big(\lambda_*-\frac{3+o(1)}{\lambda_*}\Big)\frac{\mathcal{A}}{\mathcal{B}}$. Using the fact that $\lambda_*\mu_n\to\infty$, some direct calculations enable us to obtain

$$\Big(\lambda_*-\frac{3+o(1)}{\lambda_*}\Big)\mathcal{A}+\mathcal{B}=(-1+o(1))\lambda_*\mu_n^2,\quad \mathcal{B}=\mu_n\lambda_*^2(1+o(1)).$$

Therefore, the expression $F(\lambda_*)$ in Lemma 5 can be simplified to

$$\begin{aligned}F(\lambda_*)=&\epsilon_n\mu_n^2+\frac{4(1-\epsilon_n)\phi(\lambda_*)}{\lambda_*^3}\cdot\Big[-6\lambda_*^{-2}+O(\lambda_*^{-4})-\frac{\mu_n}{\lambda_*}(1+o(1))\Big]\\=&\epsilon_n\mu_n^2-\frac{(4+o(1))\mu_n\phi(\lambda_*)}{\lambda_*^4}.\end{aligned}$$

Finally, Lemma 4 implies that $\lambda_*=(1+o(1))\frac{\log\epsilon_n^{-1}}{\mu_n}$. Replacing λ_* by this rate in the above equation gives us the result in Proposition 1.

2.5.4 Proof of Proposition 2

Define the supremum risk of optimally tuned hard thresholding estimator as

$$R_H(\Theta(k_n,\tau_n),\sigma_n)=\inf_{\lambda>0}\sup_{\theta\in\Theta(k_n,\tau_n)}\mathbb{E}_\theta\big\|\hat\eta_H(y,\lambda)-\theta\big\|_2^2,$$

where $y_i = \theta_i + \sigma_n z_i$, with $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. It is straightforward to verify that

$$R_H(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R_H(\Theta(k_n, \mu_n), 1). \tag{2.34}$$

Without loss of generality, let $\sigma_n = 1$ in the model. We first obtain the lower bound, by calculating the risk at a specific value of $\underline{\theta}$ such that $\underline{\theta}_i = \mu_n$ for $i \in \{1, 2, \ldots, k_n\}$ and $\underline{\theta}_i = 0$ for $i > k_n$:

$$R_H(\Theta(k_n, \mu_n), 1) \geq \inf_{\lambda > 0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2. \tag{2.35}$$

Denote the one-dimensional risk:

$$r_H(\lambda, \mu) := \mathbb{E}\left(\hat{\eta}_H(\mu + z, \lambda) - \mu\right)^2, \quad z \sim \mathcal{N}(0,1), \quad \forall \mu \in \mathbb{R}, \lambda \geq 0.$$

It is then direct to confirm that

$$\mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 = n\Big[(1 - \epsilon_n) r_H(\lambda, 0) + \epsilon_n r_H(\lambda, \mu_n)\Big]. \tag{2.36}$$

Let λ_n^* be the optimal choice of λ in $\mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|^2$ so that

$$\inf_{\lambda > 0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 = \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda_n^*) - \underline{\theta}\|_2^2.$$

To evaluate the lower bound in (2.35), we consider two scenarios for the optimal choice λ_n^* and in each one we obtain a lower bound for $\mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda_n^*) - \underline{\theta}\|^2$. But before considering these two cases, we use the integration by part to find the following more explicit forms for $r_H(\lambda, 0)$ and $r_H(\lambda, \mu)$:

$$\begin{aligned}
r_H(\lambda, 0) &= 2 \int_{\lambda}^{\infty} z^2 \phi(z) dz = 2\lambda\phi(\lambda) + 2(1 - \Phi(\lambda)), \\
r_H(\lambda, \mu) &= \mu^2 \int_{-\lambda-\mu}^{\lambda-\mu} \phi(z) dz + \int_{-\infty}^{-\lambda-\mu} z^2 \phi(z) dz + \int_{\lambda-\mu}^{\infty} z^2 \phi(z) dz \\
&= (\mu^2 - 1)\Big[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)\Big] + 1 + (\lambda - \mu)\phi(\lambda - \mu) + (\lambda + \mu)\phi(\lambda + \mu),
\end{aligned} \tag{2.37}$$

where we recall that $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and CDF of $\mathcal{N}(0,1)$ respectively. Now we consider two cases for the optimal choice λ_n^* and in each case find a lower bound for the risk.

- **Case I** $\lambda_n^* = O(1)$**:** we have $\lambda_n^* \le c$ for some constant $c > 0$. Hence, from (2.36) we obtain

$$\begin{aligned}
&\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda) - \underline{\theta}\|_2^2 = \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda_n^*) - \underline{\theta}\|_2^2 \\
\ge\ & n(1-\epsilon_n) r_H(\lambda_n^*, 0) = n(1-\epsilon_n)\Big[2\lambda_n^*\phi(\lambda_n^*) + 2(1-\Phi(\lambda_n^*))\Big] \\
\ge\ & n(1-\epsilon_n)\Big[2(1-\Phi(\lambda_n^*))\Big] \ge n(1-\epsilon_n)\Big[2(1-\Phi(c))\Big] \ge n\epsilon_n\mu_n^2.
\end{aligned}$$

  The last inequality is because $\epsilon_n\mu_n^2 = o(1)$ and $(1-\epsilon_n)[2(1-\Phi(c))] = \Theta(1)$.

- **Case II** $\lambda_n^* = \omega(1)$**:** then $\lambda_n^* \to \infty$ as $n \to \infty$. From (2.36) and (2.37), we have

$$\begin{aligned}
&\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda) - \underline{\theta}\|_2^2 = \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda_n^*) - \underline{\theta}\|_2^2 \ge k_n r_H(\lambda_n^*, \mu_n) \\
=\ & k_n(\mu_n^2 - 1)\left[1 - \int_{\lambda_n^*-\mu_n}^{\infty} \phi(z)dz - \int_{\lambda_n^*+\mu_n}^{\infty} \phi(z)dz\right] + k_n \\
& + k_n(\lambda_n^* - \mu_n)\phi(\lambda_n^* - \mu_n) + k_n(\lambda_n^* + \mu_n)\phi(\lambda_n^* + \mu_n) \\
\overset{(a)}{=}\ & k_n\mu_n^2 + k_n(\lambda_n^* - \mu_n + o(1))\phi(\lambda_n^* - \mu_n) + k_n(\lambda_n^* + \mu_n + o(1))\phi(\lambda_n^* + \mu_n) \\
\ge\ & k_n\mu_n^2 = n\epsilon_n\mu_n^2,
\end{aligned}$$

  where to obtain (a), we have used the Gaussian tail bound in Lemma 1 under the scaling $\lambda_n^* \to \infty$ and $\mu_n \to 0$.

Note that since the two cases we have discussed above cover all the ranges of λ_n^*, we conclude that

$$R_H(\Theta(k_n,\mu_n), 1) \ge \inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda) - \underline{\theta}\|_2^2 \ge n\epsilon_n\mu_n^2,$$

for all sufficiently large n. To obtain the matching upper bound, we have

$$R_H(\Theta(k_n,\mu_n), 1)$$

$$
\begin{aligned}
&= \inf_{\lambda>0} \sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \left\|\hat{\eta}_H(y,\lambda) - \theta\right\|_2^2 \\
&\le \lim_{\lambda\to\infty} \sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \left\|\hat{\eta}_H(y,\lambda) - \theta\right\|_2^2 \\
&\le \lim_{\lambda\to\infty} \Big(\sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y,\lambda)\|_2^2 + \sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \langle -2\hat{\eta}_H(y,\lambda), \theta\rangle + \sup_{\theta\in\Theta(k_n,\mu_n)} \|\theta\|_2^2 \Big) \\
&\le n\epsilon_n\mu_n^2 + \lim_{\lambda\to\infty} \Big(\sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y,\lambda)\|_2^2 + 2\sqrt{n\epsilon_n\mu_n^2} \sqrt{\sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y,\lambda)\|_2^2} \Big). \qquad (2.38)
\end{aligned}
$$

To obtain the last inequality, we have used Cauchy–Schwarz inequality and $\sup_{\theta\in\Theta(k_n,\mu_n)} \|\theta\|_2^2 = k_n\mu_n^2$. From (2.38), to show $R_H(\Theta(k_n,\mu_n),1) \le n\epsilon_n\mu_n^2$, it is sufficient to prove

$$
\lim_{\lambda\to\infty} \sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y,\lambda)\|_2^2 = 0.
$$

Define $f_\lambda(\mu) := \mathbb{E}|\hat{\eta}_H(\mu+z,\lambda)|^2, z\sim\mathcal{N}(0,1)$. It is not hard to verify that $f_\lambda(\mu)$, as a function of μ, is symmetric around zero and increasing over $[0,\infty)$ for all $\lambda > 0$. As a result,

$$
\begin{aligned}
\lim_{\lambda\to\infty} \sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y,\lambda)\|_2^2 &\le \lim_{\lambda\to\infty} \left[(n-k_n) f_\lambda(0) + k_n f_\lambda(\sqrt{k_n}\mu_n)\right] \\
&= (n-k_n) \lim_{\lambda\to\infty} f_\lambda(0) + k_n \lim_{\lambda\to\infty} f_\lambda(\sqrt{k_n}\mu_n) \\
&= 0 + 0 = 0.
\end{aligned}
$$

The last line holds because $\lim_{\lambda\to\infty} f_\lambda(\mu) = 0, \forall\mu\in\mathbb{R}$ from dominated convergence theorem. The dominated convergence theorem can be used since $|\hat{\eta}_H(\mu+z,\lambda)|^2 \le |\mu+z|^2$ and $\lim_{\lambda\to\infty} |\hat{\eta}_H(\mu+z,\lambda)|^2 = 0$.

2.5.5 Proof of Theorem 6

Recall the scale invariance property in Section 3.4.1: $R(\Theta(k_n,\tau_n),\sigma_n) = \sigma_n^2 \cdot R(\Theta(k_n,\mu_n),1)$, where $\mu_n = \tau_n/\sigma_n$. Moreover, the estimator $\hat{\eta}_E(y,\lambda,\gamma) := \frac{1}{1+\gamma}\hat{\eta}_S(y,\lambda)$ defined in Equation (2.10) also preserves an invariance: $t\cdot\hat{\eta}_E(y,\lambda,\gamma) = \hat{\eta}_E(ty,t\lambda,\gamma), \forall t\ge 0$. Therefore, to prove both the

upper and lower bounds, in this section, it is sufficient to consider the simpler unit-variance model:

$$y_i = \theta_i + z_i, \quad i = 1, \ldots, n, \tag{2.39}$$

where $(z_i) \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$. We find an upper bound for the minimax risk by calculating the supremum risk of $\eta_E(y, \lambda, \gamma)$ with proper tuning. The lower bound is obtained by using Theorem 11 and considering the independent block prior again. Both steps are more challenging than the corresponding steps in the proof of Theorem 5.

Upper bound

To analyze the supremum risk of $\hat{\eta}_E(y, \lambda, \gamma)$, it is important to understand its risk in one dimension. Define the one-dimensional risk function as:

$$r_e(\mu; \lambda, \gamma) = \mathbb{E}\Big(\frac{1}{1+\gamma}\hat{\eta}_S(\mu + z, \lambda) - \mu\Big)^2, \quad z \sim \mathcal{N}(0, 1). \tag{2.40}$$

The following property of the risk function plays a pivotal role in our analysis.

Lemma 6. *For any given tuning parameters $\lambda > 0$, $\gamma \in [0, +\infty]$, it holds that*

(i) $r_e(\mu; \lambda, \gamma)$, as a function of μ, is symmetric, and increasing over $\mu \in [0, +\infty)$.

(ii) $\max_{(x,y):x^2+y^2=c^2} [r_e(x; \lambda, \gamma) + r_e(y; \lambda, \gamma)] = 2r_e(c/\sqrt{2}; \lambda, \gamma), \quad \forall c > 0.$

Proof. (i) Proving the symmetry of $r_e(\mu; \lambda, \gamma)$ is straightforward and is hence skipped. To prove the monotonicity of $r_e(\mu; \lambda, \gamma)$, we will calculate its derivative and show that it is positive for all $\mu > 0$. To this end, we first decompose $r_e(\mu; \lambda, \gamma)$ into three terms:

$$r_e(\mu; \lambda, \gamma) = \frac{1}{(1+\gamma)^2}\mathbb{E}(\hat{\eta}_S(\mu + z, \lambda) - \mu)^2 + \frac{\gamma^2\mu^2}{(1+\gamma)^2} + \frac{2\gamma\mu}{(1+\gamma)^2}\mathbb{E}(\mu - \hat{\eta}_S(\mu + z, \lambda)).$$

Accordingly, the derivative of $r_e(\mu;\lambda,\gamma)$ takes the form:

$$\begin{aligned}\frac{\partial r_e(\mu;\lambda,\gamma)}{\partial\mu} =& \frac{1}{(1+\gamma)^2}\frac{\partial\mathbb{E}(\hat{\eta}_S(\mu+z,\lambda)-\mu)^2}{\mu} + \frac{2\gamma^2\mu}{(1+\gamma)^2} \\ & -\frac{2\gamma}{(1+\gamma)^2}\left[\mu\frac{\partial\mathbb{E}(\hat{\eta}_S(\mu+z,\lambda)-\mu)}{\partial\mu} + \mathbb{E}(\hat{\eta}_S(\mu+z,\lambda)-\mu)\right]. \end{aligned} \tag{2.41}$$

Using the explicit expression $\hat{\eta}_S(\mu+z,\lambda) = \text{sign}(\mu+z)(|\mu+z|-\lambda)_+$, we can calculate

$$\begin{aligned}&\frac{\partial\mathbb{E}(\hat{\eta}_S(\mu+z,\lambda)-\mu)}{\partial\mu} = \frac{\partial}{\partial\mu}\mathbb{E}\Big[(-\mu)I_{(|\mu+z|\le\lambda)} + (z-\lambda)I_{(z+\mu>\lambda)} + (z+\lambda)I_{(z+\mu<-\lambda)}\Big] \\ =& -\mathbb{P}(|z+\mu|\le\lambda) - \mu[-\phi(\lambda-\mu)+\phi(-\lambda-\mu)] - \mu\phi(\lambda-\mu) + \mu\phi(-\lambda-\mu) \\ =& -\mathbb{P}(|z+\mu|\le\lambda),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial\mathbb{E}(\hat{\eta}_S(\mu+z,\lambda)-\mu)^2}{\mu} =& \frac{\partial}{\partial\mu}\mathbb{E}\left[\mu^2 I_{(|\mu+z|\le\lambda)} + (z-\lambda)^2 I_{(z+\mu>\lambda)}(z+\lambda)^2 I_{(z+\mu<-\lambda)}\right] \\ =& 2\mu\mathbb{P}(|\mu+z|\le\lambda) + \mu^2[-\phi(\lambda-\mu)+\phi(-\lambda-\mu)] + \mu^2\phi(\lambda-\mu) - \mu^2\phi(-\lambda-\mu) \\ =& 2\mu\mathbb{P}(|\mu+z|\le\lambda).\end{aligned}$$

Putting the above two results into (2.41), we obtain, $\forall\mu>0$,

$$\begin{aligned}\frac{\partial r_e(\mu;\lambda,\gamma)}{\partial\mu} =& \frac{2\mu}{(1+\gamma)^2}\mathbb{P}(|z+\mu|\le\lambda) + \frac{2\gamma^2\mu}{(1+\gamma)^2} \\ & + \frac{2\gamma}{(1+\gamma)^2}\left[\mu\mathbb{P}(|z+\mu|\le\lambda) + \mathbb{E}(\mu-\hat{\eta}_S(\mu+z,\lambda))\right] > 0,\end{aligned} \tag{2.42}$$

where the derivative is positive as all the terms on the right-hand side are non-negative and at least one of them is positive for all $\mu>0$. To verify this, all others are obvious and only the last term $\mathbb{E}(\mu-\hat{\eta}_S(\mu+z,\lambda))$ needs be checked: this term is positive because it is an odd function and has positive derivative.

(ii) Since the case where $\gamma=+\infty$ is trivial, we consider $\gamma\in[0,\infty)$ in the rest of the proof. Let

$H(x) := r_e(x;\lambda,\gamma) + r_e(\sqrt{c^2-x^2};\lambda,\gamma)$ and consider $\max_{0\le x\le c} H(x)$. Since $H(x)$ is continuous over $[0,c]$, we find the maximum by evaluating the derivative of $H(x)$ over $(0,c)$. Using the derivative calculation (2.42), we have

$$H^{'}(x) = r_e^{'}(x;\lambda,\gamma) - \frac{x}{\sqrt{c^2-x^2}} r_e^{'}(\sqrt{c^2-x^2};\lambda,\gamma) = \frac{2x}{1+\gamma} f_1(x) + \frac{2\gamma x}{(1+\gamma)^2} f_2(x),$$

where

$$\begin{aligned} f_1(x) &:= \mathbb{P}(|x+z| \le \lambda) - \mathbb{P}(|\sqrt{c^2-x^2}+z| \le \lambda), \\ f_2(x) &:= \frac{1}{\sqrt{c^2-x^2}} \mathbb{E}\hat{\eta}_S(\sqrt{c^2-x^2}+z,\lambda) - \frac{1}{x}\mathbb{E}\hat{\eta}_S(x+z,\lambda). \end{aligned}$$

We now show that $H^{'}(x) > 0$ for $x \in (0, \frac{c}{\sqrt{2}})$, $H^{'}(\frac{c}{\sqrt{2}}) = 0$, and $H^{'}(x) < 0$ for $x \in (\frac{c}{\sqrt{2}}, c)$. It is straightforward to confirm that $H^{'}(\frac{c}{\sqrt{2}}) = 0$. Hence, it is sufficient to show both $f_1(x)$ and $f_2(x)$ are positive over $(0, \frac{c}{\sqrt{2}})$ and negative over $(\frac{c}{\sqrt{2}}, c)$. This can be proved if we show that both $f_1(x)$ and $f_2(x)$ are strictly decreasing over $(0,c)$, given that $f_1(c/\sqrt{2}) = f_2(c/\sqrt{2}) = 0$.

Regarding $f_1(x)$, it is direct to verify that $\mathbb{P}(|x+z| \le \lambda)$ is strictly decreasing over $(0,c)$, and accordingly $\mathbb{P}(|\sqrt{c^2-x^2}+z| \le \lambda)$ is strictly increasing over $(0,c)$. Hence $f_1(x)$ is strictly decreasing over $(0,c)$. It remains to prove the monotonicity of $f_2(x)$. By the structure of $f_2(x)$, it is sufficient to show $\mathbb{E}[\frac{1}{x}\hat{\eta}_S(x+z,\lambda)]$ is a strictly increasing function of x for $x > 0$. We compute the derivative:

$$\begin{aligned} &\frac{\partial \mathbb{E}\frac{1}{x}\hat{\eta}_S(x+z,\lambda)}{\partial x} = -\frac{1}{x^2}\mathbb{E}\hat{\eta}_S(x+z,\lambda) + \frac{1}{x}\mathbb{P}(|x+z| > \lambda) \\ =& -\frac{1}{x^2}\left(\mathbb{E}\left[(x+z-\lambda)I_{(x+z>\lambda)} + (x+z+\lambda)I_{(x+z<-\lambda)}\right] - x\int_{\lambda-x}^{\infty}\phi(z)dz - x\int_{\lambda+x}^{\infty}\phi(z)dz\right) \\ =& -\frac{1}{x^2}\underbrace{\left[\phi(\lambda-x) - \lambda\int_{\lambda-x}^{\infty}\phi(z)dz + \lambda\int_{\lambda+x}^{\infty}\phi(z)dz - \phi(\lambda+x)\right]}_{h(x)}. \end{aligned}$$

Therefore, for $x > 0$, $\frac{\partial \mathbb{E}\frac{1}{x}\hat{\eta}_S(x+z,\lambda)}{\partial x} > 0$ if and only if $h(x) < 0$. In fact,

$$\begin{aligned} h^{'}(x) &= (\lambda - x)\phi(\lambda - x) + (\lambda + x)\phi(\lambda + x) - \lambda\phi(\lambda - x) - \lambda\phi(\lambda + x) \\ &= x(\phi(\lambda + x) - \phi(\lambda - x)) < 0, \quad \forall x > 0. \end{aligned}$$

Also, it is straightforward to confirm that $h(0) = 0$. Thus $h(x) < 0$ for $x > 0$. □

The one-dimensional risk function properties in Lemma 6 will enable us to locate the parameter value at which the supremum risk of $\hat{\eta}_E(y, \lambda, \gamma)$ over the parameter space $\Theta(k_n, \mu_n)$ is achieved. The following lemma provides the detailed supremum risk calculation for a carefully-picked choice of the tuning.

Lemma 7. *Consider model* (2.39). *Suppose* $\epsilon_n = k_n/n \to 0$, $\mu_n \to \infty$, *and* $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, *as* $n \to \infty$. *Then the estimator* $\hat{\eta}_E(y, \lambda_n, \gamma_n) = \frac{1}{1+\gamma_n}\hat{\eta}_S(y, \lambda_n)$, *with* $\gamma_n = (2\epsilon_n \mu_n^2 e^{\frac{3}{2}\mu_n^2})^{-1} - 1$ *and* $\lambda_n = 2\mu_n$, *has supremum risk:*

$$\sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 = k_n \mu_n^2 - (\sqrt{2/\pi} + o(1)) \cdot \frac{k_n^2}{n} \mu_n e^{\mu_n^2}.$$

Proof. Using the one-dimensional risk function in (2.40), we can write:

$$\sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 = \sup_{\theta \in \Theta(k_n, \mu_n)} \sum_{i=1}^{n} r_e(\theta_i; \lambda_n, \gamma_n).$$

According to the properties proved in Lemma 6, it is clear that the above supremum is attained at the parameter vector θ in which there are k_n non-zero components and they are all equal to μ_n (it occurs at a particular boundary of the parameter space $\Theta(k_n, \mu_n)$). Therefore, the supremum risk can be simplified to

$$\begin{aligned} &\sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 = n\Big[(1 - \epsilon_n) r_e(0; \lambda_n, \gamma_n) + \epsilon_n r_e(\mu_n; \lambda_n, \gamma_n)\Big] \qquad (2.43) \\ =& n\left[\frac{1 - \epsilon_n}{(1 + \gamma_n)^2} \mathbb{E}\hat{\eta}_S^2(z, \lambda_n) + \frac{\epsilon_n}{(1 + \gamma_n)^2} \mathbb{E}\hat{\eta}_S^2(\mu_n + z, \lambda_n) - \frac{2\epsilon_n \mu_n}{1 + \gamma_n} \mathbb{E}\hat{\eta}_S(\mu_n + z, \lambda_n) + \epsilon_n \mu_n^2\right]. \end{aligned}$$

To further calculate the supremum risk, we evaluate the three expectations in the above expression, using the Gaussian tail bound $\int_t^\infty \phi(z)dz = \left(\frac{1}{t} - \frac{1}{t^3} + \frac{3+o(1)}{t^5}\right)\phi(t)$ as $t \to \infty$. For the particular choice $\lambda_n = 2\mu_n \to \infty$, we have

$$\mathbb{E}\hat{\eta}_S^2(z, \lambda_n) = 2\left[(1+\lambda_n^2)\int_{\lambda_n}^\infty \phi(z)dz - \lambda_n\phi(\lambda_n)\right] = \frac{1+o(1)}{2\mu_n^3}\phi(2\mu_n). \tag{2.44}$$

Furthermore,

$$\begin{aligned}\mathbb{E}\hat{\eta}_S^2(\mu_n + z, \lambda_n) =& \left[(1+(\mu_n-\lambda_n)^2)\int_{\lambda_n-\mu_n}^\infty \phi(z)dz - (\lambda_n-\mu_n)\phi(\lambda_n-\mu_n)\right] \\ &+ \left[(1+(\mu_n+\lambda_n)^2)\int_{\lambda_n+\mu_n}^\infty \phi(z)dz - (\lambda_n+\mu_n)\phi(\lambda_n+\mu_n)\right] \\ =& \frac{2+o(1)}{(\lambda_n-\mu_n)^3}\phi(\lambda_n-\mu_n) + \frac{2+o(1)}{(\lambda_n+\mu_n)^3}\phi(\lambda_n+\mu_n) = \frac{2+o(1)}{\mu_n^3}\phi(\mu_n),\end{aligned} \tag{2.45}$$

and

$$\begin{aligned}\mathbb{E}\hat{\eta}_S(\mu_n + z, \lambda_n) =& \phi(\lambda_n-\mu_n) - (\lambda_n-\mu_n)\int_{\lambda_n-\mu_n}^\infty \phi(z)dz - \phi(\lambda_n+\mu_n) + (\mu_n+\lambda_n)\int_{\lambda_n+\mu_n}^\infty \phi(z)dz \\ =& \frac{1+o(1)}{(\lambda_n-\mu_n)^2}\phi(\lambda_n-\mu_n) - \frac{1+o(1)}{(\lambda_n+\mu_n)^2}\phi(\lambda_n+\mu_n) = \frac{1+o(1)}{\mu_n^2}\phi(\mu_n).\end{aligned} \tag{2.46}$$

Plugging (2.44)-(2.46) into (2.43) with the particular choice $\gamma_n = (2\epsilon_n\mu_n^2 e^{\frac{3}{2}\mu_n^2})^{-1} - 1$ considered in the lemma, we obtain

$$\begin{aligned}&\sup_{\theta\in\Theta(k_n,\mu_n)} \mathbb{E}_\theta\|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 \\ &= k_n\mu_n^2 + (2+o(1))\cdot n\epsilon_n^2\mu_n e^{\frac{3}{2}\mu_n^2}\phi(\mu_n) \\ &+(8+o(1))\cdot n\epsilon_n^3\mu_n e^{3\mu_n^2}\phi(\mu_n) - (4+o(1))\cdot n\epsilon_n^2\mu_n e^{\frac{3}{2}\mu_n^2}\phi(\mu_n) \\ &= k_n\mu_n^2 - (2+o(1))\cdot n\epsilon_n^2\mu_n e^{\frac{3}{2}\mu_n^2}\phi(\mu_n).\end{aligned}$$

The last equation holds because $\mu_n = o(\sqrt{\log\epsilon_n^{-1}})$ implies $\epsilon_n e^{\frac{3}{2}\mu_n^2} = o(1)$, so that the third term on

the right-hand side of the first equation is negligible. □

Now we can combine the preceding results we proved to obtain an upper bound for the minimax risk: with $\gamma_n = (2\epsilon_n \mu_n^2 e^{\frac{3}{2}\mu_n^2})^{-1} - 1$ and $\lambda_n = 2\mu_n$, it holds that

$$
\begin{aligned}
&R(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R(\Theta(k_n, \mu_n), 1) \\
&\leq \sigma_n^2 \cdot \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 = \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \sigma_n \lambda_n, \gamma_n) - \theta\|_2^2 \\
&= \sigma_n^2 \left(k_n \mu_n^2 - (\sqrt{2/\pi} + o(1)) \cdot \frac{k_n^2}{n} \mu_n e^{\mu_n^2} \right) = n\sigma_n^2 \Big(\epsilon_n \mu_n^2 - (\sqrt{2/\pi} + o(1)) \epsilon_n^2 \mu_n e^{\mu_n^2} \Big).
\end{aligned}
$$

Lower bound

The derivation of the lower bound follows the same roadmap of proof for the lower bound in Theorem 5. It relies on the independent block prior constructed in Section 2.5.2. According to Equation (2.13), the key step is to calculate the Bayes risk $B(\pi_S^{\mu,m})$ of the symmetric spike prior $\mu_S^{\mu,m}$ for $(\mu \in (0, \mu_n])$, in the regime $m = n/k_n \to \infty, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$. It turns out that setting $\mu = \mu_n$ will lead to a sharp lower bound. We summarize the result in the next lemma.

Lemma 8. *As $m = n/k_n \to \infty, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, the Bayes risk $B(\pi_S^{\mu_n,m})$ satisfies*

$$
B(\pi_S^{\mu_n,m}) \geq \mu_n^2 \left[1 - \frac{e^{\mu_n^2}}{2m} (1 + o(1)) \right].
$$

Proof. The result is an analog of Lemma 3 in Regime (II). Adopt the same notation from the proof of Lemma 3: $p_m = \frac{e^{\mu y_1} - e^{-\mu y_1}}{\sum_{i=1}^m (e^{\mu y_i} + e^{-\mu y_i})}$. In light of Lemma 2, it is sufficient to show that

(i) $\mathbb{E}_{\mu_n e_1} (p_m - 1)^2 \geq 1 - \frac{1}{m} e^{\mu_n^2} (1 + o(1))$,

(ii) $(m-1)\mathbb{E}_{\mu_n e_2} p_m^2 \geq \frac{1}{2m} e^{\mu_n^2} (1 + o(1))$.

Regarding Part (i), we have

$$
\mathbb{E}_{\mu_n e_1} [p_m - 1]^2
$$

$$\geq 1 - 2 \cdot \mathbb{E}\left(\frac{e^{\mu_n(\mu_n+z_1)} - e^{-\mu_n(\mu_n+z_1)}}{\sum_{j\neq 1}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n+z_1)} + e^{-\mu_n(\mu_n+z_1)}} \right)$$
$$\geq 1 - 2 \cdot \mathbb{E}\left(\frac{e^{\mu_n(\mu_n+z_1)}}{\sum_{j\neq 1}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n+z_1)} + e^{-\mu_n(\mu_n+z_1)}} \right).$$

Thus, (i) will be proved by showing that

$$\mathbb{E}\left(\frac{e^{\mu_n(\mu_n+z_1)}}{\sum_{j\neq 1}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n+z_1)} + e^{-\mu_n(\mu_n+z_1)}} \right) \leq \frac{1}{2m} e^{\mu_n^2}(1 + o(1)).$$

The expectation on the left-hand side of the above can be splitted into a summation of two truncated expectations according to the following condition:

$$e^{\mu_n(\mu_n+z_1)} + e^{-\mu_n(\mu_n+z_1)} \geq e^{\mu_n z_1} + e^{-\mu_n z_1}$$
$$\Leftrightarrow (e^{\mu_n^2} - 1)\left(e^{\mu_n z_1} - e^{-\mu_n z_1 - \mu_n^2}\right) \geq 0 \Leftrightarrow \mu_n z_1 \geq -\mu_n z_1 - \mu_n^2 \Leftrightarrow z_1 \geq -\frac{1}{2}\mu_n.$$

In the first case,

$$\mathbb{E}\left(\frac{e^{\mu_n(\mu_n+z_1)} I_{(z_1 \geq -\frac{1}{2}\mu_n)}}{\sum_{j\neq 1}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n+z_1)} + e^{-\mu_n(\mu_n+z_1)}} \right)$$
$$\leq \mathbb{E}\left(\frac{e^{\mu_n(\mu_n+z_1)} I_{(z_1 \geq -\frac{1}{2}\mu_n)}}{\sum_{j=1}^{m}(e^{\mu_n z_j} + e^{-\mu_n z_j})} \right) \leq e^{\mu_n^2}\mathbb{E}\left(\frac{e^{\mu_n z_1}}{\sum_{j=1}^{m}(e^{\mu_n z_j} + e^{-\mu_n z_j})} \right)$$
$$= \frac{e^{\mu_n^2}}{2}\mathbb{E}\left(\frac{e^{\mu_n z_1} + e^{-\mu_n z_1}}{\sum_{j=1}^{m}(e^{\mu_n z_j} + e^{-\mu_n z_j})} \right) = \frac{e^{\mu_n^2}}{2m},$$

where in the last two equations we have used the symmetry and exchangeability of i.i.d. standard normal variables $\{z_i\}_{j=1}^{m}$. In the second case,

$$\mathbb{E}\left(\frac{e^{\mu_n(\mu_n+z_1)} I_{(z_1 \leq -\frac{1}{2}\mu_n)}}{\sum_{j\neq 1}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n+z_1)} + e^{-\mu_n(\mu_n+z_1)}} \right)$$
$$\leq e^{\mu_n^2}\mathbb{E}\left(\frac{e^{\mu_n z_1} I_{(z_1 \leq -\frac{1}{2}\mu_n)}}{\sum_{j=1}^{m} e^{\mu_n z_j}} \right) = \frac{e^{\mu_n^2}}{m}\mathbb{E}\left(\frac{\sum_{j=1}^{m} e^{\mu_n z_j} \mathbb{1}_{(z_j \leq -\frac{1}{2}\mu_n)}}{\sum_{j=1}^{m} e^{\mu_n z_j}} \right), \tag{2.47}$$

where the last equality is again due to exchangeability of $\{z_j\}_{j=1}^m$. Denoting

$$Y_n := \frac{1}{me^{\frac{1}{2}\mu_n^2}} \sum_{j=1}^m e^{\mu_n z_j}, \quad Z_n := \frac{1}{me^{\frac{1}{2}\mu_n^2}} \sum_{j=1}^m e^{\mu_n z_j} I_{(z_j \le -\frac{1}{2}\mu_n)},$$

then the last expectation in (2.47) can be written as $\mathbb{E}(Z_n/Y_n)$, and it remains to show $\mathbb{E}(Z_n/Y_n) = o(1)$. It is straightforward to check that $\mathbb{E}Y_n = 1$. Furthermore, since $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, it is direct to verify that $\mathrm{Var}(Y_n) \le \frac{m}{m^2 e^{\mu_n^2}} e^{2\mu_n^2} = o(1)$. Hence, $Y_n \to 1$ in probability. In addition,

$$\begin{aligned}\mathbb{E}(Z_n) =& \mathbb{E}\left(e^{\mu_n z_1} I_{z_1 \le -\frac{1}{2}\mu_n} \cdot e^{-\frac{1}{2}\mu_n^2}\right) = \int_{-\infty}^{-\frac{1}{2}\mu_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2} + \mu_n z - \frac{1}{2}\mu_n^2} dz \\ =& \int_{-\infty}^{-\frac{\mu_n}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\mu_n)^2} dz = \int_{-\infty}^{-\frac{3}{2}\mu_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = o(1).\end{aligned}$$

Thus, $Z_n \to 0$ in probability. As a result, $Z_n/Y_n \to 0$ in probability. Since $|Z_n/Y_n| \le 1$, dominated convergence theorem guarantees that $\mathbb{E}(Z_n/Y_n) \to 0$.

To prove Part (ii), it is equivalent to prove

$$\mathbb{E}\frac{(e^{\mu_n z_1} - e^{-\mu_n z_1})^2}{[\sum_{j\neq 2}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n + z_2)} + e^{-\mu_n(\mu_n+z_2)}]^2} \ge \frac{1}{2m^2} e^{\mu_n^2}(1 + o(1)).$$

Towards this goal, we have

$$\begin{aligned}&\mathbb{E}\frac{(e^{\mu_n z_1} - e^{-\mu_n z_1})^2}{[\sum_{j\neq 2}(e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n(\mu_n+z_2)} + e^{-\mu_n(\mu_n+z_2)}]^2} \\ \overset{\text{(a)}}{\ge}& \mathbb{E}\frac{(e^{\mu_n z_1} - e^{-\mu_n z_1})^2}{[2(m-2)e^{\frac{\mu_n^2}{2}} + e^{\frac{3}{2}\mu_n^2} + e^{-\frac{\mu_n^2}{2}} + e^{\mu_n z_1} + e^{-\mu_n z_1}]^2} \overset{\text{(b)}}{\ge} \mathbb{E}\frac{(e^{\mu_n z_1} - e^{-\mu_n z_1})^2 I_{(|z_1| \le 3\mu_n)}}{[2(m-2)e^{\frac{\mu_n^2}{2}} + 4\sqrt{m}e^{\frac{\mu_n^2}{2}}]^2} \\ \overset{\text{(c)}}{=}& \frac{2}{e^{\mu_n^2}(2m - 4 + 4\sqrt{m})^2} \cdot \left[\mathbb{E}e^{2\mu_n z_1} I_{|z_1| \le 3\mu_n} - \mathbb{P}(|z_1| \le 3\mu_n)\right] \\ =& \frac{2}{e^{\mu_n^2}(2m - 4 + 4\sqrt{m})^2} \cdot \left(e^{2\mu_n^2} \int_{-5\mu_n}^{\mu_n} \phi(z)dz - \int_{-3\mu_n}^{3\mu_n} \phi(z)dz\right) = \frac{1}{2m^2} e^{\mu_n^2}(1 + o(1)).\end{aligned}$$

Here, Inequality (a) is by applying the Jensen's inequality with respect to $z_2, \ldots, z_m$ (conditioned on z_1), as $1/(x+c)^2$ $(c > 0)$ is a convex function of $x > 0$. Inequality (b) holds because $e^{\frac{3}{2}\mu_n^2} +$

$e^{-\frac{\mu_n^2}{2}} + e^{\mu_n z_1} + e^{-\mu_n z_1} \le 4e^{3\mu_n^2}$ when $|z_1| \le 3\mu_n$, and $e^{3\mu_n^2} \le \sqrt{m}e^{\frac{\mu_n^2}{2}}$ (for large n) under the condition $\mu_n = o(\sqrt{\log m})$. Equality (c) is due to the symmetry of $z_1 \sim \mathcal{N}(0,1)$. □

Our goal now is to use Lemma 8 to finish the proof of the lower bound in Theorem 6:

$$\begin{aligned} R(\Theta(k_n, \tau_n), \sigma_n)) =& \sigma_n^2 \cdot R(\Theta(k_n, \mu_n), 1) \ge k_n \sigma_n^2 \cdot B(\pi_S^{\mu_n, m}) \\ \ge& k_n \sigma_n^2 \mu_n^2 \left[1 - \frac{e^{\mu_n^2}}{2m}(1 + o(1)) \right] = n\sigma_n^2 \left[\epsilon_n \mu_n^2 - \frac{1}{2}\epsilon_n^2 \mu_n^2 e^{\mu_n^2}(1 + o(1)) \right]. \end{aligned}$$

2.5.6 Proof of Theorem 7

Like in the proof of Theorems 5 and 6, we calculate the minimax risk by deriving matching upper and lower bounds. However, a notable difference of the proof of Theorem 7 is that the tight upper bound is obtained not by analyzing the supremum risk of a given estimator, but rather by a Bayesian approach. In this approach, we establish a uniform upper bound for the Bayes risk of an arbitrary distribution supported *on average* on the parameter space, and use the minimax theorem (i.e. Theorem 9) to connect the result to the matching upper bound of the minimax risk. We present the details of the upper and lower bounds in Sections 2.5.6 and 2.5.6, respectively.

Upper bound

Consider the univariate Gaussian model:

$$Y = \theta + Z, \tag{2.48}$$

where $\theta \in \mathbb{R}$ and $Z \sim \mathcal{N}(0,1)$. For a given constant $A > 1$, define a class of priors for θ:

$$\Gamma^A(\epsilon, \mu) := \left\{ \pi \in \mathcal{P}(\mathbb{R}) : \pi(\{0\}) \ge 1 - \epsilon,\ \mathbb{E}_\pi \theta^2 \le \epsilon \mu^2,\ \ \mathrm{supp}(\pi) \in [-A\mu, A\mu] \right\}, \tag{2.49}$$

where $\mathcal{P}(\mathbb{R})$ denotes the class of all probability measures defined on $\mathbb{R}$, and $\epsilon \in [0,1], \mu > 0$. Note that $\pi \in \Gamma^A(\epsilon, \mu)$ implies that $\pi = (1-\epsilon)\delta_0 + \epsilon G$, for some distribution G satisfying $\mathbb{E}_G \theta^2 \le \mu^2$ and

$\text{supp}(G) \subseteq [-A\mu, A\mu]$. The worst-case Bayes risk (i.e., the one of the least favorable distribution), under this univariate Gaussian model with squared error loss, is defined as

$$B^A(\epsilon, \mu, 1) := \sup \left\{ B(\pi) : \pi \in \Gamma^A(\epsilon, \mu) \right\}, \tag{2.50}$$

where

$$B(\pi) = \mathbb{E}(\mathbb{E}(\theta|Y) - \theta)^2, \;\; \theta \sim \pi, \; Y \mid \theta \sim \mathcal{N}(\theta, 1).$$

The following lemma allows us to obtain an upper bound for $R(\Theta^A(k_n, \tau_n), \sigma_n)$ in terms of $B^A(\epsilon, \mu, 1)$.

Lemma 9. *The minimax risk satisfies the following inequality:*

$$R(\Theta^A(k_n, \tau_n), \sigma_n) \leq n\sigma_n^2 \cdot B^A(\epsilon_n, \mu_n, 1).$$

Proof. The proof closely follows the arguments in the proof of Theorem 8.21 of [3]. However, since the parameter space we consider is different, we cover a full proof here for completeness. For notational simplicity, let $\Theta_n := \Theta^A(k_n, \tau_n)$. Consider the class of priors

$$\mathcal{M}_n := \mathcal{M}(k_n, \tau_n, A) = \left\{ \pi \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_\pi \|\theta\|_0 \leq k_n, \; \mathbb{E}_\pi \|\theta\|_2^2 \leq k_n \tau_n^2, \;\; \text{supp}(\pi) \subseteq [-A\tau_n, A\tau_n]^n \right\},$$

where $\mathcal{P}(\mathbb{R}^n)$ denotes the set of all probability measures on $\mathbb{R}^n$. Let $\mathcal{M}_n^e := \mathcal{M}^e(k_n, \tau_n, A) \subseteq \mathcal{M}(k_n, \tau_n, A)$ be its exchangeable subclass, consisting of the distributions $\pi \in \mathcal{M}_n$ that are permutation invariant over the n coordinates. Using notation $B(\pi, \mathcal{M}) := \sup_{\pi \in \mathcal{M}} B(\pi)$, we will show that

$$R(\Theta_n, \sigma_n) \leq B(\pi, \mathcal{M}_n) = B(\pi, \mathcal{M}_n^e) \leq n\sigma_n^2 \cdot B^A(\epsilon_n, \mu_n, 1). \tag{2.51}$$

We start with equality in (2.51).

$$R(\Theta_n, \sigma_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta_n} \mathbb{E} \|\hat{\theta} - \theta\|_2^2 \overset{(a)}{\leq} \inf_{\hat{\theta}} \sup_{\pi \in \mathcal{M}_n} \mathbb{E}_\pi \|\hat{\theta} - \theta\|_2^2 \overset{(b)}{=} \sup_{\pi \in \mathcal{M}_n} \inf_{\hat{\theta}} \mathbb{E}_\pi \|\hat{\theta} - \theta\|_2^2 = B(\pi, M_n).$$

Inequality (a) is due to the fact that $\mathcal{M}_n$ contains all point mass priors δ_θ, for every $\theta \in \Theta_n$. To obtain Equality (b) we have used the minimax theorem, i.e. Theorem 9, as $\mathcal{M}_n$ is a convex set of probability measures. To prove the second inequality in (2.51), note that for any $\pi \in \mathcal{M}_n$, we can construct a corresponding prior:

$$\pi^e = \frac{1}{n!} \sum_{\sigma:[n]\to[n]} \pi \circ \sigma,$$

where σ denotes a permutation of the coordinates of θ, and $\pi \circ \sigma$ is the distribution after permutation. In other words, π^e is the distribution averaged over all the permutations, thus $\pi^e \in \mathcal{M}_n^e$. Given that $B(\pi)$ is a concave function (it is the infimum of linear functions), we have $B(\pi, \mathcal{M}_n) \leq B(\pi, \mathcal{M}_n^e)$ which implies $B(\pi, \mathcal{M}_n) = B(\pi, \mathcal{M}_n^e)$ since $\mathcal{M}_n^e \subseteq \mathcal{M}_n$.

To show the last inequality in (2.51), for any exchangeable prior $\pi \in \mathcal{M}_n^e$, let π_1 be its univariate marginal distribution. Using the constraints on π from $\mathcal{M}_n$ and the fact that π is symmetric over its n coordinates, we have

$$\pi_1(\theta_1 = 0) \geq 1 - \epsilon_n,\ \mathbb{E}_{\pi_1}\theta_1^2 \leq \epsilon_n \tau_n^2,\ \ \operatorname{supp} \pi_1 \subseteq [-A\tau_n, A\tau_n]$$

Hence $\pi_1 \in \Gamma^A(\epsilon_n, \tau_n)$ defined in (2.49). Furthermore, according to Theorem 10, the product prior π_1^n is less favorable than π^e, namely, $B(\pi) \leq B(\pi_1^n) = nB(\pi_1)$. Rescaling the noise level to one and maximizing over $\pi_1 \in \Gamma^A(\epsilon_n, \mu_n)$ completes the proof. □

Lemma 9 reduces the problem of obtaining the upper bound for frequentist minimax risk (under Gaussian sequence model) to the problem of upper bounding the worst-case Bayes risk (under a univariate Gaussian model). Our next goal is to find an upper bound for $B^A(\epsilon_n, \mu_n, 1)$. Towards this end, we first state a useful lemma.

Lemma 10. *Under model* (2.48)*, consider prior* $\pi = (1-\epsilon)\delta_0 + \epsilon G \in \Gamma^A(\epsilon, \mu)$*, as defined in* (2.49)*. Then,*

$$\mathbb{E}(\mathbb{E}(\theta|Y))^2 = \int \frac{\epsilon^2 (\int t e^{tz - \frac{t^2}{2}} dG(t))^2}{1 - \epsilon + \epsilon \int e^{tz - \frac{t^2}{2}} dG(t)} \phi(z) dz,$$

where $\phi(\cdot)$ *denotes the density function of standard normal random variable.*

Proof. Given the prior $\pi = (1-\epsilon)\delta_0 + \epsilon G$, the posterior mean of θ is given by

$$\mathbb{E}(\theta|Y=y) = \frac{\epsilon \int \theta\phi(y-\theta)dG(\theta)}{(1-\epsilon)\phi(y) + \epsilon\int \phi(y-\theta)dG(\theta)}.$$

Thus,

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(\theta|Y))^2 =& (1-\epsilon)\int \left[\frac{\epsilon\int t\phi(z-t)dG(t)}{(1-\epsilon)\phi(z)+\epsilon\int\phi(z-t)dG(t)}\right]^2 \phi(z)dz \\
&+ \epsilon \iint \left[\frac{\epsilon\int t\phi(\theta+\tilde{z}-t)dG(t)}{(1-\epsilon)\phi(\theta+\tilde{z})+\epsilon\int\phi(\theta+\tilde{z}-t)dG(t)}\right]^2 \phi(\tilde{z})d\tilde{z}dG(\theta) \\
=& \int \left[\frac{\epsilon\int t\phi(z-t)dG(t)}{(1-\epsilon)\phi(z)+\epsilon\int\phi(z-t)dG(t)}\right]^2 \cdot \left[(1-\epsilon)\phi(z) + \epsilon\int \phi(z-\theta)dG(\theta)\right]dz \\
=& \int \left[\frac{\epsilon\int te^{tz-\frac{t^2}{2}}dG(t)}{(1-\epsilon)+\epsilon\int e^{tz-\frac{t^2}{2}}dG(t)}\right]^2 \cdot \left[(1-\epsilon) + \epsilon\int e^{tz-\frac{t^2}{2}}dG(t)\right]\phi(z)dz \\
=& \int \frac{\epsilon^2(\int te^{tz-\frac{t^2}{2}}dG(t))^2}{1-\epsilon+\epsilon\int e^{tz-\frac{t^2}{2}}dG(t)}\phi(z)dz,
\end{aligned}$$

where the second equality is by a simple change of variable. □

We can now obtain a sharp upper bound for $B^A(\epsilon_n, \mu_n, 1)$.

Lemma 11. *Consider* $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$. *Under model* (2.48)*, the worst-case Bayes risk* $B^A(\epsilon_n, \mu_n, 1)$ *defined in* (2.50) *satisfies that for any* $A > 1$*,*

$$B^A(\epsilon_n, \mu_n, 1) \le \epsilon_n\mu_n^2 - \frac{1+o(1)}{2}\epsilon_n^2\mu_n^2 e^{\mu_n}.$$

Proof. For prior $\pi \in \Gamma^A(\epsilon, \mu)$, using the law of total expectation,

$$\mathbb{E}(\mathbb{E}(\theta|Y) - \theta)^2 = \mathbb{E}\theta^2 - \mathbb{E}(\mathbb{E}(\theta|Y))^2. \tag{2.52}$$

We first obtain a lower bound for the term $\mathbb{E}(\mathbb{E}(\theta|Y))^2$. We start with the expression derived in

Lemma 10 and develop a series of lower bounds,

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(\theta|Y))^2 &= \int \frac{\epsilon^2(\int te^{tz-\frac{t^2}{2}}dG(t))^2}{1-\epsilon+\epsilon\int e^{tz-\frac{t^2}{2}}dG(t)}\phi(z)dz \geq \int_{|z|\leq\sqrt{\log 1/\epsilon}} \frac{\epsilon^2(\int te^{tz-\frac{t^2}{2}}dG(t))^2}{1-\epsilon+\epsilon\int e^{tz-\frac{t^2}{2}}dG(t)}\phi(z)dz \\
&\overset{\text{(a)}}{\geq} \frac{\epsilon^2}{1-\epsilon+\epsilon^{\frac{1}{2}}}\int_{|z|\leq\sqrt{\log 1/\epsilon}}\left(\int te^{tz-\frac{t^2}{2}}dG(t)\right)^2\phi(z)dz \\
&\overset{\text{(b)}}{=} \frac{\epsilon^2}{1-\epsilon+\epsilon^{\frac{1}{2}}}\iint\left[tt'e^{tt'}\int_{-\sqrt{\log 1/\epsilon}-(t+t')}^{\sqrt{\log 1/\epsilon}-(t+t')}\phi(z)dz\right]dG(t)dG(t') \\
&= \frac{\epsilon^2}{1-\epsilon+\epsilon^{\frac{1}{2}}}\iint_{tt'\geq 0}\left[tt'e^{tt'}\int_{-\sqrt{\log 1/\epsilon}-(t+t')}^{\sqrt{\log 1/\epsilon}-(t+t')}\phi(z)dz\right]dG(t)dG(t') \\
&\quad + \frac{\epsilon^2}{1-\epsilon+\epsilon^{\frac{1}{2}}}\iint_{tt'<0}\left[tt'e^{tt'}\int_{-\sqrt{\log 1/\epsilon}-(t+t')}^{\sqrt{\log 1/\epsilon}-(t+t')}\phi(z)dz\right]dG(t)dG(t') \\
&\overset{\text{(c)}}{\geq} \frac{\epsilon^2}{1-\epsilon+\epsilon^{\frac{1}{2}}}\left(\iint_{tt'\geq 0}\left[tt'e^{tt'}\int_{-\sqrt{\log 1/\epsilon}-(t+t')}^{\sqrt{\log 1/\epsilon}-(t+t')}\phi(z)dz\right]dG(t)dG(t') - |A\mu|^2\right) \\
&\overset{\text{(d)}}{\geq} \frac{\epsilon^2}{1-\epsilon+\epsilon^{\frac{1}{2}}}\left(\int_{-\sqrt{\log 1/\epsilon}-2A\mu}^{\sqrt{\log 1/\epsilon}-2A\mu}\phi(z)dz\cdot\iint_{tt'\geq 0}tt'e^{tt'}dG(t)dG(t') - |A\mu|^2\right). \qquad (2.53)
\end{aligned}$$

Inequality (a) holds because for $|z| \leq \sqrt{\log 1/\epsilon}$,

$$\epsilon\int e^{tz-\frac{t^2}{2}}dG(t) = \epsilon e^{\frac{1}{2}z^2}\int e^{-\frac{1}{2}(z-t)^2}dG(t) \leq \epsilon e^{\frac{1}{2}z^2} \leq \epsilon^{\frac{1}{2}}.$$

To obtain Equality (b) we do the following simple calculations:

$$\begin{aligned}
&\int_{|z|\leq\sqrt{\log 1/\epsilon}}\left(\int te^{tz-\frac{t^2}{2}}dG(t)\right)^2\phi(z)dz \\
&= \int_{|z|\leq\sqrt{\log 1/\epsilon}}\left[\iint tt'e^{zt-t^2/2}e^{zt'-t'^2/2}dG(t)dG(t')\right]\phi(z)dz \\
&= \iint\left[tt'e^{tt'}\int_{|z|\leq\sqrt{\log 1/\epsilon}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-(t+t'))^2}dz\right]dG(t)dG(t') \\
&= \iint\left[tt'e^{tt'}\int_{-\sqrt{\log 1/\epsilon}-(t+t')}^{\sqrt{\log 1/\epsilon}-(t+t')}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz\right]dG(t)dG(t')
\end{aligned}$$

Inequality (c) holds because $e^{-|tt'|} \leq 1$ and $\operatorname{supp} G \subseteq [-A\mu, A\mu]$. Inequality (d) is due to

the fact that $\operatorname{supp} G \subseteq [-A\mu, A\mu]$ and $\int_{-\sqrt{\log 1/\epsilon}-a}^{\sqrt{\log 1/\epsilon}-a} \phi(z)dz$ (as a function of a) is symmetric and decreasing over $[0,\infty)$. To continue from (2.53), we further lower bound $\iint_{tt'\geq 0} tt'e^{tt'}dG(t)dG(t')$. To simplify notation, define two random variables $t, t' \overset{i.i.d.}{\sim} G$. We have

$$
\begin{aligned}
&\iint_{tt'\geq 0} tt'e^{tt'}dG(t)dG(t') = \mathbb{E}[tt'e^{tt'}I_{(tt'\geq 0)}] \\
=&\sum_{k=0}^{\infty} \mathbb{E}\frac{1}{k!}(tt')^{k+1}(I_{(t>0,t'>0)} + I_{(t<0,t'<0)}) \\
=&\sum_{k=0}^{\infty} \frac{1}{k!}\left(\mathbb{E}[t^{k+1}I_{(t>0)}]\cdot\mathbb{E}[(t')^{k+1}I_{(t'>0)}] + \mathbb{E}[t^{k+1}I_{(t<0)}]\cdot\mathbb{E}[(t')^{k+1}I_{(t'<0)}]\right) \\
=&\sum_{k=0}^{\infty} \frac{1}{k!}\left((\mathbb{E}t^{k+1}I_{(t>0)})^2 + (\mathbb{E}t^{k+1}I_{(t<0)})^2\right) \\
=&\sum_{k=0}^{\infty} \frac{1}{k!}\left((\mathbb{E}|t|^{k+1}I_{(t>0)})^2 + (\mathbb{E}|t|^{k+1}I_{(t<0)})^2\right) \\
\overset{(a)}{\geq}&\sum_{k=0}^{\infty} \frac{1}{k!}\frac{1}{2}\left(\mathbb{E}|t|^{k+1}I_{(t>0)} + \mathbb{E}|t|^{k+1}I_{(t<0)}\right)^2 \\
=&\frac{1}{2}\sum_{k=0}^{\infty} \frac{1}{k!}\left(\mathbb{E}|t|^{k+1}\right)^2 \overset{(b)}{\geq} \frac{1}{2}(\mathbb{E}|t|)^2 + \frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\mathbb{E}|t|^2\right)^{k+1} \geq \frac{1}{2}\left(\mathbb{E}|t|^2e^{\mathbb{E}|t|^2} - \mathbb{E}|t|^2\right),
\end{aligned}
$$

where (a) is due to the basic inequality $2(x^2+y^2) \geq (x+y)^2$, and (b) is by Hölder's inequality $(\mathbb{E}|t|^2)^{k+1} \leq (\mathbb{E}|t|^{k+1})^2, k \geq 1$. Combining the above inequality with (2.52) and (2.53) gives

$$
\begin{aligned}
B^A(\epsilon_n, \mu_n, 1) =& \sup_{\pi\in\Gamma^A(\epsilon_n,\mu_n)} \mathbb{E}(\mathbb{E}(\theta|Y) - \theta)^2 \\
\leq& \sup_{\mathbb{E}|t|^2\leq\mu_n^2} \left(\epsilon_n + \frac{\epsilon_n^2\Delta_n}{2(1-\epsilon_n+\sqrt{\epsilon_n})}\right)\mathbb{E}|t|^2 - \frac{\epsilon_n^2\Delta_n}{2(1-\epsilon_n+\sqrt{\epsilon_n})}\mathbb{E}|t|^2e^{\mathbb{E}|t|^2} + \frac{\epsilon^2A^2\mu_n^2}{1-\epsilon+\sqrt{\epsilon}},
\end{aligned} \tag{2.54}
$$

where $\Delta_n = \int_{-\sqrt{\log 1/\epsilon_n}-2\mu_nA}^{\sqrt{\log 1/\epsilon_n}-2\mu_nA} \phi(z)dz$. The results we obtained so far are non-asymptotic. We now make use of the conditions $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ to derive the final asymptotic result. Under such scaling conditions, it is straightforward to confirm that the expression on the right-hand side of (2.54) is increasing in $\mathbb{E}|t|^2$ when n is sufficiently large (by calculating its deriva-

tive). As a result,

$$
\begin{aligned}
B^A(\epsilon_n, \mu_n, 1) &\le \Big(\epsilon_n + \frac{\epsilon_n^2 \Delta_n}{2(1-\epsilon_n+\sqrt{\epsilon_n})}\Big)\mu_n^2 - \frac{\epsilon_n^2 \Delta_n}{2(1-\epsilon_n+\sqrt{\epsilon_n})}\mu_n^2 e^{\mu_n^2} + \frac{\epsilon^2 A^2 \mu_n^2}{1-\epsilon+\sqrt{\epsilon}} \\
&= \epsilon_n \mu_n^2 + \frac{1+o(1)}{2}\epsilon_n^2 \mu_n^2 - \frac{1+o(1)}{2}\epsilon_n^2 \mu_n^2 e^{\mu_n^2} + O(\epsilon_n^2 \mu_n^2) \\
&= \epsilon_n \mu_n^2 - \frac{1}{2}\epsilon_n^2 \mu_n^2 e^{\mu_n^2}(1+o(1)).
\end{aligned}
$$

□

Combing Lemmas 9 and 11 provides the upper bound for the minimax risk:

$$
R(\Theta^A(k_n, \tau_n), \sigma_n) \le n\sigma_n^2 \Big(\epsilon_n \mu_n^2 - \frac{1}{2}\epsilon_n^2 \mu_n^2 e^{\mu_n^2}(1+o(1))\Big).
$$

Lower bound

Recall that in the lower bound derivation for Theorem 6, in Section 2.5.5, the proof is based on the independent block prior π^{IB} with single spike distribution $\pi_S^{\mu_n, m}$ which is first introduced in Section 2.5.2. Since the spike locations are at $\pm\mu_n$, which are contained in $[-A\mu_n, A\mu_n]$ for any $A > 1$, this implies that $\operatorname{supp} \pi^{IB} \subseteq \Theta^A(k_n, \mu_n)$ as well. As a result, the proof in Section 2.5.5 also works for the new parameter space $\Theta^A(k_n, \mu_n)$ and it yields the same lower bound:

$$
R(\Theta^A(k_n, \tau_n), \sigma_n) \ge n\sigma_n^2 \Big(\epsilon_n \mu_n^2 - \frac{1}{2}\epsilon_n^2 \mu_n^2 e^{\mu_n^2}(1+o(1))\Big).
$$

2.5.7 Proof of Proposition 3

Comparing the results in Propositions 1 and 3, we can see that the supremum risk of optimally tuned soft thresholding has the same second-order asymptotic approximation in Regimes (I) and (II). Thus, the proof of Proposition 3 shares a lot of similarity with that of Proposition 1. For simplicity we will not repeat every detail. Referring to the proof of Proposition 1 in Section 2.5.3, the key is to obtain the accurate order of the optimal tuning λ_* and evaluate the function value

$F(\lambda_*)$, where we recall the definitions: $\lambda_* = \arg\min_{\lambda \geq 0} F(\lambda)$, $z \sim \mathcal{N}(0,1)$ and

$$F(\lambda) = (1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z,\lambda) + \epsilon_n \mathbb{E}(\hat{\eta}_S(\mu_n + z, \lambda) - \mu_n)^2.$$

We first address the order of λ_*.

Lemma 12. *Consider $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = o\left(\sqrt{\log \epsilon_n^{-1}}\right)$, as $n \to \infty$. It holds that*

$$\log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2} - 2\log\log\frac{2}{\epsilon_n} < \lambda_*\mu_n < \log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2}, \tag{2.55}$$

for sufficiently large n.

Proof. This lemma is an analog of Lemma 4 (comparing Equation (2.18) with (2.55)). The proof is thus similar too. We will skip equivalent calculations and only highlight the differences.

First, we show that $\lambda_*\mu_n^{-1} \to \infty$. Otherwise, $\lambda_*\mu_n^{-1} \leq C$ for some constant $C > 0$ (take a subsequence if necessary). Then when n is large,

$$\begin{aligned} F(\lambda_*) &\geq (1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z,\lambda_*) \geq (1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z, C\mu_n) \\ &= 2(1-\epsilon_n)\left[(1+(C\mu_n)^2)\int_{C\mu_n}^{\infty}\phi(z)dz - C\mu_n\phi(C\mu_n)\right] \\ &\overset{(a)}{=} \frac{4+o(1)}{\mu_n^3}\phi(C\mu_n) \overset{(b)}{>} \epsilon_n\mu_n^2 = F(+\infty), \end{aligned}$$

where (a) is by the Gaussian tail bound, and (b) is due to $\mu_n = o\left(\sqrt{\log \epsilon_n^{-1}}\right)$. The result $F(\lambda_*) > F(+\infty)$ contradicts with the optimality of λ_*.

Second, we utilize the derivative equation $F'(\lambda_*) = 0$ in Equation (2.23) to obtain more accurate order information of λ_*. The results $\mu_n \to \infty, \lambda_*\mu_n^{-1} \to \infty$ imply that $\lambda_* \to \infty, \lambda_* - \mu_n \to \infty, \lambda_*\mu_n \to \infty$. This is all needed to obtain Equation (2.24) and Equations (2.27)-(2.28). As a result, Equation (2.29) holds here as well:

$$2 + o(1) = \epsilon_n\mu_n\lambda_*\exp(\lambda_*\mu_n - \mu_n^2/2). \tag{2.56}$$

To reach (2.55) under the scaling $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$, the rest of the argument is exactly the same as the one in the proof of Lemma 4 . □

The next lemma characterizes $F(\lambda_*)$.

Lemma 13. *Consider $\epsilon_n \to 0, \mu_n \to \infty, \mu_n = (\sqrt{\log \epsilon_n^{-1}})$, as $n \to \infty$. It holds that*

$$F(\lambda_*) = \epsilon_n \mu_n^2 - \exp\left[-\frac{1}{2}\frac{1}{\mu_n^2}\left(\log\frac{1}{\epsilon_n}\right)^2\left(1+o(1)\right)\right].$$

Proof. This proof deviates a bit from the one of Lemma 5. We will more directly utilize the order information of λ_* proved in Lemma 12 to calculate $F(\lambda_*)$. Before that, we need a refinement of (2.56). This is achieved by refining Equation (2.24) and Equations (2.27)-(2.28) with higher-order approximations:

$$\begin{aligned}
-\lambda_* \int_{\lambda_*}^{\infty} \phi(z)dz + \phi(\lambda_*) &= \frac{1+O(\lambda_*^{-2})}{\lambda_*^2}\phi(\lambda_*),\\
-\phi(\lambda_*-\mu_n) + \lambda_* \int_{\lambda_*-\mu_n}^{\infty} \phi(z)dz &= \left[\frac{\mu_n}{\lambda_*-\mu_n} - \frac{\lambda_* + O(\lambda_*^{-1})}{(\lambda_*-\mu_n)^3}\right]\phi(\lambda_*-\mu_n),\\
-\phi(\lambda_*+\mu_n) + \lambda_* \int_{\lambda_*+\mu_n}^{\infty} \phi(z)dz &= o\left(\frac{1}{\lambda_*^4}\right)\phi(\lambda_*-\mu_n).
\end{aligned}$$

Plugging the above into Equation (2.23) and arranging terms gives

$$\begin{aligned}
e^{\lambda_*\mu_n - \frac{\mu_n^2}{2}}\frac{\epsilon_n\mu_n\lambda_*^2}{2(\lambda_*-\mu_n)} - 1 =& \frac{(1-\epsilon_n)(1+O(\lambda_*^{-2}))\mu_n}{\mu_n - (\lambda_*-\mu_n)^{-2}(\lambda_* + O(\lambda_*^{-1}))} - 1\\
=& \frac{\lambda_*(\lambda_*-\mu_n)^{-2} + O(\lambda_*^{-2}\mu_n)}{\mu_n - (\lambda_*-\mu_n)^{-2}(\lambda_* + O(\lambda_*^{-1}))} = \frac{1+o(1)}{\lambda_*\mu_n},
\end{aligned} \tag{2.57}$$

where in the second equality we have used $\epsilon_n\lambda_*^2 = o(1)$ and $\lambda_*^{-1}\mu_n = o(1)$ which are implied by the order of λ_* from Lemma 12.

Now we are ready to evaluate $F(\lambda_*)$. We first use Gaussian tail bound to approximate the three

expectations (i.e. Equations (2.20)-(2.22)) in the expression of $F(\lambda_*)$ (i.e. Equation (2.19)):

$$\begin{aligned}
\mathbb{E}\hat{\eta}_S^2(z,\lambda_*) &= \frac{4+O(\lambda_*^{-2})}{\lambda_*^3}\phi(\lambda_*),\\
\mathbb{E}\hat{\eta}_S(\mu_n+z,\lambda_*) &= \frac{1+O(\lambda_*^{-2})}{(\lambda_*-\mu_n)^2}\phi(\lambda_*-\mu_n),\\
\mathbb{E}\hat{\eta}_S^2(\mu_n+z,\lambda_*) &= \frac{2+O(\lambda_*^{-2})}{(\lambda_*-\mu_n)^3}\phi(\lambda_*-\mu_n).
\end{aligned}$$

Using these three approximations in Equation (2.19), we obtain

$$\begin{aligned}
F(\lambda_*) &= (1-\epsilon_n)\frac{4+O(\lambda_*^{-2})}{\lambda_*^3}\phi(\lambda_*) + \epsilon_n\mu_n^2 - 2\epsilon_n\mu_n\frac{1+O(\lambda_*^{-2})}{(\lambda_*-\mu_n)^2}\phi(\lambda_*-\mu_n) + \epsilon_n\frac{2+O(\lambda_*^{-2})}{(\lambda_*-\mu_n)^3}\phi(\lambda_*-\mu_n)\\
&= \epsilon_n\mu_n^2 - \phi(\lambda_*)\left[\frac{-4+O(\epsilon_n+\lambda_*^{-2})}{\lambda_*^3} + \frac{2\epsilon_n\mu_n}{(\lambda_*-\mu_n)^2}\cdot e^{\lambda_*\mu_n-\frac{\mu_n^2}{2}}\left(1+O\left(\frac{1}{\lambda_*\mu_n}\right)\right)\right].
\end{aligned}$$

We further replace $e^{\lambda_*\mu_n-\frac{\mu_n^2}{2}}$ in the above with the result from (2.57) to have

$$\begin{aligned}
F(\lambda_*) &= \epsilon_n\mu_n^2 - \phi(\lambda_*)\cdot\left[\frac{-4+O(\epsilon_n+\lambda_*^{-2})}{\lambda_*^3} + \frac{4}{\lambda_*^2(\lambda_*-\mu_n)}\left(1+O\left(\frac{1}{\lambda_*\mu_n}\right)\right)\right]\\
&\overset{(a)}{=} \epsilon_n\mu_n^2 - \phi(\lambda_*)\frac{4\mu_n}{\lambda_*^3(\lambda_*-\mu_n)}\left(1+O\left(\frac{1}{\mu_n^2}\right)\right) = \epsilon_n\mu_n^2 - \frac{4+o(1)}{\sqrt{2\pi}}e^{-\frac{\lambda_*^2}{2}}\cdot\frac{\mu_n}{\lambda_*^4}\\
&\overset{(b)}{=} \epsilon_n\mu_n^2 - \exp\left[-\frac{1}{2}\frac{1}{\mu_n^2}\left(\log\frac{1}{\epsilon_n}\right)^2\left(1+o(1)\right)\right].
\end{aligned}$$

Here, to obtain (a) we have used $\epsilon_n\lambda_*^2 = o(1)$ and $\lambda_*^{-1}\mu_n = o(1)$ implied by Lemma 12; (b) is due to the order $\lambda_* = \mu_n^{-1}\log\epsilon_n^{-1}(1+o(1))$ again from Lemma 12. □

Lemma 13 readily leads to the supremum risk of optimally tuned soft thresholding:

$$\inf_{\lambda}\sup_{\theta\in\Theta(k_n,\tau_n)}\mathbb{E}_\theta\|\hat{\eta}_S(y,\lambda)-\theta\|_2^2 = n\sigma_n^2F(\lambda_*) = n\sigma_n^2\left(\epsilon_n\mu_n^2 - \exp\left[-\frac{1}{2}\frac{1}{\mu_n^2}\left(\log\frac{1}{\epsilon_n}\right)^2\left(1+o(1)\right)\right]\right).$$

2.5.8 Proof of Proposition 4

The proof of this proposition is similar to the proof of Proposition 2 presented in Section 2.5.4. Hence, for the sake of brevity we adopt the same notation from Section 2.5.4 and only discuss the differences. If $R_H(\Theta(k_n, \tau_n), \sigma_n)$ denotes the supremum risk of optimally tuned hard thresholding estimator, then we will have

$$R_H(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R_H(\Theta(k_n, \mu_n), 1). \tag{2.58}$$

Without loss of generality, let $\sigma_n = 1$ in the model. As in the proof of Proposition 2, we obtain a lower bound by calculating the risk at the following specific value of $\underline{\theta}$ such that $\underline{\theta}_i = \mu_n$ for $i \in \{1, 2, \ldots, k_n\}$ and $\underline{\theta}_i = 0$ for $i > k_n$. We have

$$\mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 = n\Big[(1 - \epsilon_n) r_H(\lambda, 0) + \epsilon_n r_H(\lambda, \mu_n)\Big]. \tag{2.59}$$

To evaluate $\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|^2$, we consider three scenarios for the optimal choice of λ_n, denoted by λ_n^*.

- **Case I** $\lambda_n^* = O(1)$**:** In this case, $\lambda_n^* \leq c$ for some constant $c > 0$. Using the same argument as the one presented for Case I in the proof of Proposition 2, we have

$$\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 \geq 2n(1 - \epsilon_n)(1 - \Phi(c)).$$

  Since $\epsilon_n \mu_n^2 \to 0$ and $(1-\epsilon_n)2(1-\Phi(c)) = \Theta(1)$, we conclude that $\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 = \omega(n\epsilon_n \mu_n^2)$.

- **Case II** $\lambda_n^* = \omega(1)$ and $\lambda_n^* = O(\mu_n)$: Let c_1 be a fixed number larger than 1. There exists c_2 such that for large enough n, $c_1 < \lambda_n^* \leq c_2 \mu_n$. We thus obtain

$$\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 = \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda_n^*) - \underline{\theta}\|_2^2$$

$$
\begin{aligned}
&= n\Big[(1-\epsilon_n) r_H(\lambda_n^*, 0) + \epsilon_n r_H(\lambda_n^*, \mu_n)\Big] \\
&\geq n(1-\epsilon_n) r_H(\lambda_n^*, 0) \\
&= n(1-\epsilon_n)\Big[2\lambda_n^* \phi(\lambda_n^*) + 2(1-\Phi(\lambda_n^*))\Big] \\
&\geq 2n(1-\epsilon_n)\lambda_n^* \phi(\lambda_n^*) \\
&\geq 2n(1-\epsilon_n)\frac{c_1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{c_2^2\mu_n^2}{2}} \geq n\epsilon_n \mu_n^2,
\end{aligned}
$$

where the last inequality is due to the scaling $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ in the current regime.

- **Case III** $\lambda_n^* = \omega(\mu_n)$**:** In a similar way as in the proof of Case II of Proposition 2, we can conclude that

$$
\begin{aligned}
&\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda) - \underline{\theta}\|_2^2 \\
\geq & k_n \mu_n^2 + k_n(\lambda_n^* - \mu_n + o(\lambda_n^*)) \cdot \phi(\lambda_n^* - \mu_n) + k_n(\lambda_n^* + \mu_n + o(\lambda_n^*)) \cdot \phi(\lambda_n^* + \mu_n) \\
\geq & k_n \mu_n^2 = n\epsilon_n \mu_n^2.
\end{aligned}
$$

Note that since the three cases we have discussed above cover all the ranges of λ_n^*, we conclude that

$$
R_H(\Theta(k_n, \mu_n), 1) \geq \inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y,\lambda) - \underline{\theta}\|_2^2 \geq n\epsilon_n \mu_n^2.
$$

The proof of the upper bound is the same as the proof of the upper bound for Proposition 2 and is hence skipped here.

2.5.9 Proof of Theorem 8

Based on the scale invariance property of minimax risk mentioned in Section 3.4.1, it is equivalent to prove

$$R(\Theta(k_n, \mu_n), 1) = 2n\epsilon_n \log \epsilon_n^{-1} - 2n\epsilon_n \nu_n \sqrt{2 \log \nu_n} \left(1 + o(1)\right),$$

where $\nu_n = \sqrt{2 \log \epsilon_n^{-1}}$. As in the proof of Theorems 5 and 6, we first obtain an upper bound by analyzing the supremum risk of hard thresholding, and then develop a matching lower bound via the Bayesian approach. Before proceeding with the proof, we cover a few properties of the one-dimensional risk function of hard thresholding that becomes useful in the proof of Theorem 8.

Properties of the risk of hard thresholding estimator

Consider the one-dimensional risk of hard thresholding for $\mu \in \mathbb{R}$ and $\lambda > 0$,

$$r_H(\lambda, \mu) := \mathbb{E}\left(\hat{\eta}_H(\mu + z, \lambda) - \mu\right)^2, \quad z \sim \mathcal{N}(0, 1).$$

The following lemma from [3] gives simple and yet accurate bounds for $r_H(\lambda, \mu)$. Let

$$\bar{r}_H(\lambda, \mu) = \begin{cases} \min\{r_H(\lambda, 0) + 1.2\mu^2, 1 + \mu^2\} & 0 \le \mu \le \lambda \\ 1 + \mu^2(1 - \Phi(\mu - \lambda)) & \mu \ge \lambda, \end{cases}$$

where $\Phi(\cdot)$ is the CDF of standard normal random variable.

Lemma 14 (Lemma 8.5 in [3])**.**

(a) For $\lambda > 0$ and $\mu \in \mathbb{R}$,

$$(5/12)\bar{r}_H(\lambda, \mu) \le r_H(\lambda, \mu) \le \bar{r}_H(\lambda, \mu).$$

(b) The large μ component of $\bar{r}_H$ has the bound

$$\sup_{\mu \geq \lambda} \mu^2 (1 - \Phi(\mu - \lambda)) \leq \begin{cases} \lambda^2/2 & \text{if } \lambda \geq \sqrt{2\pi} \\ \lambda^2 & \text{if } \lambda \geq 1. \end{cases}$$

Our main goal in this section is to derive accurate approximations for $\sup_{\mu \geq 0} r_H(\lambda, \mu)$. The next lemma provides an accurate characterization of the risk for two different choices of μ. The importance of these choices becomes clear when we analyze $\sup_{\mu \geq 0} r_H(\lambda, \mu)$ later in this section.

Lemma 15. *As $\lambda \to \infty$, the risk of the hard thresholding, $r_H(\lambda, \mu)$, satisfies*

$$r_H(\lambda, \lambda) = \frac{1 + o(1)}{2} \lambda^2, \quad r_H(\lambda, \lambda - \sqrt{2 \log \lambda}) = \lambda^2 - (2\sqrt{2} + o(1)) \lambda \sqrt{\log \lambda}.$$

Proof. First note that the risk of hard thresholding can be written as

$$\begin{aligned} r_H(\lambda, \mu) =& \mu^2 \left[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu) \right] + \int_{|z+\mu| > \lambda} z^2 \phi(z) dz \\ =& (\mu^2 - 1) \left[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu) \right] + 1 + (\lambda - \mu) \phi(\lambda - \mu) + (\lambda + \mu) \phi(\lambda + \mu). \end{aligned} \tag{2.60}$$

Let $\mu = \lambda - \sqrt{2 \log \lambda}$. As $\lambda \to \infty$, we analyze the order of each term in the above expression:

$$\begin{aligned} & r_H(\lambda, \lambda - \sqrt{2 \log \lambda}) \\ =& \left[(\lambda - \sqrt{2 \log \lambda})^2 - 1 \right] \cdot \left(1 - \frac{1 + o(1)}{\sqrt{2 \log \lambda}} \phi(\sqrt{2 \log \lambda}) \right) + 1 \\ & + \sqrt{2 \log \lambda} \cdot \phi(\sqrt{2 \log \lambda}) + (2\lambda - \sqrt{2 \log \lambda}) \phi(2\lambda - \sqrt{2 \log \lambda}) \\ =& \left(\lambda - \sqrt{2 \log \lambda} \right)^2 + O\left(\frac{\lambda}{\sqrt{\log \lambda}} \right) = \lambda^2 - (2\sqrt{2} + o(1)) \lambda \sqrt{\log \lambda}, \end{aligned}$$

where in the first equality we have applied the Gaussian tail bound: $1 - \Phi(x) = (1 + o(1)) x^{-1} \phi(x)$

as $x \to \infty$. To prove the first part of the lemma, let $\mu = \lambda$. From (2.60) we have

$$r_H(\lambda, \lambda) = (\lambda^2 - 1)\left(\frac{1}{2} - \Phi(-2\lambda)\right) + 1 + 2\lambda\phi(2\lambda) = \lambda^2/2\left(1 + o(1)\right).$$

□

We now obtain the asymptotic approximation of $\sup_{\mu\geq 0} r_H(\lambda, \mu)$ in the next lemma.

Lemma 16. *As $\lambda \to \infty$, the supremum risk satisfies*

$$\sup_{\mu\geq 0} r_H(\lambda, \mu) = \lambda^2 - 2\sqrt{2}\lambda\sqrt{\log\lambda} + o(\lambda\sqrt{\log\lambda}).$$

Proof. Define

$$\mu^* = \arg\max_{\mu\geq 0} r_H(\lambda, \mu).$$

Comparing the upper bounds from Lemma 14 and the risk at $\lambda - \sqrt{2\log\lambda}$ in Lemma 15, we can conclude that the superemum risk is attained at $\mu = \mu^* \leq \lambda$ (when λ is large). To evaluate $r_H(\lambda, \mu^*)$, it is important to derive an accurate approximation for μ^*. We first claim that $\mu^*/\lambda \to 1$. Suppose this is not true. Then $\mu^* \leq c\lambda$ for some constant $c \in [0, 1)$ (take a sequence if necessary). According to Lemma 14 (a), for large enough values of λ, we have

$$r_H(\lambda, \mu^*) \leq \bar{r}_H(\lambda, \mu^*) \leq 1 + (\mu^*)^2 \leq \tilde{c}\lambda^2, \quad \tilde{c} \in (0, 1).$$

However, the above upper bound is strictly smaller than the risk $r_H(\lambda, \lambda - \sqrt{2\log\lambda})$ calculated in Lemma 15, contradicting with the definition of μ^*.

Second, we show that $\lambda - \mu^* \to \infty$, while $(\lambda - \mu^*)/\lambda \to 0$. Otherwise, it satisfies $0 \leq \lambda - \mu^* \leq c$ for some finite constant $c \geq 0$ (take a sequence if necessary). Then from (2.60) we have

$$r_H(\lambda, \mu^*) \leq \Phi(c)\lambda^2\left(1 + o(1)\right).$$

Comparing this with $r_H(\lambda, \lambda - \sqrt{2\log\lambda})$ from Lemma 15 leads to the same contradiction.

Third, we prove that for any given $c > 1$, $\lambda - \mu^* \le c\sqrt{2\log\lambda}$ for sufficiently large λ. Otherwise, there exists some constant $c > 1$ such that $\lambda_n - \mu_n^* > c\sqrt{2\log\lambda_n}$ for a sequence $\lambda_n \to \infty$ as $n \to \infty$. As a result, using Equation (2.60), and the result proved earlier that $\lambda_n - \mu_n^* \to \infty$, we obtain that for large n,

$$\begin{aligned} r_H(\lambda_n, \mu_n^*) &\le (\mu_n^*)^2 + 1 + (\lambda_n - \mu_n^*)\phi(\lambda_n - \mu_n^*) + (\lambda_n + \mu_n^*)\phi(\lambda_n + \mu_n^*) \\ &\le \left(\lambda_n - c\sqrt{2\log\lambda_n}\right)^2 + O(1) = \lambda_n^2 - (2c + o(1))\lambda_n\sqrt{2\log\lambda_n}. \end{aligned}$$

Again, comparing the above with $r_H(\lambda_n, \lambda_n - \sqrt{2\log\lambda_n}) = \lambda_n^2 - (2 + o(1))\lambda_n\sqrt{2\log\lambda_n}$ in Lemma 15, we see that $r_H(\lambda_n, \mu_n^*) < r_H(\lambda_n, \lambda_n - \sqrt{2\log\lambda_n})$ when n is large, which is a contradiction.

Finally, we prove that $(\lambda - \mu^*)/\sqrt{2\log\lambda} \to 1$ as $\lambda \to \infty$. Suppose this is not true. Given the result proved in the last paragraph, then there exists some constant $c < 1$ such that $\lambda_n - \mu_n^* < c\sqrt{2\log\lambda_n}$ for a sequence $\lambda_n \to \infty$ as $n \to \infty$. Using Equation (2.60) and Gaussian tail bound $1 - \Phi(x) = \frac{1+o(1)}{x}\phi(x)$ as $x \to \infty$, we have

$$\begin{aligned} r_H(\lambda_n, \mu_n^*) &= (\mu_n^*)^2\left[\Phi(\lambda_n - \mu_n^*) - \Phi(-\lambda_n - \mu_n^*)\right] + O(1) \\ &\le (\mu_n^*)^2\Phi(\lambda_n - \mu_n^*) + O(1) \\ &= (\mu_n^*)^2\left[1 - \frac{1 + o(1)}{\lambda_n - \mu_n^*}\phi(\lambda_n - \mu_n^*)\right] + O(1). \end{aligned}$$

Because $\phi(\lambda_n - \mu_n^*) \ge 1/\sqrt{2\pi} \cdot \exp\left(-\frac{2c^2\log\lambda_n}{2}\right) = 1/(\sqrt{2\pi}\lambda_n^{c^2})$, we continue with

$$\begin{aligned} r_H(\lambda_n, \mu_n^*) &\le (\mu_n^*)^2 - \frac{(\lambda_n - c\sqrt{2\log\lambda_n})^2}{c\sqrt{2\log\lambda_n}}\frac{1}{\sqrt{2\pi}\lambda_n^{c^2}} \cdot \big(1 + o(1)\big) + O(1) \\ &\le \lambda_n^2 - \frac{\lambda_n^{2-c^2}}{\sqrt{\log\lambda_n}} \cdot \Big(\frac{1}{2c\sqrt{\pi}} + o(1)\Big). \end{aligned}$$

Note that for $c < 1$, $\lambda_n^{2-c^2}/\sqrt{\log\lambda_n} = \omega(\lambda_n\sqrt{\log\lambda_n})$. Hence $r_H(\lambda_n, \mu_n^*) < r_H(\lambda_n, \lambda_n - \sqrt{2\log\lambda_n})$ when n is sufficiently large. The same contradiction arises.

Having the precise order that $\mu^* = \lambda - (1+o(1))\sqrt{2\log\lambda}$, we can easily evaluate $\sup_{\mu\geq 0} r_H(\lambda,\mu)$ from (2.60): as $\lambda \to \infty$,

$$\begin{aligned} r_H(\lambda, \lambda - \sqrt{2\log\lambda}) &\leq \sup_{\mu\geq 0} r_H(\lambda,\mu) = r_H(\lambda,\mu^*) \\ &= (\mu^*)^2(\Phi(\lambda-\mu^*) - \Phi(-\lambda-\mu^*)) + O(1) \\ &\leq (\mu^*)^2 + O(1) = (\lambda - (1+o(1))\sqrt{2\log\lambda})^2 + O(1) \\ &= \lambda^2 - 2\sqrt{2}\lambda\sqrt{\log\lambda} + o(\lambda\sqrt{\log\lambda}). \end{aligned}$$

Combining this result with Lemma 15 completes the proof. □

Upper bound

We are in the position to compute the supremum risk of $\hat{\eta}_H(y,\lambda_n)$ with $\lambda_n = \sigma_n\sqrt{2\log\epsilon_n^{-1}}$ in Theorem 8. First of all, due to the scale invariance of hard thresholding, the supremum risk can be written in the form:

$$\sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda_n) - \theta\right\|_2^2 = \sigma_n^2\left[(n-k_n)r_H(\nu_n,0) + \sup_{\|\tilde{\theta}\|_2^2\leq k_n\mu_n^2}\sum_{i=1}^{k_n} r_H(\nu_n,\tilde{\theta}_i)\right],$$

where $\tilde{\theta}\in\mathbb{R}^{k_n}$ and $\nu_n = \sqrt{2\log\epsilon_n^{-1}}$. Given that the one-dimensional risk function $r_H(\nu_n,\tilde{\theta}_i)$ is symmetric in $\tilde{\theta}_i$, if its maximizer satisfies $\arg\max_{\tilde{\theta}_i\geq 0} r_H(\nu_n,\tilde{\theta}_i) \leq \mu_n$, then we will have

$$\sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda_n) - \theta\right\|_2^2 = \sigma_n^2\left[(n-k_n)r_H(\nu_n,0) + k_n\sup_{\mu\geq 0} r_H(\nu_n,\mu)\right]. \tag{2.61}$$

This will allow us to focus on finding the supremum risk of hard thresholding in the univariate setting that we discussed in the last section. In the proof of Lemma 16, we already showed that $\arg\max_{\tilde{\theta}_i\geq 0} r_H(\nu_n,\tilde{\theta}_i) \leq \nu_n$ when n is large. It is then clear that in the current regime $\mu_n = \omega(\sqrt{2\log\epsilon_n^{-1}})$, it holds that $\arg\max_{\tilde{\theta}_i\geq 0} r_H(\nu_n,\tilde{\theta}_i) \leq \mu_n$ for large n. Therefore, the supremum risk of hard thresholding over $\Theta(k_n,\tau_n)$ can be simplified as in (2.61). We can apply Lemma 16 to

continue from (2.61):

$$\begin{aligned}
&\sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda_n)-\theta\right\|_2^2 \\
&= n\sigma_n^2\left[(1-\epsilon_n)r_H(\nu_n,0)+\epsilon_n\sup_{\mu\geq 0} r_H(\nu_n,\mu)\right] \\
&= n\sigma_n^2\left[(1-\epsilon_n)r_H(\nu_n,0)+\epsilon_n\left(\nu_n^2-2\nu_n\sqrt{2\log\nu_n}+o(\nu_n\sqrt{\log\nu_n})\right)\right], \qquad (2.62)
\end{aligned}$$

where $\nu_n = \sqrt{2\log\epsilon_n^{-1}}$. We now identify the dominating terms in the above expression. First,

$$r_H(\nu_n,0) = 2\int_{\nu_n}^{\infty} z^2\phi(z)dz = 2\nu_n\phi(\nu_n)+2(1-\Phi(\nu_n)) = (2+o(1))\nu_n\phi(\nu_n) = O(\epsilon_n\nu_n), \qquad (2.63)$$

where the last two equations are due to the Gaussian tail bound $1-\Phi(x) = \frac{1+o(1)}{x}\phi(x)$ as $x\to\infty$ and $\nu_n = \sqrt{2\log\epsilon_n^{-1}}$. Therefore, from (2.62) we obtain

$$\begin{aligned}
&\sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda_n)-\theta\right\|_2^2 \\
&= n\sigma_n^2\left[\epsilon_n\nu_n^2-2\epsilon_n\nu_n\sqrt{2\log\nu_n}+o(\epsilon_n\nu_n\sqrt{\log\nu_n})\right] \\
&= n\sigma_n^2\epsilon_n\left(2\log\epsilon_n^{-1}-(2+o(1))\nu_n\sqrt{2\log\nu_n}\right).
\end{aligned}$$

This completes our proof of the upper bound in Theorem 8.

The sharp upper bound we have derived is from the hard thresholding estimator $\hat{\eta}_H(y,\lambda_n)$ with tuning $\lambda_n = \sigma_n\nu_n$. To shed more light on the performance of hard thresholding, we provide a discussion on the optimal choices of λ_n. The lemma below characterizes the possible choices of λ_n that leads to optimal supremum risk (up to second order).

Lemma 17. *Consider model* (2.1)*, and parameter space* (2.6) *under Regime (III), in which* $\epsilon_n\to 0$*,* $\mu_n\to\infty$*,* $\mu_n = \omega(\sqrt{\log\epsilon_n^{-1}})$*, as* $n\to\infty$*. Let* $\nu_n = \sqrt{2\log\epsilon_n^{-1}}$*. Consider the tuning regime*

$\lambda_n\sigma_n^{-1} \to \infty$ *and* $\lambda_n\sigma_n^{-1} \le \mu_n$. *If* λ_n *satisfies:*

$$(\nu_n^2 - c_1 \log\log \nu_n) \le \lambda_n^2\sigma_n^{-2} \le (\nu_n^2 + c_2\nu_n\sqrt{2\log \nu_n})$$

when n is large, for some constant $c_1 < 1$ *and every* $c_2 > 0$*, then*

$$\inf_{\lambda} \sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda) - \theta\right\|_2^2 = \sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda_n) - \theta\right\|_2^2 + o\left(n\sigma_n^2\epsilon_n\nu_n\sqrt{\log \nu_n}\right). \quad (2.64)$$

On the other hand, if $(\nu_n^2 - c_1 \log\log \nu_n) \ge \lambda_n^2\sigma_n^{-2}$ *for a constant* $c_1 \ge 1$ *or if* $\lambda_n^2\sigma_n^{-2} \ge (\nu_n^2 + c_2\nu_n\sqrt{2\log \nu_n})$ *for some* $c_2 > 0$*, then the conclusion* (2.64) *will not hold.*

Proof. Denote $\tilde{\lambda}_n = \lambda_n\sigma_n^{-1}$. Given that we focus on the tuning regime $\tilde{\lambda}_n \to \infty$ and $\tilde{\lambda}_n \le \mu_n$, the result (2.62) continues to hold here:

$$\sup_{\theta\in\Theta(k_n,\tau_n)} \mathbb{E}_\theta\left\|\hat{\eta}_H(y,\lambda_n) - \theta\right\|_2^2 = n\sigma_n^2 \cdot \left[(1-\epsilon_n)r_H(\tilde{\lambda}_n, 0) + \epsilon_n\left(\tilde{\lambda}_n^2 - 2\tilde{\lambda}_n\sqrt{2\log \tilde{\lambda}_n} + o(\tilde{\lambda}_n\sqrt{\log \tilde{\lambda}_n})\right)\right].$$

Hence, we define

$$A(\lambda) := (1-\epsilon_n)r_H(\lambda, 0) + \epsilon_n\left[\lambda^2 - 2\lambda\sqrt{2\log\lambda} + o\left(\lambda\sqrt{\log\lambda}\right)\right], \quad (2.65)$$

where the notation $o(\cdot)$ is understood as $\lambda \to \infty$. We proved before that $A(\nu_n) = \epsilon_n(\nu_n^2 - (2 + o(1))\nu_n\sqrt{2\log \nu_n})$. Now we consider four different regions for $\tilde{\lambda}_n$ (when n is large):

- Case $\tilde{\lambda}_n^2 \le \nu_n^2 - 2c\log(\nu_n/\sqrt{2\pi})$ for some constant $c > 1$. Equation (2.63) implies

$$\begin{aligned} A(\tilde{\lambda}_n) \ge& (1-\epsilon_n)r_H(\tilde{\lambda}_n, 0) \ge (1-\epsilon_n)r_H\left(\left(\nu_n^2 - 2c\log(\nu_n/2\pi)\right)^{1/2}, 0\right) \\ =& (2+o(1))\left(\nu_n^2 - 2c\log(\nu_n/2\pi)\right)^{1/2} \cdot \phi\left(\left(\nu_n^2 - 2c\log(\nu_n/\sqrt{2\pi})\right)^{1/2}\right) \\ =& \frac{2+o(1)}{\sqrt{2\pi}}\nu_n \exp\left(-\frac{\nu_n^2 - 2c\log\frac{\nu_n}{\sqrt{2\pi}}}{2}\right) = \Theta\left(\epsilon_n(\nu_n)^{1+c}\right). \end{aligned}$$

Note that $A(\tilde{\lambda}_n) = \omega(A(\nu_n))$, and hence $\tilde{\lambda}_n$ does not satisfy (2.64).

- Case $\nu_n^2 - 2c_1 \log(\nu_n/\sqrt{2\pi}) \le \tilde{\lambda}_n^2 \le \nu_n^2 - c_2 \log\log \nu_n$ for any constant $c_1 \le 1$ and some constant $c_2 \ge 1$. Since $\tilde{\lambda}_n^2 \le \nu_n^2 - c_2 \log\log \nu_n$, the same argument as in the previous case gives

$$(1-\epsilon_n) r_H(\tilde{\lambda}_n, 0) \ge \frac{2+o(1)}{\sqrt{2\pi}} \left(\epsilon_n \nu_n \left(\sqrt{\log \nu_n} \right)^{c_2} \right). \tag{2.66}$$

Moreover, using the upper and lower bounds we set for $\tilde{\lambda}_n$, we obtain

$$\begin{aligned} &\epsilon_n \left(\tilde{\lambda}_n^2 - 2\tilde{\lambda}_n \sqrt{2\log \tilde{\lambda}_n} + o\left(\tilde{\lambda}_n \sqrt{\log \tilde{\lambda}_n} \right) \right) \\ \ge &\epsilon_n \left[\nu_n^2 - 2c_1 \log \frac{\nu_n}{\sqrt{2\pi}} - 2\nu_n \sqrt{2\log \nu_n} + o\left(\nu_n \sqrt{\log \nu_n} \right) \right] \\ = &\epsilon_n \left[\nu_n^2 - 2\nu_n \sqrt{2\log \nu_n} + o\left(\nu_n \sqrt{\log \nu_n} \right) \right]. \end{aligned} \tag{2.67}$$

Combining (2.66)-(2.67) yields

$$A(\tilde{\lambda}_n) \ge \epsilon_n \Big[\nu_n^2 + \nu_n \sqrt{2\log \nu_n} \Big(-2 + o(1) + \frac{2+o(1)}{2\sqrt{\pi}} (\sqrt{\log \nu_n})^{c_2 - 1} \Big) \Big].$$

Since $c_2 \ge 1$, it is clear that $A(\tilde{\lambda}_n) - A(\nu_n) = \Omega(\epsilon_n \nu_n \sqrt{\log \nu_n})$. Therefore, this choice of $\tilde{\lambda}_n$ does not satisfy (2.64).

- Case $\nu_n^2 - c_1 \log\log \nu_n \le \tilde{\lambda}_n^2 \le \nu_n^2 + c_2 \nu_n \sqrt{2\log \nu_n}$ for some constant $c_1 < 1$ and every $c_2 > 0$. With the lower bound of $\tilde{\lambda}_n$, similar calculations as in the previous two cases lead to

$$(1-\epsilon_n) r_H(\tilde{\lambda}_n, 0) \le r_H \left(\left(\nu_n^2 - c_1 \log\log \nu_n \right)^{1/2}, 0 \right) = \Theta \left(\epsilon_n \nu_n \left(\sqrt{\log \nu_n} \right)^{c_1} \right).$$

Furthermore, the upper and lower bounds of $\tilde{\lambda}_n$ for some $c_1 < 1$ and every $c_2 > 0$ imply that

$\tilde{\lambda}_n^2 - \nu_n^2 = o(\nu_n\sqrt{\log \nu_n})$. Thus,

$$\epsilon_n\left(\tilde{\lambda}_n^2 - 2\tilde{\lambda}_n\sqrt{2\log\tilde{\lambda}_n} + o\left(\tilde{\lambda}_n\sqrt{2\log\tilde{\lambda}_n}\right)\right) \leq \epsilon_n\left(\nu_n^2 - 2\nu_n\sqrt{2\log\nu_n} + o\left(\nu_n\sqrt{\log\nu_n}\right)\right).$$

Putting together the above two results into (2.65), we have

$$\begin{aligned} A(\tilde{\lambda}_n) &\leq \Theta\left(\epsilon_n\nu_n\left(\sqrt{\log\nu_n}\right)^{c_1}\right) + \epsilon_n\left(\nu_n^2 - 2\nu_n\sqrt{2\log\nu_n} + o\left(\nu_n\sqrt{\log\nu_n}\right)\right) \\ &= \epsilon_n\left(\nu_n^2 - (2+o(1))\nu_n\sqrt{2\log\nu_n}\right). \end{aligned}$$

Thus, $A(\tilde{\lambda}_n) \leq A(\nu_n) + o(\epsilon_n\nu_n\sqrt{\log\nu_n})$, and $\tilde{\lambda}_n$ satisfies (2.64).

- Case $\tilde{\lambda}_n^2 \geq \nu_n^2 + c\nu_n\sqrt{2\log\nu_n}$ for some constant $c > 0$. We only need consider $\tilde{\lambda}_n = (1 + o(1))\nu_n$, because for larger values of λ_n, (2.65) implies that $A(\tilde{\lambda}_n)/A(\nu_n) > 1$ for large n. When $\tilde{\lambda}_n = (1+o(1))\nu_n$, we have

$$\begin{aligned} A(\tilde{\lambda}_n) &\geq \epsilon_n\left(\tilde{\lambda}_n^2 - 2\tilde{\lambda}_n\sqrt{2\log\tilde{\lambda}_n} + o\left(\tilde{\lambda}_n\sqrt{2\log\tilde{\lambda}_n}\right)\right) \\ &\geq \epsilon_n\left(\nu_n^2 - (2-c)\nu_n\sqrt{2\log\nu_n} + o\left(\nu_n\sqrt{2\log\nu_n}\right)\right). \end{aligned}$$

Since $c > 0$, the above implies that $A(\tilde{\lambda}_n) - A(\nu_n) = \Omega(\epsilon_n\nu_n\sqrt{\log\nu_n})$. Hence $\tilde{\lambda}_n$ does not satisfy (2.64).

□

Lower bound

As in the proof of lower bound in Theorems 5-7, we will apply Theorem 11 and utilize the independent block prior that is first described in Section 2.5.2. To simplify the calculations a bit here, we will use the block prior with one minor modification: adopting the notation from Section

2.5.2, the spike prior $\pi_S^{\mu,m}$ in use is now changed to a one-sided spike prior:

$$\pi_S^{\mu,m}(\theta^{(j)} = \mu e_i) = \frac{1}{m}, \quad 1 \le i \le m, \tag{2.68}$$

where $\mu \in (0, \mu_n]$. The key is to calculate the Bayes risk $B(\pi_S^{\mu,m})$ and obtain a result like Lemma 3. To this end, we first mention a lemma that will become useful later in the proof.

Lemma 18. *Let* $z_1, \ldots, z_m \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ *and* $\nu_m = \sqrt{2\log m}$. *Suppose* $2\mu > \nu_m$ *and* $\delta < \Phi(\nu_m - \mu)$. *Then*

$$\mathbb{P}\left(m^{-1} e^{-\frac{1}{2}\mu^2} \sum_{j=1}^m e^{\mu z_j} \le \delta \right) \le \frac{1}{\sqrt{2\pi}\nu_m} + \frac{1}{\sqrt{2\pi}} \frac{1}{\left[\Phi(\nu_m - \mu) - \delta\right]^2} \frac{1}{2\mu - \nu_m} e^{-(\mu - \nu_m)^2}.$$

Proof. Define the notation:

$$\begin{aligned} &X_{mj} = e^{\mu z_j}, \quad \bar{X}_{mj} = X_{mj} I_{(X_{mj} \le e^{\mu\nu_m})}, \\ &S_m = \sum_{j=1}^m X_{mj}, \quad \bar{S}_m = \sum_{j=1}^m \bar{X}_{mj}, \\ &a_m = \mathbb{E}\bar{S}_m = m e^{\mu^2/2} \Phi(\nu_m - \mu). \end{aligned}$$

Then

$$\mathbb{P}\Big(m^{-1} e^{-\frac{1}{2}\mu^2} \sum_{j=1}^m e^{\mu z_j} \le \delta \Big) = \mathbb{P}\left\{ a_m - S_m \ge \left[\Phi(\nu_m - \mu) - \delta\right] \cdot m e^{\frac{1}{2}\mu^2} \right\} = \mathbb{P}\left(\frac{a_m - S_m}{e^{\mu\nu_m}} \ge t \right),$$

where $t := \left[\Phi(\nu_m - \mu) - \delta\right] \cdot m e^{\frac{1}{2}\mu^2 - \mu\nu_m}$. Clearly,

$$\mathbb{P}\left(\frac{a_m - S_m}{e^{\mu\nu_m}} \ge t \right) \le \mathbb{P}\left(S_m \ne \bar{S}_m \right) + \mathbb{P}\left(\left| \frac{\bar{S}_m - a_m}{e^{\mu\nu_m}} \right| > t \right).$$

For the following calculation, we will use Gaussian tail bound $1 - \Phi(x) \le x^{-1}\phi(x)$ for $x > 0$. To

obtain a proper upper bound for the first term, we note that

$$\begin{aligned}\mathbb{P}\left(S_m \neq \bar{S}_m\right) \leq & \mathbb{P}\left(\cup_{j=1}^m \{\bar{X}_{mj} \neq X_{mj}\}\right) \leq \sum_{j=1}^m \mathbb{P}\left(X_{mj} > e^{\mu\nu_m}\right) \\ = & \sum_{j=1}^m \mathbb{P}\left(e^{\mu z_j} > e^{\mu\nu_m}\right) = m(1-\Phi(\nu_m)) \leq \frac{m}{\nu_m}\phi(\nu_m) = \frac{1}{\sqrt{2\pi}\nu_m}.\end{aligned}$$

For the second term, we use Chebyshev's inequality and the fact that $a_m = \mathbb{E}\bar{S}_m$ and $\mathrm{Var}(X) \leq \mathbb{E}X^2$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\bar{S}_m - a_m}{e^{\mu\nu_m}}\right| > t\right) \leq & t^{-2}e^{-2\mu\nu_m}\mathbb{E}(\bar{S}_m - a_m)^2 \leq (te^{\mu\nu_m})^{-2}\sum_{j=1}^m \mathbb{E}\bar{X}_{mj}^2 \\ \leq & \frac{1}{\left[\Phi(\nu_m - \mu) - \delta\right]^2}\frac{1}{\sqrt{2\pi}}\frac{1}{2\mu - \nu_m}e^{-(\mu-\nu_m)^2}.\end{aligned}$$

The last inequality is based on the following calculation:

$$\begin{aligned}\mathbb{E}\bar{X}_{mj}^2 = & \mathbb{E}\left(e^{\mu z_j}I_{(e^{\mu z_j} \leq e^{\mu\nu_m})}\right)^2 = \int_{z\leq\nu_m} e^{2\mu z}\phi(z)dz = e^{2\mu^2}(1-\Phi(2\mu-\nu_m)) \\ \leq & \frac{1}{\sqrt{2\pi}}\frac{1}{2\mu-\nu_m}e^{2\mu^2-\frac{1}{2}(2\mu-\nu_m)^2} = \frac{1}{\sqrt{2\pi}}\frac{1}{2\mu-\nu_m}e^{-\frac{1}{2}\nu_m^2+2\mu\nu_m},\end{aligned}$$

and

$$\begin{aligned}& (te^{\mu\nu_m})^{-2}m \cdot \frac{1}{\sqrt{2\pi}}\frac{1}{2\mu-\nu_m}e^{-\frac{1}{2}\nu_m^2+2\mu\nu_m} \\ = & \frac{1}{\left[\Phi(\nu_m-\mu)-\delta\right]^2}\frac{1}{m^2}e^{-\mu^2}m\frac{1}{\sqrt{2\pi}}\frac{1}{2\mu-\nu_m}e^{-\frac{1}{2}\nu_m^2+2\mu\nu_m} \\ = & \frac{1}{\left[\Phi(\nu_m-\mu)-\delta\right]^2}\frac{1}{\sqrt{2\pi}}\frac{1}{2\mu-\nu_m}e^{-(\mu-\nu_m)^2}.\end{aligned}$$

□

We are now ready to calculate the Bayes risk $B(\pi_S^{\mu,m})$ in the following lemma.

Lemma 19. *Let $\nu_m = \sqrt{2\log m}$ and $\mu = \nu_{m-1} - \sqrt{2\log\nu_{m-1}}$. As $m \to \infty$, the Bayes risk $B(\pi_S^{\mu,m})$*

satisfies

$$B(\pi_S^{\mu,m}) \geq \nu_m^2 - 2\nu_m\sqrt{2\log\nu_m}\left(1+o(1)\right).$$

Proof. For the one-sided spike prior $\pi_S^{\mu,m}$ introduced in (2.68), doing similar calculations as in the proof of Lemma 2, we can obtain the expression for the Bayes risk:

$$B(\pi_S^{\mu,m}) = \mu^2\mathbb{E}_{\mu e_1}(p_m-1)^2 + (m-1)\mu^2\mathbb{E}_{\mu e_2}p_m^2 \geq \mu^2 - 2\mu^2\mathbb{E}_{\mu e_1}p_m, \tag{2.69}$$

where $p_m = \frac{e^{\mu y_1}}{\sum_{j=1}^m e^{\mu y_j}}$; $\mathbb{E}_{\mu e_1}(\cdot)$ is taken with respect to $y \sim \mathcal{N}(\mu e_1, I)$ and $\mathbb{E}_{\mu e_2}(\cdot)$ for $y \sim \mathcal{N}(\mu e_2, I)$. Now the goal is to upper bound $\mathbb{E}_{\mu e_1}p_m$. We have

$$\mathbb{E}_{\mu e_1}p_m = \mathbb{E}\frac{e^{\mu(\mu+z_1)}}{\sum_{j\neq 1}e^{\mu z_j} + e^{\mu(\mu+z_1)}} = \mathbb{E}\frac{(m-1)^{-1}e^{\frac{1}{2}\mu^2+\mu z_1}}{(m-1)^{-1}e^{\frac{1}{2}\mu^2+\mu z_1} + (m-1)^{-1}e^{-\frac{1}{2}\mu^2}\sum_{j\neq 1}e^{\mu z_j}}, \tag{2.70}$$

where $z_1,\ldots,z_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Define the following two events:

$$\mathcal{F}_1 = \left\{(m-1)e^{-\frac{1}{2}\mu^2-\mu z_1} \geq M\right\}, \quad \mathcal{F}_2 = \left\{(m-1)^{-1}e^{-\frac{1}{2}\mu^2}\sum_{j\neq 1}e^{\mu z_j} \geq \delta\right\},$$

where δ and M are two positive constants to be determined later. Since the ratio inside the expectation of (2.70) is smaller than one, and on the event $\mathcal{F}_1\cap\mathcal{F}_2$ it is smaller than $\frac{1}{M\delta}$, we can continue from (2.70) to obtain

$$\mathbb{E}_{\mu e_1}p_m \leq \frac{1}{M\cdot\delta} + \mathbb{P}(\mathcal{F}_1^c) + \mathbb{P}(\mathcal{F}_2^c). \tag{2.71}$$

Hence, we aim to find upper bounds for $\mathbb{P}(\mathcal{F}_1^c)$ and $\mathbb{P}(\mathcal{F}_2^c)$. For the first probability, using Gaussian tail bound that $1-\Phi(x) \leq \frac{1}{x}\phi(x)$ for $x>0$, and that $e^{\nu_{m-1}^2/2} = m-1$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{F}_1^c) &= \mathbb{P}\left((m-1)e^{-\frac{1}{2}\mu^2-\mu z} < M\right) \\
&= \mathbb{P}\left(z > -\frac{1}{2}\mu - \frac{1}{\mu}\log\frac{M}{m-1}\right) = 1-\Phi\left(-\frac{1}{\mu}\log M + \frac{1}{2\mu}(\nu_{m-1}^2-\mu^2)\right) \\
&\leq \frac{1}{-\frac{1}{\mu}\log M + \frac{1}{2\mu}(\nu_{m-1}^2-\mu^2)}\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2\mu^2}\left[\frac{1}{2}(\nu_{m-1}^2-\mu^2) - \log M\right]^2\right) := U_1,
\end{aligned}$$

as long as $\nu_{m-1}^2 - \mu^2 > 2\log M$. Regarding $\mathbb{P}(\mathcal{F}_2^c)$, if we limit our choice of $0 < \delta < \Phi(\nu_{m-1} - \mu)$, then from Lemma 18,

$$\mathbb{P}(\mathcal{F}_2^c) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{\nu_{m-1}} + \frac{1}{\sqrt{2\pi}} \frac{1}{\left[\Phi(\nu_{m-1} - \mu) - \delta\right]^2} \cdot \frac{1}{2\mu - \nu_{m-1}} e^{-(\mu - \nu_{m-1})^2} := U_2.$$

Now we set $M = \nu_{m-1}$ and recall $\mu = \nu_{m-1} - \sqrt{2\log \nu_{m-1}}$. We will show that $U_1 = o(\nu_{m-1}^{-1})$ and $U_2 = O(\nu_{m-1}^{-1})$. First, for U_1,

$$\begin{aligned}
&\frac{1}{2\mu^2}\left[\frac{1}{2}(\nu_{m-1}^2 - \mu^2) - \log M\right]^2 \\
&= \frac{1}{2\mu^2}\left[\frac{1}{2}(2\nu_{m-1}\sqrt{2\log \nu_{m-1}} - 2\log \nu_{m-1}) - \log \nu_{m-1}\right]^2 \\
&= \frac{1}{2\mu^2}\left[\nu_{m-1}\sqrt{2\log \nu_{m-1}} - 2\log \nu_{m-1}\right]^2 \\
&= \frac{\nu_{m-1}^2}{\mu^2}\log \nu_{m-1} - \frac{2\sqrt{2}\nu_{m-1}}{\mu^2}\left(\log \nu_{m-1}\right)^{3/2} + \frac{2}{\mu^2}\left(\log \nu_{m-1}\right)^2 \geq \log \nu_{m-1} + o(1),
\end{aligned}$$

where in the last inequality we used $\mu^2 < \nu_{m-1}^2$ (for large m). Therefore,

$$e^{-\frac{1}{2\mu^2}\left[\frac{1}{2}(\nu_{m-1}^2 - \mu^2) - \log M\right]^2} \leq \nu_{m-1}^{-1}\left(1 + o(1)\right),$$

and

$$\begin{aligned}
&\frac{1}{-\frac{1}{\mu}\log M + \frac{1}{2\mu}(\nu_{m-1}^2 - \mu^2)} = \frac{1}{\frac{1}{\mu}\cdot\left(\nu_{m-1}\sqrt{2\log \nu_{m-1}} - 2\log \nu_{m-1}\right)} \\
&\leq \left(\sqrt{2\log \nu_{m-1}} - \frac{2\log \nu_{m-1}}{\nu_{m-1}}\right)^{-1} = o(1).
\end{aligned}$$

In combination,

$$U_1 \leq o(1)\cdot \nu_{m-1}^{-1}\left(1 + o(1)\right) = o(\nu_{m-1}^{-1}). \tag{2.72}$$

For U_2, we set δ to be any fixed constant between $(0, 1)$. Since $\nu_{m-1} - \mu \to +\infty$, it holds that $\Phi(\nu_{m-1} - \mu) - \delta > \delta'$ for some constant $\delta' > 0$, when m is large. Also, we have the identity

$e^{-(\mu-\nu_{m-1})^2} = e^{-2\log\nu_{m-1}} = \nu_{m-1}^{-2}$. So the second term in U_2 is of order $O(\nu_{m-1}^{-3})$. Thus,

$$U_2 = \frac{1+o(1)}{\sqrt{2\pi}\nu_{m-1}}. \tag{2.73}$$

Note that we have set $M = \nu_{m-1}$. Hence, $1/(M\cdot\delta) = O(1/\nu_{m-1})$. Combining (2.71)-(2.73), we have

$$\mathbb{E}_{\mu e_1} p_m \le O(1/\nu_{m-1}).$$

Finally, the above together with (2.69) shows that

$$\begin{aligned} B(\pi_S^{\mu,m}) &\ge \mu^2 - 2\mu^2 O\left(\nu_{m-1}^{-1}\right) \\ &= \nu_{m-1}^2 - 2\nu_{m-1}\sqrt{2\log\nu_{m-1}}\left(1+o(1)\right) \\ &= \nu_m^2 - 2\nu_m\sqrt{2\log\nu_m}\left(1+o(1)\right). \end{aligned}$$

□

Now, we aim to apply Lemma 19 to derive the minimax lower bound. First note that in the current regime $\epsilon_n \to 0, \mu_n = \omega(\sqrt{\log\epsilon_n^{-1}})$, the choice of μ with $m = n/k_n = \epsilon_n^{-1}$ in Lemma 19 satisfies $\mu < \mu_n$ when n is large. Thus, the constructed block prior is supported on the parameter space $\Theta(k_n, \mu_n)$ so that we can use Equation (2.13) and Lemma 19 to conclude

$$\begin{aligned} R(\Theta(k_n,\tau_n),\sigma_n) &= \sigma_n^2 \cdot R(\Theta(k_n,\mu_n),1) \ge k_n\sigma_n^2 \cdot B(\pi_S^{\mu,m}) \\ &\ge k_n\sigma_n^2 \cdot \left(\nu_m^2 - 2\nu_m\sqrt{2\log\nu_m}(1+o(1))\right) \\ &= n\sigma_n^2\left(2\epsilon_n\log\epsilon_n^{-1} - 2\epsilon_n\nu_m\sqrt{2\log\nu_m}(1+o(1))\right), \end{aligned}$$

where $\nu_m = \sqrt{2\log m} = \sqrt{2\log\epsilon_n^{-1}}$.

2.5.10 Proof of Proposition 6

Roadmap of the proof

Propositions 1 and 3 have derived the supremum risk of optimally tuned soft thresholding in Regimes (I) and (II) respectively. Proposition 6 continues to obtain it in Regime (III). Hence, we will use some existing results from the proof of Propositions 1 and 3 to simplify the present proof. First of all, referring to Equations (2.58)-(2.17) in the proof of Proposition 1, the supremum risk can be expressed as

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \left\| \hat{\eta}_S(y, \lambda) - \theta \right\|_2^2 = n\sigma_n^2 \cdot \inf_{\lambda} \underbrace{\left[(1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z, \lambda) + \epsilon_n \mathbb{E}(\hat{\eta}_S(z+\mu_n, \lambda) - \mu_n)^2 \right]}_{:=F(\lambda)},$$

with $z \sim \mathcal{N}(0,1)$. Define the optimal tuning $\lambda_* = \arg\min_{\lambda \geq 0} F(\lambda)$. Then it is equivalent to prove

$$F(\lambda_*) = 2\epsilon_n \log \epsilon_n^{-1} - (6 + o(1))\epsilon_n \log \nu_n,$$

where $\nu_n = \sqrt{2 \log \epsilon_n^{-1}}$. To reach the above, we will first find the tight upper bound for $F(\lambda_*)$ in Section 2.5.10, and then obtain the matching lower bound in Section 2.5.10. Before we do these two parts, let us prove a lemma that provides an approximation for $F(\lambda)$. This approximation will help us in the calculation of both the upper and lower bounds.

Lemma 20. *Consider $\epsilon_n \to 0$, $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$, as $n \to \infty$. If $\lambda \to \infty$ and $\mu_n - \lambda \to +\infty$, then*

$$F(\lambda) = 2(1-\epsilon_n)\left[(1+\lambda^2)(1-\Phi(\lambda)) - \lambda\phi(\lambda) \right] + \epsilon_n \left[\lambda^2 + 1 - \frac{(2+o(1))\mu_n}{(\mu_n - \lambda)^2} \phi(\mu_n - \lambda) \right].$$

Furthermore, when λ is large, it holds that

$$C(\lambda) \leq F(\lambda) \leq D(\lambda),$$

where

$$C(\lambda) := 2(1-\epsilon_n)\cdot\left(\frac{2}{\lambda^3}-\frac{12}{\lambda^5}\right)\frac{1}{\sqrt{2\pi}}\epsilon_n\cdot e^{\frac{1}{2}(v_n^2-\lambda^2)} \tag{2.74}$$
$$+\epsilon_n\left[\lambda^2+1-\frac{(2+o(1))\mu_n}{(\mu_n-\lambda)^2}\phi(\mu_n-\lambda)\right],$$

and

$$D(\lambda) := \epsilon_n\left\{(1-\epsilon_n)\frac{4}{\sqrt{2\pi}\lambda^3}e^{\frac{1}{2}(v_n^2-\lambda^2)}+\lambda^2+1\right\}. \tag{2.75}$$

Proof. Throughout the proof, we will use the Gaussian tail bound in Lemma 1 to do calculations. With the expression of $F(\lambda)$ calculated in Equations (2.19)-(2.22), we have that as $\lambda\to\infty, \mu_n-\lambda\to+\infty$,

$$\begin{aligned}
F(\lambda) &= 2(1-\epsilon_n)\cdot\left[(1+\lambda^2)(1-\Phi(\lambda))-\lambda\phi(\lambda)\right]\\
&\quad+\epsilon_n\cdot\left\{(\lambda^2+1)+\left[(\mu_n^2-\lambda^2-1)(1-\Phi(\mu_n-\lambda))-(\mu_n+\lambda)\phi(\mu_n-\lambda)\right]\right.\\
&\quad\left.-\left[(\mu_n^2-\lambda^2-1)\cdot(1-\Phi(\mu_n+\lambda))-(\mu_n-\lambda)\phi(\mu_n+\lambda)\right]\right\}\\
&= 2(1-\epsilon_n)\left[(1+\lambda^2)(1-\Phi(\lambda))-\lambda\phi(\lambda)\right]+\epsilon_n\left[\lambda^2+1-\frac{(2+o(1))\mu_n}{(\mu_n-\lambda)^2}\phi(\mu_n-\lambda)\right],
\end{aligned}$$

where in the last equation we have used $1-\Phi(x)=\left(\frac{1}{x}-\frac{1+o(1)}{x^3}\right)\phi(x)$ as $x\to\infty$.

As $\lambda\to\infty$, we obtain

$$\begin{aligned}
&(1+\lambda^2)(1-\Phi(\lambda))-\lambda\phi(\lambda)\\
=&\left[(1+\lambda^2)\left(\frac{1}{\lambda}-\frac{1}{\lambda^3}+\frac{3}{\lambda^5}-\frac{15}{\lambda^7}+\frac{105}{\lambda^9}\right)-\lambda\right]\phi(\lambda)+O\left(\frac{\phi(\lambda)}{\lambda^9}\right)\\
=&\left(\frac{2}{\lambda^3}-\frac{12}{\lambda^5}+\frac{90}{\lambda^7}\right)\phi(\lambda)+O\left(\frac{\phi(\lambda)}{\lambda^9}\right).
\end{aligned}$$

Thus,

$$F(\lambda) = 2(1-\epsilon_n)\cdot\left(\frac{2}{\lambda^3}-\frac{12}{\lambda^5}+\frac{90}{\lambda^7}+O\left(\frac{1}{\lambda^9}\right)\right)\cdot\frac{1}{\sqrt{2\pi}}\epsilon_n\cdot e^{\frac{1}{2}(\nu_n^2-\lambda^2)}+\epsilon_n\Big[\lambda^2+1-\frac{(2+o(1))\mu_n}{(\mu_n-\lambda)^2}\phi(\mu_n-\lambda)\Big].$$

As a result, it is straightforward to verify that $C(\lambda)$ and $D(\lambda)$ defined in (2.74)-(2.75) provide lower and upper bounds for $F(\lambda)$. □

Upper bound

Consider $\lambda=\sqrt{\nu_n^2-6\log\nu_n}$, then $\lambda\to\infty$ and $\mu_n-\lambda\to\infty$. From Lemma 20,

$$\begin{aligned}
&F(\lambda_*)\le F(\lambda)\le D(\lambda)\\
=&\epsilon_n\left\{(1-\epsilon_n)\frac{4}{\sqrt{2\pi}}e^{\frac{1}{2}\left[(\nu_n^2-\lambda^2)-6\log\lambda\right]}+\lambda^2+1\right\}\\
=&\epsilon_n\left\{\frac{4+o(1)}{\sqrt{2\pi}}+\lambda^2+1\right\}=\epsilon_n\nu_n^2-6\epsilon_n\log\nu_n\left(1+o(1)\right).
\end{aligned}\tag{2.76}$$

Lower bound

We now derive a matching lower bound for $F(\lambda_*)$. This requires a careful analysis of the order of the optimal tuning λ_*. We break it down in several steps:

Step 1. First, we show that $\lambda_*\to\infty, \mu_n-\lambda_*\to+\infty$. We will need the following lemma.

Lemma 21 (Lemma 8.3 in [3])**.** *Define* $r_S(\lambda,\mu)=\mathbb{E}(\hat\eta_S(\mu+z,\lambda)-\mu)^2$, *and* $\bar r_S(\lambda,\mu)=\min\{r_S(\lambda,0)+\mu^2, 1+\lambda^2\}$. *For all* $\lambda>0$ *and* $\mu\in\mathbb{R}$,

$$\frac{1}{2}\bar r_S(\lambda,\mu)\le r_S(\lambda,\mu)\le\bar r_S(\lambda,\mu).$$

Suppose $\lambda_*\to\infty$ is not true. Then $\lambda_*\le c$ for some finite constant $c\ge 0$ (take a subsequence

if necessary). Then, from the definition of $F(\lambda_*)$ we have

$$F(\lambda_*) \geq (1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z,\lambda_*) \geq (1-\epsilon_n)\mathbb{E}\hat{\eta}_S^2(z,c) = \Omega(1) = \omega(\epsilon_n \nu_n^2),$$

which contradicts with (2.76). Further suppose $\mu_n - \lambda_* \to +\infty$ is not true. Then $\lambda_* \geq \mu_n - c$ for some finite constant c (take a subsequence if necessary). From Lemma 21 we obtain for large n,

$$F(\lambda_*) \geq \epsilon_n r_S(\lambda_*, \mu_n) \geq \frac{1}{2}\epsilon_n \min(\mu_n^2, \lambda_*^2) \geq \frac{1}{4}\epsilon_n \mu_n^2 = \omega(\epsilon_n \nu_n^2),$$

where we used $\mu_n = \omega(\sqrt{2\log \epsilon_n^{-1}}) = \omega(\nu_n)$. The same contradiction arises.

Step 2. We next claim that $\lambda_* = (1+o(1))\nu_n$. Otherwise, $\lambda_* = (c+o(1))\nu_n$ for some constant $c \neq 1$ (take a subsequence if necessary). For $c > 1$, given that we have proved $\lambda_* \to \infty, \mu_n - \lambda_* \to +\infty$, we can apply Lemma 20 to reach

$$F(\lambda_*) \geq \epsilon_n \left[\lambda_*^2 + 1 - \frac{(2+o(1))\mu_n}{(\mu_n - \lambda_*)^2}\phi(\mu_n - \lambda_*)\right] = \epsilon_n \lambda_*^2 (1+o(1)) = (c^2 + o(1)) \cdot \epsilon_n \nu_n^2.$$

This contradicts with (2.76). For $c < 1$, we have the same contradiction by applying Lemma 20 again:

$$F(\lambda_*) = \frac{4+o(1)}{\lambda_*^3}\phi(\lambda_*) + \epsilon_n \lambda_*^2 (1+o(1)) = \omega(\epsilon_n \nu_n^2).$$

Here, the last inequality holds because $\lambda_* \leq (1-\gamma)\nu_n$ for some constant $\gamma \in (0,1)$ when n is large, so that

$$\frac{1}{\lambda_*^3} e^{-\frac{\lambda_*^2}{2}} \geq \frac{1}{(1-\gamma)^3 \nu_n^3} e^{-\frac{(1-\gamma)^2}{2}\nu_n^2} = \epsilon_n \frac{1}{(1-\gamma)^3 \nu_n^3} e^{(\gamma - \frac{\gamma^2}{2})\nu_n^2} = \omega(\epsilon_n \nu_n^2).$$

Step 3. Finally, we prove that $\nu_n^2 - \lambda_*^2 = (6+o(1))\log \nu_n$. Suppose this is not true. Then $\nu_n^2 - \lambda_*^2 = (c+o(1))\log \nu_n$ for some $c \neq 6$ (take a subsequence if necessary). Since we have proved

$\lambda_* = (1 + o(1))\nu_n$, we can use the lower bound in Lemma 20 and simplify it to

$$F(\lambda_*) \geq C(\lambda_*) = \frac{(4 + o(1))\epsilon_n}{\sqrt{2\pi}} \frac{e^{\frac{1}{2}(\nu_n^2 - \lambda_*^2)}}{\lambda_*^3} + \epsilon_n\Big(\lambda_*^2 + 1 + o(1)\Big). \tag{2.77}$$

For the case $c > 6$, since

$$\frac{1}{\lambda_*^3} e^{\frac{1}{2}(\nu_n^2 - \lambda_*^2)} = e^{\frac{1}{2}(\nu_n^2 - \lambda_*^2 - 6\log\nu_n) + 3\log\frac{\nu_n}{\lambda}} = \nu_n^{\tilde{c}},$$

with $\tilde{c} = \frac{c-6+o(1)}{2} > 0$, (2.77) implies that

$$F(\lambda_*) \geq \Theta(\epsilon_n \nu_n^{\tilde{c}}) + \epsilon_n \nu_n^2 - (c + o(1))\epsilon_n \log \nu_n,$$

contradicting with (2.76). Regarding the case $c < 6$, (2.77) directly leads to

$$F(\lambda_*) \geq \epsilon_n \nu_n^2 - (c + o(1))\epsilon_n \log \nu_n + (1 + o(1))\epsilon_n.$$

No mater what value $c \in [-\infty, 6)$ takes, the above lower bound is larger than the upper bound in (2.76), resulting in the same contradiction.

Now that we have derived the accurate order information for λ_*: $\lambda_*^2 = \nu_n^2 - (6 + o(1)) \log \nu_n$, we can plug it into (2.77) to obtain the sharp lower bound:

$$F(\lambda_*) \geq \epsilon_n\Big(\nu_n^2 - (6 + o(1)) \log \nu_n\Big).$$

2.5.11 Proof of Proposition 7

Using the simple form of $\hat{\eta}_L(y, \lambda)$, the calculation is straightforward:

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_L(y, \lambda) - \theta\|_2^2 = \inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \sum_{i=1}^{n} \left(\frac{1}{1 + \lambda} y_i - \theta_i\right)^2$$

$$= \inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \sum_{i=1}^{n} \left[\left(\frac{\lambda}{1+\lambda} \right)^2 \theta_i^2 + \left(\frac{1}{1+\lambda} \right)^2 \sigma_n^2 \right] = \inf_{\lambda} \frac{\lambda^2 k_n \tau_n^2 + n\sigma_n^2}{(1+\lambda)^2} = \frac{n\sigma_n^2 \epsilon_n \mu_n^2}{1 + \epsilon_n \mu_n^2}.$$

Chapter 3: SNR-aware minimaxity in sparse linear regression

3.1 Introduction

Consider the linear regression model

$$y_i = x_i^T \beta + \sigma z_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

in which $y_i \in \mathbb{R}$ denotes the response, $x_i \in \mathbb{R}^p$ represents the feature or covariate vector, $\beta \in \mathbb{R}^p$ is the unknown signal vector to be estimated, and finally $z_1, \ldots, z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ are standard normal errors. We are interested in studying this problem for broad range of p considering p comparable with n, or even larger than n. To ease one of the major concerns that linear regression procedures remain inconsistent unless $p/n \to 0$, following the rich literature of sparse linear regression [24, 32, 34, 5, 6], we consider the sparsity structure of the signal in this paper. Specifically, we assume that the true regression coefficients are k-sparse:

$$\beta \in \Theta(k) := \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq k\}, \tag{3.2}$$

where $\|\beta\|_0$ denotes the number of non-zero components of β. In evaluating the performance of estimators, the minimax framework has been one of the most popular approaches, aiming to obtain an optimal estimator which has the best worst-case performance among all estimators. In other words, estimators are measured by the minimax risk:

$$R(\Theta(k), \sigma) := \inf_{\hat{\beta}} \sup_{\beta \in \Theta(k)} \mathbb{E}_\beta \|\hat{\beta} - \beta\|^2. \tag{3.3}$$

However, obtaining the exact minimax risk is mathematically challenging and has remained open. Hence, researchers have explored approaches that aim to "approximate" the minimax risk. One of these approaches is known as the rate optimal minimaxity. To witness the existing results and clarify the limitations, let us assume that the feature vectors satisfy $\{x_i\}_{i=1}^n \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{1}{n} I_p)$ and are independent with the noise errors $\{z_i\}_{i=1}^n$. The noise level $\sigma > 0$ may vary with the sample size n. By translating the result of [14] in this setting we obtain

$$R(\Theta(k), \sigma) \sim \sigma^2 k \log(p/k),$$

where the notation "$\sim$" means that as $n, p \to \infty$ and $(k \log(p/k))/n \to 0$, the ratio $R(\Theta(k), \sigma)/(k \log(p/k))$ remains bounded. Furthermore, it has been shown in the literature [13, 7, 14, 15] that many estimators, such as best subset selection [16, 17], Dantzig selector [18] and LASSO [19] achieve this rate-optimal minimax criteria, meaning that their risks (under optimal tuning) divided by $k \log(p/k)$ remain bounded[1].

Despite the rate-optimal minimaxity of the aforementioned estimators, extensive simulation results reported in [8, 20] have confirmed that when the signal-to-noise ratio (SNR) is low, all these estimators exhibit suboptimal performance and adding an ℓ_2-squared regularizer can improve the performance of the estimators. Hence, the rate-optimal minimax results lead to misleading guidelines for practitioners.

There could be two explanations for the mismatch between the rate-optimal minimax framework and the simulation studies:

- Explanation 1: As is clear, the rate optimal minimax result does not evaluate the minimax risk exactly. It ignores the constant in the minimax risk approximation and only captures the rate behavior in view of k and p for mathematical simplicity. It is possible that if we calculate the exact maximum risk for estimators, the differences between constants can explain the discrepancies between the simulation studies and the rate-optimal minimax results.

[1] In some of these results, the risk is stated with high probability and the rate is $k \log p$ instead of $k \log(p/k)$.

- Explanation 2: It could be that since the minimax framework only focuses on the spots of the parameter space that are hard for the estimation problem, its theoretical implications will be different from the simulation studies. Hence, the framework needs to be amended to provide more informative results.

To pinpoint the correct explanation for the discrepancy between the minimax studies and simulation results, [21] considered the asymptotic framework $n, p \to \infty$, $k/p \to 0$ and $(k \log p)/n \to 0$ and tried to find a better approximation of the minimax risk. The result in [21] is based on the Sorted L-One Penalized Estimator (SLOPE) introduced in [35]. For $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, the SLOPE is defined as the solution of

$$\hat{\beta}_{SLOPE} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|^2 + \lambda_{(1)}|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)},$$

where $|b|_{(1)} \geq |b|_{(2)} \geq \ldots |b|_{(p)}$ are the order statistics of $|b_1|, |b_2|, \ldots, |b_p|$. The following result from [21] aims to provide a better approximation of the minimax risk.

Theorem 12 (Theorem 1.2 & 1.3 in [21])**.** *Assume model* (3.1) *with random Gaussian designs* $\{x_i\}_{i=1}^n \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{1}{n} I_p)$ *and parameter space* (3.2)*. Suppose* $k/p \to 0$ *and* $(k \log p)/n \to 0$*. For any* $\epsilon > 0$*,*

$$\inf_{\hat{\beta}} \sup_{\beta \in \Theta(k)} P\left(\frac{\|\hat{\beta} - \beta\|^2}{2\sigma^2 k \log(p/k)} > 1 - \epsilon \right) \to 1.$$

In addition, fix $0 < q < 1$ *and set* $\lambda_i = \sigma(1 + \epsilon)\Phi^{-1}(1 - iq/2p)$*, where* Φ *is the CDF of standard Gaussian. Then, the SLOPE achieves the above asymptotic minimax risk in that:*

$$\sup_{\beta \in \Theta(k)} P\left(\frac{\|\hat{\beta}_{SLOPE} - \beta\|^2}{2\sigma^2 k \log(p/k)} > 1 + 3\epsilon \right) \to 0.$$

In view of the above theorem, we have the following comments.

Remark 2. *Theorem 12 does not exactly characterize the minimax risk and the minimax estimator. However, in spirit, it is similar to the minimax result. Intuitively speaking, it can be interpreted*

in that as $n, p \to \infty$ and $k/p \to 0$, the minimax risk is approximately $2\sigma^2 k \log(p/k)$, and that SLOPE achieves the minimax risk. Hence, as the first contribution of this paper, we prove in the following theorem for one part of this intuitive statement. Besides, compared to the probabilistic statement in Theorem 12, we proved in the conventional form of minimax risk as in (3.3)*, which alleviates some of the concern that under unrealized rare events, risks of estimators become unbounded.*

Theorem 13. *Assume model* (3.1) *and parameter space* (3.2)*. Suppose $n, p \to \infty$. If $k/p \to 0$ and $\big(k \log p\big)/n \to 0$, then the minimax risk defined in* (3.3) *satisfies*

$$R(\Theta(k), \sigma) = 2\sigma^2 \cdot k \log(p/k)\Big(1 + o(1)\Big).$$

Remark 3. *Compared to the rate-optimal minimax results, Theorem 12 has the advantage of characterizing the constant of the minimax risk accurately. This is confirmed by Theorem 13 that the constant is indeed for the minimax risk defined in* (3.3)*. However, it still suffers from the same issue as the rate-optimal minimax risk. The same estimator is optimal irrespective of the signal-to-noise ratio. This implies that Explanation 1 is not the proper reasoning.*

As will be clarified in this paper, there are two main issues causing the discrepancy between the theoretical and simulation results: (1) Since we do not impose any constraint on the signal strength, the minimax framework only focuses on a particular signal-to-noise ratio that makes the estimation problem the hardest. Hence, the factor of SNR affecting practical results is masked by the minimax framework. (2) The approximations we obtain for the minimax risk in rate-optimal minimax framework, and even in Theorem 12 are not accurate enough for distinguishing performances of different estimators and hence more accurate approximations are required for this purpose.

To address the first issue, we incorporate the notion of SNR into the minimax framework, and introduce the notion of SNR-aware minimaxity. We will discuss this framework in Section 3.2.

In view of the second issue of current minimax results, we will consider and analyze a higher-order expansion of the minimax risk. As will be clarified later, these two changes create a more

insightful minimax framework that can offer results consistent with the simulation studies performed elsewhere.

3.2 SNR-aware minimaxity

As discussed in the previous section, one of the main reasons of the minimax framework to produce misguidance for practitioners, is that the signal strength is not controlled and hence the minimax framework sets the signal strength to a level that makes the estimation problem the hardest. As a result, the framework in an indirect way becomes blind to the changes in the signal-to-noise ratio. To develop the SNR-aware minimax framework, we start by inserting a notion of signal-to-noise ratio in the minimax setting. To this end, we consider the following SNR-aware parameter space:

$$\Theta(k, \tau) := \left\{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq k,\ \|\beta\|_2^2 \leq k\tau^2\right\}. \tag{3.4}$$

The new parameter introduced in this model, i.e. τ is a measure of signal strength. Compared to the basic sparse parameter space in (3.2), $\Theta(k, \tau)$ can monitor the changes in SNR. Hence, the minimax framework we develop with this parameter space can reveal the impact of the SNR on the sparse linear regression problem.

Given this new parameter space, the corresponding minimax risk is defined as

$$\mathbb{R}(\Theta(k, \tau), \sigma) := \inf_{\hat{\beta}} \sup_{\beta \in \Theta(k,\tau)} \mathbb{E}_\beta \|\hat{\beta} - \beta\|^2. \tag{3.5}$$

As discussed in the last section, characterizing the exact minimax risk for $R(\Theta(k), \sigma)$ is mathematically hard. It is even more challenging to obtaining exact $\mathbb{R}(\Theta(k, \tau), \sigma)$. Hence, we aim to find accurate approximations for this quantity. Following the approach proposed for sparse linear regression problems [12, 21], we consider the sparsity parameter defined as

$$\epsilon := \frac{k}{p},$$

and assume that $\epsilon \to 0$ as $n, p \to \infty$. Given that we have also introduced the notion of signal strength to our framework, we expect the SNR level, defined as

$$\mu := \frac{\tau}{\sigma},$$

to affect the final results as well. Specifically, we aim to study $R(\Theta(k, \tau), \sigma)$ for different values of (ϵ, μ). Due to the mathematical challenges in identifying exact minimax risk, we focus on obtaining asymptotic minimaxity, and consider the following regimes: as $n, p \to \infty$,

Regime (I) Low signal-to-noise ratio: $\mu \to 0$, $\epsilon \to 0$;

Regime (II) Moderate signal-to-noise ratio: $\mu \to \infty$, $\epsilon \to 0$, $\mu = o(\sqrt{\log \epsilon^{-1}})$;

Regime (III) High signal-to-noise ratio: $\epsilon \to 0$, $\mu = \omega(\sqrt{\log \epsilon^{-1}})$.

As will be discussed later in the paper, each regime exhibits unique minimaxity, and distinct minimax estimators emerge in different regimes. But before that, we first derive first-order asymptotic result similar as the classical one and reveal its limitations in the SNR-aware minimax setting.

3.2.1 First-order asymptotics

Theorem 14. *Assume model* (3.1) *and parameter space* (3.4)*. The following hold:*

- *Regime* (I)*: When* $k/p \to 0$*,* $(k \log(p/k))/n \to 0$ *and* $\mu = \tau/\sigma \to 0$*,*

$$R(\Theta(k, \tau)) = k\tau^2\Big(1 + o(1)\Big).$$

- *Regime* (II)*: When* $k/p \to 0$*,* $(k \log(p/k))/n \to 0$ *and* $\mu = \tau/\sigma = o\Big(\sqrt{\log(p/k)}\Big)$*,*

$$R(\Theta(k, \tau)) = k\tau^2\Big(1 + o(1)\Big).$$

- *Regime (III): When* $k/p \to 0$, $(k\log(p/k))/n \to 0$ *and* $\mu = \tau/\sigma = \omega\Big(\sqrt{\log(p/k)}\Big)$,

$$R(\Theta(k,\tau)) = k\sigma^2 \cdot 2\log(p/k)\Big(1 + o(1)\Big).$$

One of the main issues in the above theorem is that the first-order asymptotic approximation of minimax risk does not seem to always offer accurate information. For example, as the signal-to-noise ratio significantly increases from Regime (I) to Regime (II), the first-order analysis falls short of capturing any difference and continues to generate the naive zero estimator as the optimal one.[2] Moreover, in Regime (III), the result is indistinguishable with the minimax result unconscious to the SNR as in Theorem 12. In the next section, we push the analysis one step further to develop second-order asymptotics. This refined version of the SNR-aware minimax analysis will provide a much more accurate approximation of the minimax risk, and can provide more useful information and resolve the confusing aspects of the first-order results presented above.

3.2.2 Second-order asymptotics

We first demonstrate the result in Regime (I). As discussed in previous section, the first order approximation of $R(\Theta(k,\tau))$ is $k\tau^2$. Indeed, this is the exact superemum risk of zero estimator achieved at the boundary of $\Theta(k,\tau)$. This seems to suggest when the signal-to-noise ratio is low, no other estimator can outperform the naïve estimator. However, we will show this conclusion is hasty when we go to higher order analysis. In fact, consider the ridge estimator [36] defined as: for $\lambda > 0$, let

$$\hat{\beta}^R(\lambda) := \operatorname*{arg\,min}_{b\in\mathbb{R}^p} \|y - Xb\|_2^2 + \lambda\|b\|_2^2.$$

The following theorem indicates that up to second order approximation, the ridge estimator is asymptotically minimax.

Theorem 15. *Assume model* (3.1) *and parameter space* (3.4). *Suppose* $n, p \to \infty$ *and* $k/p \to 0$

[2]The zero estimator has the exact risk of $k\tau^2$, referring to the proof in Section 3.4.4.

and $k/n \to 0$. *In Regime* (I) *where* $\mu = \tau/\sigma \to 0$, *the minimax risk defined in* (3.5) *satisfies*

$$R(\Theta(k,\tau),\sigma) = k\tau^2\Big(1 - \frac{k\mu^2}{p}\big(1+o(1)\big)\Big).$$

In addition, the ridge estimator $\hat{\beta}^R$ *with tuning* $\lambda = p/(k\mu^2)$ *is asymptotically minimax up to the second order term, i.e.*

$$\sup_{\beta\in\Theta(k,\tau)} \mathbb{E}_\beta\|\hat{\beta}^R(\lambda) - \beta\|^2 = k\tau^2\Big(1 - \frac{k\mu^2}{p}\big(1+o(1)\big)\Big).$$

The proof of this theorem can be found in Section 3.4.5. The direct result of this theorem implies that the naïve zero estimator is sub-optimal because its exact superemum risk only corresponds with the first order of the minimax. In addition, note that the Gaussian sequence model is a special case of the linear regression model, the simulation results in Chapter 2 Section 2.3 is relevant to the discussion here. As indicated by Figure 2.2 and 2.3, when the SNR level is low, the ridge estimator (equals the linear estimator in Gaussian sequence model case) outperforms other estimators on the plot. This corresponds with the conclusion of Theorem 15 here.

Along the discussion, the next theorem aims to obtain second-order approximation of $R(\Theta(k,\tau))$ in Regime (II). However, as we present of proof of Theorem 13, even the first-order upper bound of the classical $R(\Theta(k))$ is not trivial to be obtained. Obtaining the upper bound up to second-order of the extra constrained $R(\Theta(k,\tau))$ is even more challenging. We leave the upper bound proof in the following theorem to future work of studies.

Theorem 16. *Assume model* (3.1) *and parameter space* (3.4). *Suppose* $n, p \to \infty$ *and* $k/p \to 0$ *and* $(k\log(p/k))/n \to 0$. *In Regime* (II) *where* $\mu = \tau/\sigma \to \infty$ *and* $\mu = o(\sqrt{\log(p/k)})$, *additionally assuming* $\mu^4/n \to 0$, *then the minimax risk defined in* (3.5) *satisfies*

$$R(\Theta(k,\tau)) \geq k\tau^2\Big(1 - \frac{k\mu^2}{2p}\cdot e^{\mu^2}\big(1+o(1)\big)\Big).$$

The proof of this theorem can be found in Section proof:sec:linear-model-second-order-med-

snr. Combining this theorem with Theorem 14 in Regime (II), we obtain that $k\tau^2\Big(1 - \frac{k\mu^2}{2p} \cdot e^{\mu^2}\big(1 + o(1)\big)\Big) \leq R(\Theta(k,\tau)) \leq k\tau^2\big(1 + o(1)\big)$. As we expected, the SNR-aware minimax result up to second-order approximation are consistent with the results in Gaussian sequence model in Chapter 2. This confirms our method in studying minimax problem of sparse estimation. The obtained results already show that the underlying SNR level intrinsically affects the minimax result and the optimality of estimators under which.

Remark 4.

In Theorem 16, besides the asymptotic setting in Regime (II)*, we made another assumption that* $\mu^4/n \to 0$*. Note that this additional assumption does not exclude too much region of* (ϵ, μ) *from Regime* (II)*. This is because by original assumption of Regime* (II)*,* $\mu = o(\sqrt{\log(p/k)})$ *already implies that* $\mu^2 \ll \log(p/k) \ll n$*, it is of large possibility that* $\mu^4 \ll n$ *is also satisfied in practical setting.*

3.3 Discussions

3.3.1 Summary

The estimation problem in sparse linear regression is more challenging compared to Gaussian sequence model. Along the studies, researchers have developed results mostly stated in rate-minimaxity. The accurate constant of the classical minimax is still hard to be characterized. Based on this, many estimators including Lasso and best subset are proved to achieve the rate-optimality. However, as revealed by empirical studies in extensive research, Lasso and best subset exhibits sub-optimally in different SNR settings compared to each other. This raises the discrepancy between the theoretical results and the simulation implications. To mitigate this gap, we first provide the first-order approximation of the classical minimax with accurate constant. However, this is still insufficient to explain the discrepancy as the the classical minimax framework will output the same minimax estimator irrespective of the SNR levels, which is against the empirical findings. This calls for the enhancement of the current minimax framework to let it incorporate the

information of the important factor – SNR into the framework. Along this line, we introduce the SNR-aware minimaxity which adds additional control of the SNR level in the parameter space and monitors the minimax risk accordingly. As we introduced in Chapter 2, we split the SNR level in three different regimes. We first obtain the first-order approximations of the SNR-aware minimax in all three regimes. Then we show that in low and moderate SNR regimes, the first-order approximations are the same and can be achieved by the zero estimator. Then, we go to the second-order analysis. We show that the asymptotic minimax estimator in low SNR regime is actually ridge estimator up to second-order minimax. However, the second-order analysis in moderate and high SNR regimes are still yet to be completed. The remaining difficulty is to find the minimax estimators in these regimes to obtain the upper bounds. So far, we have shown that the low SNR regime and the moderate SNR regime lower bound results are correspondence with that in Gaussian sequence model. The obtained results already demonstrate that the SNR level intrinsically affects the minimax results in sparse linear regression problem and the corresponding optimality of estimators. The analysis of the SNR-aware minimax framework provides new perspectives of the sparse estimation and more practical guidance for empirical studies.

3.3.2 Future research

- This thesis provides the second-order approximations for the SNR-aware minimax in low SNR regime and lower bound second-order approximation in moderate SNR regime. The second-order upper bound in moderate SNR regime and the approximation in high SNR regime are still missing. It will complete the work if these results are obtained.

- This thesis analyzes the linear regression model under the Gaussian random design of feature matrix X. It will be interesting to study the minimax as well as the SNR-aware minimax problem for other common designs, e.g., the fixed design under additional assumptions, the correlated design with the correlation of i-th and j-th column of X in the form $\rho^{|i-j|}$ and e.t.c..

- The classical minimax result in Theorem 13 does not reveal a single estimator to be asymptically minimax such as Lasso or best subset. This does not eliminate the possibility that one of them can attain the asymptotic minimax. It leaves us to think about more efficient proof method to obtain more accurate risk upper bound for each estimator.

- The current results rely on the assumptions that $k/p \to 0$ and $(k \log p)/n \to 0$. This characterizes the sparse signals and high dimensional asymptotic when $p \ll e^n$. It will be interesting to explore the topic when the signal is denser $k/p \geq c$ and ultra high dimensional setting $n = O(\log p)$.

3.4 Proofs of the main results

Throughout the proof sections, we adopt the following notations: We use the uppercase alphabets for matrices and lowercase alphabets for vectors. $\|\cdot\|_l$ is the L^l norm on the vector space. If without any subscript, the default $\|\cdot\|$ stands for the L^2 norm. Sometimes, we denote $[p] = \{1, \ldots, p\}$ for tidiness. In the proof sections of lower bounds, for $i \in [p]$, let x_i denote the i-th column of the matrix $X \in \mathbb{R}^{n\times p}$; let $x_{i,1} \in \mathbb{R}$ and $x_{i,-1}\mathbb{R}^{n-1}$ denote the first and the rest of coordinates of column vector x_i.

3.4.1 Preliminaries

Scale invariance

The minimax risks defined in Equations (3.3) and (3.5) of the main text have the following scale invariance property

$$R(\Theta(k), \sigma) = \sigma^2 \cdot R(\Theta(k), 1),$$
$$R(\Theta(k, \tau), \sigma) = \sigma^2 \cdot R(\Theta(k, \mu), 1),$$

where we recall that $\mu = \tau/\sigma$. This can be easily verified by rescaling the linear regression model to have unit variance.

Preliminary probability results

Lemma 22 (Weak law for triangular arrays, Theorem 2.2.6 in [37])**.** *For each n let $X_{n,k}$, $1 \le k \le n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{X}_{n,k} = X_{n,k}\mathbb{1}_{(|X_{n,k}| \le b_n)}$. Suppose that as $n \to \infty$*

(i) $\sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0$, *and*

(ii) $b_n^{-2} \sum_{k=1}^{n} \mathbb{E}\bar{X}_{n,k}^2 \to 0$.

If we let $S_n = X_{n,1} + \ldots + X_{n,n}$ and put $a_n = \sum_{k=1}^{n} \mathbb{E}\bar{X}_{n,k}$ then

$$(S_n - a_n)/b_n \to 0 \text{ in probability.}$$

The following lemma is stated in Corollary 4.2.13 in [38].

Lemma 23 (Covering number of the unit sphere)**.** *The covering numbers of the unit Euclidean sphere S^{n-1} satisfy, for any $\epsilon \in (0, 1]$, we have*

$$\mathcal{N}(S^{n-1}, \epsilon) \le \left(\frac{3}{\epsilon}\right)^n.$$

For $\epsilon > 1$, the unit sphere can be covered by just one ϵ-ball, so $\mathcal{N}(S^{n-1}, \epsilon) = 1$.

The following lemma states a simple concentration for χ distribution. The proof follows from the concentration of the Lipschitz function of Gaussians (Theorem 2.26 in [5]) and that the ℓ_2 norm is 1-Lipschitz function of a Gaussian vector.

Lemma 24. *Let $z_1, \ldots, z_n \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$, then for every $t \ge 0$,*

$$P\left(\|z\| \le (1 + t)\sqrt{n}\right) \ge 1 - e^{-\frac{nt^2}{2}}.$$

The following result is Lemma 2 of [39].

Lemma 25 (χ^2-concentration)**.** *Fix $\tau > 0$, and let $Z_i \sim \mathcal{N}(0,1)$, $i = 1, \ldots, d$. Then,*

$$P\Big(\sum_{i=1}^{d} Z_i^2 < d(1-\tau)\Big) \leq e^{\frac{d}{2}\left(\tau+\log(1-\tau)\right)},$$

and

$$P\Big(\sum_{i=1}^{d} Z_i^2 > d(1+\tau)\Big) \leq e^{-\frac{d}{2}\left(\tau-\log(1+\tau)\right)}.$$

Let $\chi_p^2(\lambda)$ denote the non-central chisquare of degrees of freedom p and the noncentrality parameter λ, we have:

Lemma 26 (Non-central chisquare, Theorem 3 & 4 in [40])**.** *Suppose $X \sim \chi_p^2(\lambda)$. Then*

(i) for $c > 0$, $P(X > p + \lambda + c) \leq \exp\left[-\frac{pc^2}{4(p+2\lambda)(p+2\lambda+c)}\right]$;

(ii) for $0 < c < p + \lambda$, $P(X < p + \lambda - c) \leq \exp\left[-\frac{pc^2}{4(p+2\lambda)^2}\right]$.

The following lemma can be proved following Exercise 2.5.10 in [38].

Lemma 27 (L^1 bound of the maximum of sub-gaussians)**.** *Let $X_1, X_2, \ldots,$ be an infinite sequence of sub-gaussian random variables which are note necessarily independent. Let $K = \max_i \|X_i\|_{\psi_2}$ be the maximum sub-gaussian norm. Then for every $N \geq 2$ we have*

$$\mathbb{E} \max_{i \leq N} |X_i| \leq CK\sqrt{\log N}.$$

The following lemma states the Gaussian maxima result in tail probability.

Lemma 28 (Concentration of the maximum of Gaussians)**.** *Let $\zeta_1, \ldots, \zeta_p \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. For all $u > 0$, we have*

$$P\Big(\max_{1 \leq i \leq p} |\zeta_i| \geq \sqrt{2\log p} + u\Big) \leq e^{-\frac{u^2}{2}}.$$

Proof. Using union bound and Gaussian tail bound, for $u \geq 1$,

$$P\big(\max_{1\leq i\leq p}|\zeta_i| \geq u\big) \leq \sum_{i=1}^{p} P\big(|\zeta_i| \geq u\big) \leq 2p\frac{1}{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \leq \frac{2p}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \leq pe^{-\frac{u^2}{2}}.$$

Let $u = \sqrt{2(\log p + t)}$, then $\sqrt{2(t+\log p)} \geq 1$ for all $t > 0$ since $p \geq 2$,

$$P\Big(\max_{1\leq i\leq p}|\zeta_i| \geq \sqrt{2(t+\log p)}\Big) \leq e^{-t}.$$

Thus,

$$P\Big(\max_{1\leq i\leq p}|\zeta_i| \geq \sqrt{2\log p} + u\Big) \leq e^{-\frac{u^2}{2}-u\sqrt{2\log p}} \leq e^{-\frac{u^2}{2}}.$$

□

In proving most of the upper bound of our results, we will constantly use the concentration of the Gaussian order statisics. We construct a Bernstein-type tail bound based on the exponential Efron-Stein inequality for order-statistics. The following result is from [41].

Lemma 29 (Theorem 2.9 in [41])**.** *Let $X_1, \ldots, X_p$ be independently distributed according to F, let $X_{(1)} \geq \ldots \geq X_{(p)}$ be the order statistics and let $\Delta_k = X_{(k)} - X_{(k+1)}$ be the k^{th} spacing. Then for $t \geq 0$ and $1 \leq k \leq p/2$,*

$$\log \mathbb{E}e^{t(X_{(k)}-\mathbb{E}X_{(k)})} \leq t\frac{k}{2}\mathbb{E}[\Delta_k(e^{t\Delta_k}-1)].$$

Based on the above lemma, we show the following properties of Gaussian order statistics concentration:

Lemma 30. *Let $X_1, \ldots, X_p \overset{i.i.d}{\sim} \mathcal{N}(0,1)$.*

(i) For all $2 \leq k \leq p$, $\mathbb{E}|X|_{(i)} \leq \sqrt{2\log(2p/i)}$.

(ii) For all $u > 0$, $1 \le k \le p/2$, and some constant $c > 0$,

$$P\Big(|X|_{(k)} - \mathbb{E}|X|_{(k)} \ge u\Big) \le \exp\Big(-k\Big(\frac{u}{c} \wedge \Big(\frac{u}{c}\Big)^2\Big)\Big).$$

Proof. Prove (i). Since $\frac{1}{m} \le \int_{m-1}^{m} \frac{1}{x} dx = \log m - \log(m-1)$, $\sum_{m=i}^{p} \frac{1}{m} \le \log p - \log(i-1) \le \log(p/i)$.

$$\mathbb{E}|X|_{(i)}^2 = \mathbb{E}\big[\Phi^{-1}(1 - U_{(i)}/2)\big] \le \mathbb{E}\big[2\log\frac{2}{U_{(i)}}\big] = 2\log 2 + 2\sum_{j=i}^{p}\frac{1}{j} \le 2\log\frac{2p}{i},$$

where $U_{(i)}$ is the i-th largest among p i.i.d. uniform random variables. Then (i) follows from $\big(\mathbb{E}|X|_{(i)}\big)^2 \le \mathbb{E}|X|_{(i)}^2$.

Prove (ii). Let the C.D.F of the absolute value of the standard normal be $\tilde{\Phi}$, then $\tilde{\Phi}(x) = 2\Phi(x) - 1$. The C.D.F. of Exp(1) is $1 - e^{-x}$. Let $Y_1, \ldots, Y_p \overset{i.i.d}{\sim} \text{Exp}(1)$ and $Y_{(1)} \ge Y_{(2)} \ge \ldots, Y_{(p)}$. Then

$$|X|_i \sim \tilde{\Phi}^{-1}(1 - e^{-Y_i}) = \Phi^{-1}(1 - \frac{1}{2}e^{-Y_i}) \Rightarrow |X|_{(i)} \sim \Phi^{-1}(1 - \frac{1}{2}e^{-Y_{(i)}}).$$

Let $\tilde{U}(t) := \Phi^{-1}(1 - \frac{1}{2t})$, $|X|_{(i)} \sim \tilde{U}(e^{Y_{(i)}})$. The spacing of exponentials satisfies $Y_{(k)} - Y_{(k+1)} \sim \frac{E_1}{k}$, where $E_1 \sim \text{Exp}(1)$ is independent of all Y_i's and X_i's. Thus,

$$\Delta_k := |X|_{(k)} - |X|_{(k+1)} \sim \tilde{U}\big(e^{\frac{E_1}{k} + Y_{(k+1)}}\big) - \tilde{U}\big(e^{Y_{(k+1)}}\big).$$

By Proposition 4.1 in [41], the property of $\tilde{U} \circ \exp$ satisfies

$$\Delta_k \le \frac{\sqrt{2}E_1}{k\sqrt{\log 2 + Y_{(k+1)}}}.$$

The integration of the exponential has, for $0 \le \mu < 1/2$,

$$\int_0^\infty \mu x(e^{\mu x} - 1)e^{-x}dx = \frac{\mu^2(2-\mu)}{(1-\mu)^2} \le \frac{2\mu^2}{1-2\mu}.$$

Thus,

$$\mathbb{E}\left[t\Delta_k(e^{t\Delta_k}-1)\,|Y_{(k+1)}\right] \le \frac{4t^2}{k^2(\log 2+Y_{(k+1)})}\frac{1}{1-\frac{2\sqrt{2}t}{k\sqrt{\log 2+Y_{(k+1)}}}} \le \frac{4t^2}{k^2\log 2}\frac{1}{1-\frac{2\sqrt{2}t}{k\sqrt{\log 2}}}.$$

Therefore,

$$\log\left(\mathbb{E}e^{t(|X|_{(k)}-\mathbb{E}|X|_{(k)})}\right) \le \frac{tk}{2}\mathbb{E}[\Delta_k(e^{t\Delta_k}-1)] \le \frac{2t^2}{k\log 2}\frac{1}{1-\frac{2\sqrt{2}t}{k\sqrt{\log 2}}}.$$

Let $v_k = \frac{4}{\log 2\cdot k}$, then

$$\log\left(\mathbb{E}e^{t(|X|_{(k)}-\mathbb{E}|X|_{(k)})}\right) \le \frac{v_k t^2}{2(1-t\sqrt{2v_k/k})}.$$

The Bernstein inequality follows

$$P\left(|X|_{(k)}-\mathbb{E}|X|_{(k)} \ge \sqrt{2v_k t}+\sqrt{2v_k/k}t\right) \le e^{-t}.$$

Write it in another form, we obtain (ii). □

3.4.2 Proof of lower bound in Theorem 13

As discussed in Section 3.4.1, the minimax risk in (3.3) is scale-invariant of noise variance. Hence, without loss of generality, we prove the theorem under $\sigma = 1$ in model (3.1).

Suppose that we have a prior distribution π on the regression coefficients whose support is contained in $\Theta(k)$. For any estimator $\hat{\beta}$, it is straightforward to see that

$$\mathbb{E}_\pi\|\hat{\beta}-\beta\|^2 \le \sup_{\beta\in\Theta(k)}\mathbb{E}\|\hat{\beta}-\beta\|^2, \tag{3.6}$$

where the expectation on the left is with respect to the randomness in (X, z, β), while the expectation on the right is with respect to (X, z) only. Let $B(\pi)$ be the Bayes risk of π for squared loss.

By taking the infimum with respect to $\hat{\beta}$, we conclude that

$$B(\pi) \leq \inf_{\hat{\beta}} \sup_{\beta \in \Theta(k)} \mathbb{E}\|\hat{\beta} - \beta\| = R(\Theta(k), 1). \tag{3.7}$$

Hence, in order to obtain a lower bound for the minimax risk, one can use a particular prior distribution and calculate the Bayes risk for that specific distribution.

The proof of lower bound of Theorem 13 relies on the independent block prior, which was once introduced in Chapter 8.6 of [3]. The independent block prior $\pi_{IB}(\lambda; p, k)$ [3, 33] is constructed in the following way: divide $(1, \ldots, p)$ into k blocks, for block j, randomly select an index $I_j \in \{(j-1)m + 1, \ldots, j \cdot m\}$, $m = \lceil p/k \rceil$ and set $\beta^{(j)} = (\beta_{(j-1)m+1}, \ldots, \beta_{jm}) = \lambda e_{I_j}$. The selection between different blocks are independent. Note that the spike choice $\lambda = \lambda_{n,p,k}$ can depend on n, p, k. Sometimes for notational simplicity, throughout the proof, we drop the dependency of λ on the asymptotic parameters without ambiguity.

The following proposition states a lower bound of the Bayes risk under π_{IB} if the spike λ is below some threshold. This proposition is sufficient to provide a lower bound matching with the upper bound up to first order asymptotics. In fact, let $\lambda = \sqrt{2\log(p/k)} \cdot \big(1 + o(1)\big)$. Then Proposition 1 indicates that

$$\mathbb{E}_\pi\|\hat{\beta}_\pi - \beta\|^2 \geq 2k\sigma^2 \log(p/k)\big(1 + o(1)\big).$$

Indeed, the condition of choice of λ in Proposition 1 is weaker, covering a broader bandwidth of λ from near zero to as large as $\sqrt{2\log(p/k)} \to \infty$. We will discuss in Section 3.4.4 that this relaxation will provide a more general lower bound for even SNR-aware minimax.

Proposition 1. *Assume model* (3.1) *and suppose* $(\log(p/k))/n \to 0$ *and* $p/k \to \infty$. *Let* $\lambda > 0$ *and* $\pi := \pi_{IB}(\lambda; p, k)$ *be the independent block prior of* β. *Denote* $\hat{\beta}_\pi$ *as the Bayesian estimator (posterior mean) under* π. *For* $\lambda > 0$ *satisfying* $\sqrt{2\log(p/k)} - \lambda \to +\infty$, *we have*

$$\mathbb{E}_\pi\|\hat{\beta}_\pi - \beta\|^2 \geq k\sigma^2\lambda^2\left(1 + o(1)\right).$$

Proof. Because of the scalability of the risk function, described in Section 3.4.1, we can prove the conclusion for $\sigma = 1$ without loss of generality. Hence, in the rest of the proof, we assume that $\sigma = 1$. Given that we have used the independent block prior, the Bayes risk satisfies the following property:

$$\mathbb{E}_\pi \|\hat{\beta}_\pi - \beta\|^2 = k\mathbb{E}_\pi \|\hat{\beta}_\pi^{(1)} - \beta^{(1)}\|^2, \tag{3.8}$$

where $\beta^{(1)}$ denotes a part of β whose coordinates belong to the first block. Similarly, $\hat{\beta}_\pi^{(1)}$ denotes a part of the posterior mean $\hat{\beta}_\pi$ whose indices correspond to the first block.

In the rest of the proof, we will also use the notation $\beta^{(-1)}$ to denote the part of β whose indices do not belong to the first block. As a result, we have $\beta = (\beta^{(1)}, \beta^{(-1)})$. We will also use the notation $\tilde{y} = y - X^{(-1)}\beta^{(-1)}$ in which $X^{(-1)}$ is a subset of matrix X whose column indices do not belong to the first block. We have

$$\tilde{y} = y - X^{(-1)}\beta^{(-1)} = X^{(1)}\beta^{(1)} + z.$$

Hence, we can conclude that $(\tilde{y}|X^{(1)}) \sim \mathcal{N}(X^{(1)}\beta^{(1)}, I_n)$. As is clear from (3.8), to obtain a lower bound for the Bayes risk, we need to find a lower bound for $\mathbb{E}_\pi \|\hat{\beta}_\pi^{(1)} - \beta^{(1)}\|^2$. We have

$$\begin{aligned}
& \mathbb{E}_\pi \|\mathbb{E}_\pi \left(\beta^{(1)} \middle| y, X\right) - \beta^{(1)}\|^2 \\
= \; & \mathbb{E}_\pi \|\mathbb{E}_{\beta^{(-1)}} \left[\mathbb{E}_\pi \left(\beta^{(1)} \middle| y, X, \beta^{(-1)}\right)\right] - \beta^{(1)}\|^2 \\
\overset{(a)}{\geq} \; & \mathbb{E}_\pi \|\mathbb{E}_\pi \left(\beta^{(1)} \middle| y, X, \beta^{(-1)}\right) - \beta^{(1)}\|^2 \\
\overset{(b)}{=} \; & \mathbb{E}_\pi \|\mathbb{E}_\pi \left(\beta^{(1)} \middle| \tilde{y}, X^{(1)}, X^{(-1)}, \beta^{(-1)}\right) - \beta^{(1)}\|^2 \\
\overset{(c)}{=} \; & \mathbb{E}_\pi \|\mathbb{E}_\pi \left(\beta^{(1)} \middle| \tilde{y}, X^{(1)}\right) - \beta^{(1)}\|^2.
\end{aligned} \tag{3.9}$$

To obtain Inequality (a), we note that if allowing additional condition on $\beta^{(-1)}$, the posterior mean $\mathbb{E}_\pi(\beta^{(1)} \; | y, X, \beta^{(-1)})$ minimizes the expected squared loss compared to other functions of $(y, X, \beta^{(-1)})$. Equality (b) is due to the fact that by knowing the values of $y, X, \beta^{(-1)}$ we can calculate $\tilde{y}, X^{(1)}, X^{(-1)}, \beta^{(-1)}$ and vice versa. Finally, Equality (c) is due to the fact that $(X^{(-1)}, \beta^{(-1)})$ are independent of $\beta^{(1)}$ and $(\tilde{y}, X^{(1)})$.

As is clear from (3.9) to obtain a lower bound, we have to find a lower bound for $\mathbb{E}_\pi \|\mathbb{E}_\pi\left(\beta^{(1)}\middle|\tilde{y}, X^{(1)}\right) - \beta^{(1)}\|^2$. Focusing on the first block for $\beta^{(1)} \in \mathbb{R}^m$, the independent block prior is reduced to the single spike prior $\pi_S(\lambda; m)$ [3] defined as: select an index $I \in \{1, \ldots, m\}$ uniformly at random and set $\beta = \lambda e_I \in \mathbb{R}^m$. The following lemma provides a lower bound for $\mathbb{E}_\pi \|\mathbb{E}_\pi\left(\beta^{(1)}\middle|\tilde{y}, X^{(1)}\right) - \beta^{(1)}\|^2$ based on the single spike prior.

Lemma 31. *Consider model* (3.1) *with* $\sigma = 1$ *and* $\beta \in \mathbb{R}^m$. *Suppose* $n, m \to \infty$ *and* $(\log m)/n \to 0$. *Let* $\pi = \pi_S(\lambda; m)$ *be the single spike prior of* β. *Denote* $\hat{\beta}_\pi$ *as the Bayesian estimator under* π. *For* $\lambda > 0$ *satisfying* $\sqrt{2\log m} - \lambda \to \infty$, *we have*

$$\mathbb{E}_\pi \|\hat{\beta}_\pi - \beta\|^2 \geq \lambda^2(1 + o(1)).$$

We can then conclude from Lemma 31 that

$$\mathbb{E}_\pi \|\mathbb{E}_\pi\left(\beta^{(1)}\middle|\tilde{y}, X^{(1)}\right) - \beta^{(1)}\|^2 \geq \lambda^2(1 + o(1)).$$

Thus, combining this equation with (3.8) and (3.9) proves

$$\mathbb{E}_\pi \|\hat{\beta}_\pi - \beta\|^2 \geq k\lambda^2\left(1 + o(1)\right).$$

□

Proof of Lemma 31

Under 1-spike prior, denote I by the index of the spike coordinate and the posterior probability by $p_i(y, X) = P(I = i|y, X)$, $i \in [p]$, we have

$$\begin{aligned} \mathbb{E}_\pi \|\hat{\beta}_\pi - \beta\|_2^2 &= \lambda^2 \mathbb{E}_{\lambda e_1}(p_1(y, X) - 1)^2 + \lambda^2(p-1)\mathbb{E}_{\lambda e_2}(p_1(y, X))^2 \\ &\geq \lambda^2 \mathbb{E}_{\lambda e_1}(p_1(y, X) - 1)^2. \end{aligned} \tag{3.10}$$

Note that in the above equation, the notation $\mathbb{E}_{\lambda e_i}$ means that we are taking the expecatation assuming that the spike has happened at the i^{th} coordinate of β, and that the posterior probability $p_1(y, X)$ is given by

$$p_1(y, X) = P(I = 1 \mid y, X) = \frac{\exp(\lambda x_1^T y - \lambda^2 \|x_1\|^2/2)}{\exp(\lambda x_1^T y - \lambda^2 \|x_1\|^2/2) + \sum_{i=2}^m \exp(\lambda x_i^T y - \lambda^2 \|x_i\|^2/2)}.$$

For notational simplicity, in the rest of the proof we use the simplified notation p_1 instead of $p_1(y, X)$. Also, the notation $P_{\lambda e_1}$ denotes the joint probability of X, y assuming that the first coordinate of β is equal to λ and the rest are zero. Since $0 \leq p_1 \leq 1$, if we can show that

$$p_1 \to 0 \quad \text{in } P_{\lambda e_1}\text{- probability}, \tag{3.11}$$

then by combining the continuous mapping and the dominated convergence theorems with (3.10), we will obtain

$$\mathbb{E}_\pi \|\hat{\beta}_\pi - \beta\|^2 \geq \lambda^2(1 + o(1)).$$

Let $p_i^{(\lambda e_1)}$ denote the expression for $p_i(y, X)$ under the assumption that $\beta = \lambda e_1$. Also, use the notation $x_{i,1}$ and $x_{i,-1}$ for the first coordinate of x_i and the vector that has all the elements of x_i except for $x_{i,1}$. Then, we have

$$\begin{aligned} p_1^{(\lambda e_1)} &= \left[1 + \frac{\sum_{i=2}^m \exp\left(\lambda x_i^T(\lambda x_1 + z) - \lambda^2 \|x_i\|^2/2\right)}{\exp\left(\lambda^2 \|x_1\|^2/2 + \lambda x_1^T z\right)}\right]^{-1} \\ &\overset{d}{=} \left[1 + \frac{\sum_{i=2}^m \exp\left(\|\lambda x_1 + z\| \lambda x_{i,1} - \lambda^2 x_{i,1}^2/2 - \lambda^2 \|x_{i,-1}\|^2/2\right)}{\exp\left(\lambda^2 \|x_1\|^2/2 + \lambda x_1^T z\right)}\right]^{-1}, \end{aligned}$$

where we have used the notation $A \overset{d}{=} B$ to denote the fact that random varibales A and B have exactly the same distributions. Also, to obtain the second equality, we have used the fact that since x_i is independent of $\lambda x_1 + z$, we have $\lambda x_i^T(\lambda x_1 + z) \overset{d}{=} \|\lambda x_1 + z\| \lambda x_{i,1}$ (this can be easily

confirmed by for instance conditioning on x_1 and seeing that the distribution of $\tau x_i^T(\tau x_1 + z)$ is $\mathcal{N}(0, \lambda^2\|\lambda x_1 + z\|^2)$. Write

$$p_1 = \left(1 + \mathcal{A}_{n,m}\mathcal{B}_{n,m}\right)^{-1},$$

where

$$\mathcal{A}_{n,m} = \frac{\sum_{i=2}^m \exp\left(\|\lambda x_1 + z\|\lambda x_{i,1} - \frac{\lambda^2}{2}x_{i,1}^2 - \frac{\lambda^2}{2}\|x_{i,-1}\|^2\right)}{(m-1)(1+\frac{\lambda^2}{n})^{-\frac{n}{2}}\exp\left(\frac{1}{2}\frac{1}{1+n/\lambda^2}\|z + \lambda x_1\|^2\right)}, \tag{3.12}$$

$$\mathcal{B}_{n,m} = \frac{(m-1)(1+\frac{\lambda^2}{n})^{-\frac{n}{2}}\exp\left(\frac{1}{2}\frac{1}{1+n/\lambda^2}\|z + \lambda x_1\|^2\right)}{\exp\left(\frac{\lambda^2}{2}\|x_1\|^2 + \lambda x_1^T z\right)}. \tag{3.13}$$

In order to show (3.11), our goal is to first show that $\mathcal{A}_{n,m} \xrightarrow{p} 1$ and $\mathcal{B}_{n,m} \xrightarrow{p} \infty$. This will be done in the next two lemmas.

Lemma 32. *Consider model* (3.1) *with* $\sigma = 1$ *and* $\beta \in \mathbb{R}^m$. *Suppose that* $(\log m)/n \to 0$. *Consider the random variable* $\mathcal{B}_{n,m}$ *defined in* (3.13). *If* $\lambda > 0$ *and* $\sqrt{2\log m} - \lambda \to +\infty$, *then*

$$\mathcal{B}_{n,m} \xrightarrow{p} \infty.$$

Proof. Throughout this proof, for notational simplicity we use the notation τ instead of $\tau_{n,p}$. We can rewrite $\mathcal{B}_{n,p}$ in the following form:

$$\mathcal{B}_{n,m} = (m-1)\left(1 + \frac{\lambda^2}{n}\right)^{-\frac{n}{2}}\exp\left[\frac{1}{2(1+n/\lambda^2)}\|z\|^2 - \left(1 - \frac{1}{1+n/\lambda^2}\right)\left(\lambda x_1^T z + \frac{\lambda^2}{2}\|x_1\|^2\right)\right].$$

Using the central limit theorem, we have the following estimate:

- $\|z\|^2 = n + O_p(\sqrt{n})$,
- $\frac{\lambda^2}{2}\|x_1\|^2 + \lambda x_1^T z = \frac{\lambda^2}{2}\left(1 + O_p(\frac{1}{\sqrt{n}})\right) + \lambda \cdot O_p(1)$.

Then, we have

$$\begin{aligned}\mathcal{B}_{n,m} &= \big(1+o_p(1)\big)\exp\bigg[\log(m-1)-\frac{\lambda^2}{2}+\frac{1}{2(1+n/\lambda^2)}\Big(n+O_p(\sqrt{n})\Big)\\ &\quad -\frac{n/\lambda^2}{1+n/\lambda^2}\Big(\frac{\lambda^2}{2}+O_p\Big(\sqrt{\log m}+\frac{\log m}{\sqrt{n}}\Big)\Big)\bigg]\\ &= \big(1+o_p(1)\big)\exp\bigg[\log(m-1)-\frac{\lambda^2}{2}+O_p\Big(\sqrt{\log m}\Big)\bigg],\end{aligned}$$

where in the second equality, we have used that $(\log m)/n \to 0$ implies $(\log m)/\sqrt{n} = o(\sqrt{\log m})$. Then to show $\mathcal{B}_{n,m} \xrightarrow{p} +\infty$, using the above expression and the continuous mapping theorem, it's sufficient to argue

$$\log(m-1)-\frac{\lambda^2}{2} = \Big(\sqrt{\log(m-1)}-\frac{\lambda}{\sqrt{2}}\Big)\Big(\sqrt{\log(m-1)}+\frac{\lambda}{\sqrt{2}}\Big) \to +\infty.$$

Note that the assumption $\sqrt{2\log m}-\lambda \to +\infty$ implies $\log(m-1)-\frac{\lambda^2}{2} = \omega(\sqrt{\log(m-1)})$. Then as $m \to \infty$, $\log(m-1)-\frac{\lambda^2}{2} = \omega(\sqrt{\log(m-1)}) \to +\infty$. Therefore,

$$\mathcal{B}_{n,m} \xrightarrow{p} +\infty.$$

□

Lemma 33. *Consider model* (3.1) *with* $\sigma = 1$ *and* $\beta \in \mathbb{R}^m$. *Suppose that* $(\log m)/n \to 0$. *Consider the random variable* $\mathcal{A}_{n,p}$ *defined in* (3.12). *If* $\lambda > 0$ *and* $\sqrt{2\log m}-\lambda \to +\infty$, *then*

$$\mathcal{A}_{n,m} \xrightarrow{p} 1.$$

Proof. The proof is based on the weak law of triangular arrays, one version of which is stated in Lemma 22. Define

$$S_{n,m} := \sum_{i=2}^{m} \exp\Big[\|\lambda x_1 + z\|\lambda x_{i,1} - \frac{\lambda^2}{2}x_{i,1}^2 - \frac{\lambda^2}{2}\|x_{i,-1}\|^2\Big].$$

Let

$$Y_{m,i} := \exp\Big[\|\lambda x_1 + z\|\lambda x_{i,1} - \frac{\lambda^2}{2}x_{i,1}^2 - \frac{\lambda^2}{2}\|x_{i,-1}\|^2\Big].$$

Note that $\{Y_{m,i} : i = 2, \ldots, n\}$ are independent only if conditioning on $\|\lambda x_1 + z\|$. Hence, instead, we will prove, for certain $b_{n,m}$ to be determined,

(i) $\sum_{i=2}^m P\Big(Y_{m,i} > b_{n,m} \,\Big|\|\lambda x_1 + z\|\Big) \to 0$ a.s.

(ii) $b_{n,m}^{-2}\sum_{i=2}^m \mathbb{E}\Big[Y_{m,i}^2 \mathbb{1}_{(Y_{m,i}\le b_{n,m})} \,\Big|\|\lambda x_1 + z\|\Big] \to 0$ a.s.

For simplicity, without ambiguity, we write $b = b_{n,m}$ in the following proof.

To prove condition (i), using Lemma 34 (i) and applying the non-central χ^2 inequality in Lemma 26,

$$\begin{aligned}
&\sum_{i=2}^m P\Big(Y_{m,i} > b \,\Big|\|\lambda x_1 + z\|\Big)\\
=\;& (m-1)P\Big[\Big(\sqrt{n}x_{i,1} - \frac{\sqrt{n}}{\lambda}\|\lambda x_1 + z\|\Big)^2 + \Big(\sqrt{n}\|x_{i,-1}\|\Big)^2 < \frac{n}{\lambda^2}\Big(\|\lambda x_1 + z\|^2 - 2\log b\Big) \,\Big|\|\lambda x_1 + z\|\Big]\\
=\;& (m-1)P\Big(\chi_n^2(\gamma_1) < \frac{n}{\lambda^2}\big(\|\lambda x_1 + z\|^2 - 2\log b\big) \,\Big|\|\lambda x_1 + z\|\Big)\\
\le\;& (m-1)\exp\Big[-\frac{nc_1^2}{4(n+2\gamma_1)^2}\Big]\\
=\;& \exp\Big[-\frac{n}{4(n+2\gamma_1)^2}\Big(c_1 - \frac{2(n+2\gamma_1)}{\sqrt{n}}\sqrt{\log(m-1)}\Big)\Big(c_1 + \frac{2(n+2\gamma_1)}{\sqrt{n}}\sqrt{\log(m-1)}\Big)\Big]. \quad (3.14)
\end{aligned}$$

where $c_1 = n + n/\lambda^2 \cdot 2\log b$ and $\gamma_1 = n/\lambda^2 \cdot \|\lambda x_1 + z\|^2$. Since $\lambda^2 = o(n)$, by strong law of large numbers,

$$\frac{1}{n}\|\lambda x_1 + z\|^2 \to 1 \text{ a.s..}$$

Thus, $n + 2\gamma_1 = n \cdot \frac{n}{\lambda^2}\big(1 + o_p(1)\big)$. A sufficient condition for the upper bound in (3.14) goes to zero a.s. is

$$\Big[\frac{n}{4(n+2\gamma_1)^2}\frac{2(n+\gamma_1)}{\sqrt{n}}\sqrt{\log(m-1)}\Big]^{-1} = o_p\Big(c_1 - \frac{2(n+2\gamma_1)}{\sqrt{n}}\sqrt{\log(m-1)}\Big),$$

or

$$4\sqrt{\frac{n}{\log(m-1)}} = o_p\left(2\log b - \frac{2(\lambda^2 + 2\|\lambda x_1 + z\|^2)}{\sqrt{n}}\sqrt{\log(m-1)}\right). \tag{3.15}$$

There exists such choice of b since

$$\frac{2(\lambda^2+2\gamma_1)}{\sqrt{n}}\sqrt{\log(m-1)}\frac{\lambda^2}{n} = \frac{2(\lambda^2 + 2\|\lambda x_1 + z\|^2)}{\sqrt{n}}\sqrt{\log(m-1)} = 4\sqrt{n\log(m-1)}\big(1+o_p(1)\big) \tag{3.16}$$

as $\lambda^2 = o(n)$ and $\log(m-1) \to \infty$. Therefore, condition (i) is satisfied.

To prove condition (ii), using Lemma 34 (ii) and applying Lemma 26, we have

$$\begin{aligned}
& b^{-2}\sum_{i=2}^{m}\mathbb{E}\Big[Y_{m,i}^2\mathbb{1}_{(Y_{m,i}\le b)}\,\Big|\|\lambda x_1 + z\|\Big] \\
= \;& b^{-2}(m-1)\Big(1+\frac{2\lambda^2}{n}\Big)^{-n/2}\exp\left(\frac{2\|\lambda x_1+z\|^2}{2+n/\lambda^2}\right) \\
& \cdot P\left[\Big(Z - \frac{n/\lambda^2}{\sqrt{2+n/\lambda^2}}\|z+\lambda x_1\|\Big)^2 + \chi^2_{n-1} \ge \Big(2+n/\lambda^2\Big)\Big(\|\lambda x_1+z\|^2 - 2\log b\Big)\,\Big|\|\lambda x_1+z\|\right] \\
\le \;& b^{-2}(m-1)\Big(1+\frac{2\lambda^2}{n}\Big)^{-n/2}\exp\left(\frac{2\|\lambda x_1+z\|^2}{2+n/\lambda^2}\right) \\
= \;& \exp\left[-2\log b + \log(m-1) - \lambda^2 + \frac{2\|\lambda x_1+z\|^2}{2+n/\lambda^2}\right].
\end{aligned}$$

Since $\lambda = O(\sqrt{\log m})$ and $\log m = o(n)$, we have $\frac{\|\lambda x_1+z\|^2}{2+n/\lambda^2} = \lambda^2\big(1+o_p(1)\big)$, $\lambda^2 = o(\sqrt{n\cdot\log m})$ and $\log(m-1) = o(\sqrt{n\log m})$. For b satisfying (3.15), (3.16), $\sqrt{n\cdot\log m} = O_p(2\log b)$. Therefore,

$$\exp\left[-2\log b + \log(m-1) - \lambda^2 + \frac{2\|\lambda x_1+z\|^2}{2+n/\lambda^2}\right] \to 0 \text{ a.s..}$$

Thus, condition (ii) is satisfied.

Finally, we calibrate

$$a_{n,m} := \sum_{i=2}^{m}\mathbb{E}\Big[Y_{m,i}\mathbb{1}_{(Y_{m,i}\le b)}\,\Big|\|\lambda x_1+z\|\Big].$$

From Lemma 34 (iii), we have

$$a_{n,m} = (m-1)\Big(1+\lambda^2/n\Big)^{-n/2} \exp\left(\frac{\|\lambda x_1 + z\|^2}{2(1+n/\lambda^2)}\right)\left(1 - P\Big[\chi_n^2(\gamma_2) \le n + \gamma_2 - c_2 \,\Big|\, \|\lambda x_1 + z\|\Big]\right),$$

where $c_2 = (1+n/\lambda^2)\cdot 2\log b + n - \frac{1+2n/\lambda^2}{1+n/\lambda^2}\|\lambda x_1 + z\|^2$ and $\gamma_2 = \frac{(n/\lambda^2)^2}{1+n/\lambda^2}\|\lambda x_1 + z\|^2$. We will show that

$$P\Big[\chi_n^2(\gamma_2) \le n + \gamma_2 - c_2 \,\Big|\, \|\lambda x_1 + z\|\Big] \to 0 \text{ a.s..}$$

Using Lemma 26, we have

$$P\Big[\chi_n^2(\gamma_2) \le n + \gamma_2 - c_2 \,\Big|\, \|\lambda x_1 + z\|\Big] \le \exp\left[-\left(\frac{\sqrt{n}c_2}{2(n+\gamma_2)}\right)^2\right].$$

As $\frac{1}{n}\|\lambda x_1 + z\|^2 \to 1$ a.s., $\gamma_2 = \frac{n}{\lambda^2}\big(1+o_p(1)\big)$. Using the selection of b in (3.15) and (3.16), $\frac{n}{\lambda^2}\sqrt{n\log m} = O_p(c_2)$. Thus, $\sqrt{\log m} = O_p\Big(\frac{\sqrt{n}c_2}{2(n+\gamma_2)}\Big)$. As $m \to \infty$, we have

$$\exp\left[-\left(\frac{\sqrt{n}c_2}{2(n+\gamma_2)}\right)^2\right] \to 0 \text{ a.s..}$$

Thus,

$$a_{n,m} = (m-1)\left(1+\frac{\lambda^2}{n}\right)^{-\frac{n}{2}} \exp\left(\frac{\|\lambda x_1 + z\|^2}{2(1+n/\lambda^2)}\right)\cdot\big(1+o_p(1)\big).$$

Now, we use Lemma 22 with the bounded convergence theorem to obtain, $\forall \epsilon > 0$,

$$P\left(\left|\frac{S_{n,m} - a_{n,m}}{b_{n,m}}\right| > \epsilon\right) = \mathbb{E}P\left(\left|\frac{S_{n,m} - a_{n,m}}{b_{n,m}}\right| > \epsilon \,\middle|\, \|\lambda x_1 + z\|\right) \to 0.$$

Therefore, we conclude

$$\mathcal{A}_{n,m} \xrightarrow{p} 1.$$

□

Lemma 34. *Assume* $x_1, \ldots, x_m \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{1}{n}I_n)$, $z \sim \mathcal{N}(0, I_n)$ *and* $\{x_i : i = 1, \ldots, m\}$ *being*

independent with z. Let

$$Y_{m,i} := \exp\left[\|\lambda x_1 + z\|\lambda x_{i,1} - \frac{\lambda^2}{2}x_{i,1}^2 - \frac{\lambda^2}{2}\|x_{i,-1}\|^2\right], \quad i = 2, \ldots, n.$$

Let $Z \sim \mathcal{N}(0,1)$, χ^2_{n-1} *denote the chi-squared variable with degrees of freedom* $n-1$. (Z, χ^2_{n-1}) *are independent with* (x_1, z). *Then for* $\forall \lambda > 0$ *and* $\forall b > 0$,

(i)
$$\sum_{i=2}^{m} P(Y_{m,i} > b | \|\lambda x_1 + z\|) = (m-1)$$
$$\cdot P\left(\left(\sqrt{n}x_{2,1} - \frac{\sqrt{n}}{\lambda}\|\lambda x_1 + z\|\right)^2 + (\sqrt{n}\|x_{2,-1}\|)^2 < \frac{n}{\lambda^2}(\|\lambda x_1 + z\|^2 - 2\log b)\middle|\|\lambda x_1 + z\|\right).$$

(ii)
$$\sum_{i=2}^{m} \mathbb{E}\left[Y_{m,i}^2 \mathbb{1}_{(Y_{m,i}\le b)} | \|\lambda x_1 + z\|\right]$$
$$= (m-1)\left(1 + \frac{2\lambda^2}{n}\right)^{-\frac{n}{2}} \exp\left(\frac{2\|\lambda x_1 + z\|^2}{2 + n/\lambda^2}\right) \cdot$$
$$P\left[\left(Z - \frac{n/\lambda^2}{\sqrt{2 + n/\lambda^2}}\|z + \lambda x_1\|\right)^2 + \chi^2_{n-1} \ge (2 + \frac{n}{\lambda^2})(\|\lambda x_1 + z\|^2 - 2\log b)\middle| \|\lambda x_1 + z\|\right].$$

(iii)
$$\sum_{i=2}^{m} \mathbb{E}\left[Y_{m,i} \mathbb{1}_{(Y_{m,i}\le b)}\middle| \|z + \lambda x_1\|\right]$$
$$= (m-1)\left(1 + \lambda^2/n\right)^{-\frac{n}{2}} \exp\left(\frac{\|z + \lambda x_1\|^2}{2(1 + n/\lambda^2)}\right) \cdot$$
$$P\left[\left(Z - \frac{n/\lambda^2\|z + \lambda x_1\|}{\sqrt{1 + n/\lambda^2}}\right)^2 + \chi^2_{n-1} \ge \left(1 + \frac{n}{\lambda^2}\right)\left(\|z + \lambda x_1\|^2 - 2\log b\right)\right].$$

Proof. $Y_{m,i} \le b$ is equivalent to

$$\exp\left[-\frac{1}{2}(\lambda x_{i,1} - \|\lambda x_1 + z\|)^2 - \frac{1}{2}\lambda^2\|x_{i,-1}\|^2 + \frac{1}{2}\|\lambda x_1 + z\|^2\right] \le b$$
$$\Leftrightarrow \quad (\lambda x_{i,1} - \|\lambda x_1 + z\|)^2 + (\lambda\|x_{i,-1}\|)^2 \ge \|\lambda x_1 + z\|^2 - 2\log b.$$

Thus, Equality (i) follows.

To show Equality (ii), we first take extra condition on $x_{i,-1}$ and integrate with $x_{i,1}$. By combin-

ing the exponential function with the normal density function of $x_{i,1}$, we obtain

$$\begin{aligned}
& \mathbb{E}\left[Y_{m,i}^2 \mathbb{1}_{Y_{m,i}\le b} \middle| x_{i,-1}, \|z+\lambda x_1\| \right] \\
= \; & \mathbb{E}\left[\exp\left(2\lambda\|\lambda x_1+z\|x_{i,1} - \lambda^2 x_{i,1}^2 - \lambda^2\|x_{i,-1}\|^2\right) \right. \\
& \left. \cdot \mathbb{1}_{\left\{(\lambda x_{i,1}-\|\lambda x_1+z\|)^2+(\lambda\|x_{i,-1}\|)^2 \ge \|\lambda x_1+z\|^2 - 2\log b\right\}} \middle| x_{i,-1}, \|z+\lambda x_1\| \right] \\
= \; & \frac{1}{\sqrt{1+2\lambda^2/n}} \exp\left(\frac{2\|z+\lambda x_1\|^2}{2+n/\lambda^2}\right) \cdot \exp\left(-\lambda^2\|x_{i,-1}\|^2\right) \\
& \cdot \mathbb{1}_{\left\{\left(Z-\frac{n/\lambda^2}{\sqrt{2+n/\lambda^2}}\|z+\lambda x_1\|\right)^2 \ge \left(2+\frac{n}{\lambda^2}\right)\left(\|\lambda x_1+z\|^2 - 2\log b - \lambda^2\|x_{i,-1}\|^2\right)\right\}},
\end{aligned} \tag{3.17}$$

where $Z \sim \mathcal{N}(0,1)$ and χ^2_{n-1} denote the Chi-squared random variable with degrees of freedom $n-1$. (Z, χ^2_{n-1} are independent with $(x_1, x_{i,-1}, z)$ for each $i = 1, \ldots n$. Then, taking the expectation of $x_{i,-1}$ and using $n\|x_{i,-1}\|^2 \sim \chi^2_{n-1}$,

$$\begin{aligned}
& \mathbb{E}\left[e^{-\lambda^2\|x_{i,-1}\|^2} \cdot \mathbb{1}_{\left\{\left(Z-\frac{n/\lambda^2}{\sqrt{2+n/\lambda^2}}\|z+\lambda x_1\|\right)^2 \ge \left(2+\frac{n}{\lambda^2}\right)\left(\|\lambda x_1+z\|^2 - 2\log b - \lambda^2\|x_{i,-1}\|^2\right)\right\}} \middle| \|\lambda x_1+z\| \right] \\
= \; & \mathbb{E}\left[e^{-\frac{\lambda^2}{n}\cdot n\|x_{i,-1}\|^2} \cdot \mathbb{1}_{\left\{n\|x_{i,-1}\|^2 \ge \frac{n}{\lambda^2}\left(\|\lambda x_1+z\|^2 - 2\log b\right) - \frac{n/\lambda^2}{2+n/\lambda^2}\left(Z-\frac{n/\lambda^2}{\sqrt{2+n/\lambda^2}}\|\lambda x_1+z\|\right)^2\right\}} \middle| \|\lambda x_1+z\| \right] \\
= \; & P\left\{ \chi^2_{n-1} \ge \left(1+\frac{2\lambda^2}{n}\right)\left[\frac{n}{\lambda^2}\left(\|\lambda x_1+z\|^2 - 2\log b\right) \right.\right. \\
& \left.\left. -\frac{n/\lambda^2}{2+n/\lambda^2}\left(Z - \frac{n/\lambda^2}{\sqrt{2+n/\lambda^2}}\|z+\lambda x_1\|\right)^2 \right] \middle| \|\lambda x_1+z\| \right\} \\
= \; & P\left\{ \left(Z - \frac{n/\lambda^2}{\sqrt{2+n/\lambda^2}}\|z+\lambda x_1\|\right)^2 + \chi^2_{n-1} \ge \left(2+\frac{n}{\lambda^2}\right)\left(\|\lambda x_1+z\|^2 - 2\log b\right) \middle| \|\lambda x_1+z\| \right\}
\end{aligned} \tag{3.18}$$

Equality (ii) follows by taking expectation with $x_{i,-1}$ of equation (3.17) and using (3.18).

The argument of Equality (iii) follows similarly from Equality (ii). □

3.4.3 Proof of upper bound in Theorem 13

As we discussed in Section 3.4.2, using the scale invariance property 3.4.1 of the minimax risk, it's equivalent to prove for the case of $\sigma = 1$ in model (3.1). In the following proof, we will make this unit variance assumption without loss of generality. Throughout the proof, for notational simplicity, everytime we use constant notation $C > 0$, we mean it is a constant that does not depend on any variable or parameter, but whose value may change at each occurrence. For $X \in \mathbb{R}^{n\times p}$, let $i \in \{1, \ldots, p\}$ and X_i stand for the i-th column of X; let $T \subseteq [p]$ and X_T stand for the submatirx consists of the columns with indices contained in T.

Proof of upper bound in Theorem 13

Obtaining constant-sharp upper bounds for the minimax risk of the sparse linear regression is quite challenging and for that reason it has remained open, despite the existing extensive literature on the topic. To obtain an upper bound matching the lower bound we derived in the last section, we aim to construct an estimator that combines the maximum likelihood estimator and LASSO. More specifically, consider the following two estimates:

- LASSO:

$$\hat{\beta}^L(\lambda) := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1. \tag{3.19}$$

- Maximum likelihood estimator:

$$\hat{\beta}^M := \arg\min_{b\in\Theta(k)} \|y - Xb\|_2^2. \tag{3.20}$$

Here, LASSO depends on the regularization parameter $\lambda > 0$. We will clarify later in the proof of the selection of λ. At this point, we assume λ may vary with the asymptotic parameters as $n \to \infty$. For notational simplicity, throughout the proof, we write $\hat{\beta}^L(\lambda) =: \hat{\beta}^L$ in appropriate context without ambiguity.

Furthermore, define the cone

$$C_{SRE}(k, c_0) := \{\Delta \in \mathbb{R}^p : \ \|\Delta\|_1 \leq (1 + c_0)\sqrt{k}\|\Delta\|_2\},$$

with c_0 determined later in the proof. Consider the observable event

$$\mathcal{A} := \mathcal{A}(\delta_0) := \left\{X \in \mathbb{R}^{n\times p} : \ \max_{j=1,\ldots,p} \|Xe_j\|_2^2 \leq 1 + \delta_0, \ \inf_{\Delta\in C_{SRE}(k,c_0):\Delta\neq 0} \frac{\|X\Delta\|_2}{\|\Delta\|_2} \geq \sqrt{1-\delta_0}\right\}, \tag{3.21}$$

where $0 < \delta_0$ is a constant that will be determined later. Define the new estimator

$$\hat{\hat{\beta}} := \hat{\beta}^L \mathbb{1}_{\mathcal{A}} + \hat{\beta}^M \mathbb{1}_{\mathcal{A}^c}. \tag{3.22}$$

Our goal is to show that

$$\sup_{\beta\in\Theta(k)} \mathbb{E}\|\hat{\hat{\beta}} - \beta\|_2^2 = 2k\log(p/k)(1 + o(1)).$$

From the construction of $\hat{\hat{\beta}}$, the risk consists of two parts

$$\mathbb{E}\|\hat{\hat{\beta}} - \beta\|^2 = \mathbb{E}\|\hat{\beta}^L - \beta\|^2 \mathbb{1}_{\mathcal{A}} + \mathbb{E}\|\hat{\beta}^M - \beta\|^2 \mathbb{1}_{\mathcal{A}^c}. \tag{3.23}$$

We will show that: (1) $\mathbb{E}\|\hat{\beta}^M - \beta\|^2 \mathbb{1}_{\mathcal{A}^c} = o\Big(k\log(p/k)\Big)$; (2) $\mathbb{E}\mathbb{E}\|\hat{\beta}^L - \beta\|^2 \mathbb{1}_{\mathcal{A}} = 2k\log(p/k)\Big(1 + o(1)\Big)$. To see (1), we use the Cauchy-Schwartz inequality and obtain

$$\mathbb{E}\|\hat{\beta}^M - \beta\|^2 \leq \left(\mathbb{E}\|\hat{\beta}^M - \beta\|_2^r\right)^{\frac{2}{r}} P(\mathcal{A}^c). \tag{3.24}$$

We have the following proposition:

Proposition 2. *Assume model* (3.1) *with* $\sigma = 1$. *Suppose* $k/p \to 0$ *and* $(k\log(p/k))/n \to 0$.

Then, for $\forall m \geq 2$, there exists a constant $C = C(m) > 0$, such that

$$\left(\mathbb{E}\|\beta - \hat{\beta}\|_2^m\right)^{\frac{2}{m}} \leq Ck\log(p/k).$$

Besides, Lemma 45 indicates that $\mathcal{A}$ holds with probability tending to one. Then, from (3.24), we have

$$\mathbb{E}\|\hat{\beta}^M - \beta\|^2 \leq o(k\log(p/k)). \tag{3.25}$$

To demonstrate (1), note that it is not trivial to directly bound the expected loss of LASSO with accurate constant not assuming additional structures of the design matrix X such as the mutual incoherence condition [42] and the restricted eigenvalue condition [43]. Largely motivated by [21], we resort to an oracle estimator to draw out the accurate constant of the minimax risk and show that it is close to LASSO with further analysis. For LASSO defined in (3.19), given $\varepsilon > 0$, choose the regularization parameter as

$$\lambda = \lambda_\varepsilon := (1+\varepsilon)\sqrt{2\log(p/k)}. \tag{3.26}$$

Let $\eta_\lambda(y) := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|y-b\|_2^2 + \lambda\|b\|_1$ and consider the oracle estimator

$$\tilde{\beta}^O := \eta_\lambda\big(\beta + X^T z\big), \tag{3.27}$$

with the same choice of $\lambda = \lambda_\varepsilon$.

We will show that as $\varepsilon \to 0$, the expected loss of the oracle estimator can be as close to $2k\log(p/k)$ as we want. In fact, we have obtained a mean-squared error (MSE) upper bound for the oracle estimator under orthogonal design of X. [3] in a previous study in [44] (Theorem 1). With this, consider an event

$$\mathcal{B}_1 := \left\{\|z\|_2 \leq (1+\varepsilon)/\sqrt{n}\right\}.$$

[3]Under orthogonal design, the model is simplified as the Gaussian sequence model, and the oracle estimator in (3.27) becomes a real estimator not depending on any latent variables as z anymore.

Then for every $0 < c \le 1+\varepsilon$, conditional on $\|z\| = c\sqrt{n}$, $X^T z \sim \mathcal{N}(0, c^2 I_p)$ and

$$\|\eta_{\lambda_\varepsilon}(\beta + X^T z) - \beta\| \stackrel{d}{=} c\|\eta_{\lambda_{\varepsilon'}}(\beta/c + \mathcal{N}(0, I_p)) - \beta/c\|,$$

where $\varepsilon' = (1+\varepsilon)/c - 1 \ge 0$. Hence, based on Theorem 1 in [44], for arbitrarily small $\delta_1 > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{B}_1}\right] &\le \mathbb{E}\left[\frac{\|z\|^2}{n}\mathbb{E}\Big(\|\eta_{\varepsilon'}\big(\beta/c + \mathcal{N}(0, I_p)\big) - \beta/c\|^2 \,\Big|\, \|z\| = c\sqrt{n}\Big)\mathbb{1}_{\{\|z\| \le (1+\varepsilon)\sqrt{n}\}}\right] \\
&\le \mathbb{E}\left[\frac{\|z\|^2}{n} 2(1+\varepsilon')^2 k \log(p/k) \cdot \Big(1 + o(1)\Big)\mathbb{1}_{\{\|z\| \le (1+\varepsilon)\sqrt{n}\}}\right] \\
&\le (1+\varepsilon)^2(1+\delta_1) \cdot 2k\log(p/k), \qquad (3.28)
\end{aligned}$$

where the last inequality is due to $(1+\varepsilon)^2/c^2 = (1+\varepsilon')^2$. Thus, under $\mathcal{B}_1$, we have the MSE of the oracle estimator upper bounded by $2k\log(p/k)\cdot(1+o(1))$, as $\varepsilon, \delta_1 \to 0$. In addition, $\mathcal{B}_1$ holds with probability at least $1 - \exp(-n\varepsilon^2/2)$.

Then, we show that the oracle estimator falls close to LASSO in a negligible distance compared to $2k\log(p/k)$. We have this lemma:

Lemma 35. *Let $S_* \subseteq \{1, \dots, p\}$ be a subset of columns assumed to contain the supports of $\hat{\beta}^L, \tilde{\beta}^O$ and β, i.e. $S_* \supseteq \operatorname{supp}(\hat{\beta}^L) \cup \operatorname{supp}(\tilde{\beta}^O) \cup \operatorname{supp}(\beta)$. Suppose for some $\delta_2 < 1/2$, all the eigenvalues of $X_{S_*}^T X_{S_*}$ lie in $[1-\delta_2, 1+\delta_2]$. Then*

$$\|\tilde{\beta}^O - \hat{\beta}^L\|_2^2 \le \frac{3\delta_2}{1-2\delta_2}\|\tilde{\beta}^O - \beta\|_2^2.$$

The proof of this lemma follows by exact mirroring the proof of Lemma 4.2 in [21], so we skip the argument. Then assume there exists some $S_* \supseteq \operatorname{supp}(\hat{\beta}^L) \cup \operatorname{supp}(\tilde{\beta}^O) \cup \operatorname{supp}(\beta)$ and let

$$\begin{aligned}
\mathcal{B}_2 := \mathcal{B}_2(\delta_2) := \Big\{ &\operatorname{supp}(\hat{\beta}^L) \cup \operatorname{supp}(\tilde{\beta}^O) \cup \operatorname{supp}(\beta) \subseteq S_*, \\
&\text{all eigenvalues of } X_{S_*}^T X_{S_*} \text{ lie in } [1-\delta_2, 1+\delta_2]\Big\}. \qquad (3.29)
\end{aligned}$$

Lemma 35 implies that

$$\mathbb{E}\|\hat{\beta}^L - \tilde{\beta}^O\|^2 \mathbb{1}_{\mathcal{B}_1 \cap \mathcal{B}_2} \le \frac{3\delta_2}{1-2\delta_2} \mathbb{E}\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{B}_1} \le \frac{3\delta_2}{1-2\delta_2}(1+\delta_1)(1+\varepsilon)^2 \cdot 2k\log(p/k). \quad (3.30)$$

The right end is $2k\log(p/k)\cdot(1+o(1))$ as $\delta_2, \varepsilon \to 0$. Then we wonder if such S_* exists to satisfy the condition of Lemma 35.

Following [21], we consider a constructed resolvent set developed by [21]:

Definition 1 (Resolvent set of Lasso)**.** *Fix $S = \mathrm{supp}(\beta)$ of cardinality at most k, and an integer k^* obeying $k < k^* < p$. The set $S_* = S_*(S, k^*)$ is said to be a resolvent set if it is the union of S and the $k^* - k$ indicies with the largest values of $|X_i^T z|$ among all $i \in \{1, \dots, p\} \setminus S$.*

Though designed for SLOPE in [21], it turns out the resolvent set and the following procedure in [21] works for LASSO as well. Denote $S_\diamond := \mathrm{supp}(\beta) \cup \mathrm{supp}(\hat{\beta}^L) \cup \mathrm{supp}(\tilde{\beta}^O)$, we have

Proposition 3. *Assume model* (3.1)*. Consider the LASSO estimator $\hat{\beta}^L$* (3.19) *and the oracle estimator $\tilde{\beta}^O$* (3.27) *with the same regularization $\lambda > 0$ as* (3.26)*. Let S_* be the resolvent set in Definition 1 and $k < k^* < p$ be its cardinality. Suppose $k^* \ge 2k$, $k^*/p \to 0$ and $(k^*\log p)/n \to 0$. Then for arbitrary small $\delta_2 \in (0, 1/2)$,*

$$\text{(3.29) } \textit{holds w.h.p. } 1 - \exp\Big[-c_1\varepsilon k\sqrt{2\log(p/k)}\Big] - c_5 e^{-c_6 n\varepsilon^2},$$

for some constants $c_1, c_4, c_5, c_6 > 0$.

At this point, let $\mathcal{B} := \mathcal{B}_1 \cap \mathcal{B}_2$, we have proved that

$$P(\mathcal{B}) \ge 1 - \exp(-n\varepsilon^2/2) - \exp\Big[-c_1\varepsilon k\sqrt{2\log(p/k)}\Big] - c_5 e^{-c_6 n\varepsilon^2} \to 1, \quad (3.31)$$

and (3.28) and (3.30) imply that

$$\mathbb{E}\|\hat{\beta}^L - \beta\|^2 \mathbb{1}_{\mathcal{A}} = \mathbb{E}\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}} + 2\mathbb{E}(\hat{\beta}^L - \tilde{\beta}^O)^T(\tilde{\beta}^O - \beta)\mathbb{1}_{\mathcal{A}\cap\mathcal{B}} \quad (3.32)$$

$$
\begin{aligned}
&+ \mathbb{E}\|\hat{\beta}^L - \tilde{\beta}^O\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}} + \mathbb{E}\|\hat{\beta}^L - \beta\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}^c} \\
\leq\ & \mathbb{E}\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}} + 2\sqrt{\mathbb{E}\|\hat{\beta}^L - \tilde{\beta}^O\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}}} \cdot \sqrt{\mathbb{E}\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}}} \\
&+ \mathbb{E}\|\hat{\beta}^L - \tilde{\beta}^O\|^2 \mathbb{1}_{\mathcal{A}\cap\mathcal{B}} + \left(\mathbb{E}\|\hat{\beta}^L - \beta\|^q \mathbb{1}_{\mathcal{A}}\right)^{\frac{2}{q}} P(\mathcal{B}^c) \\
\leq\ & \mathbb{E}\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{B}_1} + 2\sqrt{\mathbb{E}\|\hat{\beta}^L - \tilde{\beta}^O\|^2 \mathbb{1}_{\mathcal{B}_2}} \cdot \sqrt{\mathbb{E}\|\tilde{\beta}^O - \beta\|^2 \mathbb{1}_{\mathcal{B}_1}} \\
&+ \mathbb{E}\|\hat{\beta}^L - \tilde{\beta}^O\|^2 \mathbb{1}_{\mathcal{B}_2} + \left(\mathbb{E}\|\hat{\beta}^L - \beta\|^q \mathbb{1}_{\mathcal{A}}\right)^{\frac{2}{q}} P(\mathcal{B}^c) \\
=\ & 2k\log(p/k) \cdot \left(1 + O\left(\varepsilon + \delta_1 + \sqrt{\delta_2}\right)\right) + \left(\mathbb{E}\|\hat{\beta}^L - \beta\|^q \mathbb{1}_{\mathcal{A}}\right)^{\frac{2}{q}} P(\mathcal{B}^c); \qquad (3.33)
\end{aligned}
$$

we have used Hölder's inequality with a power factor $q \geq 2$. Therefore, as $\varepsilon, \delta_1, \delta_2 \to 0$, the first term in the last line is the dominating term.

It remains to derive an upper bound for . We refer to [15] that it has been shown that under a condition similar to the restricted eigenvalue condition that LASSO can achieve the sparse minimax risk up to rate optimality. The condition introduced in [15] is called the Strong Eigenvalue Condition defined below. This leads us to consider event $\mathcal{A}$.

Definition 2. *Let $c_0 > 0$ and $s \in \{1, \ldots, p\}$. We call a design matrix $X \in \mathbb{R}^{n\times p}$ satisfying the $SRE(s, c_0)$ condition if $\|Xe_j\|_2 \leq 1$ for all $j = 1, \ldots, p$, and*

$$
\theta(s, c_0) := \min_{\Delta \in C_{SRE}(s,c_0):\Delta\neq 0} \frac{\|X\Delta\|_2}{\|\Delta\|_2} > 0,
$$

where $C_{SRE}(s, c_0) := \{\Delta \in \mathbb{R}^p :\ \|\Delta\|_1 \leq (1 + c_0)\sqrt{s}\|\Delta\|_2\}$ is a cone in $\mathbb{R}^p$.

However, the result (Theorem 4.2 in [15]) requires the regularization of LASSO satisfying $\lambda \geq C\sqrt{2\log(p/k)}$ for a constant $C > 4 + \sqrt{2} > (1 + \varepsilon)$. Compared to our choice in (3.26), larger λ will cause the constant in the risk upper bound of the oracle estimator (3.28) being larger than 2. From (3.33), since the risk from the oracle estimator is the dominating term, it then results in an upper bound of the minimax risk with worse constant (larger than 2). Therefore, in the following proposition, we intend to amend this result such that it can adapt to the choice of λ in (3.26).

Consider $\mathcal{A}$ (3.21), let the constant in $\mathcal{A}$ be $c_0 = c_0(\delta_3, \delta_4, \delta_0)$ defined by (3.67), we have

Proposition 4. *Assume model* (3.1) *and* $\varepsilon > 0$. *Consider LASSO estimator* $\hat{\beta}^L$ (3.19) *with regularization* λ_ε *as in* (3.26). *Let* $\mathcal{A}(c_0, \delta_0)$ *be the event in* (3.21). *If* δ_3, $\delta_4 > 0$ *such that* $\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3) < 1+\varepsilon$. *Then, for* $q > 2$ *and some constant* $C_q > 0$, *we have*

$$\mathbb{E}\Big[\|\hat{\beta}^L - \beta\|_2^q \, \mathbb{1}_{\mathcal{A}}\Big] \leq \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)^q \Big(\frac{\lambda_\varepsilon\sqrt{k}}{\delta_4^2}\Big)^q \Big[\frac{1}{(1-\delta_0)^q} + \frac{C_q}{(\delta_4^2 k\log(p/k))^q}\Big].$$

To demonstrate $\Big(\mathbb{E}\|\hat{\beta}^L - \beta\|^q \mathbb{1}_{\mathcal{A}}\Big)^{\frac{2}{q}} P(\mathcal{B}^c) \leq o(k\log(p/k))$, we have the following argument: First, fix $\varepsilon, \delta_0 > 0$. Second, select and fix δ_3, $\delta_4 > 0$ such that $\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3) < 1+\varepsilon$. Then, from (3.31) and Proposition 4, if $n \geq \varepsilon^{-3}$ and $k\sqrt{2\log(p/k)} \geq \varepsilon^{-2}$,

$$\begin{aligned}
&\Big(\mathbb{E}\|\hat{\beta}^L - \beta\|^q \mathbb{1}_{\mathcal{A}}\Big)^{\frac{2}{q}} P(\mathcal{B}^c) \\
\leq\ & \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)^2 \Big(\frac{\lambda_\varepsilon\sqrt{k}}{\delta_4^2}\Big)^2 \Big[\frac{1}{(1-\delta_0)^2} + \frac{C_q^{2/q}}{(\delta_4^2 k\log(p/k))^2}\Big] \cdot P(\mathcal{B}^c) \\
\leq\ & Ck\log(p/k)\cdot\varepsilon^{-4}\cdot\Big[c_2\exp\Big(-c_3 n\varepsilon^2\Big) + \exp\Big(-c_1\varepsilon k\sqrt{2\log(p/k)}\Big)\Big] \\
\leq\ & o\Big(k\log(p/k)\Big). \qquad (3.34)
\end{aligned}$$

Review (3.23), (3.25), (3.33) and (3.34), we have completed the proof.

Proof of Proposition 2

Proof of Proposition 2. Recall the definition of the maximum likelihood estimator (MLE). For $k = |\operatorname{supp}(\beta)|$,

$$\hat{\beta}^M = \underset{b\in\Theta(k)}{\arg\min}\, \|y - Xb\|_2^2. \qquad (3.35)$$

Note that the MLE is the minimizer of (3.35), for $\forall\beta \in \Theta(k)$,

$$\|y - X\beta\|_2^2 \leq \|y - X\hat{\beta}^M\|_2^2$$

With $y = X\beta + z$, this implies

$$\|X(\beta - \hat{\beta}^M)\|_2^2 \le 2z^T X(\hat{\beta}^M - \beta). \tag{3.36}$$

Fix $s \in \{1, \dots, p\}$, let

$$V_s := \inf\{\|X\Delta\|_2^2 : \ \|\Delta\|_2 = 1, \ \|\Delta\|_0 = s\}. \tag{3.37}$$

Since both β and $\hat{\beta}^M$ are in $\Theta(k)$, from (3.36) we have

$$V_{2k} \cdot \|\beta - \hat{\beta}^M\|_2^2 \le \|X(\beta - \hat{\beta}^M)\|_2^2 \le 2z^T X(\hat{\beta}^M - \beta) \le 2\|\hat{\beta}^M - \beta\|_2 \cdot \sup_{\substack{\|u\|_2=1 \\ \|u\|_0=2k}} z^T Xu.$$

Hence,

$$0 \le \|\beta - \hat{\beta}^M\|_2 \le \frac{2}{V_{2k}} \cdot \sup_{\substack{\|u\|_2=1 \\ \|u\|_0=2k}} z^T Xu.$$

Then, using Hölder's inequality, we have

$$\mathbb{E}\|\beta - \hat{\beta}\|_2^m \le 2^m \left(\mathbb{E}\frac{1}{V_{2k}^r}\right)^{\frac{m}{r}} \left(\mathbb{E}\Big(\sup_{\substack{\|u\|_2=1 \\ \|u\|_0=2k}} z^T Xu\Big)^q\right)^{\frac{m}{q}}, \tag{3.38}$$

where $r, q > 0$ and $\frac{m}{r} + \frac{m}{q} = 1$. Hence, we need to bound the two terms on the right hand side of (3.38). First, we have the following bound for the first term.

Lemma 36. *Suppose the Gaussian random design $X \in \mathbb{R}^{n\times p}$ in model* (3.1)*. For $s \in \{1, \dots, p\}$, let V_s be defined as in* (3.37)*. If $(s\log(p/s))/n \to 0$, then, for $\forall r > 0$,*

$$\mathbb{E}\frac{1}{V_s^r} = O(1).$$

In addition, we have the following upper bound for $\left[\mathbb{E}\Big(\sup_{\substack{\|u\|_2=1 \\ \|u\|_0=2k}} z^T Xu\Big)^q\right]^{\frac{1}{q}}$.

Lemma 37. *Suppose the standard normal vector $z \in \mathbb{R}^n$ and the Gaussian design $X \in \mathbb{R}^{n\times p}$ in model* (3.1)*. Denote $g \sim \mathcal{N}(0, I_p)$. For $s \in \{1, \dots, p\}$, suppose $p/s \to \infty$ and $s\log(p/s) \to \infty$.*

Define $T(s) = \{u \in \mathbb{R}^p : \|u\|_2 = 1, \|u\|_0 \leq s\}$. *Then, for* $\forall q \geq 1$,

$$\left[\mathbb{E}\Big(\sup_{u\in T(s)} z^T Xu\Big)^q\right]^{2/q} = \left[\mathbb{E}\left(\frac{\|z\|_2}{\sqrt{n}}\right)^q\right]^{2/q} \cdot \left(\mathbb{E}\Big(\sup_{u\in T(s)} \langle g, u\rangle\Big)^q\right)^{2/q} \leq Cs\log(p/s),^4$$

for some constant $C > 0$.

The proof of these two lemmas are presented after this proposition. Then, combining (3.38) with the results of Lemmas 36 and 37 completes the proof.

□

Proof of Lemma 36. Throughout the proof, we fix $s \in \{1, \ldots, p\}$ and let $V := V_s$ for notational simplicity. For some $0 < x = O(1)$ whose exact value will be determined later, we have

$$\mathbb{E}\frac{1}{V^r} = \mathbb{E}\left(\frac{1}{V^r}\mathbb{1}_{(V\leq x)}\right) + \mathbb{E}\left(\frac{1}{V^r}\mathbb{1}_{(V>x)}\right). \tag{3.39}$$

Since x is bounded, we have

$$\mathbb{E}\left(\frac{1}{V^r}\mathbb{1}_{(V>x)}\right) < \frac{1}{x^r}. \tag{3.40}$$

Hence, in the rest of the proof, we aim to obtain an upper bound for $\mathbb{E}\left(\frac{1}{V^r}\mathbb{1}_{(V\leq x)}\right)$.

Towards this goal, we first construct a left-tail probabilistic bound. For $\forall t \in (0, 1/2)$, using the union bound, we have

$$P(V \leq 1-t) = P\Big(\min_{\substack{S\subseteq[p]\\|S|=s}} \inf_{\|\Delta\|_2=1} \|X_S\Delta\|_2^2 \leq 1-t\Big) \leq \binom{p}{s} P\Big(\inf_{\|\Delta\|_2=1} \|X_S\Delta\|_2^2 \leq 1-t\Big),$$

where the right end of inequality is for some fixed $S \subseteq [p]$, $|S| = s$. Consider $\mathcal{N}(\varepsilon)$ to be the ε-net of $S^{s-1} = \{\Delta \in \mathbb{R}^s : \|\Delta\|_2 = 1\}$. Note that ε can be set according to t. Then for $\forall \Delta \in S^{s-1}$, there exists a $\Delta' \in \mathcal{N}(\varepsilon)$ such that $\|\Delta - \Delta'\|_2 \leq \varepsilon$ and

$$\|X_S\Delta\|_2^2 = \|X_S\Delta'\|_2^2 + \langle X_S(\Delta-\Delta'), X_S(\Delta+\Delta')\rangle \geq \inf_{\Delta\in\mathcal{N}(\varepsilon)} \|X_S\Delta\|_2^2 - \sqrt{2}\varepsilon \sup_{\Delta\in S^{s-1}} \|X_S\Delta\|_2^2. \tag{3.41}$$

[4] We can drop $|\cdot|$ in expectation since for symmetric $T(s)$, $\sup_{u\in T(s)} z^T Xu \geq 0$ and $\sup_{u\in T(s)} \langle g, u\rangle \geq 0$.

Consider $A := \{\sup_{\Delta \in S^{s-1}} \|X_S\Delta\|_2^2 \leq \frac{1}{1-t}\}$. Let $\sqrt{2}\varepsilon = (1-t)^2$, we have

$$\begin{aligned}
& P\Big(\inf_{\Delta \in S^{s-1}} \|X_S\Delta\|_2^2 \leq 1-t\Big) \\
\leq\ & P\Big(\inf_{\Delta \in \mathcal{N}(\varepsilon)} \|X_S\Delta\|_2^2 \leq 2(1-t),\ A\Big) + P(A^c) \\
\leq\ & \frac{\Big(3\sqrt{2}\Big)^s}{(1-t)^{2s}} P\Big(\|X_S\Delta\|_2^2 \leq 2(1-t)\Big) + P(A^c),
\end{aligned}$$

where the last inequality uses the union bound and that $|\mathcal{N}(\varepsilon)| \leq (3/\varepsilon)^s = \big(3/((1-t)^2/\sqrt{2})\big)^s$ from Lemma 23. Note that the last expression holds for some fixed $S \subseteq [p]$, $|S| = s$, and some certain $\Delta \in S^{s-1}$.

Thus, using the integrated tail bound expression for expectation, we have

$$\begin{aligned}
& \mathbb{E}\frac{1}{V^r}\mathbb{1}_{(V \leq x)} = \int_{0<1-t\leq x} P(V \leq 1-t) r(1-t)^{-r-1} dt \\
\leq\ & \int_{0<1-t\leq x} \binom{p}{s} P\Big(\inf_{\Delta \in S^{s-1}} \|X_S\Delta\|_2^2 \leq 1-t\Big) r(1-t)^{-r-1} dt \\
\leq\ & \binom{p}{s} \int_{0<1-t\leq x} \frac{\Big(3\sqrt{2}\Big)^s}{(1-t)^{2s}} P\Big(\|X_S\Delta\|_2^2 \leq 2(1-t)\Big) r(1-t)^{-r-1} dt \\
+\ & \binom{p}{s} \int_{0<1-t\leq x} P\Big(\sup_{\Delta \in S^{s-1}} \|X_S\Delta\|_2^2 > \frac{1}{1-t}\Big) r(1-t)^{-r-1} dt. \qquad (3.42)
\end{aligned}$$

Let I_1 denote the first line of the last expression, I_2 denote the second line.

Consider $0 < x \leq e^{-2}$. We first calculate I_1. Note that $n\|X_S\Delta\|_2^2 \sim \chi_n^2$, applying the deviation in Lemma 25, we have

$$\begin{aligned}
I_1 &\leq \binom{p}{s} \int_{0<1-t\leq x} \frac{\Big(3\sqrt{2}\Big)^s}{(1-t)^{2s}} r(1-t)^{-r-1} \exp\Big[\frac{n}{2}\Big(2t - 1 + \log(2(1-t))\Big)\Big] dt \\
&= \binom{p}{s} \int_{0<1-t\leq x} \Big(3\sqrt{2}\Big)^s r(1-t)^{\frac{n}{2}-r-2s-1} \exp\Big[\frac{n}{2}\Big(2t - 1 + \log 2\Big)\Big] dt \\
&\leq \binom{p}{s} \Big(3\sqrt{2}\Big)^s r e^{\frac{n}{2}(1+\log 2)} \int_{1-x}^{1} (1-t)^{\frac{n}{2}-r-2s-1} dt \\
&= \binom{p}{s} \Big(3\sqrt{2}\Big)^s r e^{\frac{n}{2}(1+\log 2)} \frac{x^{\frac{n}{2}-r-2s}}{\frac{n}{2}-r-2s}
\end{aligned}$$

$$\leq \binom{p}{s}\Big(3\sqrt{2}\Big)^s r\frac{e^{\frac{n}{2}(-1+\log 2)+r+4s}}{\frac{n}{2}-r-2s}$$
$$\leq e^{C_1(s\log(p/s))-C_2 n} = o(1), \tag{3.43}$$

where $C_1, C_2 > 0$ are constants. The last equality is by the assumption $(s\log(p/s))/n \to 0$.

Then we calculate I_2. Consider another ε'-net $\mathcal{N}(\varepsilon')$ for S^{s-1}. Following the same argument as in (3.41), for certain $S \subseteq [p]$, $|S| = s$, we have

$$\sup_{\Delta\in S^{s-1}} \|X_S\Delta\|_2^2 \leq \sup_{\Delta\in\mathcal{N}(\varepsilon')} \|X_S\Delta\|_2^2 + \sqrt{2}\varepsilon'\Big(\sup_{\Delta\in S^{s-1}} \|X_S\Delta\|_2^2\Big).$$

Thus,

$$\sup_{\Delta\in S^{s-1}} \|X_S\Delta\|_2^2 \leq \frac{1}{1-\sqrt{2}\varepsilon'} \sup_{\Delta\in\mathcal{N}(\varepsilon')} \|X_S\Delta\|_2^2.$$

Let $1-\sqrt{2}\varepsilon' = \sqrt{1-t}$. For $0 < 1-t \leq x$, by Lemmas 23 and 25, we have

$$\begin{aligned}
&P\Big(\sup_{\Delta\in S^{s-1}} \|X_S\Delta\|_2^2 > \frac{1}{1-t}\Big) \leq P\Big(\sup_{\Delta\in\mathcal{N}(\varepsilon')} \|X_S\Delta\|_2^2 > \frac{1-\sqrt{2}\varepsilon'}{1-t}\Big)\\
\leq\ & |\mathcal{N}(\varepsilon')|\cdot P\Big(\|X_S\Delta\|_2^2 > \frac{1}{\sqrt{1-t}}\Big) \leq \Big(\frac{3}{\varepsilon'}\Big)^s \cdot P\Big(\|X_S\Delta\|_2^2 > \frac{1}{\sqrt{1-t}}\Big)\\
\leq\ & \Big(\frac{3\sqrt{2}}{1-\sqrt{x}}\Big)^s \exp\Big\{-\frac{n}{2}\Big(\frac{1}{\sqrt{1-t}} - 1 - \log\frac{1}{\sqrt{1-t}}\Big)\Big\}.
\end{aligned}$$

Then let $\mathcal{B} := \mathcal{B}_1 \cap \mathcal{B}_2$,

Note that for $\forall t \in [1-x, 1)$, $\sqrt{x}\Big(\frac{1}{\sqrt{1-t}} - 1\Big) \geq \log\frac{1}{\sqrt{1-t}}$. Then, selecting x small enough such that $1 - \sqrt{x}(1 + \frac{4(r+1)}{n}) > 0$, we have

$$\begin{aligned}
I_2 &\leq \binom{p}{s}\Big(\frac{3\sqrt{2}}{1-\sqrt{x}}\Big)^s \int_{0<1-t\leq x} r(1-t)^{-r-1}\exp\Big\{-\frac{n}{2}\Big(\frac{1}{\sqrt{1-t}} - 1 - \log\frac{1}{\sqrt{1-t}}\Big)\Big\}dt\\
&= \binom{p}{s}\Big(\frac{3\sqrt{2}}{1-\sqrt{x}}\Big)^s r\int_{0<1-t\leq x} \exp\Big\{-\frac{n}{2}\Big[\frac{1}{\sqrt{1-t}} - 1 - \Big(1+\frac{4(r+1)}{n}\Big)\log\frac{1}{\sqrt{1-t}}\Big]\Big\}dt\\
&\leq \binom{p}{s}\Big(\frac{3\sqrt{2}}{1-\sqrt{x}}\Big)^s r\int_{0<1-t\leq x} \exp\Big\{-\frac{n}{2}\Big[1-\sqrt{x}\Big(1+\frac{4(r+1)}{n}\Big)\Big]\cdot\Big(\frac{1}{\sqrt{1-t}} - 1\Big)\Big\}dt
\end{aligned}$$

$$\leq \binom{p}{s}\left(\frac{3\sqrt{2}}{1-\sqrt{x}}\right)^s rx \cdot \exp\Big\{-\frac{n}{2}\Big[1-\sqrt{x}\Big(1+\frac{4(r+1)}{n}\Big)\Big]\cdot\Big(\frac{1}{\sqrt{x}}-1\Big)\Big\}$$
$$\leq e^{C_1(s\log(p/s))-C_2 n} = o(1), \tag{3.44}$$

where $C_1, C_2 > 0$ are constants. The last equality is by the assumption $(s\log(p/s))/n \to 0$. Thus, combing (3.42) with (3.43) and (3.44), for $x > 0$ being small enough constant, we have

$$\mathbb{E}\Big(\frac{1}{V^r}\mathbb{1}_{(V\leq x)}\Big) = o(1).$$

Therefore, from (3.39) and (3.40), we conclude that

$$\mathbb{E}\Big(\frac{1}{V^r}\Big) \leq O(1).$$

□

Proof of Lemma 37. We first construct the upper bound for $\Big[\mathbb{E}\Big(\sup_{u\in T(s)}\langle g,u\rangle\Big)^q\Big]^{2/q}$. Using Minkowski's inequality,

$$\Big[\mathbb{E}\Big(\sup_{u\in T(s)}\langle g,u\rangle\Big)^q\Big]^{2/q} \leq \Big(\mathbb{E}\Big|\sup_{u\in T(s)}\langle g,u\rangle - \mathbb{E}\sup_{u\in T(s)}\langle g,u\rangle\Big|^q\Big)^{2/q} + \Big(\mathbb{E}\sup_{u\in T(s)}\langle g,u\rangle\Big)^2. \tag{3.45}$$

Note that the second term above is the Gaussian complexity on $T(s)$. Following Exercise 5.7 in [5], for some constant $C > 0$, we have

$$\mathbb{E}\sup_{u\in T(s)}\langle g,u\rangle \leq \sqrt{Cs\log(p/s)}. \tag{3.46}$$

To bound the first term in (3.45), let $F(g) := \sup_{u\in T(s)}\langle g,u\rangle$. Then, for any $g, g' \in \mathbb{R}^p$,

$$\langle g,u\rangle = \langle g-g',u\rangle + \langle g',u\rangle \leq \|g-g'\|_2 \cdot \Big(\sup_{u\in T(s)}\|u\|_2\Big) + F(g').$$

Thus, F is a 1-Lipschitz function. Using the concentration of Lipshitz function of Gaussians (e.g.,

Theorem 2.2.6 in [5]), we obtain

$$\begin{aligned}
&\mathbb{E}\Big|\sup_{u\in T(s)}\langle g,u\rangle-\mathbb{E}\sup_{u\in T(s)}\langle g,u\rangle\Big|^q=\int_{t>0}qt^{q-1}P\Big(|F(g)-\mathbb{E}F(g)|>t\Big)dt\\
\leq\ &\int_{t>0}2qt^{q-1}e^{-\frac{t^2}{2}}dt=2^{\frac{q}{2}}q\Gamma(\frac{q}{2}):=C_q. \qquad (3.47)
\end{aligned}$$

Here $\Gamma(\cdot)$ is the Gamma function. $C_q>0$ is a constant only depend on q. Then, based on (3.45), (3.46), (3.47), for some constant $C>0$, we have

$$\left[\mathbb{E}\Big(\sup_{u\in T(s)}\langle g,u\rangle\Big)^{2q}\right]^{1/q}\leq(C_q)^{\frac{2}{q}}+Cs\log(p/s)\leq Cs\log(p/s),$$

as $s\log(p/s)\to\infty$.

Finally, from the moments of the χ^2 distribution, for $q<n$,

$$\mathbb{E}\left(\frac{\|z\|_2}{\sqrt{n}}\right)^q=n^{-\frac{q}{2}}\cdot\frac{2^{\frac{q}{2}}\cdot\Gamma(\frac{n+q}{2})}{\Gamma(\frac{n}{2})}\leq 2^{\frac{q}{2}}=O(1).$$

Therefore, for some constant $C>0$,

$$\left(\mathbb{E}\Big(\sup_{u\in T(s)}z^TXu\Big)^q\right)^{2/q}\leq Cs\log(p/s).$$

□

Proof of Proposition 3

Proof of Proposition 3. We first bound the eigenvalues. Considering Lemma 39, let $t>0$ in the condition of Lemma 39 be small enough such that, for the $\delta_2\in(0,1/2)$ stated in Proposition 3,

$$1-\delta_2\leq\sigma_{\min}(X_{S_*}^TX_{S_*})\leq\sigma_{\max}(X_{S_*}^TX_{S_*})\leq 1+\delta_2. \qquad (3.48)$$

For example, select $t = \sqrt{k^* \log(p/k^*)/n}$ in Lemma 39, then (3.48) holds with probability at least $1 - (Cek^*/p)^{k^*}$ as $k^*/p \to 0$ and $k^* \log(p/k^*)/n \to 0$.

To show the second statement in Proposition 3, we resort to a lemma whose proof mostly follows [21].

Lemma 38. *Suppose $k^* \geq 2k$, $k^*/p \to 0$ and $(k^* \log p)/n \to 0$. Then*

$$\inf_{\|\beta\|_0 \leq k} P(S_\diamond \subseteq S_*) \geq 1 - \exp\big[-c_4 \varepsilon k \sqrt{2\log(p/k)}\big] - c_5 e^{-c_6 n\varepsilon^2},$$

for some constants $c_4, c_5, c_6 > 0$.

The proof of this lemma is demonstrated after the current proof.

Then, combining (3.48) with Lemma 38, we conclude the statement in Proposition 3. □

Proof of Lemma 38. By construction, $\mathrm{supp}(\beta) \subseteq S_*$ so we only need to show (1) $\mathrm{supp}(\hat{\beta}^L) \subseteq S_*$ and (2) $\mathrm{supp}(\tilde{\beta}^O) \subseteq S_*$.

We first demonstrate (1). Suppose $\hat{b}_{S_*}$ is the solution to the reduced problem:

$$\arg\min_{b \in \mathbb{R}^{|S_*|}} \frac{1}{2}\|y - X_{S_*} b\|^2 + \lambda_\varepsilon \|b\|_1. \tag{3.49}$$

The KKT condition of LASSO implies that if $\|X_{S_*^c}^T (y - X_{S_*} \hat{b}_{S_*})\|_\infty \leq \lambda_\varepsilon$, then $\hat{\beta}^L_{S_*} = \hat{b}_{S_*}$ and $\hat{\beta}^L_{S_*^c} = \mathbf{0}$. Hence, it's sufficient to prove the following two conditions:

$$\|X_{S_*^c}^T X_{S_*} (\beta_{S_*} - \hat{b}_{S_*})\|_\infty \leq \frac{\varepsilon}{2}\sqrt{2\log(p/k)}, \tag{3.50}$$

and

$$\|X_{S_*^c}^T z\|_\infty \leq \Big(1 + \frac{\varepsilon}{2}\Big)\sqrt{2\log(p/k)}. \tag{3.51}$$

Before analyzing (3.50) and (3.51), we first illustrate a property of the resolvent set. Let $Q \in \mathbb{R}^{n \times n}$

be an orthogonal matrix such that

$$Qz = \big(\|z\|, 0, \ldots, 0\big).$$

In the proofs, Q can further be set to be measurable with respect to z. Q is independent of X. Let $\tilde{\omega} \in \mathbb{R}^{1\times p}$, $\tilde{W} \in \mathbb{R}^{(n-1)\times p}$ and

$$W = \begin{bmatrix} \tilde{\omega} \\ \tilde{W} \end{bmatrix} := QX.$$

Q being independent of X implies that W is still a Gaussian random matrix, with i.i.d. $\mathcal{N}(0, 1/n)$ entries. Note that

$$X_i^T z = (QX_i)^T(Qz) = \|z\|(QX_i)_1 = \|z\|\tilde{\omega}_i. \tag{3.52}$$

This indicates that S_* is composed of the union of S and $k^* - k$ indices in $\{1, \ldots, p\} \setminus S$ of the largest $|\tilde{\omega}_i|$. Since $\tilde{\omega}$ and $\tilde{W}$ are independent, $\tilde{W}$ and S_* are also independent. Thus, $\tilde{W}_{S_*^c}$ and $\tilde{W}_{S_*}$ are both Gaussian random matrices.

Show (3.50). Rearrange the objective term as

$$\begin{aligned} &X_{S_*^c}^T X_{S_*}(\beta_{S_*} - \hat{b}_{S_*}) = X_{S_*^c}^T X_{S_*}(X_{S_*}^T X_{S_*})^{-1}\big(X_{S_*}^T(y - X_{S_*}\hat{b}_{S_*}) - X_{S_*}^T z\big) \\ &= X_{S_*^c}^T Q^T \underbrace{QX_{S_*}(X_{S_*}^T X_{S_*})^{-1}\big(X_{S_*}^T(y - X_{S_*}\hat{b}_{S_*}) - X_{S_*}^T z\big)}_{:=\xi}. \end{aligned}$$

We first derive the bound for $\|\xi\|$. Since $\hat{b}_{S_*}$ is the solution to the reduced Lasso problem (3.49), its KKT condition implies that

$$\|X_{S_*}(y - X_{S_*}\hat{b}_{S_*})\|_\infty \le \lambda_\varepsilon. \tag{3.53}$$

Then by Lemma 39, let $t = 1/2$ in the condition, we obtain

$$\|X_{S_*}(X_{S_*}^T X_{S_*})^{-1}\| \le \left(\sqrt{1 - 1/n} - \sqrt{k^*/n} - 1/2\right)^{-1} < 2.01 \tag{3.54}$$

with probability at least $1 - e^{-n/8}$ for sufficiently large p, where in the last step we have used $k^*/n \to 0$. In addition, we use Lemma 40 to bound $\|X_{S_*}^T z\|$. Hence, from (3.53), (3.54) and Lemma 40, we have

$$\begin{aligned}
\|\xi\| &\le \|X_{S_*}(X_{S_*}^T X_{S_*})^{-1}\| \cdot \|X_{S_*}^T(y - X_{S_*}\hat{b}_{S_*}) - X_{S_*}^T z\| \\
&\le 2.01\Big(\sqrt{k^*}\lambda_\varepsilon + 4\sqrt{2k^*\log(p/k^*)}\Big) \\
&\le 2.01\Big((1+\varepsilon)\sqrt{2} + 4\sqrt{2}\Big)\sqrt{k^*\log(p/k^*)} \\
&\le C \cdot \sqrt{k^*\log(p/k^*)} \qquad (3.55)
\end{aligned}$$

with probability at least $1 - e^{-n/2} - (\sqrt{2}ek^*/p)^{k^*} - e^{-n/8} =: 1 - P_\xi$; where in the third line we have used $k^* \ge k$.

Now, write

$$X_{S_*^c}^T X_{S_*}(\beta_{S_*} - \hat{b}_{S_*}) = X_{S_*^c}^T Q^T \xi = (\tilde{\omega}_{S_*^c}^T, 0)\xi + (0, \tilde{W}_{S_*^c}^T)\xi. \qquad (3.56)$$

For the first term, we have

$$\|(\tilde{\omega}_{S_*^c}^T, 0)\xi\|_\infty = \xi_1 \cdot \|\tilde{\omega}_{S_*^c}^T\|_\infty \le \|\xi\|_2 \cdot \|\tilde{\omega}_{S_*^c}^T\|_\infty. \qquad (3.57)$$

Based on (3.52), we recognize $\tilde{\omega}_{S_*^c}$ is a Gaussian vector excludes the $k * -k$ largest (in absolute value) coordinates from $p - k$ indices of $\{1, \ldots, p\} \setminus S$. Thus, suppose $\zeta_1, \ldots, \zeta_{p-k}$ are i.i.d. $\mathcal{N}(0,1)$ variables and $|\zeta|_{(1)} \ge \ldots \ge |\zeta|_{(p-k)}$. We have

$$\sqrt{n}\|\tilde{\omega}_{S_*^c}\|_\infty = |\zeta|_{(k^*-k+1)} \le \Big(1 + \frac{\varepsilon}{4}\Big) \cdot \sqrt{2\log\frac{2(p-k)}{k^*-k+1}} \le \Big(1 + \frac{\varepsilon}{3}\Big)\sqrt{2\log\frac{p}{k}}, \qquad (3.58)$$

where we used Lemma 30 and note that it holds with probability at least

$$1 - \exp\big[-C\varepsilon k\sqrt{2\log(p/k)}\big] =: 1 - P_1,$$

for some constant $C > 0$; we used that $k \le k^* - k + 1$ and $Cp/k \le (p-k)/k \le p/k$. Thus, from (3.57) and (3.58) and $k^* \ge k$, we obtain

$$\|(\tilde{\omega}_{S_*^c}^T, 0)\xi\|_\infty \le C \cdot \sqrt{\frac{k^*}{n} \log \frac{p}{k^*}} \cdot \sqrt{2 \log p} \le C \cdot \sqrt{\frac{k^* \log p}{n}} \sqrt{2 \log \frac{p}{k}}, \tag{3.59}$$

We continue to bound the second term in (3.56). From the discussion about (3.52), we note that ξ is independent of $\tilde{W}_{S_*^c}$, thus

$$\|(0, \tilde{W}_{S_*^c}^T)\xi\|_\infty \overset{d}{=} \left\| \sqrt{\frac{\xi_2^2 + \cdots + \xi_n^2}{n}} (\zeta_1, \ldots, \zeta_{p-k^*}) \right\|_\infty \le \|\xi\|_2 \frac{1}{\sqrt{n}} |\zeta|_{(1)},$$

with $\zeta_1, \ldots, \zeta_{p-k^*} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ and $|\zeta|_{(1)}$ denote the largest one in absolute value. From Lemma 28, we have $|\zeta|_{(1)} \le C \cdot \sqrt{2 \log(p-k^*)}$, holds with probability at least $1 - \exp[-C\sqrt{\log p}] =: 1 - P_2$, with some constant $C > 0$; we used $k^*/p \to 0$. Thus, with (3.55), we have

$$\|(0, \tilde{W}_{S_*^c}^T)\xi\|_\infty \le C \sqrt{\frac{k^*}{n} \log \frac{p}{k^*}} \sqrt{2 \log(p-k^*)} \le C \sqrt{\frac{k^* \log p}{n}} \sqrt{2 \log \frac{p}{k}}. \tag{3.60}$$

Combining (3.59) and (3.60), from (3.56) we have

$$\|X_{S_*^c}^T X_{S_*} (\beta_{S_*} - \hat{\beta}_{S_*}^L)\|_\infty \le \|(\tilde{\omega}_{S_*^c}^T, 0)\xi\|_\infty + \|(0, \tilde{W}_{S_*^c}^T)\xi\|_\infty \le C \sqrt{\frac{k^* \log p}{n}} \sqrt{2 \log(p/k)}, \tag{3.61}$$

with probability at least $1 - P_\xi - P_1 - P_2$ for (3.55), (3.57), (3.60) to hold.

Show (3.51). Following the discussion about (3.52), under event (3.57) and that $\|z\|_2/\sqrt{n} \le (1 + \varepsilon/3)$, we have

$$\|X_{S_*^c}^T z\|_\infty = \|z\|_2 \cdot \|\tilde{\omega}_{S_*^c}\|_\infty \le \left(1 + \frac{\varepsilon}{3}\right)^2 \sqrt{2 \log \frac{2(p-k)}{k^* - k + 1}} \le \left(1 + \frac{\varepsilon}{2}\right) \sqrt{2 \log \frac{p}{k}} \tag{3.62}$$

holds with probability at least $1 - P_z - P_\xi$ where

$$P\left(\|z\|_2/\sqrt{n} \ge 1 + \varepsilon/3 \right) \le e^{-n\varepsilon^2/18} := P_z.$$

As a result, we have (1):

$$\mathrm{supp}(\hat{\beta}^L) \subseteq S_*$$

holds with high probability $1 - P_1 - P_2 - P_\xi - P_z$ under events (3.55), (3.57), (3.60) and (3.62).

Back to show (2). In order to show $\mathrm{supp}(\tilde{\beta}^O) \subseteq S_*$, using the KKT condition for the convex problem defined for the oracle esimator in (3.27), it is sufficient to see that

$$\|X_{S_*^c}^T z\|_\infty \leq \lambda_\varepsilon. \tag{3.63}$$

We mark this event holding under (3.62) with corresponding high probability $1 - P_z - P_\xi$.

As a conclusion of (3.61), (3.62) and (3.63), we have that for some constants c_4, c_5, $c_6 > 0$, $\mathrm{supp}(\hat{\beta}^L) \subseteq S_*$ holds with probability at least

$$1 - P_\xi - P_1 - P_2 - P_z = 1 - \exp\Big[- c_4 \varepsilon k \sqrt{2\log(p/k)} \Big] - c_5 e^{-c_6 n \varepsilon^2}.$$

□

In the previous proof, we have cited several properties of the random Gaussian matrix $X \in \mathbb{R}^{n\times p}$. We state them as the following lemmas, which are auxiliaries proved by [21], so we skip their proofs.

Lemma 39 (Lemma A.11 in [21])**.** *Let $k < k^* < \min\{n, p\}$ be any (deterministic) integer. Denote by $\sigma_{\min}$ and $\sigma_{\max}$, respectively, the smallest and the largest singular value of X_{S_*}. Then for any $t > 0$,*

$$\sigma_{\min} > \sqrt{1 - 1/n} - \sqrt{k^*/n} - t$$

holds with probability at least $1 - e^{-nt^2/2}$. Furthermore,

$$\sigma_{\max} < \sqrt{1 - 1/n} + \sqrt{k^*/n} + \sqrt{8k^* \log(p/k^*)/n} + t$$

holds with probability at least $1 - e^{-nt^2/2} - (\sqrt{2}ek^*/p)^{k^*}$.

Lemma 40 (Lemma A.7 in [21])**.** *Let $1 \le k < p$ be any (deterministic) integer, then*

$$\sup_{|T|=k^*} \|X_T^T z\| \le \sqrt{32k^* \log(p/k^*)}$$

with probability at least $1 - e^{-n/2} - (\sqrt{2}ek^*/p)^{k^*}$. Above, the superemum is taken over all the subsets of $\{1, \ldots, p\}$ with cardinality k^*.

Proof of Proposition 4

The basic proof idea of this section comes from [15]. Throughout the proof, we will use the following notation: Let $\Big(|u|_{(1)}, \ldots, |u|_{(p)}\Big)$ denote the non-increasing rearrangement of $(|u_1|, \ldots, |u_p|)$. For given $\delta_3, \delta_4, \delta_5 \in (0, 1)$ and any $u = (u_1, \ldots, u_p) \in \mathbb{R}^p$, define

$$H(u) := (1 + \delta_4)\Big(\sum_{j=1}^{k} |u|_{(j)} 4\sqrt{\log(2p/j)} + (1 + \delta_3) \sum_{j=k+1}^{p} |u|_{(j)} \sqrt{2\log(p/k)}\Big), \tag{3.64a}$$

$$G(u) := (1 + \delta_4)\delta_5^{-1}\sqrt{\log(1/\delta_5)}\|Xu\|_2. \tag{3.64b}$$

In addition, let

$$\delta(\lambda) := e^{-\frac{\lambda^2}{2}}, \;\; \Leftrightarrow \;\; \sqrt{2\log(1/\delta(\lambda))} = \lambda. \tag{3.65}$$

We will use the following inequality: by using Stirling's formula, for any $s \in [p]$, $s\log(s/e) \le \log(s!) \le s\log(s)$. Hence,

$$s\log(2p/s) \le \sum_{j=1}^{s} \log(2p/j) = s\log(2p) - \log(s!) \le s\log(2ep/s). \tag{3.66}$$

Proof of Proposition 4. Following [15], we first show a bound for the realization of the random matrix X, i.e. fix $X \in \mathbb{R}^{n\times p}$. We have the following lemma:

Lemma 41. *Assume model* (3.1) *and let $\|\beta\|_0 \le k$. Given $\varepsilon, \delta_0 \in (0, 1)$, let δ_3, δ_4 be any positive*

numbers such that $\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3) < 1+\varepsilon$. *Assume the* $SRE(k, c_0)$ *condition holds with*

$$c_0 := \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)\Big/\left(1 - \frac{\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3)}{1+\varepsilon}\right). \tag{3.67}$$

Consider the LASSO estimator with tuning parameter λ_ε *satisfying* (3.26). *Then, on the event* (3.78)*, we have*

$$\|\hat{\beta}^L - \beta\|_2 \le C(k, \lambda_\varepsilon, \delta_5)\lambda_\varepsilon\sqrt{k}, \tag{3.68}$$

where $\delta_5 \in (0,1)$ *is a parameter in* $G(u)$ *in event* (3.78) *and*

$$C(k, \lambda_\varepsilon, \delta_5) := \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)\cdot\left(\frac{1}{\delta_4^2}\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))} \vee \frac{1}{\theta^2(k, c_0)}\right). \tag{3.69}$$

The proof of this lemma is presented later in the section.

The above lemma relies on the event (3.78) which sets a limit for the randomness from z, though it still assumes X is fixed. Then we refer to Lemma 44, which analyzes the probability for the condition of the above lemma to hold, for every $\delta_5 \in (0,1)$. By integrating the tail bound, we obtain the following lemma:

Lemma 42. *Assume model* (3.1) *with* $\|\beta\|_0 \le k$. *Let* $c_0 > 0$ *be as in* (3.67). *Consider the LASSO estimator* $\hat{\beta}^L$ (3.19) *with tuning parameter* λ_ε *in* (3.26). *Suppose the* $SRE(k, c_0)$ *holds. Then, for any* $q > 2$*,*

$$\mathbb{E}\|\hat{\beta}^L - \beta\|_2^q \le \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)^q\left(\frac{\lambda_\varepsilon\sqrt{k}}{\delta_4^2}\right)^q\left[\frac{1}{\theta^{2q}(k, c_0)} + \frac{C_q}{(\delta_4^2 k\log(p/k))^q}\right],$$

where $C_q = q/2\cdot\Gamma(q)$ *with Gamma function* $\Gamma(\cdot)$.

The proof of this lemma is presented later in the section.

Lemma 42 can be viewed as an uniform upper bound for the conditional expectation $\mathbb{E}\Big(\|\hat{\beta} -$

$\beta\|_2^q \mid X\Big)$ for all $X \in \mathcal{A}$. Then, applying the tower property, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\|\hat{\beta}-\beta\|_2^q \, \mathbb{1}_{\mathcal{A}}\Big] = \mathbb{E}\Big[\mathbb{E}\Big(\|\hat{\beta}-\beta\|_2^q \mid X\Big)\mathbb{1}_{\mathcal{A}}\Big] \\
\leq\ & \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}}+1\right)^q\Big(\frac{\lambda_\varepsilon\sqrt{k}}{\delta_4^2}\Big)^q\Big[\frac{1}{(1-\delta_0)^q}+\frac{C_q}{(\delta_4^2 k\log(p/k))^q}\Big].
\end{aligned}$$

□

The next lemma referring to Lemma A.2 in [15] describes an inequality useful for simplification of LASSO problem.

Lemma 43. *Let $h : \mathbb{R}^p \to \mathbb{R}$ be a convex function. Let $z \in \mathbb{R}^n$, X be any $n \times p$ matrix, and $y = X\beta + z$. If $\hat{\beta}^L$ is a solution of the minimization problem $\min_{\beta\in\mathbb{R}^p}\Big(\frac{1}{2}\|X\beta - y\|_2^2 + h(\beta)\Big)$, then $\hat{\beta}^L$ satisfies for all $\beta \in \mathbb{R}^p$*

$$\|X(\hat{\beta}^L-\beta)\|_2^2 \leq z^T X(\hat{\beta}^L-\beta) + h(\beta) - h(\hat{\beta}^L).$$

Proof of Lemma 41. From Lemma 43, letting $h(\cdot) = \lambda_\varepsilon\|\cdot\|_1$ in the condition of Lemma 43, we have that for all $\beta \in \mathbb{R}^p$, the following holds almost surely:

$$\|X(\hat{\beta}^L-\beta)\|_2^2 \leq \triangle^*, \tag{3.70}$$

where

$$\triangle^* := z^T X(\hat{\beta}^L-\beta) + \lambda_\varepsilon\|\beta\|_1 - \lambda_\varepsilon\|\hat{\beta}^L\|_1.$$

In the remaining proof, let $u = \hat{\beta}^L - \beta$ and define

$$\tilde{H}(u) := \sqrt{1+\delta_0}(1+\delta_4)\Big(4\|u\|_2\Big(\sum_{j=1}^{k}\log(2p/j)\Big)^{1/2} + (1+\delta_3)\sqrt{2\log(p/k)}\sum_{j=k+1}^{p}|u|_{(j)}\Big).$$

Using the Cauchy-Schwarz inequality and (3.66),

$$\sqrt{1+\delta_0}H(u) \le \tilde{H}(u) \le \frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}}\lambda_\varepsilon\sqrt{k}\|u\|_2 + \frac{\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3)}{1+\varepsilon}\lambda_\varepsilon\sum_{j=k+1}^{p}|u|_{(j)} := F(u), \tag{3.71}$$

where $H(\cdot)$ is defined in (3.64), and the last inequality follows from (3.26), (3.66) and $4\sqrt{2\log(2ep/k)} \le (4+\sqrt{2})/\sqrt{2}\cdot\sqrt{2\log(p/k)}$. Let $S := \mathrm{supp}(\beta)$,

$$\|\beta\|_1 - \|\hat{\beta}^L\|_1 = \|\beta\|_1 - \|\beta+u\|_1 = \|\beta\|_1 - \|\beta_S + u_S\|_1 - \|u_{S^c}\|_1 \le \|u_S\|_1 - \|u_{S^c}\|_1. \tag{3.72}$$

Then, on the event (3.78), using (3.71) and (3.72) we have

$$\triangle^* \le \lambda_\varepsilon\Big(\sqrt{k}\|u\|_2 - \sum_{j=k+1}^{p}|u|_{(j)}\Big) + \max\big(F(u), \tilde{G}(u)\big), \tag{3.73}$$

where $G(u)$ is defined in (3.64) and $\tilde{G}(u) := \sqrt{1+\delta_0}G(u)$. Transforming $\tilde{G}(u)$ using (3.65), we have

$$\tilde{G}(u) = \sqrt{1+\delta_0}\lambda_\varepsilon\sqrt{k}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\frac{1+\delta_4}{\delta_4}\|Xu\|_2.$$

From (3.73), we have the following discussion:

- If $\tilde{G}(u) > F(u)$, this implies

$$\|u\|_2 \le \frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\frac{1}{\delta_4}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\|Xu\|_2. \tag{3.74}$$

  Therefore,

$$\begin{aligned}\triangle^* &\le \lambda_\varepsilon\sqrt{k}\|u\|_2 + \tilde{G}(u)\\ &\le \lambda_\varepsilon\sqrt{k}\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\frac{1}{\delta_4}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\|Xu\|_2\end{aligned}$$

$$+\lambda_\varepsilon\sqrt{k}\sqrt{1+\delta_0}(1+\delta_4)\frac{1}{\delta_4}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\|Xu\|_2$$
$$=\left[\sqrt{1+\delta_0}(1+\delta_4)+\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\right]\frac{1}{\delta_4}\lambda_\varepsilon\sqrt{k}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\|Xu\|_2.$$

Combining the above with (3.70),

$$\|Xu\|_2\le\left[\sqrt{1+\delta_0}(1+\delta_4)+\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\right]\frac{1}{\delta_4}\lambda_\varepsilon\sqrt{k}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}.$$

Then, connecting the above with (3.74), we have

$$\begin{aligned}\|\hat\beta^L-\beta\|_2&\le\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\frac{1}{\delta_4}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\\&\quad\cdot\left[\sqrt{1+\delta_0}(1+\delta_4)+\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\right]\frac{1}{\delta_4}\lambda_\varepsilon\sqrt{k}\sqrt{\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}}\\&=\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\left[\sqrt{1+\delta_0}(1+\delta_4)+\frac{\sqrt{2}}{4+\sqrt{2}}(1+\varepsilon)\right]\frac{1}{\delta_4^2}\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))}\lambda_\varepsilon\sqrt{k}.\end{aligned}\tag{3.75}$$

- If $\tilde G(u)\le F(u)$, replacing the maximum with $F(u)$, we obtain

$$\begin{aligned}\triangle^*&\le\lambda_\varepsilon\Big(\sqrt{k}\|u\|_2-\sum_{j=k+1}^{p}|u|_{(j)}\Big)\\&\quad+\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}}\lambda_\varepsilon\sqrt{k}\|u\|_2+\frac{\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3)}{1+\varepsilon}\lambda_\varepsilon\sum_{j=k+1}^{p}|u|_{(j)}\\&=\left(1+\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}}\right)\lambda_\varepsilon\sqrt{k}\|u\|_2-\left(1-\frac{\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3)}{1+\varepsilon}\right)\lambda_\varepsilon\sum_{j=k+1}^{p}|u|_{(j)}\\&=:\triangle.\end{aligned}$$

From (3.70), we know $\triangle\ge\triangle^*\ge 0$ almost surely. As a result, from the above equations,

$u = \hat{\beta}^L - \beta \in C_{SRE}(k, c_0)$, where

$$c_0 := \left(1 + \frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}}\right) / \left(1 - \frac{\sqrt{1+\delta_0}(1+\delta_4)(1+\delta_3)}{1+\varepsilon}\right).$$

Note that by the assumption of this lemma, we have selected δ_3, δ_4 such that the above $c_0 > 0$.Thus, we can apply $SRE(k, c_0)$ and have

$$\|\hat{\beta}^L - \beta\|_2 \leq \frac{\|Xu\|_2}{\theta(k, c_0)} \leq \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)\lambda_\varepsilon\sqrt{k}\frac{\|Xu\|_2}{\theta^2(k, c_0)}, \tag{3.76}$$

where the second inequality is due to

$$\|Xu\|_2^2 \leq \triangle^* \leq \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)\lambda_\varepsilon\sqrt{k}\frac{\|Xu\|_2}{\theta(k, c_0)}.$$

Combining (3.75) and (3.76), we have

$$\|\hat{\beta}^L - \beta\|_2 \leq C(k, \lambda_\varepsilon, \delta_5)\lambda_\varepsilon\sqrt{k}, \tag{3.77}$$

with

$$C(k, \lambda_\varepsilon, \delta_5) := \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right) \cdot \left(\frac{1}{\delta_4^2}\frac{\log(1/\delta_5)}{k\log(1/\delta(\lambda_\varepsilon))} \vee \frac{1}{\theta^2(k, c_0)}\right).$$

□

Lemma 44 (Bound on the stochastic error)**.** *Let* $\delta_5 \in (0, 1)$ *and* $z \sim \mathcal{N}(0, I_n)$ *and* $X \in \mathbb{R}^{n\times p}$ *be a fixed matrix such that* $\max_{j\in[p]} \|Xe_j\|_2 \leq \sqrt{1+\delta_0}$*, then*

$$\left\{z^T Xu \leq \sqrt{1+\delta_0} \cdot \max\left(H(u), G(u)\right),\ \forall u \in \mathbb{R}^p\right\} \tag{3.78}$$

holds with probability at least $1 - \delta_5/2$.

Proof. Because of the scalability, we prove it for $\max_{j\in[p]} \|Xe_j\|_2 \le 1$. For $\delta_3 \in (0,1)$, let

$$N(u) := \sum_{j=1}^{k} |u|_{(j)} 4\sqrt{\log(2p/j)} + (1+\delta_3) \sum_{j=k+1}^{p} |u|_{(j)} \sqrt{2\log(p/k)}.$$

Let $g_j = z^T X e_j$, $j = 1 \ldots, p$. We have

$$\begin{aligned}
\sup_{N(u)\le 1} z^T X u &\le \sup_{N(u)\le 1} \Bigg[\sum_{j=1}^{k} 4\sqrt{\log(2p/j)}\, |u|_{(j)}\, \frac{|g|_j}{4\sqrt{\log(2p/j)}} \\
&\quad + \sum_{j=k+1}^{p} (1+\delta_3)\sqrt{2\log(p/k)}\, |u|_{(j)}\, \frac{|g|_{(j)}}{(1+\delta_3)\sqrt{2\log(p/k)}} \Bigg] \\
&\le \Big(\max_{1\le j\le k} \frac{|g|_{(j)}}{4\sqrt{\log(2p/j)}} \Big) \vee \Big(\max_{k+1\le j\le p} \frac{|g|_{(j)}}{(1+\delta_3)\sqrt{2\log(p/k)}} \Big) \\
&= \Big(\max_{1\le j\le k} \frac{|g|_{(j)}}{4\sqrt{\log(2p/j)}} \Big) \vee \frac{|g|_{(k+1)}}{(1+\delta_3)\sqrt{2\log(p/k)}}.
\end{aligned}$$

For $L > 0$ to be determined, let $T := \{u \in \mathbb{R}^p : \max\left(N(u), \frac{1}{L}\|Xu\|_2\right) \le 1\}$. Note that $f(v) := \sup_{u\in T} v^T X u$ is a Lipschitz function with Lipschitz constant L. By the concentration of the Lipschitz function of Gaussian distribution around the median (for example, Inequality (1.4) in [45]), we have with probability at least $1 - \delta_5/2$,

$$\begin{aligned}
\sup_{u\in T} z^T X u &\le \mathrm{Med}\left[\sup_{u\in T} z^T X u \right] + L\sqrt{2\log(1/\delta_5)} \\
&\le \mathrm{Med}\left[\sup_{N(u)\le 1} z^T X u \right] + L\sqrt{2\log(1/\delta_5)} \\
&\le 1 + \delta_4,
\end{aligned}$$

where in the last inequality we have let $L = \delta_4/\sqrt{2\log(1/\delta_5)}$ and used Lemma 47. □

Proof of Lemma 42. From (3.69),

$$C(k, \lambda_\varepsilon, \delta_5) \ge \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon} \frac{4+\sqrt{2}}{\sqrt{2}} + 1 \right) \frac{1}{\theta^2(k, c_0)}. \tag{3.79}$$

Then $\delta_5^* := \exp\left(\frac{k\delta_4^2}{\theta^2(k,c_0)}\log(1/\delta(\lambda_\varepsilon))\right)$ is the smallest $\delta_5 > 0$ such that the equality in (3.79) holds. From (3.68), for $t \geq \log(1/\delta_5^*) := T$,

$$Z := \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)^{-1} \frac{\delta_4^2 k \log(1/\delta(\lambda_\varepsilon))}{\lambda_\varepsilon\sqrt{k}} \|\hat{\beta}^L - \beta\|_2 \leq t,$$

with probability at least $1 - (e^{-t})/2$. Thus, for any $q > 2$,

$$\mathbb{E}Z^q = \int_0^\infty qt^{q-1}P(Z > t)dt \leq \int_0^T qt^{q-1}dt + \int_T^\infty qt^{q-1}\frac{e^{-t}}{2}dt \leq T^q + \frac{q}{2}\Gamma(q) = T^q + C_q,$$

where $\Gamma(\cdot)$ is the Gamma function. Thus, under the $SRE(k, c_0)$ condition, we obtain

$$\mathbb{E}\|\hat{\beta} - \beta\|_2^q \leq \left(\frac{\sqrt{1+\delta_0}(1+\delta_4)}{1+\varepsilon}\frac{4+\sqrt{2}}{\sqrt{2}} + 1\right)^q \left(\frac{\lambda_\varepsilon\sqrt{k}}{\delta_4^2}\right)^q \left[\frac{1}{\theta^{2q}(k,c_0)} + \frac{C_q}{(\delta_4^2 k\log(p/k))^q}\right].$$

□

The following lemmas are auxiliary to prove Lemmas 41 and 42 and Proposition 4. Some of them are based on [15]. We skip those proofs to prevent duplicate work.

The following lemma guarantees that $SRE(s, c_0)$ holds for Gaussian random matrix with high probability.

Lemma 45. *Let $X \in \mathbb{R}^{n\times p}$ have the random Gaussian design with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries. Consider the $SRE(k, c_0)$ with $c_0 > 0$ and $k \in \{1, \ldots, p\}$. There exist absolute constants $C, C' > 0$ such that the following holds. For $\forall\delta_0 \in (0, 1)$, if*

$$n \geq Cc_0^2\delta_0^{-2}k\log(2ep/k) \tag{3.80}$$

then with probability at least $1 - 3\exp\left(-C'n\delta_0^2\right)$ we have

$$\max_{j=1,\ldots,p} \|Xe_j\|^2 \leq 1 + \delta_0, \qquad \inf_{\Delta\in C_{SRE}(k,c_0):\Delta\neq 0} \frac{\|X\Delta\|_2}{\|\Delta\|_2} \geq \sqrt{1-\delta_0}. \tag{3.81}$$

Proof. The statement of the second inequality in (3.81) follows from the proof of Theorem 8.3 in [15]. We derive that for the first inequality. Applying the Bernstein type of concentration of χ^2 distribution (for example, Lemma 1 in [46]), we have for some constant $C > 0$, with probability at least $1 - \exp(-Cn\delta_0^2)$,

$$\|Xe_j\|_2^2 \leq 1 + \delta_0.$$

Then using the union bound, we have $\max_{j=1,\ldots,p} \|Xe_j\|_2^2 \leq 1 + \delta_0$ holds with probability at least

$$1 - pe^{-Cn\delta_0^2} \geq 1 - e^{-Cn\delta_0^2/2},$$

if $n \geq C\delta_0^{-2} \log p$. Note that $k \log(ep/k) \geq \log p$ for all $k \in \{1, \ldots, p\}$. The conclusion holds for the second inequality in (3.81). □

Lemma 46 (Proposition E.1 in [15])**.** *Let $g_1, \ldots, g_p$ be standard Gaussian random variables. Denote by $\left(|g|_{(1)}, \ldots, |g|_{(p)}\right)$ the non-increasing rearrangement of $\left(|g_1|, \ldots, |g_p|\right)$. Then for any $s \in \{1 \ldots, p\}$ and all $t > 0$, we have*

$$P\left(\frac{1}{s}\sum_{j=1}^{s} |g|_{(j)}^2 > t \log(2p/s)\right) \leq (2p/s)^{1-3t/8}.$$

The proof of the next lemma follows that of Proposition E.2 in [15].

Lemma 47. *Under the assumptions of Lemma 46, assume $k/p \to 0$, then*

$$\begin{aligned} &P\left(\Big(\max_{1\leq j\leq k} \frac{|g|_{(j)}}{4\sqrt{\log(2p/j)}}\Big) \vee \frac{|g|_{(k+1)}}{(1+\delta_3)\sqrt{2\log(p/k)}} \leq 1\right) \\ &\geq 1 - \frac{k}{2p} - \exp\big[-C \cdot \delta_3 k\sqrt{2\log(p/k)}\big] \geq \frac{1}{2}, \end{aligned}$$

for any $\delta_3 \geq C/(k\sqrt{2\log(p/k)})$ with some $C > 0$.

Proof. Lemma 46 with $t = 16/3$ and the inequality $|g|_{(j)}^2 \leq \frac{1}{j}\sum_{l=1}^{j} |g|_{(l)}^2$ imply

$$P\left(|g|_{(j)}^2 \leq \frac{16}{3}\log(2p/j)\right) \geq 1 - \frac{j}{2p}, \quad j = 1, \ldots, p. \tag{3.82}$$

Let $q \geq 0$ be an integer such that $2^q \leq k < 2^{q+1}$. Applying (3.82) to $j = 2^l$ for $l = 0, \ldots, q-1$ and using the union bound, we obtain that the event

$$\Omega_0 := \left\{ \max_{l=0,\ldots,q-1} \frac{|g|_{(2^l)}\sqrt{3}}{4\sqrt{\log(2p/2^l)}} \leq 1 \right\}$$

satisfies $P(\Omega_0) \geq 1 - \sum_{l=0}^{q-1} \frac{2^l}{2p} = 1 - \frac{2^q - 1}{2p} \geq 1 - \frac{k}{2p}$. For any $j < 2^q$, there exists $l \in \{0, \ldots, q-1\}$ such that $2^l \leq j < 2^{l+1}$. On the event Ω_0,

$$|g|_{(j)} \leq |g|_{(2^l)} \leq \frac{4}{\sqrt{3}}\sqrt{\log\frac{2p}{2^l}} \leq \frac{4}{\sqrt{3}}\sqrt{\log\frac{4p}{j}} \leq 4\sqrt{\log\frac{2p}{j}}, \quad \forall j < 2^q.$$

And for $2^q \leq j \leq k$,

$$|g|_{(j)} \leq |g|_{2^{q-1}} \leq \frac{4}{\sqrt{3}}\sqrt{\log\frac{2p}{2^{q-1}}} < \frac{4}{\sqrt{3}}\sqrt{\log\frac{8p}{j}} \leq 4\sqrt{\log\frac{2p}{j}}.$$

Thus, on the event Ω_0 we have $|g|_{(j)} \leq 4\sqrt{\log(2p/j)}$ for all $j = 1, \ldots, k$.

In addition, using Lemma 30, we have

$$P\left(\frac{|g|_{(k+1)}}{(1+\delta_3)\sqrt{2\log\frac{p}{k}}} \geq 1\right) \leq \exp\left[- kC\delta_3\sqrt{2\log(p/k)}\right].$$

□

3.4.4 Proof of Theorem 14

As discussed in Section 3.4.1, without of loss of generality, it's equivalent to prove the theorem in the case of $\sigma = 1$ in model (3.1). We will see that up to first order approximation of the minimax risk, there is no additional proof technique for different regimes. So we will state the upper bounds of the three regimes in one subsection and state the lower bounds in the other subsection.

Upper bound

The upper bounds for Regime (I) and (II) can be simply established from the risk of the zero estimator, as such:

$$R(\Theta(k,\mu),1) = \inf_{\hat{\beta}} \sup_{\beta\in\Theta(k,\mu)} \mathbb{E}\|\hat{\beta}-\beta\|^2 \leq \sup_{\beta\in\Theta(k,\mu)} \mathbb{E}\|\mathbf{0}-\beta\|^2 \leq k\mu^2,$$

where the last equality follows naturally from the SNR constraint in $\Theta(k,\mu)$.

The upper bound for Regime (III) follows the upper bound in Theorem 13. This is obvious because

$$\Theta(k,\mu) \subseteq \Theta(k) \Rightarrow R(\Theta(k,\mu),1) \leq R(\Theta(k),1) = 2k\log(p/k)\Big(1+o(1)\Big).$$

Lower bound

The proof of lower bound follows the same roadmap of the proof of lower bound for Theorem 13: Suppose π is a prior distribution for β. Let $B(\pi)$ be the Bayes risk of π for squared loss. Based on the definition of the minimax risk in (3.3), if $\text{supp}(\pi) \subseteq \Theta(k,\tau)$, then

$$B(\pi) \leq \inf_{\hat{\beta}} \mathbb{E}_\pi\|\hat{\beta}-\beta\|^2 \leq \inf_{\hat{\beta}} \sup_{\beta\in\Theta(k,\mu)} \mathbb{E}\|\hat{\beta}-\beta\|^2 = R(\Theta(k,\mu),1).$$

Therefore, the lower bound of the minimax risk can be provided by the Bayes risk of a prior π whose support is contained in the parameter space $\Theta(k,\mu)$.

As previously introduced in Section 3.4.2, we still consider the independent block prior $\pi_{IB}(\lambda;p,k)$. From the construction steps, it already implies that $\pi_{IB}(\lambda;p,k)$ is supported on $\Theta(k)$, i.e. satisfying the sparsity constraint in $\Theta(k,\mu)$. Consider additionally:

$$\text{If } |\lambda| \leq \mu, \text{ then } \text{supp}(\pi_{IB}(\lambda;p,k)) \subseteq \Theta(k,\mu). \tag{3.83}$$

Thus, the indepedent block prior with $0 < \lambda \leq \mu$ can provide a lower bound for the minimax risk

over $\Theta(k, \mu)$.

As we discussed in Section 3.4.2, Proposition 1 provides a lower bound for the Bayes risk of the independent block prior matching with the minimax upper bound in Theorem 13. In view of $R(\Theta(k, \mu), 1)$, it turns out that Proposition 1 can also provide lower bounds for $B(\pi_{IB}(\lambda; p, k))$ with spike location satisfying (3.83). This includes the considerations of all three regimes stated in Theorem 13. In fact, from Prosposition 1, we obtain:

- For Regime (I), let $\lambda = \mu \to 0$, then

$$B(\pi_{IB}(\lambda; p, k)) \geq k\mu^2\Big(1 + o(1)\Big).$$

- For Regime (II), let $\lambda = \mu$, then $\lambda = o\Big(\sqrt{2\log(p/k)}\Big)$ and

$$B(\pi_{IB}(\lambda; p, k)) \geq k\mu^2\Big(1 + o(1)\Big).$$

- For Regime (III), let $\lambda = \sqrt{2\log(p/k)}\big(1 + o(1)\big)$ and $\sqrt{2\log(p/k)} - \lambda \to +\infty$, then

$$B(\pi_{IB}(\lambda; p, k)) \geq 2k\log(p/k)\Big(1 + o(1)\Big).$$

3.4.5 Proof of Theorem 15

Based on the scalability of the model discussed in Section (3.4.1), it is equivalent to prove the theorem for $\sigma = 1$.

Upper bound

The following lemma constructs the upper bound from the ridge estimator. Let $\hat{\beta}^R$ denote the ridge estimator:

$$\hat{\beta}^R := \arg\min_b \ \|y - Xb\|^2 + \lambda\|b\|^2$$

$$= (X^T X + \lambda I)^{-1} X^T y.$$

Lemma 48. *Assume model* (3.1)*. Suppose* $k/p \to 0$ *and* $k/n \to 0$*. As* $\mu \to 0$*, the ridge estimator with* $\lambda = (k\mu^2)^{-1}$ *has superemum risk*

$$\sup_{\beta \in \Theta(k,\mu)} \mathbb{E}\|\hat{\beta}^R - \beta\|^2 \leq k\mu^2 \Big(1 - \frac{k\mu^2}{p} + o\Big(\frac{k\mu^2}{p}\Big)\Big).$$

Proof. The ridge estimator risk is

$$\begin{aligned}
& \mathbb{E}\|\hat{\beta}^R - \beta\|^2 = \mathbb{E}\|(X^T X + \lambda I)^{-1} X^T (X\beta + z) - \beta\|^2 \\
= \;& \mathbb{E}\|(X^T X + \lambda I)^{-1}(X^T X + \lambda I)\beta - (X^T X + \lambda I)^{-1}\lambda\beta + (X^T X + \lambda I)^{-1} X^T z - \beta\|^2 \\
= \;& \mathbb{E}\| - (X^T X + \lambda I)^{-1}\lambda\beta + (X^T X + \lambda I)^{-1} X^T z\|^2 \\
= \;& \mathbb{E}\|(X^T X + \lambda I)^{-1}\lambda\beta\|^2 + \mathbb{E}\|(X^T X + \lambda I)^{-1} X^T z\|^2, \qquad (3.84)
\end{aligned}$$

where the last step used $\mathbb{E}\lambda\beta^T (X^T X + \lambda I)^{-2} X^T z = 0$. To deal with the first term, we assume $X^T X = Q^T \Lambda Q$, where $Q \in \mathbb{R}^{p\times p}$ is orthogonal and $\Lambda = \operatorname{diag}\Big(\sigma_1(X^T X), \ldots, \sigma_p(X^T X)\Big)$. Here, $\sigma_1 \geq \ldots \geq \sigma_p$ denote the eigenvalues of $X^T X$. Using that the function $f(x) := \frac{1}{(1+x)^2} - (1 - 2x + 3x^2) \leq f(0) = 0$, $\forall x > 0$, we have

$$\begin{aligned}
& \left(\frac{1}{\lambda} X^T X + I\right)^{-2} - \left(I - \frac{2}{\lambda} X^T X + \frac{3}{\lambda^2}(X^T X)^2\right) \\
= \;& Q^T \left[\left(\frac{1}{\lambda}\Lambda + I\right)^{-2} - \left(I - \frac{2}{\lambda}\Lambda + \frac{3}{\lambda^2}\Lambda^2\right)\right] Q \\
= \;& Q^T \operatorname{diag}\left[f\Big(\frac{\sigma_1}{\lambda}\Big), \;\ldots\;, f\Big(\frac{\sigma_p}{\lambda}\Big)\right] Q \leq \mathbf{0}_{p\times p}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\|(X^T X + \lambda I)^{-1}\lambda\beta\|^2 \leq \mathbb{E}\left[\|\beta\|^2 - \frac{2\|X\beta\|^2}{\lambda} + \frac{3\beta^T (X^T X)^2 \beta}{\lambda^2}\right]$$

$$
\begin{aligned}
&= \|\beta\|^2 \cdot \left[1 - \frac{2}{\lambda} + \frac{3\mathbb{E}\beta^T (X^T X)^2 \beta}{\lambda^2 \|\beta\|^2}\right] \\
&= \|\beta\|^2 \cdot \left[1 - 2\frac{k\mu^2}{p} + \frac{3k\mu^2}{p} \cdot \frac{(k\mu^2/p) \cdot \mathbb{E}\beta^T (X^T X)^2 \beta}{\|\beta\|^2}\right] \\
&= \|\beta\|^2 \cdot \left[1 - 2\frac{k\mu^2}{p} + 3\Big(\frac{k\mu^2}{p}\Big)^2 \cdot \Big(1 + \frac{p+1}{n}\Big)\right] \\
&\le k\mu^2 \left[1 - 2\frac{k\mu^2}{p} + o\Big(\frac{k\mu^2}{p}\Big)\right]. \qquad (3.85)
\end{aligned}
$$

The first equality uses that $(n/\|\beta\|^2) \cdot \|X\beta\|^2 \sim \chi^2_n$. In the second equality, we adopt $\lambda = (k\mu^2/p)^{-1}$ and we will show $\frac{k\mu^2}{p\|\beta\|^2}\, \mathbb{E}\beta^T (X^T X)^2 \beta = o(1)$. By directly calculating the element of $(X^T X)^2$, $\mathbb{E}(X^T X)^2 = (1 + \frac{p+1}{n}) I_p$.

$$
\mathbb{E}\beta^T (X^T X)^2 \beta = (1 + \frac{p+1}{n}) \cdot \|\beta\|^2.
$$

Then, the second term in (3.84) can be calculated by $\frac{n}{\|z\|^2}\|X^T z\|^2 \sim \chi^2_p$.

$$
\mathbb{E}\|(X^T X + \lambda I)^{-1} X^T z\|^2 \le \frac{1}{\lambda^2}\mathbb{E}\|X^T z\|^2 = \frac{1}{\lambda^2} \cdot \frac{p}{n}\mathbb{E}\|z\|^2 = k\mu^2 \cdot \frac{k\mu^2}{p}. \qquad (3.86)
$$

Combining (3.85) and (3.86), we conclude

$$
\sup_{\beta \in \Theta(k,\mu)} \mathbb{E}\|\hat{\beta}^R - \beta\|^2 \le k\mu^2 \Big(1 - \frac{k\mu^2}{p} + o\Big(\frac{k\mu^2}{p}\Big)\Big).
$$

□

Lower bound

We construct the lower bound from the independent block prior described in Section 3.4.2 except that the signal can now be evenly positive or negative. Denote $\pi_{\pm IB}(\tau; p, k)$ as the symmetric independent block prior: divide $(1, \ldots, p)$ into k blocks; For each block j, let $m = \lceil p/k \rceil$ and randomly select an index $I_j \in \{(j-1)m + 1, \ldots, j \cdot m\}$; Set $\beta^{(j)} := (\beta_{(j-1)m+1}, \ldots, \beta_{jm}) = \pm\tau e_{I_j}$ evenly with probability $\frac{1}{2}$; The selection in different blocks of coordinates are independent.

Proposition 5. *Assume model* (3.1) *and parameter space* (3.4). *Suppose* $n \to \infty$ *and* $p/k \to \infty$.

If $\mu = \tau/\sigma \to 0$, then the Bayes risk of the symmetric independent block prior satisfies

$$B(\pi_{\pm IB}(\tau, p, k)) \geq k\tau^2\Big(1 - \frac{k\mu^2}{p} + o\Big(\frac{k\mu^2}{p}\Big)\Big).$$

The proof follows the argument of (3.8) and (3.9) in the proof of Proposition 1 and the following lemma. Throughout the proof, we let $m = \lceil p/k \rceil$.

Lemma 49. *Assume model* (3.1). *Consider the symmetric spike prior* $\big(\pi_S(\mu, m)\big)(\beta = \pm\mu e_j) = \frac{1}{2m}$, $j = 1, \ldots, m$. *Suppose* $n, m \to \infty$ *and* $\mu \to 0$. *Then the Bayes risk satisfies*

$$B(\pi_S(\mu, m)) \geq \mu^2 - \frac{\mu^4}{m}\Big(1 + o(1)\Big).$$

Proof. Using the symmetry of the spike prior distribution, the Bayes risk

$$\begin{aligned} B(\pi_S(\mu, m)) &= \mathbb{E}_{\mu e_1}(\hat{\beta}_1 - \mu)^2 + (m-1)\mathbb{E}_{\mu e_2}\hat{\beta}_1^2 \\ &\geq \mu^2\Big(1 - 2\mathbb{E}_{\mu e_1} p_m + (m-1)\mathbb{E}_{\mu e_2} p_m^2\Big), \end{aligned} \tag{3.87}$$

where the Bayesian estimator of β at the first coordinate is $\hat{\beta}_1 = (\hat{\beta}_\pi)_1 = \mu p_m$. Here we denote

$$p_m := \frac{\exp(\mu x_1^T y - \mu^2\|x_1\|^2/2) - \exp(-\mu x_1^T y - \mu^2\|x_1\|^2/2)}{\sum_{i=1}^m \Big[\exp(\mu x_i^T y - \mu^2\|x_i\|^2/2) + \exp(-\mu x_i^T y - \mu^2\|x_i\|^2/2)\Big]}. \tag{3.88}$$

Under $\beta = \mu e_1$,

$$p_m = \frac{\exp(\mu x_1^T(\mu x_1 + z) - \mu^2\|x_1\|^2/2) - \exp(-\mu x_1^T(\mu x_1 + z) - \mu^2\|x_1\|^2/2)}{\sum_{i=1}^m \Big[\exp(\mu x_i^T(\mu x_1 + z) - \mu^2\|x_i\|^2/2) + \exp(-\mu x_i^T(\mu x_1 + z) - \mu^2\|x_i\|^2/2)\Big]}.$$

Let $D_{n,m}$ denote the denominator of the above equation. We write

$$p_m = p_m^{(1)} + p_m^{(2)} + p_m^{(3)},$$

$$p_m^{(1)} := \frac{\exp(\mu x_1 z)\Big[\exp(\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{1}{2}\mu^2\|x_1\|^2)\Big]}{D_{n,m}}, \tag{3.89}$$

$$p_m^{(2)} := \frac{\exp(-\mu x_1 z)\Big[\exp(-\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{3}{2}\mu^2\|x_1\|^2)\Big]}{D_{n,m}}, \tag{3.90}$$

$$p_m^{(3)} := \frac{\exp(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2)}{D_{n,m}}. \tag{3.91}$$

Then from Lemmas 50 and 51, we have

$$\mathbb{E}_{\mu e_1} p_m = \mathbb{E}p_m^{(1)} + \mathbb{E}p_m^{(2)} \le \frac{\mu^2}{2m}\Big(1 + o(1)\Big) + \frac{\mu^2}{2m}\Big(1 + o(1)\Big) = \frac{\mu^2}{m}\Big(1 + o(1)\Big).$$

And from Lemma 52,

$$(m-1)\mathbb{E}_{\mu e_2} p_m^2 \ge (m-1)\cdot\frac{\mu^2}{m^2}\Big(1 + o(1)\Big) = \frac{\mu^2}{m}\Big(1 + o(1)\Big).$$

Thus, from (3.87),

$$B(\pi_S(\mu, m)) \ge \mu^2 - \frac{\mu^4}{m}\Big(1 + o(1)\Big).$$

□

Lemma 50. *Assume model* (3.1). *Suppose* $n, m \to \infty$ *and* $\mu \to 0$. *Then* $p_m^{(1)}$ *and* $p_m^{(2)}$ *defined in* (3.89) *and* (3.90) *satisfy*

$$\textit{(i)}\ \mathbb{E}p_m^{(1)} \le \frac{\mu^2}{2m}(1 + o(1)), \qquad \textit{(ii)}\ \mathbb{E}p_m^{(2)} \le \frac{\mu^2}{2m}(1 + o(1)).$$

Proof. The technique in proving (i) and (ii) will be similar. Since the numerators in both $p_m^{(1)}$ and $p_m^{(2)}$ are nonnegative, we can use

$$D_{n,m} \ge 2\sum_{i=1}^{m} \exp(-\mu^2\|x_i\|^2/2). \tag{3.92}$$

Show (i). First, we show

$$\frac{p_m^{(1)}}{\mu^2/2m} - 1 \xrightarrow{p} 0. \tag{3.93}$$

We have

$$\frac{\sum_{i=1}^m \exp(-\mu^2\|x_i\|^2/2)}{me^{\frac{\mu^2}{2}}} \xrightarrow{p} 1.$$

Because

- $\mathbb{E}\exp\Big(-\frac{\mu^2}{2}\|x_i\|^2\Big) = \Big(1-\frac{\mu^2}{n}\Big)^{-\frac{n}{2}}$.
- $\mathrm{Var}\Big(\exp\big(-\frac{\mu^2}{2}\|x_i\|^2\big)\Big) = \mathbb{E}\exp\big(-\mu^2\|x_i\|^2\big) - \Big(\mathbb{E}\exp\big(-\frac{\mu^2}{2}\|x_i\|^2\big)\Big)^2$
  $= \Big(1-2\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} - \Big(1-\frac{\mu^2}{n}\Big)^{-n} = o\big(e^{\mu^2}\big) = O(1).$

Then we can apply the weak law of large numbers. And we have

$$\frac{\exp\Big(\frac{1}{2}\mu^2\|x_1\|^2\Big) - \exp\Big(-\frac{1}{2}\mu^2\|x_i\|^2\Big)}{\mu^2} \xrightarrow{p} 1,$$
$$\exp\big(\mu x_1^T z\big) \xrightarrow{p} 1.$$

Thus, (3.93) follows.

Second, we show $\frac{2m}{\mu^2}p_m^{(1)}$ is dominated by L^1 random variable. Since $p_m^{(1)} \geq 0$,

$$\begin{aligned}
\frac{2m}{\mu^2}p_m^{(1)} &\leq \frac{2m}{\mu^2}\frac{\exp(-\mu x_1^T z)\Big[\exp(\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{1}{2}\mu^2\|x_1\|^2)\Big]}{2\sum_{i=1}^m \exp\big(-\frac{\mu^2}{2}\|x_i\|^2\big)} \\
&\overset{(a)}{\leq} \frac{2m}{\mu^2}\frac{\exp(-\mu x_1^T z)\Big[\exp(\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{1}{2}\mu^2\|x_1\|^2)\Big]}{2m\exp\Big(-\frac{1}{m}\sum_{i=1}^m \frac{\mu^2}{2}\|x_i\|^2\Big)} \\
&= \frac{1}{\mu^2}\exp\Big(\frac{1}{m}\sum_{i=1}^m \frac{\mu^2}{2}\|x_i\|^2\Big)\cdot\exp\big(-\mu x_1^T z\big)\cdot\Big[\exp(\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{1}{2}\mu^2\|x_1\|^2)\Big] \\
&= \frac{1}{\mu^2}\exp\Big[-\mu x_1^T z + \big(\frac{1}{m}+1\big)\frac{\mu^2}{2}\|x_1\|^2\Big]\cdot\exp\Big(\frac{1}{m}\frac{\mu^2}{2}\sum_{i=2}^m \|x_i\|^2\Big) \\
&- \frac{1}{\mu^2}\exp\Big[-\mu x_1^T z - \big(1-\frac{1}{m}\big)\frac{\mu^2}{2}\|x_1\|^2\Big]\cdot\exp\Big(\frac{1}{m}\frac{\mu^2}{2}\sum_{i=2}^m \|x_i\|^2\Big) := \Delta,
\end{aligned}$$

where Inequality (a) used the arithmetic-geometric inequality of the denominator. Using $x_1^T z \overset{d}{=} \|x_1\| z_1$ and by conditioning on $\|x_1\|$ first,

$$\begin{aligned}
\mathbb{E}\Delta &= \frac{1}{\mu^2}\mathbb{E}\exp\Big(\frac{\mu^2\|x_1\|^2}{2} + \big(\frac{1}{m}+1\big)\frac{\mu^2}{2}\|x_1\|^2\Big)\cdot\Big(1-\frac{\mu^2}{nm}\Big)^{-\frac{n(m-1)}{2}} \\
&\quad - \frac{1}{\mu^2}\mathbb{E}\exp\Big(\frac{\mu^2\|x_1\|^2}{2} - \big(1-\frac{1}{m}\big)\frac{\mu^2}{2}\|x_1\|^2\Big)\cdot\Big(1-\frac{\mu^2}{nm}\Big)^{-\frac{n(m-1)}{2}} \\
&= \frac{1}{\mu^2}\cdot\Big(1-\frac{\mu^2}{nm}\Big)^{-\frac{n(m-1)}{2}}\cdot\Big[\Big(1-\frac{2}{n}\big(1+\frac{1}{2m}\big)\mu^2\Big)^{-\frac{n}{2}} - \Big(1-\frac{2}{n}\frac{\mu^2}{2m}\Big)^{-\frac{n}{2}}\Big] \\
&= \frac{1}{\mu^2}\cdot(1+o(1))\Big(\mu^2+o(\mu^2)\Big) \\
&= O(1).
\end{aligned}$$

Thus, by dominated convergence theorem,

$$\mathbb{E}p_m^{(1)} \le \frac{\mu^2}{2m}(1+o(1)).$$

Show (ii). First, we show

$$\frac{p_m^{(2)}}{\mu^2/2m} - 1 \xrightarrow{p} 0. \tag{3.94}$$

Because

- $\frac{\exp\left(-\frac{1}{2}\mu^2\|x_1\|^2\right)-\exp\left(-\frac{3}{2}\mu^2\|x_1\|^2\right)}{\mu^2} \xrightarrow{p} 1.$
- $\exp(-\mu x_1^T z) \overset{d}{=} \exp(\mu x_1^T z) \xrightarrow{p} 1.$

Second, we show that $\frac{p_m^{(2)}}{\mu^2/2m}$ is dominated by a L^1 random variable. Since $p_m^{(2)} \ge 0$,

$$\begin{aligned}
\frac{2m}{\mu^2}p_m^{(2)} &\le \frac{2m}{\mu^2}\frac{\exp(-\mu x_1^T z)\Big[\exp(-\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{3}{2}\mu^2\|x_1\|^2)\Big]}{2\sum_{i=1}^m \exp\big(-\frac{\mu^2}{2}\|x_i\|^2\big)} \\
&\le \frac{2m}{\mu^2}\frac{\exp(-\mu x_1^T z)\Big[\exp(-\frac{1}{2}\mu^2\|x_1\|^2) - \exp(-\frac{3}{2}\mu^2\|x_1\|^2)\Big]}{2m\exp\Big(-\frac{1}{m}\sum_{i=1}^m \frac{\mu^2}{2}\|x_i\|^2\Big)}
\end{aligned}$$

$$= \frac{1}{\mu^2}\exp\Big(\frac{1}{m}\frac{\mu^2}{2}\sum_{i=2}^{m}\|x_i\|^2\Big)\cdot\Big[\exp\Big(\mu x_1^T z - \big(1-\frac{1}{m}\big)\frac{\mu^2}{2}\|x_1\|^2\Big)$$
$$-\exp\Big(-\mu x_1^T z - \big(3-\frac{1}{m}\big)\frac{\mu^2}{2}\|x_1\|^2\Big)\Big] := \Gamma.$$

And

$$\begin{aligned}
\mathbb{E}\Gamma &= \frac{1}{\mu^2}\Big(1-\frac{\mu^2}{nm}\Big)^{-\frac{n(m-1)}{2}}\mathbb{E}\Big[\exp\Big(\frac{\mu^2}{2}\|x_1\|^2 - \big(1-\frac{1}{m}\big)\frac{\mu^2}{2}\|x_1\|^2\Big)\\
&\quad -\exp\Big(\frac{\mu^2}{2}\|x_1\|^2 - \big(3-\frac{1}{m}\big)\frac{\mu^2}{2}\|x_1\|^2\Big)\Big]\\
&= \frac{1}{\mu^2}\Big(1-\frac{\mu^2}{nm}\Big)^{-\frac{n(m-1)}{2}}\Big[\Big(1-\frac{2}{n}\frac{\mu^2}{2m}\Big)^{-\frac{n}{2}} - \Big(1-\frac{2}{n}\big(-1+\frac{1}{2m}\big)\mu^2\Big)^{-\frac{n}{2}}\Big]\\
&= \frac{1}{\mu^2}\big(1+o(1)\big)\cdot\big(\mu^2+o(\mu^2)\big)\\
&= O(1).
\end{aligned}$$

Thus, by dominated convergence theorem,

$$\mathbb{E}p_m^{(2)} \le \frac{\mu^2}{2m}\big(1+o(1)\big).$$

□

Lemma 51. *Assume model* (3.1). *Suppose* $n, m \to \infty$ *and* $\mu \to 0$. *Then* $p_m^{(3)}$ *defined in* (3.91) *satisfies*

$$\mathbb{E}p_m^{(3)} = o\Big(\frac{\mu^2}{m}\Big).$$

Proof. Since $z \overset{d}{=} -z$, let

$$\frac{\exp(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2)}{D_{n,m}} = \frac{\exp(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2)}{A},$$
$$\frac{\exp(-\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2)}{D_{n,m}} = \frac{\exp(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2)}{B},$$

where

$$
\begin{aligned}
A &:= \exp\Big(\frac{1}{2}\mu^2\|x_1\|^2 + \mu x_1^T z\Big) + \exp\Big(-\frac{3}{2}\mu^2\|x_1\|^2 - \mu x_1^T z\Big) \\
&\quad + \sum_{i=2}^{m} \exp\Big(\mu x_i^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_i\|^2\Big) + \exp\Big(-\mu x_i^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_i\|^2\Big), \\
B &\overset{d}{:=} \exp\Big(\frac{1}{2}\mu^2\|x_1\|^2 - \mu x_1^T z\Big) + \exp\Big(-\frac{3}{2}\mu^2\|x_1\|^2 + \mu x_1^T z\Big) \\
&\quad + \sum_{i=2}^{m} \exp\Big(\mu x_i^T(\mu x_1 - z) - \frac{\mu^2}{2}\|x_i\|^2\Big) + \exp\Big(-\mu x_i^T(\mu x_1 - z) - \frac{\mu^2}{2}\|x_i\|^2\Big) \\
&= \exp\Big(-2\mu^2\|x_1\|^2\Big)\exp\Big(\frac{1}{2}\mu^2\|x_1\|^2 + \mu x_1^T z\Big) + \exp\Big(2\mu^2\|x_1\|^2\Big)\exp\Big(-\frac{3}{2}\mu^2\|x_1\|^2 - \mu x_1^T z\Big) \\
&\quad + \sum_{i=2}^{m}\Big[\exp\Big(-2\mu^2 x_i^T x_1\Big)\exp\Big(\mu x_i^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_i\|^2\Big) \\
&\quad + \exp\Big(2\mu^2 x_i^T x_1\Big)\exp\Big(-\mu x_i^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_i\|^2\Big)\Big].
\end{aligned}
$$

Then,

$$
\begin{aligned}
\mathbb{E}p_m^{(3)} &= \frac{(B-A)\exp\big(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2\big)}{BA}, \\
&= \mathbb{E}\frac{I\cdot\exp\big(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2\big) + II\cdot\exp\big(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2\big)}{BA} \\
&= \mathbb{E}\frac{\Delta_1}{BA} + \mathbb{E}\frac{\Delta_2}{BA},
\end{aligned}
$$

where

$$
\left.
\begin{aligned}
B - A = &\Big[\exp\big(-2\mu^2\|x_1\|^2\big) - 1\Big]\cdot\exp\big(\frac{1}{2}\mu^2\|x_1\|^2 + \mu x_1^T z\big) \\
&+ \Big[\exp\big(2\mu^2\|x_1\|^2\big) - 1\Big]\cdot\exp\big(-\frac{3}{2}\mu^2\|x_1\|^2 - \mu x_1^T z\big)
\end{aligned}
\right\} := I
$$

$$
\left.
\begin{aligned}
&+ \sum_{i=2}^{m}\Big[\exp\big(-2\mu^2 x_i^T x_1\big) - 1\Big]\cdot\exp\big(\mu x_i^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_i\|^2\big) \\
&+ \Big[\exp\big(2\mu^2 x_i^T x_1\big) - 1\Big]\cdot\exp\big(-\mu x_i^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_i\|^2\big)
\end{aligned}
\right\} := II
$$

and

$$\Delta_1 := I \cdot \exp\big(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2\big), \qquad \Delta_2 := II \cdot \exp\big(\mu x_1^T z - \frac{1}{2}\mu^2\|x_1\|^2\big).$$

First, we show

$$\mathbb{E}\frac{\Delta_1}{BA} = o\Big(\frac{\mu^2}{m}\Big). \tag{3.95}$$

We have

$$\begin{aligned}
\Delta_1 &= \Big[\exp\big(-2\mu^2\|x_1\|^2\big) - 1\Big] \cdot \exp\big(2\mu x_1^T z\big) + \Big[\exp\big(2\mu^2\|x_1\|^2\big) - 1\Big] \cdot \exp\big(-2\mu^2\|x_1\|^2\big) \\
&= \Big[1 - \exp\big(-2\mu^2\|x_1\|^2\big)\Big] \cdot \Big[1 - \exp\big(2\mu x_1^T z\big)\Big] \\
&\le \Big[1 - \exp\big(-2\mu^2\|x_1\|^2\big)\Big]\Big[1 - \exp\big(-2\mu^2\|x_1\|^2\big)\Big] \cdot \mathbb{1}_{(x_1^T z \le 0)}.
\end{aligned}$$

The last line is non-negative, so we can use the lower bound of BA.

$$\begin{aligned}
&\frac{m}{\mu^2}\mathbb{E}\frac{\Delta_1}{BA} \le \frac{m}{\mu^2}\mathbb{E}\frac{\big[1 - \exp(-2\mu^2\|x_1\|^2)\big] \cdot \big[1 - \exp(2\mu x_1^T z)\big] \cdot \mathbb{1}_{(x_1^T z \le 0)}}{\big[2m\exp\big(-\frac{1}{m}\frac{\mu^2}{2}\sum_{i=1}^m \|x_i\|^2\big)\big]^2} \\
= \;& \frac{1}{4m\mu^2}\mathbb{E}\exp\big(\frac{1}{m}\mu^2\sum_{i=2}^m\|x_i\|^2\big) \cdot \mathbb{E}\Big\{\Big[\exp\big(\frac{1}{m}\mu^2\|x_i\|^2\big) - \exp\big(-(2-\frac{1}{m})\mu^2\|x_1\|^2\big)\Big] \\
& \cdot \Big[1 - \exp(2\mu x_1^T z)\Big] \cdot \mathbb{1}_{(x_1^T z \le 0)}\Big\} \\
\le \;& \frac{1}{4m\mu^2} \cdot \Big[1 - \frac{2}{nm}\mu^2\Big]^{-\frac{n(m-1)}{2}} \cdot \mathbb{E}\Big\{\Big[\exp\big(\frac{1}{m}\mu^2\|x_i\|^2\big) - \exp\big(-(2-\frac{1}{m})\mu^2\|x_1\|^2\big)\Big] \\
& \cdot 2\mu\big(-x_1^T z\big) \cdot \mathbb{1}_{(x_1^T z \le 0)}\Big\},
\end{aligned}$$

where the last inequality is using $1 - e^{-t} \le t$, $\forall t \ge 0$. Let $\cos\theta = \frac{x_1^T z}{\|x_1\|\|z\|}$. $(\|x_1\|, \|z\|, \cos\theta)$ are mutually independent. Conditioned on $\{x_1^T z \le 0\}$, θ is uniformly distributed in $[-\pi, -\frac{\pi}{2}] \cup [\frac{\pi}{2}, \pi]$. The expectation in the last line is

$$\begin{aligned}
& \mathbb{E}\Big\{\Big[\exp\big(\frac{1}{m}\mu^2\|x_i\|^2\big) - \exp\big(-(2-\frac{1}{m})\mu^2\|x_1\|^2\big)\Big] \cdot 2\mu\big(-x_1^T z\big) \cdot \mathbb{1}_{(x_1^T z \le 0)}\Big\} \\
= \;& \mathbb{E}\Big\{\Big[\exp\big(\frac{1}{m}\mu^2\|x_i\|^2\big) - \exp\big(-(2-\frac{1}{m})\mu^2\|x_1\|^2\big)\Big] \cdot 2\mu\|x_1\|\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\cdot\mathbb{E}\Big[\|z\|\Big]\cdot\mathbb{E}\Big\{(-\cos\theta)\mathbb{1}_{(|\theta|\in[\frac{\pi}{2},\pi])}\Big\}\\
=\;& \frac{1}{\sqrt{n}}\frac{\sqrt{2}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\Big\{\Big[1-\frac{2\mu^2}{nm}\Big]^{-\frac{n+1}{2}}-\Big[1+\frac{2}{n}(2-\frac{1}{m})\mu^2\Big]^{-\frac{n+1}{2}}\Big\}\cdot\frac{\sqrt{2}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\cdot\frac{1}{\pi}\\
=\;& O\Big(\frac{\mu^3}{\sqrt{n}}\Big).
\end{aligned}
$$

Thus,

$$
\frac{m}{\mu^2}\mathbb{E}\frac{\Delta_1}{BA}\le O\Big(\frac{\mu}{m\sqrt{n}}\Big)=o(1).
$$

Second, we show

$$
\mathbb{E}\frac{\Delta_2}{BA}=o\Big(\frac{\mu^2}{m}\Big).
$$

We have

$$
\begin{aligned}
\Delta_2 =\;& \exp\Big(\mu x_1^Tz-\frac{1}{2}\mu^2\|x_1\|^2\Big)\cdot\sum_{i=2}^m\Big[\Big(\exp\big(-2\mu^2x_i^Tx_1\big)-1\Big)\cdot\exp\Big(\mu x_i^T(\mu x_1+z)-\frac{\mu^2}{2}\|x_i\|^2\Big)\Big]\\
&+\exp\Big(\mu x_1^Tz-\frac{1}{2}\mu^2\|x_1\|^2\Big)\cdot\sum_{i=2}^m\Big[\Big(\exp\big(2\mu^2x_i^Tx_1\big)-1\Big)\cdot\exp\Big(-\mu x_i^T(\mu x_1+z)-\frac{\mu^2}{2}\|x_i\|^2\Big)\Big].
\end{aligned}
$$

We let $\Delta_2^{(1)}$ denote the first line and $\Delta_2^{(2)}$ denote the second line of the above equation. $\Delta_2^{(1)}\overset{d}{=}\Delta_2^{(2)}$ by $x_i\overset{d}{=}-x_i$ for each $i=2,\dots,m$. Then,

$$
\begin{aligned}
&\frac{m}{\mu^2}\mathbb{E}\frac{\Delta_2}{BA}=\frac{2m}{\mu^2}\mathbb{E}\frac{\Delta_2^{(2)}}{BA}\le\frac{2m}{\mu^2}\sum_{i=2}^m\mathbb{E}\Big[\exp\Big(\mu x_1^Tz-\frac{1}{2}\mu^2\|x_1\|^2\Big)\Big(\exp\big(2\mu^2x_i^Tx_1\big)-1\Big)\\
&\cdot\exp\Big(-\mu x_i^T(\mu x_1+z)-\frac{\mu^2}{2}\|x_i\|^2\Big)/(BA)\Big]\\
\le\;&\frac{2m(m-1)}{\mu^2}\cdot\mathbb{E}\Big\{\exp\Big(\mu x_1^Tz-\frac{1}{2}\mu^2\|x_1\|^2\Big)\cdot\Big[\exp\big(2\mu^2x_2^Tx_1\big)-1\Big]\cdot\exp\Big(-\mu x_2^T(\mu x_1+z)-\frac{\mu^2}{2}\|x_2\|^2\Big)\\
&\cdot\mathbb{1}_{(x_2^Tx_1\ge0)}\Big/\Big[2m\exp\Big(-\frac{1}{m}\frac{\mu^2}{2}\sum_{i=1}^m\|x_i\|^2\Big)\Big]^2\Big\}\\
=\;&\frac{m-1}{2m\mu^2}\cdot\mathbb{E}\Big\{\exp\Big(\mu(x_1-x_2)^Tz-\frac{1}{2}\mu^2\|x_1\|^2-\mu^2x_1^Tx_2-\frac{1}{2}\mu^2\|x_2\|^2\Big)\\
&\cdot\Big[\exp(2\mu^2x_1^Tx_2)-1\Big]\cdot\exp\Big(\frac{1}{m}\mu^2\sum_{j=1}^m\|x_i\|^2\Big)\mathbb{1}_{(x_1^Tx_2\ge0)}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{m-1}{2m\mu^2}\cdot\Big[1-\frac{2}{nm}\mu^2\Big]^{-\frac{n(m-2)}{2}}\cdot\mathbb{E}\Big\{\exp\Big(\frac{\mu^2}{2}\|x_1-x_2\|^2-\frac{1}{2}\mu^2\|x_1\|^2-\mu^2x_1^Tx_2-\frac{1}{2}\mu^2\|x_2\|^2\Big)\\
&\quad\cdot\Big[\exp(2\mu^2x_1^Tx_2)-1\Big]\exp\Big(\frac{1}{m}\mu^2\|x_1\|^2+\frac{1}{m}\mu^2\|x_2\|^2\Big)\cdot\mathbb{1}_{(x_1^Tx_2\geq 0)}\Big\}\\
&= \frac{m-1}{2m\mu^2}\cdot\Big[1-\frac{2}{nm}\mu^2\Big]^{-\frac{n(m-2)}{2}}\cdot\mathbb{E}\Big\{\Big[1-\exp(-2\mu^2x_1^Tx_2)\Big]\cdot\exp\Big(\frac{1}{m}\mu^2\|x_1\|^2+\frac{1}{m}\mu^2\|x_2\|^2\Big)\cdot\mathbb{1}_{(x_1^Tx_2\geq 0)}\Big\}\\
&\leq \frac{m-1}{2m\mu^2}\cdot\Big[1-\frac{2}{nm}\mu^2\Big]^{-\frac{n(m-2)}{2}}\cdot\mathbb{E}\Big[2\mu^2\|x_1\|\cdot\|x_2\|\cdot\cos\theta\mathbb{1}_{(|\theta|\leq\frac{\pi}{2})}\cdot\exp\Big(\frac{1}{m}\mu^2\|x_1\|^2+\frac{1}{m}\mu^2\|x_2\|^2\Big)\Big]\\
&= \frac{m-1}{2m\mu^2}\cdot\Big[1-\frac{2}{nm}\mu^2\Big]^{-\frac{n(m-2)}{2}}\cdot\frac{2\mu^2}{\pi}\cdot\Big[\mathbb{E}\|x_1\|\cdot\exp\Big(\frac{\mu^2}{m}\|x_1\|^2\Big)\Big]^2\\
&= O\Big(\frac{1}{n}\Big),
\end{aligned}
$$

where in the last equality we used

$$
\begin{aligned}
&\mathbb{E}\|x_1\|\cdot\exp\Big(\frac{\mu^2}{m}\|x_1\|^2\Big)\\
&= \int_0^\infty\frac{1}{2^{n/2}\Gamma(n/2)}\frac{1}{\sqrt{n}}r^{\frac{n+1}{2}-1}\exp\Big(-\frac{1}{2}(1-\frac{2\mu^2}{nm})r\Big)dr\\
&= \frac{2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2})}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\frac{1}{\sqrt{n}}\Big(1-\frac{2\mu^2}{nm}\Big)^{-\frac{n+1}{2}}\\
&= O\Big(\frac{1}{\sqrt{n}}\Big).
\end{aligned}
$$

Thus,

$$
\frac{m}{\mu^2}\mathbb{E}\frac{\Delta_2}{BA}\leq O\Big(\frac{1}{n}\Big)=o(1).
$$

□

Lemma 52. *Assume model* (3.1)*. Suppose* $n, m\to\infty$ *and* $\mu\to 0$*. Then under* $\beta=\mu e_2$*,* p_m *defined in* (3.88) *satisfies*

$$
\mathbb{E}_{\mu e_2}p_m^2\geq\frac{\mu^2}{m^2}\Big(1+o(1)\Big).
$$

Proof. Under $\beta = \mu e_2$,

$$\mathbb{E}_{\mu e_2} p_m^2 = \mathbb{E}\frac{\Big[\exp\big(\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\big) - \exp\big(-\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\big)\Big]^2}{\Big[\sum_{i=1}^m \exp\big(\mu x_i^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_i\|^2\big) + \exp\big(-\mu x_i^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_i\|^2\big)\Big]^2}.$$

Observe that the numerator is free from $(x_3, \ldots, x_m)$. By conditioning on (x_1, x_2, z), apply Jensen's inequality on $f(x) := \frac{1}{(x+c)^2}$, $x > 0$,

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=3}^m \exp\Big(\mu x_i^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_i\|^2\Big) + \exp\Big(-\mu x_i^T(\mu x_2+z) - \frac{\mu^2}{2}\|x_i\|^2\Big)\Big|(x_1,x_2,z)\Big] \\
= \;& 2(m-2)\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\Big(\frac{1}{2}\frac{\mu^2/n}{1+\mu^2/n}\|\mu x_2 + z\|^2\Big).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\mu e_2} p_m^2 \geq\; & \mathbb{E}\Big[\exp\Big(\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big) - \exp\Big(-\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big)\Big]^2 \\
& \cdot \Big[\exp\Big(\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big) + \exp\Big(-\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big) \\
& + \exp\Big(\mu x_2^T z + \frac{1}{2}\mu^2\|x_2\|^2\Big) + \exp\Big(-\mu x_2^T z - \frac{3}{2}\mu^2\|x_2\|^2\Big) \\
& + 2(m-2)\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\Big(\frac{1}{2}\frac{\mu^2/n}{1+\mu^2/n}\|\mu x_2 + z\|^2\Big)\Big]^{-2}.
\end{aligned}$$

To further simplify the denominator, we note that the numerator depends on x_2 and z only through $\mu x_2 + z$. We construct a random variable $v := -\frac{n}{\mu}x_2 + z$ being independent of $\mu x_2 + z$, and take conditional expectation of v on other variables in the denominator by applying Jensen's inequality on function $f(x) = \frac{1}{(x+c)^2}$,

$$\begin{aligned}
\mathbb{E}_{\mu e_2} p_m^2 \geq\; & \mathbb{E}\Big\{\Big[\exp\Big(\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big) - \exp\Big(-\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big)\Big]^2 \\
& \cdot \Big(\mathbb{E}_v\Big[\exp\Big(\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big) + \exp\Big(-\mu x_1^T(\mu x_2+z) - \frac{1}{2}\mu^2\|x_1\|^2\Big) \\
& + \exp\Big(\mu x_2^T z + \frac{1}{2}\mu^2\|x_2\|^2\Big) + \exp\Big(-\mu x_2^T z - \frac{3}{2}\mu^2\|x_2\|^2\Big)
\end{aligned}$$

$$
\begin{aligned}
&+ 2(m-2)\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\Big(\frac{1}{2}\frac{\mu^2/n}{1+\mu^2/n}\|\mu x_2+z\|^2\Big)\Big]\Big)^{-2}\Big\} \\
&\overset{(a)}{=} \mathbb{E}\Big\{\Big[\exp\Big(\mu x_1^T(\mu x_2+z)-\frac{1}{2}\mu^2\|x_1\|^2\Big)-\exp\Big(-\mu x_1^T(\mu x_2+z)-\frac{1}{2}\mu^2\|x_1\|^2\Big)\Big)\Big]^2 \\
&\cdot\Big[\exp\Big(\mu x_1^T(\mu x_2+z)-\frac{1}{2}\mu^2\|x_1\|^2\Big)+\exp\Big(-\mu x_1^T(\mu x_2+z)-\frac{1}{2}\mu^2\|x_1\|^2\Big) \\
&+\Big(1+\frac{1}{1+\frac{n}{\mu^2}}\Big)^{-\frac{n}{2}}\exp\Big(\frac{3\big(\frac{n}{\mu^2}\big)^2+5\frac{n}{\mu^2}+2}{2\big(1+\frac{n}{\mu^2}\big)^2\big(2+\frac{n}{\mu^2}\big)}\|\mu x_2+z\|^2\Big) \\
&+\Big(1+\frac{1}{1+\frac{n}{\mu^2}}\Big)^{-\frac{n}{2}}\exp\Big(-\frac{1}{2\big(1+\frac{n}{\mu^2}\big)}\|\mu x_2+z\|^2\Big) \\
&+ 2(m-2)\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\Big(\frac{1}{2}\frac{\mu^2/n}{1+\mu^2/n}\|\mu x_2+z\|^2\Big)\Big]^{-2}\Big\} \\
&\overset{(b)}{\geq} \mathbb{E}\Big\{4\mu^2\Big(x_1^T(\mu x_2+z)\Big)^2\exp\Big(-\mu^2\|x_1\|^2\Big) \\
&\cdot\Big[\exp\Big(\mu x_1^T(\mu x_2+z)-\frac{1}{2}\mu^2\|x_1\|^2\Big)+\exp\Big(-\mu x_1^T(\mu x_2+z)-\frac{1}{2}\mu^2\|x_1\|^2\Big) \\
&+\Big(1+\frac{1}{1+\frac{n}{\mu^2}}\Big)^{-\frac{n}{2}}\exp\Big(\frac{3\big(\frac{n}{\mu^2}\big)^2+5\frac{n}{\mu^2}+2}{2\big(1+\frac{n}{\mu^2}\big)^2\big(2+\frac{n}{\mu^2}\big)}\|\mu x_2+z\|^2\Big) \\
&+\Big(1+\frac{1}{1+\frac{n}{\mu^2}}\Big)^{-\frac{n}{2}}\exp\Big(-\frac{1}{2\big(1+\frac{n}{\mu^2}\big)}\|\mu x_2+z\|^2\Big) \\
&+ 2(m-2)\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\Big(\frac{1}{2}\frac{\mu^2/n}{1+\mu^2/n}\|\mu x_2+z\|^2\Big)\Big]^{-2}\Big\} \\
&\geq \mathbb{E}\Big\{4\mu^2\Big(x_1^T(\mu x_2+z)\Big)^2\exp\Big(-\mu^2\|x_1\|^2\Big) \\
&\cdot\Big[\exp\Big(\mu x_1^T(\mu x_2+z)\Big)+\exp\Big(-\mu x_1^T(\mu x_2+z)\Big) \\
&+\exp\Big(\frac{3}{2}\frac{\mu^2}{n}\|\mu x_2+z\|^2\Big)+1 \\
&+ 2(m-2)\exp\Big(\frac{1}{2}\frac{\mu^2}{n}\|\mu x_2+z\|^2\Big)\Big]^{-2}\Big\} \\
&\overset{(c)}{=} \mathbb{E}\Big\{4\mu^2 w_1^2\|\mu x_2+z\|^2 e^{-\mu^2 w_1^2}e^{-\mu^2\|w_{-1}\|^2} \\
&\cdot\Big[\exp\Big(\mu w_1\|\mu x_2+z\|\Big)+\exp\Big(-\mu w_1\|\mu x_2+z\|\Big) \\
&+\exp\Big(\frac{3}{2}\frac{\mu^2}{n}\|\mu x_2+z\|^2\Big)+1 \\
&+ 2(m-2)\exp\Big(\frac{1}{2}\frac{\mu^2}{n}\|\mu x_2+z\|^2\Big)\Big]^{-2}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(d)}{\geq} \frac{4\mu^2}{M}\mathbb{E}\Big[w_1^2 e^{-\mu^2 w_1^2}\mathbb{1}_{(|w_1|\leq b/\sqrt{n})}\Big]\cdot\mathbb{E}\Big[\|\mu x_2+z\|^2\mathbb{1}_{(\|\mu x_2+z\|\leq c\sqrt{n})}\Big]\cdot\mathbb{E}\Big[e^{-\mu^2\|w_{-1}\|^2}\Big]\\
&\overset{(e)}{\geq} \frac{4\mu^2}{M}\Big(1+\frac{2}{n}\mu^2\Big)^{-\frac{n-1}{2}}\frac{1}{n\big(1+2\frac{\mu^2}{n}\big)}\Big[-\frac{2b}{\sqrt{2\pi}}e^{-\frac{1}{2}(1+\frac{2\mu^2}{n})b^2}+\frac{1}{\sqrt{1+\frac{2\mu^2}{n}}}\int_{-b\sqrt{1+\frac{2\mu^2}{n}}}^{b\sqrt{1+\frac{2\mu^2}{n}}}\phi(x)dx\Big]\\
&\cdot n\Big(1+\frac{\mu^2}{n}\Big)\cdot\Big[1-\exp\Big(-\frac{1}{2}\Big(\sqrt{\frac{c^2n}{1+\frac{\mu^2}{n}}}-\frac{d}{2}-\sqrt{\frac{d}{2}}\Big)^2\Big)\Big]\\
&\overset{(f)}{=} \frac{4\mu^2}{M}\Big(1+o(1)\Big)\\
&\overset{(g)}{=} \frac{\mu^2}{m^2}\Big(1+o(1)\Big).
\end{aligned}
$$

Equality (a) uses for $\forall a\in\mathbb{R}^n$, $b\sim\mathcal{N}(0,\sigma^2 I_n)$, ϵ_1 and ϵ_2 being two constants,

$$
\mathbb{E}_b\Big[\exp\Big(-\frac{1}{2}\epsilon_1\|b\|^2-\epsilon_2 a^T b\Big)\Big|a\Big]=\big(1+\epsilon_1\sigma^2\big)^{-\frac{n}{2}}\exp\Big(\frac{\sigma^2\epsilon_2^2}{2(1+\epsilon_1\sigma^2)}\|a\|^2\Big).
$$

Inequality (b) uses that $(e^x-e^{-x})^2\geq(2x)^2$ for all $x\in\mathbb{R}$. In Equality (c), we let $Q\in\mathbb{R}^{n\times n}$ be an orthogonal matrix such that $Q(\mu x_2+z)=(\|\mu x_2+z\|,0,\ldots,0)$. Denote $Qx_1=:w=(w_1,w_{-1})$, $w_1\in\mathbb{R}$, $w_{-1}\in\mathbb{R}^{n-1}$. Then, $x_1^T(\mu x_2+z)=w_1\|\mu x_2+z\|$, $\|x_1\|^2=\|w_1\|^2$, and $w\sim\mathcal{N}(0,\frac{1}{n}I_n)$, w is independent of $\|\mu x_2+z\|$. In Inequality (d), $M:=\Big[e^{\mu bc}+e^{-\mu bc}+e^{\frac{3}{2}\mu^2c^2}+1+2(m-2)e^{\frac{1}{2}\mu^2c^2}\Big]^2$. Inequality (e) uses that

- $\mathbb{E}w_1^2e^{-\mu^2w_1^2}\mathbb{1}_{(|w_1|\leq b/\sqrt{n})}=\frac{1}{n(1+2\frac{\mu^2}{n})}\Big[-\frac{2b}{\sqrt{2\pi}}e^{-\frac{1}{2}(1+\frac{2\mu^2}{n})b^2}+\frac{1}{\sqrt{1+\frac{2\mu^2}{n}}}\int_{-b\sqrt{1+\frac{2\mu^2}{n}}}^{b\sqrt{1+\frac{2\mu^2}{n}}}\phi(x)dx\Big]$.
- $\mathbb{E}\|\mu x_2+z\|^2\mathbb{1}_{(\|\mu x_2+z\|\leq c\sqrt{n})}=\Big(\frac{\mu^2}{n}+1\Big)n\cdot P\Big(\chi^2_{n+1}\leq\frac{c^2n}{\mu^2/n+1}\Big)$.
  $\geq n\Big(1+\frac{\mu^2}{n}\Big)\cdot\Big[1-\exp\Big(-\frac{1}{2}\Big(\sqrt{\frac{c^2n}{1+\frac{\mu^2}{n}}}-\frac{d}{2}-\sqrt{\frac{d}{2}}\Big)^2\Big)\Big]$

Equality (f) holds by setting $b=c=\mu^{-\frac{1}{4}}=\omega(1)$. Finally, Inequality (g) holds because $M\leq 4m^2e^{\sqrt{\mu}}$ for all μ being sufficiently small. □

3.4.6 Proof of Theorem 16

Lower bound

Due to the scalability discussion in Section 3.4.1, we present the proof for $\sigma = 1$. Let $\pi_{\pm IB}$ be the symmetric independent block prior described in Section 3.4.5. The following proposition states the lower bound.

Proposition 6. *Assume model* (3.1). *Suppose* $n \to \infty$, $p/k \to \infty$ *and* $\log(p/k)/n \to 0$. *Let* $\mu := \tau/\sigma \to \infty$ *and* $\mu = o\big(\sqrt{\log(p/k)}\big)$. *Additionally, assume* $\mu^4/n \to 0$. *Then the Bayes risk of the symmetric independent block prior satisfies*

$$B(\pi_{\pm IB}(\tau; p, k)) \geq k\tau^2\Big(1 - \frac{k\mu^2}{2p} \cdot e^{\mu^2}\big(1 + o(1)\big)\Big).$$

The proof directly follows the argument of (3.8) and (3.9) in the proof of Proposition 1 and the following lemma.

Lemma 53. *Assume model* (3.1). *Suppose* $n \to \infty$, $p/k \to \infty$ *and* $\log(p/k)/n \to 0$. *Let* $\mu := \tau/\sigma \to \infty$ *and* $\mu = o\big(\sqrt{\log(p/k)}\big)$. *Additionally, assume* $\mu^4/n \to 0$. *Then the Bayes risk of the symmetric spike prior* $(\pi_S(\mu, m))(\beta = \pm\mu e_j) = \frac{1}{2m}$, $j = 1, \ldots, m$ *satisfies*

$$B(\pi_S(\mu, m)) \geq \mu^2 - \frac{\mu^2 e^{\mu^2}}{2m}\Big(1 + o(1)\Big).$$

Proof. Using the symmetry of the spike prior distribution, the Bayes risk

$$\begin{aligned} B(\pi_S(\mu, m)) &= \mathbb{E}_{\mu e_1}(\hat{\beta}_1 - \mu)^2 + (m-1)\mathbb{E}_{\mu e_2}\hat{\beta}_1^2 \\ &\geq \mu^2\Big(1 - 2\mathbb{E}_{\mu e_1} p_m + (m-1)\mathbb{E}_{\mu e_2} p_m^2\Big), \end{aligned} \tag{3.96}$$

where the Bayesian estimator of β at the first coordinate is $\hat{\beta}_1 = (\hat{\beta}_\pi)_1 = \mu p_m$. Here we denote

$$p_m := \frac{\exp(\mu x_1^T y - \mu^2\|x_1\|^2/2) - \exp(-\mu x_1^T y - \mu^2\|x_1\|^2/2)}{\sum_{j=1}^m \Big[\exp(\mu x_j^T y - \mu^2\|x_j\|^2/2) + \exp(-\mu x_j^T y - \mu^2\|x_j\|^2/2)\Big]}. \tag{3.97}$$

Then from Lemmas 54 and 60, we have

$$B(\pi_S(\mu, m)) \geq \mu^2 - \frac{\mu^2 e^{\mu^2}}{2m}\Big(1 + o(1)\Big).$$

□

Lemma 54. *Assume model* (3.1). *Let* $m := \lceil p/k \rceil$. *Suppose* $n, m \to \infty$, $\mu \to \infty$ *and* $\mu^2 = o(\log m)$. *Additionally, assume* $\mu^4/n \to 0$. *Then under* $\beta = \mu e_1$, p_m *defined in* (3.97) *satisfies*

$$\mathbb{E}_{\mu e_1} p_m \leq \frac{e^{\mu^2}}{2m}\big(1 + o(1)\big).$$

Proof. Under $\beta = \mu e_1$,

$$\begin{aligned}
\mathbb{E}_{\mu e_1} p_m &= \mathbb{E}\bigg[\exp\Big(\mu x_1^T z + \frac{\mu^2}{2}\|x_1\|^2\Big) - \exp\Big(-\mu x_1^T z - \frac{3\mu^2}{2}\|x_1\|^2\Big)\bigg] \\
&\quad \cdot \bigg[\sum_{j=2}^m \bigg(\exp\Big(\mu x_j^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_j\|^2\Big) + \exp\Big(-\mu x_j^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_j\|^2\Big)\bigg) \\
&\quad + \exp\Big(\mu x_1^T z + \frac{\mu^2}{2}\|x_1\|^2\Big) + \exp\Big(-\mu x_1^T z - \frac{3\mu^2}{2}\|x_1\|^2\Big)\bigg]^{-1}. \\
&\leq \mathbb{E}\exp\Big(\mu x_1^T z + \frac{\mu^2}{2}\|x_1\|^2\Big) \cdot \bigg[\sum_{j=2}^m \bigg(\exp\Big(\mu x_j^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_j\|^2\Big) \\
&\quad + \exp\Big(-\mu x_j^T(\mu x_1 + z) - \frac{\mu^2}{2}\|x_j\|^2\Big) \\
&\quad + \exp\Big(\mu x_1^T z + \frac{\mu^2}{2}\|x_1\|^2\Big) + \exp\Big(-\mu x_1^T z - \frac{3\mu^2}{2}\|x_1\|^2\Big)\bigg]^{-1}.
\end{aligned}$$

We will show that $me^{-\mu^2} \cdot p_m$ is dominated by an L^1 integrable random variable and then prove by the dominated convergence theorem that $\mathbb{E}_{\mu e_1} p_m \leq e^{\mu^2}/(2m) \cdot \Big(1 + o(1)\Big)$. We first construct such

integrable upper bound. Observe that x_j and x_1 are tangled through $\mu x_1 + z$. Let $y = \mu x_1 + z$ and $v := \Big(-\frac{n}{\mu} x_1 + z \Big) / \sqrt{n \cdot (1 + n/\mu^2)} \sim \mathcal{N}(0, \frac{1}{n})$. y is independent of v, furthermore $(y, v, \{x_j\}_{j=2}^m)$ are mutual independent. In exchange of variables from (x_1, z) to (y, v), we have

$$\begin{aligned}
\tfrac{\mu^2}{2}\|x_1\|^2 + \mu x_1^T z &= \mu^2 \tfrac{2+\mu^2/n}{2(1+\mu^2/n)} \tfrac{\|y\|^2}{n(1+\mu^2/n)} - \tfrac{\mu}{(1+\mu^2/n)^{3/2}} v^T y - \tfrac{\mu^2}{2(1+\mu^2/n)} \|v\|^2, \\
-\mu x_1^T z - \tfrac{3\mu^2}{2}\|x_1\|^2 &= -\mu^2 \tfrac{2+3\mu^2/n}{2(1+\mu^2/n)} \tfrac{\|y\|^2}{n(1+\mu^2/n)} + \mu \tfrac{1+2\mu^2/n}{(1+\mu^2/n)^{3/2}} v^T y - \tfrac{\mu^2}{2(1+\mu^2/n)} \|v\|^2.
\end{aligned}$$

Thus,

$$\mathbb{E}_{\mu e_1} p_m \le \mathbb{E} U, \tag{3.98}$$

with

$$\begin{aligned}
U :=& \exp\Big(\mu^2 \frac{2+\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} \Big) \exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}} v^T y - \frac{\mu^2}{2(1+\mu^2/n)} \|v\|^2 \Big) \\
&\cdot \Bigg[\sum_{j=2}^m \Big(\exp\Big(\mu x_j^T y - \frac{\mu^2}{2}\|x_j\|^2 \Big) + \exp\Big(-\mu x_j^T y - \frac{\mu^2}{2}\|x_j\|^2 \Big) \Big) \\
&+ \exp\Big(\mu^2 \frac{2+\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} + \frac{\mu}{(1+\mu^2/n)^{3/2}} v^T y - \frac{\mu^2}{2(1+\mu^2/n)} \|v\|^2 \Big) \\
&+ \exp\Big(-\mu^2 \frac{2+3\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} - \mu \frac{1+2\mu^2/n}{(1+\mu^2/n)^{3/2}} v^T y - \frac{\mu^2}{2(1+\mu^2/n)} \|v\|^2 \Big) \Bigg]^{-1},
\end{aligned} \tag{3.99}$$

where we have used $v \overset{d}{=} -v$. Following Lemma 55, we obtain

$$\mathbb{E} U \le \frac{e^{\mu^2}}{m} \cdot \big(1 + o(1)\big). \tag{3.100}$$

Now, we show $\mathbb{E}_{\mu e_1} p_m = \frac{e^{\mu^2}}{2m} \cdot \big(1 + o(1)\big)$ through dominated converge theorem. Consider the truncation on the following events:

(i) $\frac{\|y\|^2}{n(1+\mu^2/n)} \ge \frac{2(1+\mu^2/n)}{2+\mu^2/n} \frac{\log 2}{\mu}$.

(ii) $\frac{v^T y}{\|y\|^2} n\sqrt{1+\frac{\mu^2}{n}} \geq -\frac{\mu}{2}\frac{2-\mu^2/n}{2(1-\mu^2/n)}$.

From (3.98), we have

$$\mathbb{E}_{\mu e_1} p_m \leq \mathbb{E}[U\mathbb{1}_{((i)\&(ii))}] + \mathbb{E}[U\mathbb{1}_{((i)^c\cup(ii)^c)}].$$

Lemma 57 indicates that

$$\mathbb{E}[U\mathbb{1}_{((i)\&(ii))}] \leq e^{\mu^2}/(2m)\cdot\big(1+o(1)\big).$$

By Lemma 59,

$$U\mathbb{1}_{((i)^c\cup(ii)^c)} \overset{d}{=} \frac{e^{\mu^2}}{m}\frac{B}{A} = o_p\Big(\frac{e^{\mu^2}}{m}\Big).$$

Then using (3.100) and the dominated convergence theorem, we have

$$\mathbb{E}[U\mathbb{1}_{((i)^c\cup(ii)^c)}] \leq o\Big(\frac{e^{\mu^2}}{m}\Big).$$

As a result,

$$\mathbb{E}_{\mu e_1} p_m \leq \frac{e^{\mu^2}}{2m}\big(1+o(1)\big).$$

□

Lemma 55. *Assume model* (3.1) *with* $\beta = \mu e_1$. *Let* $m := \lceil p/k \rceil$. *Suppose* $n, m \to \infty$ *and* $\mu^2 = o(\log m)$. *Additionally, assume* $\mu^4/n \to 0$. *Then* U *defined in* (3.99) *with* $v := \Big(-\frac{n}{\mu}x_1 + z\Big)/\sqrt{n\cdot(1+n/\mu^2)}$ *satisfies*

$$\mathbb{E}_{\mu e_1} U \leq \frac{e^{\mu^2}}{m}\Big(1+o(1)\Big).$$

Proof. Using that $e^a+e^{-a} \geq e^b+e^{-b}$ for $|a| \geq |b|$, $\exp\Big(\mu x_j^T y - \frac{\mu^2}{2}\|x_j\|^2\Big) + \exp\Big(-\mu x_j^T y - \frac{\mu^2}{2}\|x_j\|^2\Big) \geq$

$$\exp\Big(\tfrac{\mu x_j^T y}{(1+\mu^2/n)^{3/2}} - \tfrac{\mu^2}{2}\|x_j\|^2\Big) + \exp\Big(-\tfrac{\mu x_j^T y}{(1+\mu^2/n)^{3/2}} - \tfrac{\mu^2}{2}\|x_j\|^2\Big),$$

$$\begin{aligned}
\mathbb{E}_{\mu e_1} U \leq & \mathbb{E}\frac{\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)}{\sum_{j=2}^m \exp\Big(\frac{\mu x_j^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2}{2}\|x_j\|^2\Big) + \exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)} \\
\leq & \mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\max_{2\leq j\leq m}\|x_j\|^2\Big) \\
& \cdot \frac{\exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)}{\sum_{j=2}^m \exp\Big(\frac{\mu x_j^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2}{2(1+\mu^2/n)}\|x_j\|^2\Big) + \exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)} \\
\leq & \mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\max_{2\leq j\leq m}\Big(\|v\|^2\vee\|x_j\|^2\Big)\Big) \\
& \cdot \frac{\exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)}{\sum_{j=2}^m \exp\Big(\frac{\mu x_j^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2}{2(1+\mu^2/n)}\|x_j\|^2\Big) + \exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)}.
\end{aligned} \tag{3.101}$$

Let the fraction in the last line be $F(v, x_2, \ldots, x_m)$. Based on the independence and homogeneity of distributions of v and $\{x_j\}_{j=2}^m$, conditional on y and $\max\{\|v\|^2, \|x_2\|^2, \ldots, \|x_m\|^2\}$,

$$\exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^T y - \frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big) \stackrel{d}{=} \exp\Big(\frac{\mu x_j^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2}{2(1+\mu^2/n)}\|x_j\|^2\Big). \tag{3.102}$$

Hence,

$$\frac{1}{m} = \mathbb{E}\Big[F(v, x_2, \ldots, x_m)\,\Big|\Big(y, \max_{2\leq j\leq m}\Big(\|v\|^2\vee\|x_j\|^2\Big)\Big)\Big].$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\mu e_1} U &\leq \frac{1}{m}\mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\max_{2\leq j\leq m}\Big(\|v\|^2\vee\|x_j\|^2\Big)\Big) \\
&= \frac{1}{m}\mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\cdot\mathbb{E}\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\max_{2\leq j\leq m}\Big(\|v\|^2\vee\|x_j\|^2\Big)\Big) \\
&= \frac{1}{m}\Big[1-\frac{2}{n}\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\cdot\mathbb{E}\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\max_{2\leq j\leq m}\Big(\|v\|^2\vee\|x_j\|^2\Big)\Big).
\end{aligned} \tag{3.103}$$

Applying Lemma 58, we have

$$\begin{aligned}\mathbb{E}_{\mu e_1} U &\le \frac{1}{m}\cdot\Big[1-\frac{2}{n}\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\exp\Big(\frac{(1+c)\mu^4}{2n(1+\mu^2/n)}\Big)\cdot\big(1+o(1)\big)\\ &= \frac{1}{m}\exp\Big(\mu^2+O\Big(\frac{\mu^4}{n}\Big)\Big)\cdot\big(1+o(1)\big) = \frac{e^{\mu^2}}{m}\cdot\big(1+o(1)\big),\end{aligned} \tag{3.104}$$

where the last equality uses the assumption $\mu^4/n = o(1)$. □

Lemma 56. *Assume $v, y \in \mathbb{R}^n$. Let $\mu > 0$ and $\mu^2/n < 1$. Then the conditions (i) and (ii) described in Lemma 57 imply that*

$$\begin{aligned}&\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\frac{\mu}{(1+\mu^2/n)^{3/2}}v^Ty-\frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big) \qquad (3.105)\\ +\ &\exp\Big(-\mu^2\frac{2+3\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}-\mu\frac{1+2\mu^2/n}{(1+\mu^2/n)^{3/2}}v^Ty-\frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)\\ \ge\ &\exp\Big(\frac{\mu}{(1+\mu^2/n)^{3/2}}v^Ty-\frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)+\exp\Big(-\frac{\mu}{(1+\mu^2/n)^{3/2}}v^Ty-\frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big).\end{aligned}$$

Proof. We start from transforming (3.105). Multipling by $\exp\Big(\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}+\frac{\mu^2}{2(1+\mu^2/n)}\|v\|^2\Big)$ on both sides and reorganizing, we obtain

$$\begin{aligned}&\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\frac{2\mu}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-1\\ &-\Big[\exp\Big(\frac{2\mu}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-\exp\Big(-\mu^2\frac{2+3\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}-\mu\frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}}v^Ty\Big)\Big]\ge 0\\ \Leftrightarrow\ &\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\frac{2\mu}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-1\\ &-\Big[\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\frac{2\mu}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-1\\ &+1-\exp\Big(-\mu^2\frac{2\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}-\mu\frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}}v^Ty\Big)\Big]\\ &\cdot\exp\Big(-\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\ge 0\\ \Leftrightarrow\ &\Big[\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\frac{2\mu}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-1\Big]\cdot\Big[\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)-1\Big]\end{aligned}$$

$$\geq 1 - \exp\Big(- \mu^2 \frac{2\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} - \mu \frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}} v^T y \Big). \tag{3.106}$$

Then if

$$\exp\Big(\mu^2 \frac{2+\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} \Big) \geq 2, \tag{3.107}$$

a sufficient condition for (3.106) to hold is

$$\begin{aligned} &\exp\Big(\mu^2 \frac{2+\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} + \frac{2\mu}{(1+\mu^2/n)^{3/2}} v^T y \Big) - 1 \\ \geq &1 - \exp\Big(- \mu^2 \frac{2\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} - \mu \frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}} v^T y \Big). \end{aligned} \tag{3.108}$$

Suppose additionally

$$\mu^2 \frac{2-\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} + \mu \frac{2(1-\mu^2/n)}{(1+\mu^2/n)^{3/2}} v^T y \geq 0, \tag{3.109}$$

or equivalently

$$\frac{v^T y}{\|y\|^2} \geq -\frac{\mu}{2} \frac{1}{n\sqrt{1+\mu^2/n}} \frac{2-\mu^2/n}{2(1-\mu^2/n)}.$$

Note that the above implies

$$\begin{aligned} &\mu^2 \frac{2+\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} + \frac{2\mu}{(1+\mu^2/n)^{3/2}} v^T y \geq 0 \\ \Leftrightarrow &\frac{v^T y}{\|y\|^2} \geq -\frac{\mu}{2} \frac{1}{n\sqrt{1+\mu^2/n}} \frac{2+\mu^2/n}{2}. \end{aligned}$$

Under condition (3.109), if case I:

$$-\mu^2 \frac{2\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} - \mu \frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}} v^T y \geq 0,$$

then (3.108) holds naturally. If case II:

$$-\mu^2 \frac{2\mu^2/n}{2(1+\mu^2/n)} \frac{\|y\|^2}{n(1+\mu^2/n)} - \mu \frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}} v^T y < 0,$$

then combining this with condition (3.109),

$$\begin{aligned}
&\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\frac{2\mu}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-1\\
\geq\ &\exp\Big(\mu^2\frac{2\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}+\mu\frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}}v^Ty\Big)-1\\
\geq\ &1-\exp\Big(-\mu^2\frac{2\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}-\mu\frac{2\mu^2/n}{(1+\mu^2/n)^{3/2}}v^Ty\Big),
\end{aligned}$$

i.e. (3.108) holds. In summary, the sufficient conditions for (3.105) to hold are

(i) $\frac{\|y\|^2}{n(1+\mu^2/n)}\geq\frac{2(1+\mu^2/n)}{2+\mu^2/n}\frac{\log 2}{\mu}$,

(ii) $\frac{v^Ty}{\|y\|^2}n\sqrt{1+\frac{\mu^2}{n}}\geq-\frac{\mu}{2}\frac{2-\mu^2/n}{2(1-\mu^2/n)}$.

□

Lemma 57. *Assume model* (3.1) *with* $\beta=\mu e_1$. *Let* $m:=\lceil p/k\rceil$. *Suppose* $n,m\to\infty$ *and* $\mu^2=o(\log m)$. *Additionally, assume* $\mu^4/n\to 0$. *Let* U *be defined in* (3.99) *with* $v:=\Big(-\frac{n}{\mu}x_1+z\Big)/\sqrt{n\cdot(1+n/\mu^2)}$. *Consider conditions:*

(i) $\frac{\|y\|^2}{n(1+\mu^2/n)}\geq\frac{2(1+\mu^2/n)}{2+\mu^2/n}\frac{\log 2}{\mu}$,

(ii) $\frac{v^Ty}{\|y\|^2}n\sqrt{1+\frac{\mu^2}{n}}\geq-\frac{\mu}{2}\frac{2-\mu^2/n}{2(1-\mu^2/n)}$.

Then

$$\mathbb{E}[U\mathbb{1}_{((i)\&(ii))}]\leq\frac{1}{2m}e^{\mu^2}\cdot\big(1+o(1)\big).$$

Proof. From (3.99) and since Lemma 56, we have

$$\begin{aligned}
\mathbb{E}[U\mathbb{1}_{((i)\&(ii))}]\leq&\mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big)\\
&\cdot\Bigg[\sum_{j=2}^m\Big(\exp\Big(\frac{\mu x_j^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|x_j\|^2}{2}\Big)+\exp\Big(-\frac{\mu x_j^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|x_j\|^2}{2}\Big)\Big)\\
&+\exp\Big(\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big)+\exp\Big(-\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big)\Bigg]^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu^4\max_{2\leq j\leq m}(\|v\|^2\vee\|x_j\|^2)}{2n(1+\mu^2/n)}\Big)\\
&\quad\cdot\exp\Big(\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big)\\
&\quad\cdot\Big[\sum_{j=2}^m\Big(\exp\Big(\frac{\mu x_j^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|x_j\|^2}{2(1+\mu^2/n)}\Big)+\exp\Big(-\frac{\mu x_j^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|x_j\|^2}{2(1+\mu^2/n)}\Big)\Big)\\
&\quad+\exp\Big(\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big)+\exp\Big(-\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big)\Big]^{-1}\\
&\overset{(b)}{=}\mathbb{E}\exp\Big(\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu^4\max_{2\leq j\leq m}(\|v\|^2\vee\|x_j\|^2)}{2n(1+\mu^2/n)}\Big)\cdot\frac{1}{2m}\\
&\overset{(c)}{\leq}\frac{1}{2m}\Big[1-\frac{2}{n}\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\exp\Big(\frac{(1+c)\mu^4}{2n(1+\mu^2/n)}\Big)\cdot\big(1+o(1)\big)\\
&\overset{(d)}{=}\frac{1}{2m}e^{\mu^2}\cdot\big(1+o(1)\big).
\end{aligned}
$$

Inequality (a) is derived by comparing the coefficients of $\|x_j\|^2$ and $\|v\|^2$ in the exponentials in the numerator and the denominator in (3.99). To show Equality (b), note that v and $(x_2,\ldots,x_m)$ are independently and identically distributed. Define

$$
v\in\mathbb{R}^n\mapsto f_\pm(v):=\exp\Big(\pm\frac{\mu v^Ty}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v\|^2}{2(1+\mu^2/n)}\Big),
$$

Then $f_+(v)\overset{d}{=}f_-(v)$ and $f_\pm(v)\overset{d}{=}f_\pm(x_j)$, $j=2,\ldots,m$. Equality (b) follows by

$$
\mathbb{E}\frac{f_+(v)}{\sum_{j=2}^m\big(f_+(x_j)+f_-(x_j)\big)+f_+(v)+f_-(v)}=\frac{1}{2}\mathbb{E}\frac{f_+(v)+f_-(v)}{\sum_{j=2}^m\big(f_+(x_j)+f_-(x_j)\big)+f_+(v)+f_-(v)}=\frac{1}{2m}.
$$

Inequality (c) uses the independence of y and $\big(v,\{x_j\}_{j=2}^m\big)$ and the result in Lemma 58. The last Equality (d) uses the assumption $\mu^4/n\to 0$. □

Lemma 58. *Suppose* $v,x_2,\ldots,x_m\overset{i.i.d}{\sim}\mathcal{N}(0,\frac{1}{n}I_n)$. *If* $\mu^2/n\to 0$ *and* $(\log m)/n\to 0$, *then there exists some constant* $c>0$ *such that*

$$
\mathbb{E}\Big[\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\max_{2\leq j\leq m}\Big(\|v\|^2\vee\|x_j\|^2\Big)\Big)\Big]\leq\exp\Big(\frac{(1+c)\mu^4}{2n(1+\mu^2/n)}\Big)\cdot\big(1+o(1)\big).
$$

Proof. We use the integrated tail probability to calculate the expectation. For some constant $c > 0$,

$$
\begin{aligned}
& \mathbb{E}\Big[\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)} \max_{2\le j\le m}\Big(\|v\|^2 \vee \|x_j\|^2\Big)\Big) - \exp\Big(\frac{(1+c)\mu^4}{2n(1+\mu^2/n)}\Big)\Big] \\
= & \int_{t>0} P\Big(\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)} \max_{2\le j\le m}\Big(\|v\|^2 \vee \|x_j\|^2\Big)\Big) - \exp\Big(\frac{(1+c)\mu^4}{2n(1+\mu^2/n)}\Big) > t\Big) dt \\
= & \int_{x>c} P\Big(\max_{2\le j\le m}\Big(\|v\|^2 \vee \|x_j\|^2\Big) > 1+x\Big) \frac{\mu^4}{2n(1+\mu^2/n)} \exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}(1+x)\Big) dx \\
\overset{(a)}{\le} & \exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\Big) \cdot \frac{\mu^4}{2n(1+\mu^2/n)} \int_{x>c} m \exp\Big\{ -\frac{n}{2}\Big[\Big(1 - \frac{\mu^4}{n^2(1+\mu^2/n)}\Big)x - \log(1+x)\Big]\Big\} dx \\
\overset{(b)}{\le} & \exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\Big) \cdot \frac{\mu^4}{2n(1+\mu^2/n)} \int_{x>c} m \exp\Big\{ -\frac{n}{2}\cdot\frac{1}{2}\Big(1 - \frac{\mu^4}{n^2(1+\mu^2/n)}\Big)x\Big\} dx \\
= & \exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\Big) \cdot \frac{4\mu^4}{n^2(1+\mu^2/n)\big(1-\frac{\mu^4}{n^2(1+\mu^2/n)}\big)} \exp\Big\{\log m - \frac{n}{4}\Big(1 - \frac{\mu^4}{n^2(1+\mu^2/n)}\Big)c\Big\} \\
\overset{(c)}{=} & \exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)}\Big) \cdot o(1).
\end{aligned}
$$

The above Inequality (a) uses the union bound and the deviation in Lemma 25. In the above Inequality (b), we adopt $c = \Big(1+\frac{\mu^4}{n^2(1+\mu^2/n)}\Big)/\Big(1-\frac{\mu^4}{n^2(1+\mu^2/n)}\Big)$ such that $\Big(1-\frac{\mu^4}{n^2(1+\mu^2/n)}\Big)x - \log(1+x) \ge \frac{1}{2}\Big(1 - \frac{\mu^4}{n^2(1+\mu^2/n)}\Big)x$ for $\forall x > c$. The last Equality (c) is because $\frac{\mu^4}{n^2} = o(1)$ and $\log m = o(n)$. Therefore,

$$
\mathbb{E}\Big[\exp\Big(\frac{\mu^4}{2n(1+\mu^2/n)} \max_{2\le j\le m}\Big(\|v\|^2 \vee \|x_j\|^2\Big)\Big)\Big] \le \exp\Big(\frac{(1+c)\mu^4}{2n(1+\mu^2/n)}\Big) \cdot \big(1+o(1)\big).
$$

□

Lemma 59. *Suppose* $v_1, v_2, \ldots, v_m \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{1}{n} I_n)$ *and* $y \sim \mathcal{N}(0, 1+\frac{\mu^2}{n})$ *with* y *independent of* $\{v_j\}_{j=1}^m$. *Consider the conditions* (i) *and* (ii) *defined in Lemma 55. Then if* $m \to \infty$, $\mu = o\big(\sqrt{\log m}\big)$, $(\log m)/n \to 0$ *and* $\mu^4/n \to 0$, *we have*

(1) $A := \frac{1}{m}\Big(1+\frac{\mu^2}{n}\Big)^{\frac{n}{2}} \exp\Big(-\frac{\mu^2}{2}\frac{1}{(1+\mu^2/n)^2}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big) \cdot \sum_{j=1}^m \exp\Big(\frac{\mu v_j^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2\|v_j\|^2}{2}\Big) \xrightarrow{p} 1.$

(2) $B := e^{-\mu^2}\Big(1+\frac{\mu^2}{n}\Big)^{\frac{n}{2}} \exp\Big(\mu^2\Big(\frac{2+\mu^2/n}{2(1+\mu^2/n)} - \frac{1}{2(1+\mu^2/n)^2}\Big)\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)$
$\cdot \exp\Big(\frac{\mu v_1^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2\|v_1\|^2}{2(1+\mu^2/n)}\Big)\mathbb{1}\big[(i)^c \cup (ii)^c\big] \xrightarrow{p} 0.$

Proof. First, one can directly follow the proof in Lemma 33 to derive (1). We omit the proof of (1) here. Second, to prove (2), it's sufficient to prove

$$\begin{aligned}
&\mathbb{E}\exp\Big(\mu^2\Big(\frac{2+\mu^2/n}{2(1+\mu^2/n)}-\frac{1}{2(1+\mu^2/n)^2}\Big)\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu v_1^T y}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v_1\|^2}{2(1+\mu^2/n)}\Big)\mathbb{1}_{\big[(i)^c\cup(ii)^c\big]}\\
\le\ &\mathbb{E}\exp\Big(\mu^2\Big(\frac{2+\mu^2/n}{2(1+\mu^2/n)}-\frac{1}{2(1+\mu^2/n)^2}\Big)\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu v_1^T y}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v_1\|^2}{2(1+\mu^2/n)}\Big)\Big(\mathbb{1}_{[(i)^c]}+\mathbb{1}_{[(ii)^c]}\Big)\\
=\ &o\Big(e^{\mu^2}\cdot\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}}\Big).
\end{aligned}$$

We first evaluate the truncated expectation on $(i)^c$. Conditional on y,

$$\mathbb{E}\exp\Big(\mu^2\Big(\frac{2+\mu^2/n}{2(1+\mu^2/n)}-\frac{1}{2(1+\mu^2/n)^2}\Big)\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu v_1^T y}{(1+\mu^2/n)^{3/2}}-\frac{\mu^2\|v_1\|^2}{2(1+\mu^2/n)}\Big)\mathbb{1}_{[(i)^c]} \tag{3.110}$$

$$\begin{aligned}
=\ &\Big[1+\frac{\mu^2}{n(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\mathbb{E}\exp\Big[\mu^2\frac{2+\mu^2/n}{2(1+\mu^2/n)}\frac{\|y\|^2}{n(1+\mu^2/n)}\Big]\mathbb{1}_{\big[\frac{\|y\|^2}{n(1+\mu^2/n)}<\frac{2(1+\mu^2/n)}{2+\mu^2/n}\frac{\log 2}{\mu}\big]}\\
=\ &\Big[1+\frac{\mu^2}{n(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\Big[1-\frac{2\mu^2}{n}\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big]^{-\frac{n}{2}}P\Big[\frac{1}{n}\chi_n^2<\Big(1-\frac{2\mu^2}{n}\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big)\frac{2(1+\mu^2/n)}{2+\mu^2/n}\frac{\log 2}{\mu}\Big]\\
\le\ &\Big[1+\frac{\mu^2}{n(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\Big[1-\frac{2\mu^2}{n}\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\exp\Big[-\frac{n}{2}\Big(\log\frac{1}{1-t}-t\Big)\Big],
\end{aligned} \tag{3.111}$$

where in the last inequality we used the deviation in Lemma 25 and

$$1-t:=\Big(1-\frac{2\mu^2}{n}\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big)\frac{2(1+\mu^2/n)}{2+\mu^2/n}\frac{\log 2}{\mu}=o(1),$$

under the assumptions $\mu\to\infty$ and $\mu^2/n\to 0$. Thus, $\frac{n}{2}\Big(\log\frac{1}{1-t}-t\Big)\ge cn=\omega(\mu^2)$, for some constant $c>0$. Note that

$$\Big[1+\frac{\mu^2}{n(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\Big[1-\frac{2\mu^2}{n}\frac{2+\mu^2/n}{2(1+\mu^2/n)}\Big]^{-\frac{n}{2}}=\exp\Big[\frac{\mu^2}{2}+O\Big(\frac{\mu^4}{n}\Big)\Big].$$

Therefore, an upper for (3.110) is

$$o\Big(e^{\frac{\mu^2}{2}\left(1+o(1)\right)}\Big) = o\Big(e^{\mu^2}\cdot\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}}\Big).$$

Then, we evaluate the truncated expectation on $(ii)^c$. Conditional on $\|y\|$ and using $v_1^T y \overset{d}{=} v_{1,1}\|y\|$,

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\frac{\mu v_1^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2\|v_1\|^2}{2(1+\mu^2/n)}\Big)\mathbb{1}_{[(ii)^c]}\,\Big|\|y\|\Big] \\
=\;& \mathbb{E}\Big[\exp\Big(\frac{\mu v_{1,1}^T\|y\|}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2 v_{1,1}^2}{2(1+\mu^2/n)} - \frac{\mu^2\|v_{1,-1}\|^2}{2(1+\mu^2/n)}\Big)\mathbb{1}_{\left[\sqrt{n}v_{1,1}<-\frac{\|y\|}{\sqrt{n(1+\mu^2/n)}}\frac{\mu}{2}\frac{2-\mu^2/n}{2(1-\mu^2/n)}\right]}\,\Big|\|y\|\Big] \\
=\;& \Big[1+\frac{\mu^2}{n(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\exp\Big[\frac{\mu^2\|y\|^2}{2n(1+\mu^2/n)^3\big(1+\frac{\mu^2}{n(1+\mu^2/n)}\big)}\Big] \\
&\cdot P\Big[\mathcal{N}(0,1) < -\frac{\|y\|}{\sqrt{n(1+\frac{\mu^2}{n})}}\Big(\sqrt{1+\frac{\mu^2}{n(1+\frac{\mu^2}{n})}}\frac{\mu}{2}\frac{2-\mu^2/n}{2(1-\mu^2/n)} + \frac{\mu}{(1+\frac{\mu^2}{n})\big(1+\frac{\mu^2}{n(1+\mu^2/n)}\big)^{1/2}}\Big)\Big] \\
\le\;& \Big[1+\frac{\mu^2}{n(1+\mu^2/n)}\Big]^{-\frac{n}{2}}\exp\Big[\frac{\mu^2\|y\|^2}{2n(1+\mu^2/n)^3\big(1+\frac{\mu^2}{n(1+\mu^2/n)}\big)}\Big] \\
&\cdot\frac{1}{\sqrt{2\pi}}\exp\Big[-\frac{1}{2}\frac{\mu^2\|y\|^2}{n(1+\frac{\mu^2}{n})}\Big(\sqrt{1+\frac{\mu^2}{n(1+\frac{\mu^2}{n})}}\frac{2-\frac{\mu^2}{n}}{4(1-\frac{\mu^2}{n})} + \frac{1}{(1+\frac{\mu^2}{n})\big(1+\frac{\mu^2}{n(1+\mu^2/n)}\big)^{1/2}}\Big)^2\Big] \\
=\;& \exp\Big[-\frac{1}{2}\mu^2\big(1+o(1)\big) + \frac{1}{2}\mu^2\big(1+o(1)\big)\cdot\frac{\|y\|^2}{n(1+\frac{\mu^2}{n})} - \frac{9}{8}\mu^2\big(1+o(1)\big)\frac{\|y\|^2}{n(1+\frac{\mu^2}{n})}\Big] \\
=\;& \exp\Big[-\frac{1}{2}\mu^2\big(1+o(1)\big) - \frac{5}{8}\mu^2\frac{\|y\|^2}{n(1+\frac{\mu^2}{n})}\big(1+o(1)\big)\Big],
\end{aligned}$$

where the inequality uses the Gaussian tail bound. Hence, it follows by the moment generating function formula of χ^2 distribution that

$$\begin{aligned}
&\mathbb{E}\exp\Big(\mu^2\Big(\frac{2+\mu^2/n}{2(1+\mu^2/n)} - \frac{1}{2(1+\mu^2/n)^2}\Big)\frac{\|y\|^2}{n(1+\mu^2/n)}\Big)\exp\Big(\frac{\mu v_1^T y}{(1+\mu^2/n)^{3/2}} - \frac{\mu^2\|v_1\|^2}{2(1+\mu^2/n)}\Big)\mathbb{1}_{[(ii)^c]} \\
\le\;& \exp\Big[-\frac{1}{2}\mu^2\big(1+o(1)\big)\Big]\mathbb{E}\exp\Big[\Big(\frac{1}{2}\mu^2\big(1+o(1)\big) - \frac{5}{8}\mu^2\big(1+o(1)\big)\Big)\frac{\|y\|^2}{n(1+\frac{\mu^2}{n})}\Big] \\
=\;& \exp\Big[-\frac{1}{2}\mu^2\big(1+o(1)\big)\Big]\cdot\exp\Big(-\frac{1}{8}\mu^2\big(1+o(1)\big)\Big)
\end{aligned}$$

$$= o\Big(e^{\mu^2} \cdot \Big(1 + \frac{\mu^2}{n}\Big)^{-\frac{n}{2}}\Big).$$

□

Lemma 60. *Assume model* (3.1). *Let* $m := \lceil p/k \rceil$. *Suppose* $n, m \to \infty$, $\mu \to \infty$ *and* $\mu = o(\sqrt{\log m})$. *Additionally, assume* $\mu^4/n \to 0$. *Then under* $\beta = \mu e_2$, p_m *defined in* (3.97) *satisfies*

$$\mathbb{E}_{\mu e_2} p_m^2 \geq \frac{e^{\mu^2}}{2m^2}\big(1 + o(1)\big).$$

Proof. Under $\beta = \mu e_2$,

$$\begin{aligned}
\mathbb{E}_{\mu e_2} p_m^2 =& \mathbb{E}\bigg[\exp\Big(\mu x_1^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_1\|^2\Big) - \exp\Big(- \mu x_1^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_1\|^2\Big)\bigg]^2 \\
& \cdot \bigg[\sum_{j \neq 2} \Big(\exp\Big(\mu x_j^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_j\|^2\Big) + \exp\Big(- \mu x_j^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_j\|^2\Big)\Big) \\
& + \exp\Big(\mu x_2^T z + \frac{\mu^2}{2}\|x_2\|^2\Big) + \exp\Big(- \mu x_2^T z - \frac{3\mu^2}{2}\|x_2\|^2\Big)\bigg]^{-2}.
\end{aligned}$$

Let $v := -\frac{n}{\mu} x_2 + z$, then v is independent of $\mu x_2 + z$. Replace $x_2 = \frac{\mu x_2 + z - v}{\mu + \frac{n}{\mu}}$ such that the above expression of p_m^2 consists of $(x_1, v, \mu x_2 + z, x_3, \ldots, x_m)$. Conditional on $(\mu x_2 + z, x_1)$, then p_m^2 is in the form of a convex function $f(x) := \frac{1}{(x+c)^2}$, where c is a constant or depends only on $(\mu x_2 + z, x_1)$. Then applying Jensen's inequality on $\mathbb{E}[f(x) \,\big|(\mu x_2 + z, v)]$, we obtain

$$\begin{aligned}
\mathbb{E}_{\mu e_2} p_m^2 \geq& \mathbb{E}\bigg[\exp\Big(\mu x_1^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_1\|^2\Big) - \exp\Big(- \mu x_1^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_1\|^2\Big)\bigg]^2 \\
& \cdot \bigg[2(m-2)\Big(1 + \frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\Big(\frac{\mu^2}{2n}\frac{\|\mu x_2 + z\|^2}{1 + \mu^2/n}\Big) \\
& + \exp\Big(\mu x_1^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_1\|^2\Big) + \exp\Big(- \mu x_1^T(\mu x_2 + z) - \frac{\mu^2}{2}\|x_1\|^2\Big) \\
& + \Big(1 + \frac{1}{1 + \frac{n}{\mu^2}}\Big)^{-\frac{n}{2}} \exp\Big(\frac{3\big(\frac{n}{\mu^2}\big)^2 + 5\frac{n}{\mu^2} + 2}{2\big(1 + \frac{n}{\mu^2}\big)^2\big(2 + \frac{n}{\mu^2}\big)}\|\mu x_2 + z\|^2\Big) \\
& + \Big(1 + \frac{1}{1 + \frac{n}{\mu^2}}\Big)^{-\frac{n}{2}} \exp\Big(- \frac{1}{2\big(1 + \frac{n}{\mu^2}\big)}\|\mu x_2 + z\|^2\Big)\bigg]^{-2}
\end{aligned}$$

$$
\begin{aligned}
\geq & \mathbb{E}\Bigg[\exp\left(\mu x_1^T(\mu x_2+z) - \frac{\mu^2}{2}\|x_1\|^2\right) - \exp\left(-\mu x_1^T(\mu x_2+z) - \frac{\mu^2}{2}\|x_1\|^2\right)\Bigg]^2 \\
& \cdot \Bigg[2(m-2)\Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\left(\frac{\mu^2}{2n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\right) \\
& + \exp\left(\mu x_1^T(\mu x_2+z) - \frac{\mu^2}{2}\|x_1\|^2\right) + \exp\left(-\mu x_1^T(\mu x_2+z) - \frac{\mu^2}{2}\|x_1\|^2\right) \\
& + \Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\left(\frac{3}{2}\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\right) + \Big(1+\frac{\mu^2}{n}\Big)^{-\frac{n}{2}} \exp\left(-\frac{\mu^2\|\mu x_2+z\|^2}{2n(1+\mu^2/n)}\right)\Bigg]^{-2}.
\end{aligned}
$$

Let the denominator of the last expression above be D^2. Considering upper bounding the denominator constrained on the following conditions:

(i) $\sqrt{n}|x_{1,1}| \leq 3\mu\frac{\|\mu x_2+z\|}{\sqrt{n}(1+\mu^2/n)}$.

(ii) $\|x_{1,-1}\|^2 \geq 1-t$, for some constant $t \in (0,1)$.

(iii) $\frac{\|\mu x_2+z\|^2}{n(1+\mu^2/n)} \leq c$, for some constant $c = O(1)$.

Using $x_1^T(\mu x_2+z) \overset{d}{=} x_{1,1}\|\mu x_2+z\|$ and under the above conditions, the denominator is upper bounded by

$$
\begin{aligned}
D &\leq 2(m-2)\exp\left(\frac{\mu^2}{2n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} - \frac{\mu^2}{2}\right) + 4\exp\left(\frac{3\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} - \frac{\mu^2}{2}(1-t)\right) \\
&\leq 2(m-2+\sqrt{m})\exp\left(\frac{\mu^2}{2n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} - \frac{\mu^2}{2}\right),
\end{aligned}
$$

where in the first inequality, we used $\left(1+\frac{\mu^2}{n}\right)^{-\frac{n}{2}} \leq e^{-\frac{\mu^2}{2}}$ and $\|x_1\|^2 \geq \|x_{1,-1}\|^2 \geq 1-t$. In the second inequality, we assume constants c and t in conditions (ii) and (iii) satisfy $\log 2 + \frac{5}{2}\mu^2 c + \frac{\mu^2}{2}t \leq \frac{1}{2}\log m$, which is possible since $\mu^2 = o(\log m)$. Then,

$$
\begin{aligned}
\mathbb{E}_{\mu e_2} p_m^2 &\geq \mathbb{E}\frac{\left[\exp\left(\mu x_{1,1}\|\mu x_2+z\| - \frac{\mu^2}{2}\|x_1\|^2\right) - \exp\left(-\mu x_{1,1}\|\mu x_2+z\| - \frac{\mu^2}{2}\|x_1\|^2\right)\right]^2 \mathbb{1}_{[(i)\&(ii)\&(iii)]}}{\left[2(m-2+\sqrt{m})\exp\left(\frac{\mu^2}{2n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} - \frac{\mu^2}{2}\right)\right]^2} \\
&\geq \mathbb{E}\frac{2\exp\left(2\mu x_{1,1}\|\mu x_2+z\| - \mu^2\|x_1\|^2\right) - 2\exp\left(-\mu^2\|x_1\|^2\right)}{\left[2(m-2+\sqrt{m})\exp\left(\frac{\mu^2}{2n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} - \frac{\mu^2}{2}\right)\right]^2}
\end{aligned}
$$

$$- \mathbb{E}\frac{2\exp\Big(2\mu x_{1,1}\|\mu x_2+z\| - \mu^2\|x_1\|^2\Big)\mathbb{1}_{[(i)^c\cup(ii)^c\cup(iii)^c]}}{\Big[2(m-2+\sqrt{m})\exp\Big(\frac{\mu^2}{2n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}-\frac{\mu^2}{2}\Big)\Big]^2}.$$

In the last expression above, let the first expectation be E_1 and the second expectation be E_2. We will show that $E_1 = \frac{1}{2m^2}e^{\mu^2}\cdot\Big(1+o(1)\Big)$. And $E_2 = o\Big(\frac{e^{\mu^2}}{m^2}\Big)$. We first calculate E_1, in which

$$\begin{aligned}
&\mathbb{E}\exp\Big(2\mu x_{1,1}\|\mu x_2+z\| - \mu^2\|x_1\|^2 - \frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} + \mu^2\Big)\\
=\;& \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{n}{2}}\mathbb{E}\exp\Big(2\frac{\mu^2}{n}\|\mu x_2+z\|^2 - \frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} + \mu^2\Big)\\
=\;& \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{n}{2}}\Big[1-\frac{2}{n}\mu^2\Big(2(1+\frac{\mu^2}{n}) - \frac{1}{1+\mu^2/n}\Big)\Big]^{-\frac{n}{2}}e^{\mu^2}\\
=\;& e^{-\mu^2+\mu^2+O(\mu^4/n)}\cdot e^{\mu^2} = e^{\mu^2}\cdot\Big(1+o(1)\Big),
\end{aligned}$$

where the last equality uses $\mu^4/n = o(1)$. The negative term in E_1 is of higher order, since

$$\begin{aligned}
&\frac{2}{(2m-2+\sqrt{n})^2}\,\mathbb{E}\exp\Big(-\mu^2\|x_1\|^2 - \frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n} + \mu^2\Big)\\
=\;& \frac{1}{2(m-1+\sqrt{m}/2)^2}\Big[1+2\frac{\mu^2}{n}\Big]^{-\frac{n}{2}}\cdot\Big[1+\frac{2\mu^2}{n}\Big]^{-\frac{n}{2}}\cdot e^{\mu^2} = o\Big(\frac{e^{\mu^2}}{m^2}\Big).
\end{aligned}$$

Thus,

$$E_1 = \frac{e^{\mu^2}}{2m^2}\cdot\Big(1+o(1)\Big).$$

To calculate E_2, we consider the constraints on $(i)^c$, $(ii)^c$ and $(iii)^c$ one by one. First,

$$\begin{aligned}
&\mathbb{E}\exp\Big(2\mu x_{1,1}|\mu x_2+z| - \mu^2\|x_1\|^2 - \frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\mathbb{1}_{[(i)^c]}\\
=\;& \mathbb{E}\exp\Big(2\mu x_{1,1}|\mu x_2+z| - \mu^2\|x_1\|^2 - \frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\mathbb{1}_{\Big[\sqrt{n}|x_{1,1}|>3\mu\frac{\|\mu x_2+z\|}{\sqrt{n}(1+\mu^2/n)}\Big]}\\
=\;& \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{n}{2}}\mathbb{E}\exp\Big(\frac{2\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+2\mu^2/n}\Big)\cdot\exp\Big(-\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\\
&\cdot P\Big[\mathcal{N}(0,1)\le -\sqrt{1+\frac{2\mu^2}{n}}\cdot\frac{3\mu\|\mu x_2+z\|}{\sqrt{n}(1+\mu^2/n)} - \frac{2\mu}{\sqrt{n}}\frac{\|\mu x_2+z\|}{\sqrt{1+2\mu^2/n}}\Big],
\end{aligned}$$

$$
\begin{aligned}
&\text{or } \mathcal{N}(0,1) \geq \sqrt{1+\frac{2\mu^2}{n}} \cdot \frac{3\mu\|\mu x_2+z\|}{\sqrt{n}(1+\mu^2/n)} - \frac{2\mu}{\sqrt{n}}\frac{\|\mu x_2+z\|}{\sqrt{1+2\mu^2/n}}\Big] \\
\leq\ & \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{n}{2}} \mathbb{E}\exp\bigg(\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big(\frac{2(1+\mu^2/n)}{1+2\mu^2/n}-1\Big)\bigg) \\
&\cdot\frac{2}{\sqrt{2\pi}}\exp\bigg[-\frac{\mu^2}{2}\frac{\|\mu x_2+z\|^2}{n(1+\mu^2/n)}\Big(3\sqrt{1+2\mu^2/n}-\frac{2\sqrt{1+\mu^2/n}}{\sqrt{1+2\mu^2/n}}\Big)^2\bigg] \\
=\ & \frac{2}{\sqrt{2\pi}}\exp\Big[-\mu^2\Big(1+o(1)\Big)+\mu^2\Big(1+o(1)\Big)-\frac{1}{2}\mu^2\Big(1+o(1)\Big)\Big] \\
=\ & \frac{2}{\sqrt{2\pi}}e^{-\frac{\mu^2}{2}\left(1+o(1)\right)} = o(1).
\end{aligned}
$$

Second,

$$
\begin{aligned}
&\mathbb{E}\exp\Big(2\mu x_{1,1}|\mu x_2+z|-\mu^2\|x_1\|^2-\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\mathbb{1}_{[(ii)^c]} \\
=\ & \mathbb{E}\exp\Big(2\mu x_{1,1}|\mu x_2+z|-\mu^2\|x_1\|^2-\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\mathbb{1}_{\left[\|x_{1,-1}\|^2<1-t\right]} \\
=\ & \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{1}{2}}\mathbb{E}\exp\Big(\frac{2\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+2\mu^2/n}\Big)\exp\Big(-\mu^2\|x_{1,-1}\|^2\Big)\mathbb{1}_{\left[\|x_{1,-1}\|^2<1-t\right]} \\
=\ & \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{1}{2}}\Big(1-\frac{2\mu^2}{n}\frac{2(1+\mu^2/n)}{1+2\mu^2/n}\Big)^{-\frac{n}{2}}\mathbb{E}\exp\Big(-\mu^2\|x_{1,-1}\|^2\Big)\mathbb{1}_{\left[\|x_{1,-1}\|^2<1-t\right]} \\
=\ & \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{n}{2}}\Big(1-\frac{2\mu^2}{n}\frac{2(1+\mu^2/n)}{1+2\mu^2/n}\Big)^{-\frac{n}{2}}P\Big(\frac{1}{n}\chi^2_{n-1}<\Big(1+\frac{2\mu^2}{n}\Big)\cdot(1-t)\Big) \\
\overset{(a)}{\leq}\ & \Big(1+\frac{2\mu^2}{n}\Big)^{-\frac{n}{2}}\Big(1-\frac{2\mu^2}{n}\frac{2(1+\mu^2/n)}{1+2\mu^2/n}\Big)^{-\frac{n}{2}}\exp\bigg[\frac{n-1}{2}\Big(\log\Big(1-t'\Big)+t'\Big)\bigg] \\
=\ & o(1),
\end{aligned}
$$

where in (a) we denote $1-t' := \Big(1+\frac{2\mu^2}{n}\Big)\frac{n}{n-1}(1-t) < 1$, thus $\log(1-t')+t' < c' < 0$ for some constant $c' < 0$. The last equality follows by $e^{c_1\mu^2-c_2 n} \to 0$ for arbitrary constants $c_1, c_2 > 0$, since $\mu^2/n \to 0$. Third,

$$
\begin{aligned}
&\mathbb{E}\exp\Big(2\mu x_{1,1}|\mu x_2+z|-\mu^2\|x_1\|^2-\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\mathbb{1}_{[(iii)^c]} \\
=\ & \mathbb{E}\exp\Big(2\mu x_{1,1}|\mu x_2+z|-\mu^2\|x_1\|^2-\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\Big)\mathbb{1}_{\left[\frac{\|\mu x_2+z\|^2}{n(1+\mu^2/n)}>c\right]}
\end{aligned}
$$

$$
\begin{aligned}
&= \left(1+\frac{2\mu^2}{n}\right)^{-\frac{n}{2}} \mathbb{E}\exp\left(\frac{2\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+2\mu^2/n}-\frac{\mu^2}{n}\frac{\|\mu x_2+z\|^2}{1+\mu^2/n}\right)\mathbb{1}_{\left[\frac{\|\mu x_2+z\|^2}{n(1+\mu^2/n)}>c\right]} \\
&= \left(1+\frac{2\mu^2}{n}\right)^{-\frac{n}{2}} \cdot \left[1-\frac{2\mu^2}{n}\left(\frac{2(1+\mu^2/n)}{1+2\mu^2/n}-1\right)\right]^{-\frac{n}{2}} \\
&\quad \cdot P\left[\frac{1}{n}\chi^2_n > c\left(1-\frac{2\mu^2}{n}\left(\frac{2(1+\mu^2/n)}{1+2\mu^2/n}-1\right)\right)\right] \\
&\overset{(a)}{=} \exp\left(O\left(\frac{\mu^4}{n}\right)\right)\cdot\exp\left[-\frac{n}{2}\left[t''-1-\log t''\right]\right] \overset{(b)}{=} o(1),
\end{aligned}
$$

where in Equality (a), $t'' := c\left(1-\frac{2\mu^2}{n}\left(\frac{2(1+\mu^2/n)}{1+2\mu^2/n}-1\right)\right) > c' > 1$ for some constant c'. And Equality (b) uses as $n\to\infty$, $\mu^4/n = o(\mu^2) = o(n)$. □

Chapter 4: Discussions

We studied the minimax problem of two canonical models: sparse signal denoising and sparse linear regression. We showed that the minimaxity in its current form is not informative enough to reflect the important factor of the sparse estimation problem, as indicated in empirical studies, the SNR level. We have shown that the classical minimax sugguests asymptotic minimax estimator irrespective of the underlying SNR level. However, sub-optimality of these estimators is demonstrated in empirical performance under different SNRs. To interpret the results and mitigate the discrepancy, we introduced two notions that can make the minimax results more meaningful and appealing for practical purposes: (i) signal-to-noise-ratio aware minimaxity, (ii) second-order asymptotic approximation of minimax risk. We showed that these two notions can alleviate the major drawbacks of the classical minimax results. For instance, in sparse signal denoising problem in Chapter 2, while the classical results prove that the hard and soft thresholding estimators are minimax optimal, the new results reveal that in a wide range of low signal-to-noise ratios the two estimators are in fact sub-optimal. Even when the signal-to-noise ratio is high, only hard thresholding is optimal and soft thresholding remains sub-optimal. Furthermore, our refined minimax analysis identified three optimal (or nearly optimal) estimators in three regimes with varying SNR: hard thresholding $\hat{\eta}_H(y,\lambda)$ of (2.5) in high SNR; $\hat{\eta}_E(y,\lambda,\gamma)$ of (2.10) in moderate SNR; linear estimator $\hat{\eta}_L(y,\lambda)$ of (2.9) in low SNR. As is clear from the definition of the three estimators, they are induced by ℓ_0-regularization, elastic net regularization [23] and ℓ_2-regularization, respectively. These regularization techniques have been widely used in statistics and machine learning [24].

The concepts of signal-to-noise ratio aware minimaxity and higher-order asymptotic approximations introduced in this thesis may open up new venues for investigating various estimation problems. We have used the same framework to revisit the sparse estimation problem in high-dimensional linear regression and obtained new insights. However, the analysis of minimax in

sparse linear regression is more challenging in the high dimensional setting. The current progress of this line of research has been arounded the rate-minimaxity. We completed the classical minimax result by characterizing the accurate constant. And we established the SNR-aware minimax results up to first and second-order accurate. It's yet to be finished of the second-order approximations in moderate and high SNR regimes. However, the obtained results in low SNR regime already demonstrates a non-trivial estimator – ridge outperforms the zero estimator in extreme low SNR (goes to zero). This explains and verifies the finding in empirical result and provide evidence that our method of higher order approximation of SNR-aware minimax result is impactful in studying sparse estimation problem. That being said, it is important to acknowledge that the additional insights gained from this framework come with increased mathematical complexity when computing minimax estimators. Therefore, one direction we plan to explore in the future is the development of simpler and more general techniques for obtaining higher-order approximations of minimax risk or the supremum risk of well-established estimators.

References

[1] L. Le Cam, *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 1986.

[2] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 1998.

[3] I. M. Johnstone, *Gaussian estimation: Sequence and wavelet models*. 2019.

[4] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st. Springer Publishing Company, Incorporated, 2008, ISBN: 0387790519, 9780387790510.

[5] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

[6] J. Fan, R. Li, C.-H. Zhang, and H. Zou, *Statistical foundations of data science*. Chapman and Hall/CRC, 2020.

[7] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE transactions on information theory*, vol. 57, no. 10, pp. 6976–6994, 2011.

[8] T. Hastie, R. Tibshirani, and R. Tibshirani, "Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons," *Statistical Science*, vol. 35, no. 4, pp. 579–592, 2020.

[9] R. Mazumder, P. Radchenko, and A. Dedieu, "Subset selection with shrinkage: Sparse linear modeling when the snr is low," *Operations Research*, 2022.

[10] L. Zheng, A. Maleki, H. Weng, X. Wang, and T. Long, "Does ℓ_p-minimization outperform ℓ_1-minimization?" *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 6896–6935, 2017.

[11] D. Donoho and I. Johnstone, "Minimax risk over ℓ_p balls for ℓ_q losses," *Probab. Theory Related Fields*, vol. 99, pp. 277–303, 1994.

[12] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern, "Maximum entropy and the nearly black object," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 54, no. 1, pp. 41–67, 1992.

[13] P. J. BICKEL, Y. RITOV, and A. B. TSYBAKOV, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[14] N. Verzelen, "Minimax risks for sparse regressions: Ultra-high dimensional phenomenons," 2012.

[15] P. C. Bellec, G. Lecué, and A. B. Tsybakov, "Slope meets lasso: Improved oracle bounds and optimality," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3603–3642, 2018.

[16] R. R. Hocking and R. Leslie, "Selection of the best subset in regression analysis," *Technometrics*, vol. 9, no. 4, pp. 531–540, 1967.

[17] E. M. L. Beale, M. G. Kendall, and D. Mann, "The discarding of variables in multivariate analysis," *Biometrika*, vol. 54, no. 3-4, pp. 357–366, 1967.

[18] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," 2007.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[20] S. Wang, H. Weng, and A. Maleki, "Which bridge estimator is the best for variable selection?" *The Annals of Statistics*, vol. 48, no. 5, pp. 2791–2823, 2020.

[21] W. Su and E. Candes, "Slope is adaptive to unknown sparsity and asymptotically minimax," 2016.

[22] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[23] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[24] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[25] H. Hazimeh and R. Mazumder, "Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms," *Operations Research*, vol. 68, no. 5, pp. 1517–1537, 2020.

[26] H. Weng, A. Maleki, and L. Zheng, "Overcoming the limitations of phase transition by higher order analysis of regularization techniques," *The Annals of Statistics*, vol. 46, no. 6A, pp. 3099–3129, 2018.

[27] W. Jiang and C.-H. Zhang, "General maximum likelihood empirical bayes estimation of normal means," 2009.

[28] I. M. Johnstone, "On minimax estimation of a sparse normal mean vector," *The Annals of Statistics*, pp. 271–289, 1994.

[29] C.-H. Zhang, "Minimax ℓ_q risk in ℓ_p balls," *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, p. 78, 2012.

[30] C. Butucea, M. Ndaoud, N. A. Stepanova, and A. B. Tsybakov, "Variable selection with hamming loss," *The Annals of Statistics*, vol. 46, no. 5, pp. 1837–1875, 2018.

[31] O. Collier, L. Comminges, and A. B. Tsybakov, "Minimax estimation of linear and quadratic functionals on sparsity classes," *The Annals of Statistics*, vol. 45, no. 3, pp. 923–958, 2017.

[32] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[33] D. L. Donoho, I. M. Johnstone, G Kerkyacharian, and D. Picard, "Universal near minimaxity of wavelet shrinkage," in *Festschrift for Lucien Le Cam*, Springer, 1997, pp. 183–218.

[34] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[35] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès, "Slope—adaptive variable selection via convex optimization," *The annals of applied statistics*, vol. 9, no. 3, p. 1103, 2015.

[36] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[37] R. Durrett, "Probability: Theory and examples," 2013.

[38] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[39] S. Jalali, A. Maleki, and R. Baraniuk, "Minimum complexity pursuit: Stability analysis," in *2012 IEEE International Symposium on Information Theory Proceedings*, 2012, pp. 1857–1861.

[40] M. Ghosh, "Exponential tail bounds for chisquared random variables," *Journal of Statistical Theory and Practice*, vol. 15, no. 2, pp. 1–6, 2021.

[41] S. Boucheron and M. Thomas, "Concentration inequalities for order statistics," 2012.

[42] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE transactions on information theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[43] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[44] Y. Guo, H. Weng, and A. Maleki, "Signal-to-noise ratio aware minimaxity and higher-order asymptotics," *arXiv preprint arXiv:2211.05954*, 2022.

[45] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 1991, vol. 23.

[46] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of statistics*, pp. 1302–1338, 2000.