

# Signal-to-noise ratio aware minimaxity and higher-order asymptotics

Yilin Guo, Haolei Weng, Arian Maleki

**Abstract**—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 paper, 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. The first contribution of this paper is to 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. The second contribution of this paper is to showcase how one can use higher-order asymptotics to obtain accurate approximations of the SNR-aware minimax risk and discover minimax estimators. The theoretical findings obtained from this refined minimax framework provide new insights and practical guidance for the estimation of sparse signals.

**Index Terms**—Minimaxity, signal-to-noise ratio, sparsity, soft thresholding, hard thresholding, linear shrinkage, higher-order asymptotics, Gaussian sequence model.

## I. INTRODUCTION

### A. Motivation

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 for the study of optimality and the design of optimal estimators 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 best subset selection in low signal-to-noise ratio (SNR) [8]–[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 paper, 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 I-B. The rest of the paper is then devoted to the development of a much more informative minimax framework that alleviates major drawbacks of the classical one. This is made possible 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.

### B. Classical minimaxity and its limitations in sparse Gaussian sequence model

We consider the Gaussian sequence model:

$$y_i = \theta_i + \sigma_n z_i, \quad i = 1, 2, \dots, n. \quad (1)$$

Here,  $y = (y_1, \dots, y_n)$  is the vector of observations,  $\theta = (\theta_1, \dots, \theta_n)$  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  $\theta$  from the sparse parameter space

$$\Theta(k_n) = \left\{ \theta \in \mathbb{R}^n : \|\theta\|_0 \leq k_n \right\}, \quad (2)$$

where  $\|\theta\|_0$  denotes the number of non-zero components of  $\theta$ , 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

$$R(\Theta(k_n), \sigma_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(k_n)} \mathbb{E}_{\theta} \|\hat{\theta} - \theta\|_2^2, \quad (3)$$

where  $\mathbb{E}_{\theta}(\cdot)$  is the expectation taken under (1) with true parameter value  $\theta$ .

Gaussian sequence model plays a fundamental role in non-parametric and high-dimensional statistics. There exists extensive literature on the minimax estimation of  $\theta$  or its functionals over various structured parameter spaces such as Sobolev

This work is supported by NSF-DMS 2210506, and NSF-DMS 2210505.

Y. Guo is with the Department of Statistics, Columbia University, New York, USA. (e-mail: yilinguo97@gmail.com). H. Weng is with the Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA. (e-mail: wenghaol@msu.edu). A. Maleki is with the Department of Statistics, Columbia University, New York, USA. (e-mail: arian.maleki@gmail.com).

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 [3], [4], [11] 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  $1 \leq i \leq n$ ,

$$\begin{aligned} [\hat{\eta}_S(y, \lambda)]_i &= \arg \min_{\mu \in \mathbb{R}} (y_i - \mu)^2 + 2\lambda|\mu| \\ &= \text{sign}(y_i)(|y_i| - \lambda)_+, \end{aligned} \quad (4)$$

$$\begin{aligned} [\hat{\eta}_H(y, \lambda)]_i &= \arg \min_{\mu \in \mathbb{R}} (y_i - \mu)^2 + \lambda^2 I(\mu \neq 0) \\ &= y_i I(|y_i| > \lambda), \end{aligned} \quad (5)$$

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. We summarize a classical asymptotic minimax result in the following theorem.

**Theorem 1** ([3], [12], [13]). *Assume model (1) and parameter space (2) with  $k_n/n \rightarrow 0$  as  $n \rightarrow \infty$ . Then the minimax risk, defined in (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 1 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 1. As is clear from the upper 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 thresholding are outperformed by the linear estimator. The same comparison holds in the lower 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 III.

In light of Theorem 1 and Figure 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 1 implies that the signal-to-noise ratio (SNR) has a significant impact on the estimation. However, the effect

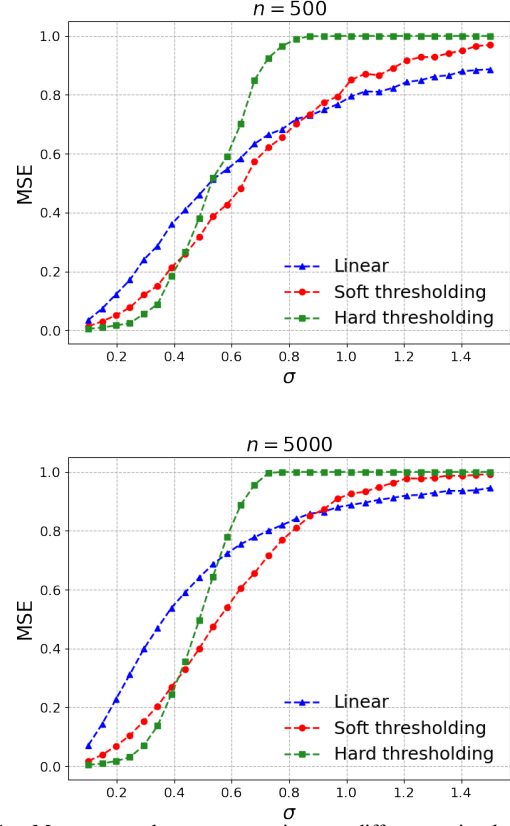


Fig. 1. Mean squared error comparison at different noise levels. Data is generated according to (1) with  $k_n = \lfloor n^{2/3} \rfloor$  and  $\theta$  having  $k_n$  components equal to 1.5. “linear” denotes the simple linear estimator  $\frac{1}{1+\lambda}y$ . All the three estimators are optimally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

of SNR is not well captured in the classical minimax results (Theorem 1).

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.

### C. Our contributions and paper structure

To overcome the limitations of the classical minimaxity discussed in Section I-B, in this paper, 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 II, 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 1, and are much more informative towards understanding the sparse estimation problem in practice.

The rest of the paper is organized as follows. Section II presents the main results from the SNR-aware minimax framework. Section III includes more simulations to support our theoretical findings. Section IV summarizes the main messages of the paper and discusses some related works. All the proofs are presented in Section V.

We collect the notations used throughout the paper 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  $n$ ,  $[n] = \{1, 2, \dots, n\}$ . We use  $I_A$  and  $I(A)$  to represent the indicator function of the set  $A$  interchangeably. For a given vector  $v = (v_1, \dots, v_p) \in \mathbb{R}^p$ ,  $\|v\|_0 = \#\{i : v_i \neq 0\}$ ,  $\|v\|_\infty = \max_i |v_i|$ , and  $\|v\|_q = (\sum_{i=1}^p |v_i|^q)^{1/q}$  for  $q \in (0, \infty)$ . We use the notation  $\delta_\mu$  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  $|a_n/b_n| \rightarrow 0$  as  $n \rightarrow \infty$ , 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  $\pi$ ,  $\text{supp}(\pi)$  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.

## II. SNR-AWARE MINIMAXITY

### A. SNR-aware minimax framework

We focus on the above-mentioned Gaussian sequence model (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:

$$\Theta(k_n, \tau_n) = \left\{ \theta \in \mathbb{R}^n : \|\theta\|_0 \leq k_n, \|\theta\|_2^2 \leq k_n \tau_n^2 \right\}. \quad (6)$$

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  $\tau_n$  can be considered as a measure of signal strength (on average) for each non-zero coordinate of  $\theta$ . 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):

$$R(\Theta(k_n, \tau_n), \sigma_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2. \quad (7)$$

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 (6) under the model (1). Denote

$$\epsilon_n = \frac{k_n}{n}, \quad \mu_n = \frac{\tau_n}{\sigma_n}. \quad (8)$$

It is clear that  $\epsilon_n$  represents the sparsity level and  $\mu_n$  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  $(\epsilon_n, \mu_n)$ . 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  $n \rightarrow \infty$ ,

- Regime (I)** Low signal-to-noise ratio:  $\mu_n \rightarrow 0, \epsilon_n \rightarrow 0$ ;
- Regime (II)** Moderate signal-to-noise ratio:  $\mu_n \rightarrow \infty, \epsilon_n \rightarrow 0, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ ;
- Regime (III)** High signal-to-noise ratio:  $\epsilon_n \rightarrow 0, \mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ .

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

### B. First order analysis of SNR-aware minimaxity and its drawbacks

Our first theorem generalizes Theorem 1, to our SNR-aware minimax framework.

**Theorem 2.** Assume model (1) and parameter space (6). The following hold:

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

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

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

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

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

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

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

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

Furthermore, both soft and hard thresholding estimators (4)-(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 3, 4, and 6 we present in Section II-C. 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 new 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 II-C 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  $\epsilon_n$  fixed and evaluate the minimax risk as a function of  $\mu_n$ , 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.

### C. Second order analysis of SNR-aware minimaxity

In this section, we discuss how the analysis provided in Section II-B can be refined to resolve the issues we raised in Section I-B.

*1) Results in Regime (I):* We start with Regime (I). As discussed in Theorem 2, 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 2 resorts to an approximation that is asymptotically accurate. However, this approximation is coarse when  $n$  is not too large and/or  $\epsilon_n$  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:

$$[\hat{\eta}_L(y, \lambda)]_i = \frac{y_i}{1 + \lambda} = \arg \min_{\mu \in \mathbb{R}} (y_i - \mu)^2 + \lambda \mu^2, \quad 1 \leq i \leq n. \quad (9)$$

**Theorem 3.** Consider model (1) and parameter space (6). For Regime (I) in which  $\epsilon_n \rightarrow 0, \mu_n \rightarrow 0$  as  $n \rightarrow \infty$ , we have

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

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

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

The proof of this theorem can be found in Section V-B. Compared with Theorem 2, Theorem 3 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 (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 (1) and parameter space (6). In Regime (I) where  $\epsilon_n \rightarrow 0, \mu_n \rightarrow 0$  as  $n \rightarrow \infty$ , the optimally tuned soft thresholding estimator  $\hat{\eta}_S(y, \lambda)$  has supremum risk:

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_S(y, \lambda) - \theta\|_2^2$$

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

The proof of this proposition can be found in Section V-C. It is straightforward to confirm that

$$\exp \left[ -\frac{1}{2} \frac{1}{\mu_n^2} \left( \log \frac{1}{\epsilon_n} \right)^2 (1 + o(1)) \right] / (\epsilon_n^2 \mu_n^4) = o(1)$$

under the scaling  $\epsilon_n \rightarrow 0, \mu_n \rightarrow 0$ . 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 (1) and parameter space (6). In Regime (I) where  $\epsilon_n \rightarrow 0, \mu_n \rightarrow 0$  as  $n \rightarrow \infty$ , the optimally tuned hard thresholding estimator  $\hat{\eta}_H(y, \lambda)$  has supremum risk:

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

The proof of this proposition is presented in Section V-D.

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 ones 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.

2) *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:

$$[\hat{\eta}_E(y, \lambda, \gamma)]_i = \frac{[\hat{\eta}_S(y, \lambda)]_i}{1 + \gamma}$$

$$= \arg \min_{u \in \mathbb{R}} (y_i - u)^2 + 2\lambda|u| + \gamma u^2, \quad 1 \leq i \leq n. \quad (10)$$

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

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

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

$$\begin{aligned} & R(\Theta(k_n, \tau_n), \sigma_n) \\ & \leq \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_{\theta} \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 \end{aligned}$$

$$= n\sigma_n^2 \left( \epsilon_n \mu_n^2 - (\sqrt{2/\pi} + o(1)) \epsilon_n^2 \mu_n e^{\mu_n^2} \right).$$

The proof of this theorem can be found in Section V-E.

**Remark 1.** Theorem 4 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  $\mu_n$  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 5) 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 5.** Consider model (1) with the following parameter space:

$$\Theta^A(k_n, \tau_n) := \left\{ \theta \in \mathbb{R}^n : \|\theta\|_0 \leq k_n, \|\theta\|_2^2 \leq k_n \tau_n^2, \|\theta\|_{\infty} \leq A \tau_n \right\}. \quad (11)$$

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

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

The theorem is proved in Section V-F.

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 4 indicates that thresholding and linear shrinkage together play an important role in estimating sparse signals in Regime (II). To shed more light on it, the following three propositions prove that neither thresholding 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 (1) and parameter space (6). In Regime (II) where  $\epsilon_n \rightarrow 0, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ , as  $n \rightarrow \infty$ , the optimally tuned soft thresholding estimator has supremum risk:

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

The proof of this proposition can be found in Section V-G.

**Proposition 4.** Consider model (1) and parameter space (6). In Regime (II) where  $\epsilon_n \rightarrow 0, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$

as  $n \rightarrow \infty$ , the optimally tuned hard thresholding estimator  $\hat{\eta}_H(y, \lambda)$  has supremum risk:

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

The proof of this proposition is presented in Section V-H.

**Proposition 5.** Consider model (1) and parameter space (6). In Regime (II) where  $\epsilon_n \rightarrow 0, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ , as  $n \rightarrow \infty$ , the optimally tuned linear estimator has supremum risk:

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

The proof of this proposition can be easily followed by the discussion in Section V-B1.

Comparing the second order term in Theorem 4 and Propositions 3-5 under the scaling condition  $\epsilon_n \rightarrow 0, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ , 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, hard 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.

3) *Results in Regime (III):* Finally, let us consider the high-SNR regime, i.e., Regime (III). As shown in Theorem 2, 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 6.** Consider model (1) and parameter space (6). For Regime (III) in which  $\epsilon_n \rightarrow 0, \mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$  as  $n \rightarrow \infty$ , we have

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

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.

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

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

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

$$\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_{\theta} \|\hat{\eta}_S(y, \lambda) - \theta\|_2^2$$

$$= n\sigma_n^2 \left( 2\epsilon_n \log \epsilon_n^{-1} - 6\epsilon_n \log \nu_n (1 + o(1)) \right),$$

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

The proof of the proposition can be found in Section V-J.

**Proposition 7.** Consider model (1) and parameter space (6). In Regime (III) where  $\epsilon_n \rightarrow 0, \mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$  as  $n \rightarrow \infty$ , the optimally tuned linear estimator achieves the supremum risk:

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

The proof of this proposition is presented in Section V-K.

Combining the above three results, we can conclude that overall in Regime (III) hard thresholding offers a better estimate than soft thresholding and linear shrinkage. 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 and linear shrinkage by producing zero coordinates while 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.

### III. NUMERICAL EXPERIMENTS

As discussed in Section I-B 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 new SNR-aware minimax framework in Section II 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  $\theta$  in the following way: for a sample size  $n$ ,  $\theta = (\theta_1, \dots, \theta_n)$  is generated by assigning  $\tau_n$  to a random choice of  $k_n$  coordinates and setting the others to zero. Then  $y = (y_1, \dots, y_n)$  and  $z = (z_1, \dots, z_n)$  are generated according to Model (1) for a certain noise level  $\sigma_n$ .

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  $\epsilon_n = k_n/n \rightarrow 0$  as  $n \rightarrow \infty$ . 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  $\mu_n$  change from small to large values, and plot the mean squared error (MSE) of different estimators as a function of  $\mu_n$ .
- 2) Let  $\sigma_n$  change from small to large values, and plot the MSE as a function of  $\sigma_n$ .

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 (9), soft thresholding  $\hat{\eta}_S$  defined in (4), hard thresholding  $\hat{\eta}_H$  defined in (5), and the soft-linear ‘‘interpolation’’ estimator

$\hat{\eta}_E$  defined in (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 2-5 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 2-3, when  $\sigma_n$  changes from small to large values, we observed that: (1) When  $\sigma_n$  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  $\sigma_n$  is in moderate area, the soft-linear ‘interpolation’ estimator  $\hat{\eta}_E$  has the minimum empirical MSE. This corresponds to Regime (II) in our theory. (3) When  $\sigma_n$  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 4-5 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  $\mu_n$  increases, the curves will correspond to Regime (II) and Regime (III). In particular, when  $\mu_n$  is large, it corresponds with the area of  $\sigma_n$  near zero in Figures 2-3. Here, it is shown more clearly that in the large SNR regime, hard thresholding has the minimum empirical MSE among all the estimators.

#### IV. DISCUSSIONS

##### A. Summary

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, 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 (5) in high SNR;  $\hat{\eta}_E(y, \lambda, \gamma)$  of (10) in moderate SNR; linear estimator  $\hat{\eta}_L(y, \lambda)$  of (9) in low SNR. As is clear from the definition of the three estimators, they are induced by  $\ell_0$ -regularization, elastic net regularization [14] and  $\ell_2$ -regularization, respectively. These regularization techniques have been widely used in statistics and machine learning [15].

The concepts of signal-to-noise ratio aware minimaxity and higher-order asymptotic approximations introduced in this

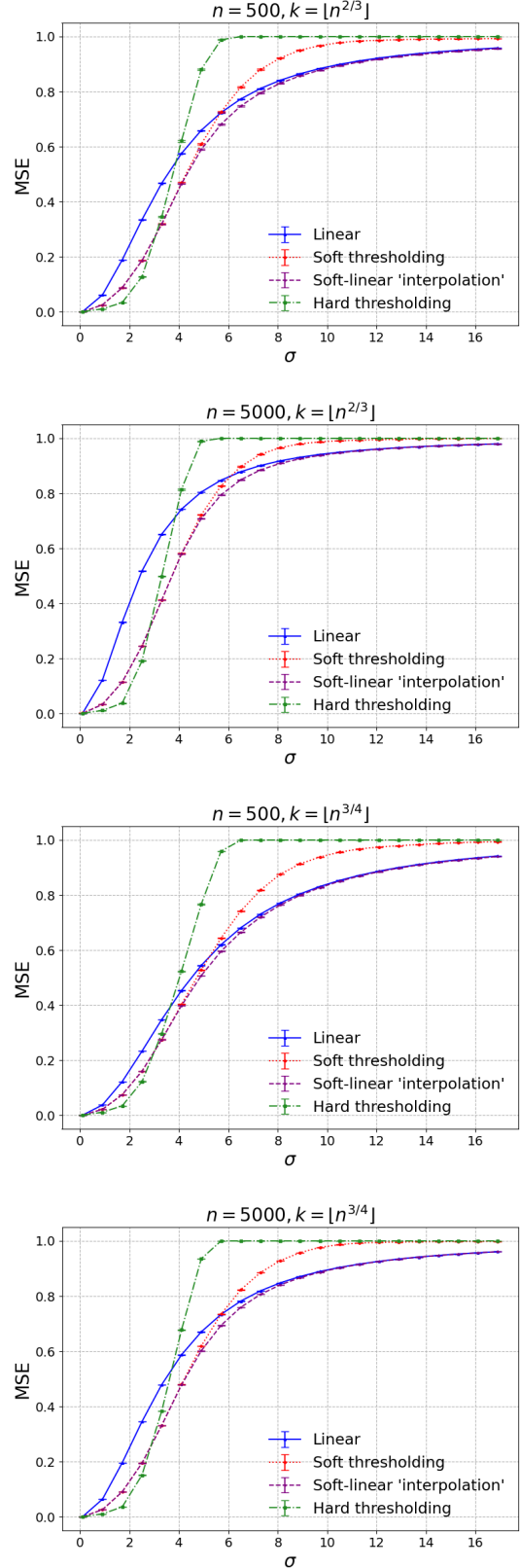


Fig. 2. Mean squared error comparison at different noise levels. On each graph, the y-axis is the scaled MSE, and the x-axis is the noise standard deviation  $\sigma_n$ . (to be continued in Fig. 3)



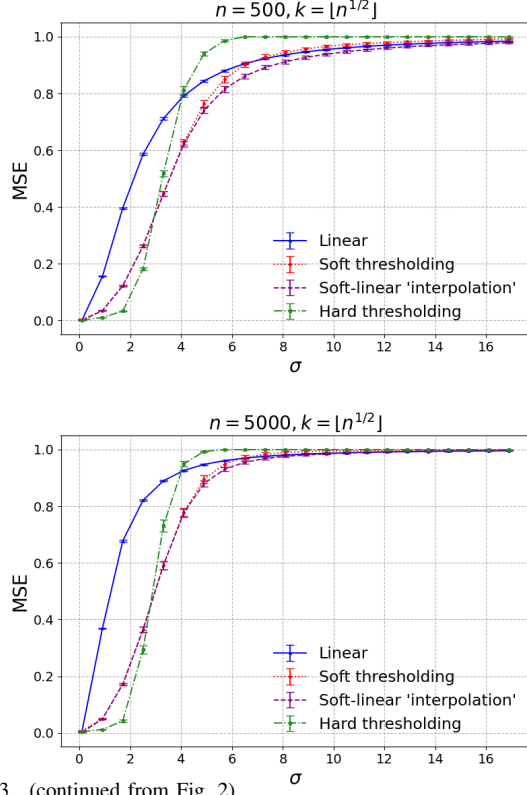


Fig. 3. (continued from Fig. 2)

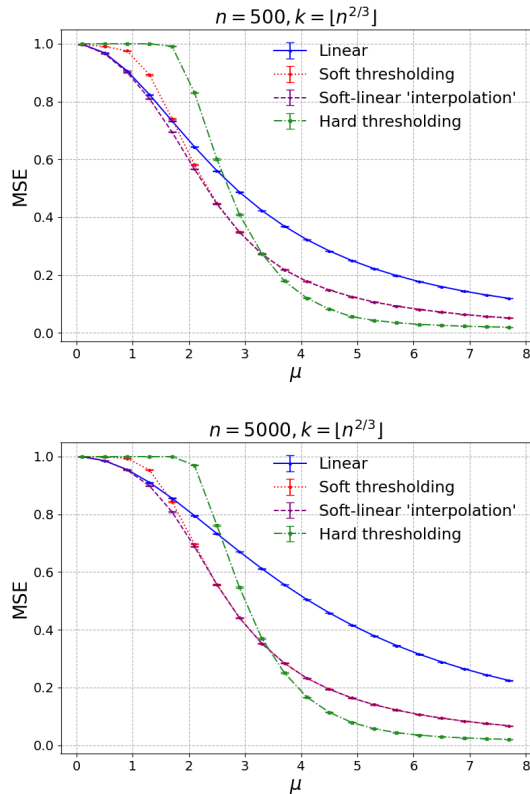
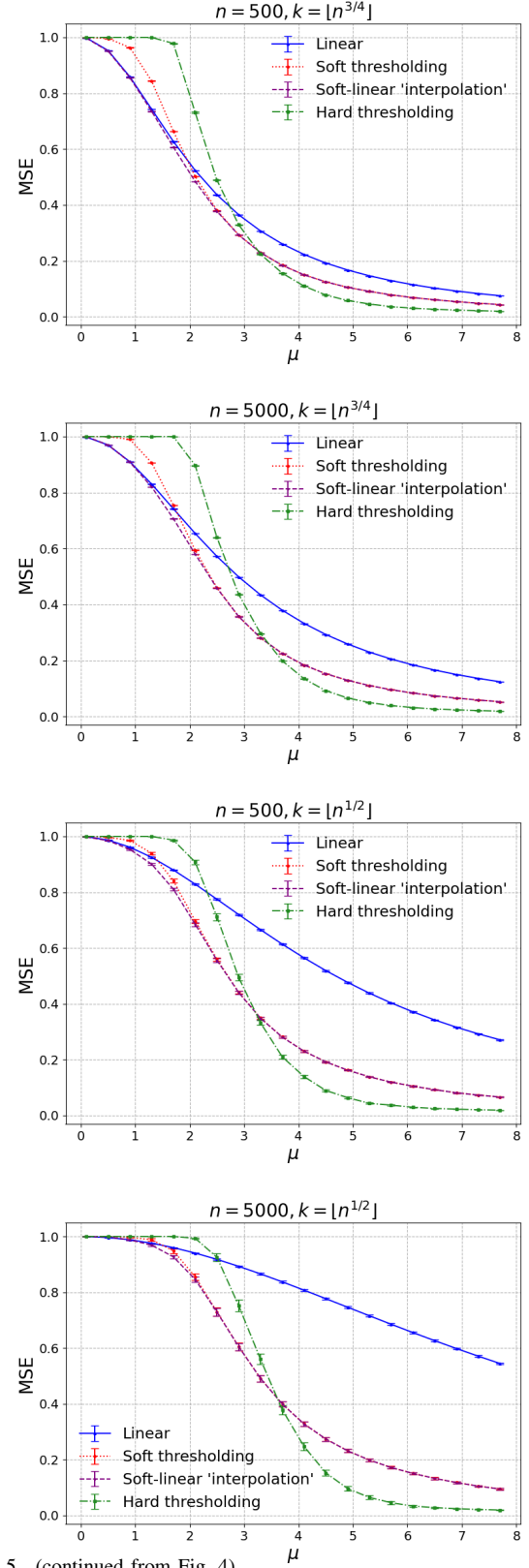
Fig. 4. Mean squared error comparison at different SNR levels. On each graph, the  $y$ -axis is the scaled MSE, and the  $x$ -axis is the SNR  $\mu_n$ . (to be continued in Fig. 5)

Fig. 5. (continued from Fig. 4)



paper may open up new venues for investigating various estimation problems. We have recently 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.

### B. 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 ( $\ell_0$ -regularization) performs better than the lasso ( $\ell_1$ -regularization) in very high SNR, while the lasso outperforms best subset selection in low SNR regimes. [9], [16] 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  $\ell_q$ -regularization under varying SNR regimes in high-dimensional sparse regression and variable selection problems [10], [17], [18]. In particular, their results revealed that among the  $\ell_q$ -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  $\ell_2$ -regularization outperforms other  $\ell_q$ -regularization when SNR is low. We obtained a stronger result that  $\ell_2$ -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 [19]. 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 [20]. 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  $\ell_p$  balls (i.e.,  $\Theta = \{\theta : \|\theta\|_p \leq C_n\}$ ) for  $p > 0$  under Gaussian sequence model has been performed in [3], [12], [21]. These works revealed that a notion of SNR involving  $C_n$  and  $\sigma_n$  plays a critical role in characterizing the asymptotic minimax risk and the optimality of linear or thresholding estimators. Finally, see [22], [23] for non-asymptotic minimax rate analysis of variable selection and functional estimation on sparse Gaussian sequence models.

### C. Future research

Several important directions are left open for future research:

- The paper considered estimating signals with sparsity  $k_n/n \rightarrow 0$ . The other denser regime where  $k_n/n \rightarrow c > 0$  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  $\tau_n$  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 paper, we have focused on the parameter spaces that imposed the exact sparsity on  $\theta$ . 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 [3], [4], [11] and references therein for a systematic treatment of this topic. An interesting future research would be to explore the implications of the SNR-aware minimaxity 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 [5], [24] 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.

## V. PROOFS

### A. Preliminaries

1) *Scale invariance*: The minimax risk defined in (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 paper. We state it without proof in the following:  $\forall \sigma > 0$ ,

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

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

2) *Gaussian tail bound*: Recall the notation that  $\phi, \Phi$  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*

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

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

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

3) *The minimax theorem*: Consider the Gaussian sequence model:

$$y_i = \theta_i + \sigma z_i, \quad i = 1, 2, \dots, n, \quad (12)$$

where  $z_1, z_2, \dots, z_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . If  $\pi$  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  $\pi$  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 7** (Theorem 4.12 in [3]). *Consider the Gaussian sequence model (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  $\pi$  is called a *least favorable distribution* (with respect to  $\mathcal{P}$ ).

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

**Theorem 8** (Lemma 4.15 in [3]).  $B(\bar{\pi}) \geq B(\pi)$ .

5) *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 9**. *Consider the minimax risk of a risk function  $r(\cdot, \cdot)$  over a parameter set  $\Theta$ :*

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

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

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

then

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

### B. Proof of Theorem 3

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 9. 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.

1) *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  $\epsilon_n = k_n/n \rightarrow 0, \mu_n = \tau_n/\sigma_n \rightarrow 0$ , and the constraint  $\|\theta\|_2^2 \leq k_n \tau_n^2, \forall \theta \in \Theta(k_n, \tau_n)$ . As a result,

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

2) *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 V-A1), 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  $z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . 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 1, 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 9, we utilize the family of *independent block priors* [3], [25]. 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  $m = n/k_n$ <sup>1</sup>:

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

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

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

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

$$R(\Theta(k_n, \mu_n), 1) \geq B(\pi^{IB}) = k_n \cdot B(\pi_S^{\mu, m}). \quad (13)$$

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.

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

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

<sup>1</sup>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.

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

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

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

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

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  $\theta = \pm \mu e_i, i = 2, \dots, m$ .  $\square$

**Lemma 3.** *As  $\mu \rightarrow 0, m \rightarrow \infty$ , 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:

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

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

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

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

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

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

To study  $E_3$ , define

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

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

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

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

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

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

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  $C \leq 4m^2 e^{2\sqrt{\mu}}$  when  $\mu$  is sufficiently small.  $\square$

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  $\epsilon_n = k_n/n \rightarrow 0, \mu_n \rightarrow 0$  guarantees that the condition  $m \rightarrow \infty, \mu \rightarrow 0$  in Lemma 3 is satisfied. We therefore combine Lemma 3 and (13) to obtain

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

### C. Proof of Proposition 1

Define the supremum risk of optimally tuned soft thresholding estimator as

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

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

$$R_s(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R_s(\Theta(k_n, \mu_n), 1). \quad (14)$$

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 (42) 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:

$$\begin{aligned}
& \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E} \sum_{i=1}^n |\hat{\eta}_S(y_i, \lambda) - \theta_i|_2^2 \\
& = (n - k_n) r_S(\lambda, 0) + k_n r_S(\lambda, \mu_n) \\
& = n \left[ (1 - \epsilon_n) r_S(\lambda, 0) + \epsilon_n r_S(\lambda, \mu_n) \right], \quad (15)
\end{aligned}$$

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

$$r_S(\lambda, \mu) = \mathbb{E}(\hat{\eta}_S(\mu + z, \lambda) - \mu)^2, \quad z \sim \mathcal{N}(0, 1). \quad (16)$$

To prove Proposition 1, we need to find the optimal  $\lambda$  that minimizes the supremum risk in (15), or equivalently, the function

$$F(\lambda) := (1 - \epsilon_n)r_S(\lambda, 0) + \epsilon_nr_S(\lambda, \mu_n). \quad (17)$$

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

$$\log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2} - 2 \log \log \frac{2}{\epsilon_n} < \lambda_* \mu_n < \log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2}, \quad (18)$$

when  $n$  is sufficiently large.

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

$$F(\lambda) = (1 - \epsilon_n)\mathbb{E}\hat{\eta}_S^2(z, \lambda) + \epsilon_n\mu_n^2 - 2\epsilon_n\mu_n\mathbb{E}\hat{\eta}_S(\mu_n + z, \lambda) + \epsilon_n\mathbb{E}\hat{\eta}_S^2(\mu_n + z, \lambda), \quad (19)$$

where the three expectations take the form

$$\mathbb{E}\hat{\eta}_S^2(z, \lambda) = 2(1 + \lambda^2) \int_{\lambda}^{\infty} \phi(z)dz - 2\lambda\phi(\lambda) \quad (20)$$

$$\begin{aligned} \mathbb{E}\hat{\eta}_S(\mu_n + z, \lambda) &= \phi(\lambda - \mu_n) + (\mu_n - \lambda) \int_{\lambda - \mu_n}^{\infty} \phi(z)dz \\ &\quad - \phi(\lambda + \mu_n) + (\mu_n + \lambda) \int_{\lambda + \mu_n}^{\infty} \phi(z)dz \end{aligned} \quad (21)$$

$$\begin{aligned} \mathbb{E}\hat{\eta}_S^2(\mu_n + z, \lambda) &= \left[ \left(1 + (\lambda - \mu_n)^2\right) \int_{\lambda - \mu_n}^{\infty} \phi(z)dz \right. \\ &\quad \left. - (\lambda - \mu_n)\phi(\lambda - \mu_n) \right] \\ &\quad + \left[ \left(1 + (\lambda + \mu_n)^2\right) \int_{\lambda + \mu_n}^{\infty} \phi(z)dz \right. \\ &\quad \left. - (\lambda + \mu_n)\phi(\lambda + \mu_n) \right]. \end{aligned} \quad (22)$$

Therefore,  $F(\lambda)$  is a differentiable function of  $\lambda$ , and as long as the infimum of  $F(\lambda)$  is not achieved at 0 or  $+\infty$ ,  $\lambda_*$  will satisfy  $F'(\lambda_*) = 0$ . From Equations (19)-(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  $\lambda$  is above a threshold. Hence, the optimal tuning  $\lambda_* \in (0, \infty)$ , and we can characterize it through the derivative equation:

$$\begin{aligned} 0 = F'(\lambda_*) &= (1 - \epsilon_n) \left[ 4\lambda_* \int_{\lambda_*}^{\infty} \phi(z)dz - 4\phi(\lambda_*) \right] \\ &\quad + \epsilon_n \left[ -2\phi(\lambda_* - \mu_n) - 2\phi(\lambda_* + \mu_n) \right. \\ &\quad \left. + 2\lambda_* \left( \int_{\lambda_* - \mu_n}^{\infty} \phi(z)dz + \int_{\lambda_* + \mu_n}^{\infty} \phi(z)dz \right) \right]. \end{aligned} \quad (23)$$

First, we show that  $\lambda_* \rightarrow \infty$ . Suppose this is not true. Then  $\lambda_* \leq C$  for some constant  $C > 0$  (take a subsequence if necessary). From (19), we have

$$F(\lambda_*) \geq (1 - \epsilon_n)r_S(C, 0)$$

$$\begin{aligned} &= 2(1 - \epsilon_n) \left[ (1 + C^2) \int_C^{\infty} \phi(z)dz - C\phi(C) \right] \\ &> \epsilon_n\mu_n^2 = F(+\infty), \end{aligned}$$

when  $n$  is large. This contradicts with the optimality of  $\lambda_*$ .

Second, we prove that  $\lambda_*\mu_n \rightarrow \infty$ . Otherwise,  $\lambda_*\mu_n = O(1)$  (take a subsequence if necessary). We will show that it leads to a contradiction in (23). Using the Gaussian tail bound  $\int_t^{\infty} \phi(z)dz = (\frac{1}{t} - \frac{1+o(1)}{t^3})\phi(t)$  as  $t \rightarrow \infty$  from Section V-A2, since  $\lambda_* \rightarrow \infty, \mu_n \rightarrow 0, \lambda_*\mu_n = O(1)$ , we obtain

$$-\lambda_* \int_{\lambda_*}^{\infty} \phi(z)dz + \phi(\lambda_*) = (1 + o(1)) \cdot \lambda_*^{-2}\phi(\lambda_*), \quad (24)$$

$$-\phi(\lambda_* + \mu_n) + \lambda_* \int_{\lambda_* + \mu_n}^{\infty} \phi(z)dz = O(\lambda_*^{-2}\phi(\lambda_*)), \quad (25)$$

$$-\phi(\lambda_* - \mu_n) + \lambda_* \int_{\lambda_* - \mu_n}^{\infty} \phi(z)dz = O(\lambda_*^{-2}\phi(\lambda_*)). \quad (26)$$

Given that  $\epsilon_n \rightarrow 0$ , combining the above results with (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  $\lambda_*\mu_n \rightarrow \infty$ , results in (25)-(26) can be strengthened:

$$-\phi(\lambda_* + \mu_n) + \lambda_* \int_{\lambda_* + \mu_n}^{\infty} \phi(z)dz = o(\mu_n\lambda_*^{-1}\phi(\lambda_* - \mu_n)), \quad (27)$$

$$\begin{aligned} &-\phi(\lambda_* - \mu_n) + \lambda_* \int_{\lambda_* - \mu_n}^{\infty} \phi(z)dz = (1 + o(1)) \\ &\quad \cdot \mu_n\lambda_*^{-1}\phi(\lambda_* - \mu_n). \end{aligned} \quad (28)$$

Plugging (24) and (27)-(28) into (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

$$2 + o(1) = \epsilon_n\mu_n\lambda_* \exp(\lambda_*\mu_n - \mu_n^2/2). \quad (29)$$

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  $2\mu_n\lambda_* \rightarrow \infty$  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

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

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 (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 (23), we have

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

$$\begin{aligned}
& + \frac{-\mu_n(\lambda + \mu_n)^2 - \lambda}{(\lambda + \mu_n)^3 \lambda^{-2}} 2\epsilon_n e^{-\lambda\mu_n - \mu_n^2/2} \Big] \\
& \geq \frac{\phi(\lambda)}{\lambda^2} \cdot \left[ -4 + 4\epsilon_n + (2 + o(1)) \cdot \epsilon_n e^{-\mu_n^2/2} \lambda \mu_n e^{\lambda\mu_n} \right],
\end{aligned}$$

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  $n$ .  $\square$

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

**Lemma 5.** *Define*

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

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

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

*Proof.* We use Gaussian tail bounds to evaluate the three expectations (20)-(22) in the expression of  $F(\lambda_*)$  in (19). Note that as shown in Lemma 4,  $\lambda_*\mu_n = \Theta(\log 2\epsilon_n^{-1})$ . The first expectation is

$$\mathbb{E}\hat{\eta}_S^2(z, \lambda_*) = 2\phi(\lambda_*) \left[ 2\lambda_*^{-3} - 12\lambda_*^{-5} + O(\lambda_*^{-7}) \right]. \quad (30)$$

Regarding the second one, we obtain

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

and

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

Therefore,

$$\begin{aligned}
\mathbb{E}\hat{\eta}_S(\mu_n + z, \lambda_*) &= \left[ (\lambda_* - \mu_n)^{-2} - (\lambda_* + \mu_n)^{-2} e^{-2\lambda_*\mu_n} \right. \\
&\left. + O\left((\lambda_* - \mu_n)^{-4}\right) \right] \phi(\lambda_* - \mu_n). \quad (31)
\end{aligned}$$

For the third expectation, we first have

$$\left( 1 + (\lambda_* - \mu_n)^2 \right) \int_{\lambda_* - \mu_n}^{\infty} \phi(z) dz - (\lambda_* - \mu_n) \phi(\lambda_* - \mu_n)$$

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

Thus,

$$\begin{aligned}
\mathbb{E}\hat{\eta}_S^2(\mu_n + z, \lambda_*) &= \left[ 2(\lambda_* - \mu_n)^{-3} + 2(\lambda_* + \mu_n)^{-3} e^{-2\lambda_*\mu_n} \right. \\
&\left. + O\left((\lambda_* - \mu_n)^{-5}\right) \right] \phi(\lambda_* - \mu_n). \quad (32)
\end{aligned}$$

Plugging (30)-(32) into (19), we have

$$\begin{aligned}
F(\lambda_*) &= 2(1 - \epsilon_n)\phi(\lambda_*) \left[ 2\lambda_*^{-3} - 12\lambda_*^{-5} \right. \\
&\left. + O(\lambda_*^{-7}) \right] + \epsilon_n \mu_n^2 \\
&- 2\epsilon_n \mu_n \left[ (\lambda_* - \mu_n)^{-2} - (\lambda_* + \mu_n)^{-2} e^{-2\lambda_*\mu_n} \right. \\
&\left. + O\left((\lambda_* - \mu_n)^{-4}\right) \right] \phi(\lambda_* - \mu_n) \\
&+ \epsilon_n \left[ 2(\lambda_* - \mu_n)^{-3} + 2(\lambda_* + \mu_n)^{-3} e^{-2\lambda_*\mu_n} \right. \\
&\left. + O\left((\lambda_* - \mu_n)^{-5}\right) \right] \phi(\lambda_* - \mu_n) \\
&= \epsilon_n \mu_n^2 + 2(1 - \epsilon_n)\phi(\lambda_*) \left[ 2\lambda_*^{-3} - 12\lambda_*^{-5} + O(\lambda_*^{-7}) \right] \\
&+ \frac{2\epsilon_n \mathcal{A} \phi(\lambda_* - \mu_n)}{(\lambda_* - \mu_n)^3}. \quad (33)
\end{aligned}$$

Next, we utilize the derivative equation (23) to further simplify (33). We first list the asymptotic approximations needed:

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

Plugging them into (23) yields

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

Obtaining the expression for  $\frac{\phi(\lambda_* - \mu_n)}{(\lambda_* - \mu_n)^3}$  from the above equation and plugging it into (33) completes the proof.  $\square$

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  $1 + \left( \lambda_* - \frac{3 + o(1)}{\lambda_*} \right) \frac{\mathcal{A}}{\mathcal{B}}$ . Using the fact that  $\lambda_*\mu_n \rightarrow \infty$ , some direct calculations enable us to obtain

$$\begin{aligned}
\left( \lambda_* - \frac{3 + o(1)}{\lambda_*} \right) \mathcal{A} + \mathcal{B} &= (-1 + o(1))\lambda_*\mu_n^2, \\
\mathcal{B} &= \mu_n\lambda_*^2(1 + o(1)).
\end{aligned}$$

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

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

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

#### D. Proof of Proposition 2

Define the supremum risk of optimally tuned hard threshold estimator as

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

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

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

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  $\theta$  such that  $\underline{\theta}_i = \mu_n$  for  $i \in \{1, 2, \dots, k_n\}$  and  $\underline{\theta}_i = 0$  for  $i > k_n$ :

$$R_H(\Theta(k_n, \mu_n), 1) \geq \inf_{\lambda > 0} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2. \quad (34)$$

Denote the one-dimensional risk:

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

It is then direct to confirm that

$$\begin{aligned} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 &= n \left[ (1 - \epsilon_n) r_H(\lambda, 0) + \epsilon_n r_H(\lambda, \mu_n) \right]. \end{aligned} \quad (35)$$

Let  $\lambda_n^*$  be the optimal choice of  $\lambda$  in  $\mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2$  so that

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

To evaluate the lower bound in (34), we consider two scenarios for the optimal choice  $\lambda_n^*$  and in each one we obtain a lower bound for  $\mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda_n^*) - \underline{\theta}\|_2^2$ . 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  $r_H(\lambda, \mu)$ :

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

$$+ (\lambda - \mu)\phi(\lambda - \mu) + (\lambda + \mu)\phi(\lambda + \mu), \quad (36)$$

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  $\lambda_n^*$  and in each case find a lower bound for the risk.

- **Case I**  $\lambda_n^* = O(1)$ : we have  $\lambda_n^* \leq c$  for some constant  $c > 0$ . Hence, from (35) we obtain

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

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  $\lambda_n^* \rightarrow \infty$  as  $n \rightarrow \infty$ . From (35) and (36), we have

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

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

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

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

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

$$\begin{aligned} R_H(\Theta(k_n, \mu_n), 1) &= \inf_{\lambda > 0} \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda) - \theta\|_2^2 \\ &\leq \lim_{\lambda \rightarrow \infty} \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda) - \theta\|_2^2 \\ &\leq \lim_{\lambda \rightarrow \infty} \left( \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda)\|_2^2 \right. \\ &\quad \left. + \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \langle -2\hat{\eta}_H(y, \lambda), \theta \rangle + \sup_{\theta \in \Theta(k_n, \mu_n)} \|\theta\|_2^2 \right) \\ &\leq n\epsilon_n \mu_n^2 + \lim_{\lambda \rightarrow \infty} \left( \sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda)\|_2^2 \right. \\ &\quad \left. + 2\sqrt{n\epsilon_n \mu_n^2} \sqrt{\sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda)\|_2^2} \right). \end{aligned} \quad (37)$$



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 (37), to show  $R_H(\Theta(k_n, \mu_n), 1) \leq n \epsilon_n \mu_n^2$ , it is sufficient to prove

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

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  $\mu$ , is symmetric around zero and increasing over  $[0, \infty)$  for all  $\lambda > 0$ . As a result,

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

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

#### E. Proof of Theorem 4

Recall the scale invariance property in Section V-A1:  $R(\Theta(k_n, \tau_n), \sigma_n) = \sigma_n^2 \cdot R(\Theta(k_n, \mu_n), 1)$ , 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 (10) also preserves an invariance:  $t \cdot \hat{\eta}_E(y, \lambda, \gamma) = \hat{\eta}_E(ty, t\lambda, \gamma), \forall t \geq 0$ . Therefore, to prove both the upper and lower bounds, in this section, it is sufficient to consider the simpler unit-variance model:

$$y_i = \theta_i + z_i, \quad i = 1, \dots, n, \quad (38)$$

where  $(z_i) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ . 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 9 and considering the independent block prior again. Both steps are more challenging than the corresponding steps in the proof of Theorem 3.

1) *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:

$$r_e(\mu; \lambda, \gamma) = \mathbb{E} \left( \frac{1}{1+\gamma} \hat{\eta}_S(\mu + z, \lambda) - \mu \right)^2, \quad z \sim \mathcal{N}(0, 1). \quad (39)$$

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  $\mu$ , is symmetric, and increases 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:

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

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

$$\begin{aligned} \frac{\partial r_e(\mu; \lambda, \gamma)}{\partial \mu} &= \frac{1}{(1+\gamma)^2} \frac{\partial \mathbb{E} (\hat{\eta}_S(\mu + z, \lambda) - \mu)^2}{\partial \mu} + \frac{2\gamma^2 \mu}{(1+\gamma)^2} \\ &\quad - \frac{2\gamma}{(1+\gamma)^2} \left[ \mu \frac{\partial \mathbb{E} (\hat{\eta}_S(\mu + z, \lambda) - \mu)}{\partial \mu} + \mathbb{E} (\hat{\eta}_S(\mu + z, \lambda) - \mu) \right]. \end{aligned} \quad (40)$$

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

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

and

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

Putting the above two results into (40), we obtain,  $\forall \mu > 0$ ,

$$\begin{aligned} \frac{\partial r_e(\mu; \lambda, \gamma)}{\partial \mu} &= \frac{2\mu}{(1+\gamma)} \mathbb{P}(|z + \mu| \leq \lambda) + \frac{2\gamma^2 \mu}{(1+\gamma)^2} \\ &\quad + \frac{2\gamma}{(1+\gamma)^2} \mathbb{E} (\mu - \hat{\eta}_S(\mu + z, \lambda)) > 0, \end{aligned} \quad (41)$$

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  $\max_{0 \leq x \leq c} H(x)$ . 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 (41), we have

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

where

$$f_1(x) := \mathbb{P}(|x+z| \leq \lambda) - \mathbb{P}(|\sqrt{c^2-x^2}+z| \leq \lambda),$$

$$f_2(x) := \frac{1}{\sqrt{c^2-x^2}} \mathbb{E} \hat{\eta}_S(\sqrt{c^2-x^2}+z, \lambda) - \frac{1}{x} \mathbb{E} \hat{\eta}_S(x+z, \lambda).$$

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  $\mathbb{P}(|x+z| \leq \lambda)$  is strictly decreasing over  $(0, c)$ , and accordingly  $\mathbb{P}(|\sqrt{c^2-x^2}+z| \leq \lambda)$  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:

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

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,

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

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

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 (38). Suppose  $\epsilon_n = k_n/n \rightarrow 0$ ,  $\mu_n \rightarrow \infty$ , and  $\mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ , as  $n \rightarrow \infty$ . 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:

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

*Proof.* Using the one-dimensional risk function in (39), we can write:

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

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

$$\begin{aligned} &\sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2 \\ &= n \left[ (1 - \epsilon_n) r_e(0; \lambda_n, \gamma_n) + \epsilon_n r_e(\mu_n; \lambda_n, \gamma_n) \right] \\ &= n \left[ \frac{1 - \epsilon_n}{(1 + \gamma_n)^2} \mathbb{E} \hat{\eta}_S^2(z, \lambda_n) + \frac{\epsilon_n}{(1 + \gamma_n)^2} \mathbb{E} \hat{\eta}_S^2(\mu_n + z, \lambda_n) \right. \\ &\quad \left. - \frac{2\epsilon_n \mu_n}{1 + \gamma_n} \mathbb{E} \hat{\eta}_S(\mu_n + z, \lambda_n) + \epsilon_n \mu_n^2 \right]. \end{aligned} \quad (42)$$

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  $t \rightarrow \infty$ . For the particular choice  $\lambda_n = 2\mu_n \rightarrow \infty$ , we have

$$\begin{aligned} \mathbb{E} \hat{\eta}_S^2(z, \lambda_n) &= 2 \left[ (1 + \lambda_n^2) \int_{\lambda_n}^{\infty} \phi(z) dz - \lambda_n \phi(\lambda_n) \right] \\ &= \frac{1 + o(1)}{2\mu_n^3} \phi(2\mu_n). \end{aligned} \quad (43)$$

Furthermore,

$$\begin{aligned} &\mathbb{E} \hat{\eta}_S^2(\mu_n + z, \lambda_n) \\ &= \left[ (1 + (\mu_n - \lambda_n)^2) \int_{\lambda_n - \mu_n}^{\infty} \phi(z) dz - (\lambda_n - \mu_n) \phi(\lambda_n - \mu_n) \right] \\ &\quad + \left[ (1 + (\mu_n + \lambda_n)^2) \int_{\lambda_n + \mu_n}^{\infty} \phi(z) dz - (\lambda_n + \mu_n) \phi(\lambda_n + \mu_n) \right] \\ &= \frac{2 + o(1)}{(\lambda_n - \mu_n)^3} \phi(\lambda_n - \mu_n) + \frac{2 + o(1)}{(\lambda_n + \mu_n)^3} \phi(\lambda_n + \mu_n) \\ &= \frac{2 + o(1)}{\mu_n^3} \phi(\mu_n), \end{aligned} \quad (44)$$

and

$$\begin{aligned} &\mathbb{E} \hat{\eta}_S(\mu_n + z, \lambda_n) = \phi(\lambda_n - \mu_n) - (\lambda_n - \mu_n) \\ &\quad \cdot \int_{\lambda_n - \mu_n}^{\infty} \phi(z) dz - \phi(\lambda_n + \mu_n) + (\mu_n + \lambda_n) \int_{\lambda_n + \mu_n}^{\infty} \phi(z) dz \\ &= \frac{1 + o(1)}{(\lambda_n - \mu_n)^2} \phi(\lambda_n - \mu_n) - \frac{1 + o(1)}{(\lambda_n + \mu_n)^2} \phi(\lambda_n + \mu_n) \\ &= \frac{1 + o(1)}{\mu_n^2} \phi(\mu_n). \end{aligned} \quad (45)$$

Plugging (43)-(45) into (42) 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

$$\sup_{\theta \in \Theta(k_n, \mu_n)} \mathbb{E}_\theta \|\hat{\eta}_E(y, \lambda_n, \gamma_n) - \theta\|_2^2$$

$$\begin{aligned}
&= k_n \mu_n^2 + (2 + o(1)) \cdot n \epsilon_n^2 \mu_n e^{\frac{3}{2} \mu_n^2} \phi(\mu_n) + (8 + o(1)) \\
&\quad \cdot n \epsilon_n^3 \mu_n e^{3 \mu_n^2} \phi(\mu_n) - (4 + o(1)) \cdot n \epsilon_n^2 \mu_n e^{\frac{3}{2} \mu_n^2} \phi(\mu_n) \\
&= k_n \mu_n^2 - (2 + o(1)) \cdot n \epsilon_n^2 \mu_n e^{\frac{3}{2} \mu_n^2} \phi(\mu_n).
\end{aligned}$$

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.  $\square$

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

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

2) *Lower bound:* The derivation of the lower bound follows the same roadmap of proof for the lower bound in Theorem 3. It relies on the independent block prior constructed in Section V-B2. According to Equation (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  $m = n/k_n \rightarrow \infty, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ . 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  $m = n/k_n \rightarrow \infty, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ , 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

$$\begin{aligned}
&\mathbb{E}_{\mu_n e_1} [p_m - 1]^2 \geq 1 - 2 \cdot \\
&\mathbb{E} \left( \frac{e^{\mu_n (\mu_n + z_1)} - e^{-\mu_n (\mu_n + z_1)}}{\sum_{j \neq 1} (e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n (\mu_n + z_1)} + e^{-\mu_n (\mu_n + z_1)}} \right) \\
&\geq 1 - 2 \cdot \\
&\mathbb{E} \left( \frac{e^{\mu_n (\mu_n + z_1)}}{\sum_{j \neq 1} (e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n (\mu_n + z_1)} + e^{-\mu_n (\mu_n + z_1)}} \right).
\end{aligned}$$

Thus, (i) will be proved by showing that

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

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

$$\begin{aligned}
&e^{\mu_n (\mu_n + z_1)} + e^{-\mu_n (\mu_n + z_1)} \geq e^{\mu_n z_1} + e^{-\mu_n z_1} \\
&\Leftrightarrow (e^{\mu_n^2} - 1) (e^{\mu_n z_1} - e^{-\mu_n z_1 - \mu_n^2}) \geq 0 \\
&\Leftrightarrow \mu_n z_1 \geq -\mu_n z_1 - \mu_n^2 \\
&\Leftrightarrow z_1 \geq -\frac{1}{2} \mu_n.
\end{aligned}$$

In the first case,

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

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

$$\begin{aligned}
&\mathbb{E} \left( \frac{e^{\mu_n (\mu_n + z_1)} I_{(z_1 \leq -\frac{1}{2} \mu_n)}}{\sum_{j \neq 1} (e^{\mu_n z_j} + e^{-\mu_n z_j}) + e^{\mu_n (\mu_n + z_1)} + e^{-\mu_n (\mu_n + z_1)}} \right) \\
&\leq e^{\mu_n^2} \mathbb{E} \left( \frac{e^{\mu_n z_1} I_{(z_1 \leq -\frac{1}{2} \mu_n)}}{\sum_{j=1}^m e^{\mu_n z_j}} \right) \\
&= \frac{e^{\mu_n^2}}{m} \mathbb{E} \left( \frac{\sum_{j=1}^m e^{\mu_n z_j} I_{(z_j \leq -\frac{1}{2} \mu_n)}}{\sum_{j=1}^m e^{\mu_n z_j}} \right), \tag{46}
\end{aligned}$$

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

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

then the last expectation in (46) 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  $\text{Var}(Y_n) \leq \frac{m}{m^2 e^{\mu_n^2}} e^{2 \mu_n^2} = o(1)$ .  
Hence,  $Y_n \rightarrow 1$  in probability. In addition,

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

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

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

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

Towards this goal, we have

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

Here, Inequality (a) is by applying the Jensen's inequality with respect to  $z_2, \dots, z_m$  (conditioned on  $z_1$ ), as  $1/(x+c)^2$  ( $c > 0$ ) is a convex function of  $x > 0$ . Inequality (b) holds because  $e^{\frac{3}{2}\mu_n^2} + e^{-\frac{\mu_n^2}{2}} + e^{\mu_n z_1} + e^{-\mu_n z_1} \leq 4e^{\frac{3}{2}\mu_n^2}$  when  $|z_1| \leq 3\mu_n$ , and  $e^{3\mu_n^2} \leq \sqrt{m}e^{\frac{\mu_n^2}{2}}$  (for large  $n$ ) under the condition  $\mu_n = o(\sqrt{\log m})$ . Equality (c) is due to the symmetry of  $z_1 \sim \mathcal{N}(0, 1)$ .  $\square$

Our goal now is to use Lemma 8 to finish the proof of the lower bound in Theorem 4:

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

#### F. Proof of Theorem 5

Like in the proof of Theorems 3 and 4, we calculate the minimax risk by deriving matching upper and lower bounds. However, a notable difference of the proof of Theorem 5 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 7) 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 V-F1 and V-F2, respectively.

1) *Upper bound:* Consider the univariate Gaussian model:

$$Y = \theta + Z, \quad (47)$$

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

$$\begin{aligned} \Gamma^A(\epsilon, \mu) &:= \left\{ \pi \in \mathcal{P}(\mathbb{R}) : \pi(\{0\}) \geq 1 - \epsilon, \mathbb{E}_\pi \theta^2 \leq \epsilon \mu^2, \right. \\ & \quad \left. \text{supp}(\pi) \in [-A\mu, A\mu] \right\}, \end{aligned} \quad (48)$$

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  $\mathbb{E}_G \theta^2 \leq \mu^2$  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

$$B^A(\epsilon, \mu, 1) := \sup \left\{ B(\pi) : \pi \in \Gamma^A(\epsilon, \mu) \right\}, \quad (49)$$

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

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

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

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

$$R(\Theta_n, \sigma_n) \leq B(\pi, \mathcal{M}_n) = B(\pi, \mathcal{M}_n^e) \leq n \sigma_n^2 \cdot B^A(\epsilon_n, \mu_n, 1). \quad (50)$$

We start with equality in (50).

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

Inequality (a) is due to the fact that  $\mathcal{M}_n$  contains all point mass priors  $\delta_\theta$ , for every  $\theta \in \Theta_n$ . To obtain Equality (b) we have used the minimax theorem, i.e. Theorem 7, as  $\mathcal{M}_n$  is a convex set of probability measures. To prove the second inequality in (50), note that for any  $\pi \in \mathcal{M}_n$ , we can construct a corresponding prior:

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

where  $\sigma$  denotes a permutation of the coordinates of  $\theta$ , and  $\pi \circ \sigma$  is the distribution after permutation. In other words,  $\pi^e$  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 (50), for any exchangeable prior  $\pi \in \mathcal{M}_n^e$ , let  $\pi_1$  be its univariate marginal distribution. Using the constraints on  $\pi$  from  $\mathcal{M}_n$  and the fact that  $\pi$  is symmetric over its  $n$  coordinates, we have

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

Hence  $\pi_1 \in \Gamma^A(\epsilon_n, \tau_n)$  defined in (48). Furthermore, according to Theorem 8, the product prior  $\pi_1^n$  is less favorable than  $\pi^e$ , 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.  $\square$

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 (47), consider prior  $\pi = (1 - \epsilon)\delta_0 + \epsilon G \in \Gamma^A(\epsilon, \mu)$ , as defined in (48). 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  $\theta$  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,

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

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

where the second equality is by a simple change of variable.  $\square$

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

**Lemma 11.** *Consider  $\epsilon_n \rightarrow 0, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$ . Under model (47), the worst-case Bayes risk  $B^A(\epsilon_n, \mu_n, 1)$  defined in (49) satisfies that for any  $A > 1$ ,*

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

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

$$\mathbb{E}(\mathbb{E}(\theta|Y) - \theta)^2 = \mathbb{E}\theta^2 - \mathbb{E}(\mathbb{E}(\theta|Y))^2. \quad (51)$$

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,

$$\begin{aligned} & \mathbb{E}(\mathbb{E}(\theta|Y))^2 = \int \frac{\epsilon^2 (\int t e^{tz - \frac{t^2}{2}} dG(t))^2}{1 - \epsilon + \epsilon \int e^{tz - \frac{t^2}{2}} dG(t)} \phi(z) dz \\ & \geq \int_{|z| \leq \sqrt{\log 1/\epsilon}} \frac{\epsilon^2 (\int t e^{tz - \frac{t^2}{2}} dG(t))^2}{1 - \epsilon + \epsilon \int e^{tz - \frac{t^2}{2}} dG(t)} \phi(z) dz \\ & \stackrel{(a)}{\geq} \frac{\epsilon^2}{1 - \epsilon + \epsilon^{\frac{1}{2}}} \int_{|z| \leq \sqrt{\log 1/\epsilon}} \left( \int t e^{tz - \frac{t^2}{2}} dG(t) \right)^2 \phi(z) dz \\ & \stackrel{(b)}{=} \frac{\epsilon^2}{1 - \epsilon + \epsilon^{\frac{1}{2}}} \iint \left[ t t' e^{t t'} \int_{-\sqrt{\log 1/\epsilon} - (t+t')}^{\sqrt{\log 1/\epsilon} - (t+t')} \phi(z) dz \right] dG(t) dG(t') \\ & = \frac{\epsilon^2}{1 - \epsilon + \epsilon^{\frac{1}{2}}} \iint_{t t' \geq 0} \left[ t t' e^{t t'} \int_{-\sqrt{\log 1/\epsilon} - (t+t')}^{\sqrt{\log 1/\epsilon} - (t+t')} \phi(z) dz \right] dG(t) dG(t') \\ & \quad + \frac{\epsilon^2}{1 - \epsilon + \epsilon^{\frac{1}{2}}} \iint_{t t' < 0} \left[ t t' e^{t t'} \int_{-\sqrt{\log 1/\epsilon} - (t+t')}^{\sqrt{\log 1/\epsilon} - (t+t')} \phi(z) dz \right] dG(t) dG(t') \\ & \stackrel{(c)}{\geq} \frac{\epsilon^2}{1 - \epsilon + \epsilon^{\frac{1}{2}}} \left( \iint_{t t' \geq 0} \left[ t t' e^{t t'} \int_{-\sqrt{\log 1/\epsilon} - (t+t')}^{\sqrt{\log 1/\epsilon} - (t+t')} \phi(z) dz \right] \right. \\ & \quad \cdot dG(t) dG(t') - |A\mu|^2 \Big) \\ & \stackrel{(d)}{\geq} \frac{\epsilon^2}{1 - \epsilon + \epsilon^{\frac{1}{2}}} \left( \int_{-\sqrt{\log 1/\epsilon} - 2A\mu}^{\sqrt{\log 1/\epsilon} - 2A\mu} \phi(z) dz \right. \\ & \quad \cdot \left. \iint_{t t' \geq 0} t t' e^{t t'} dG(t) dG(t') - |A\mu|^2 \right). \quad (52) \end{aligned}$$

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:

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

$$\begin{aligned}
&= \iint \left[ tt' e^{tt'} \int_{|z| \leq \sqrt{\log 1/\epsilon}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-(t+t'))^2} dz \right] dG(t) dG(t') \\
&= \iint \left[ tt' e^{tt'} \int_{-\sqrt{\log 1/\epsilon-(t+t')}}^{\sqrt{\log 1/\epsilon-(t+t')}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \right] dG(t) dG(t') \\
&= \iint \left[ tt' e^{tt'} \int_{-\sqrt{\log 1/\epsilon-(t+t')}}^{\sqrt{\log 1/\epsilon-(t+t')}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \right] dG(t) dG(t') \\
&= \epsilon_n \mu_n^2 + \frac{1+o(1)}{2} \epsilon_n^2 \mu_n^2 - \frac{1+o(1)}{2} \epsilon_n^2 \mu_n^2 e^{\mu_n^2} + O(\epsilon_n^2 \mu_n^2) \\
&= \epsilon_n \mu_n^2 - \frac{1}{2} \epsilon_n^2 \mu_n^2 e^{\mu_n^2} (1+o(1)).
\end{aligned}$$

Inequality (c) holds because  $e^{-|tt'|} \leq 1$  and  $\text{supp } G \subseteq [-A\mu, A\mu]$ . Inequality (d) is due to the fact that  $\text{supp } G \subseteq [-A\mu, A\mu]$  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 (52), we further lower bound  $\iint_{tt' \geq 0} tt' e^{tt'} dG(t) dG(t')$ . To simplify notation, define two random variables  $t, t' \stackrel{i.i.d.}{\sim} G$ . We have

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

where (a) is due to the basic inequality  $2(x^2 + y^2) \geq (x+y)^2$ , and (b) is by Hölder's inequality  $(\mathbb{E}|t|^{k+1})^2 \leq (\mathbb{E}|t|^{k+1})^2, k \geq 1$ . Combining the above inequality with (51) and (52) gives

$$\begin{aligned}
B^A(\epsilon_n, \mu_n, 1) &= \sup_{\pi \in \Gamma^A(\epsilon_n, \mu_n)} \mathbb{E}(\mathbb{E}(\theta|Y) - \theta)^2 \\
&\leq \sup_{\mathbb{E}|t|^2 \leq \mu_n^2} \left( \epsilon_n + \frac{\epsilon_n^2 \Delta_n}{2(1 - \epsilon_n + \sqrt{\epsilon_n})} \right) \mathbb{E}|t|^2 \\
&\quad - \frac{\epsilon_n^2 \Delta_n}{2(1 - \epsilon_n + \sqrt{\epsilon_n})} \mathbb{E}|t|^2 e^{\mathbb{E}|t|^2} + \frac{\epsilon^2 A^2 \mu_n^2}{1 - \epsilon + \sqrt{\epsilon}}, \quad (53)
\end{aligned}$$

where  $\Delta_n = \int_{-\sqrt{\log 1/\epsilon_n - 2\mu_n A}}^{\sqrt{\log 1/\epsilon_n - 2\mu_n A}} \phi(z) dz$ . The results we obtained so far are non-asymptotic. We now make use of the conditions  $\epsilon_n \rightarrow 0, \mu_n \rightarrow \infty, \mu_n = o(\sqrt{\log \epsilon_n^{-1}})$  to derive the final asymptotic result. Under such scaling conditions, it is straightforward to confirm that the expression on the right-hand side of (53) is increasing in  $\mathbb{E}|t|^2$  when  $n$  is sufficiently large (by calculating its derivative). As a result,

$$B^A(\epsilon_n, \mu_n, 1) \leq \left( \epsilon_n + \frac{\epsilon_n^2 \Delta_n}{2(1 - \epsilon_n + \sqrt{\epsilon_n})} \right) \mu_n^2$$

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

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

2) *Lower bound*: Recall that in the lower bound derivation for Theorem 4, in Section V-E2, the proof is based on the independent block prior  $\pi^{IB}$  with single spike distribution  $\pi_S^{\mu_n, m}$  which is first introduced in Section V-B2. 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  $\text{supp } \pi^{IB} \subseteq \Theta^A(k_n, \mu_n)$  as well. As a result, the proof in Section V-E2 also works for the new parameter space  $\Theta^A(k_n, \mu_n)$  and it yields the same lower bound:

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

### G. 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 V-C, the key is to obtain the accurate order of the optimal tuning  $\lambda_*$  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  $\lambda_*$ .

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

$$\log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2} - 2 \log \log \frac{2}{\epsilon_n} < \lambda_* \mu_n < \log 2\epsilon_n^{-1} + \frac{\mu_n^2}{2}, \quad (54)$$

for sufficiently large  $n$ .

*Proof.* This lemma is an analog of Lemma 4 (comparing Equation (18) with (54)). The proof is thus similar too. We will skip equivalent calculations and only highlight the differences.

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

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

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

Second, we utilize the derivative equation  $F'(\lambda_*) = 0$  in Equation (23) to obtain more accurate order information of  $\lambda_*$ . The results  $\mu_n \rightarrow \infty, \lambda_* \mu_n^{-1} \rightarrow \infty$  imply that  $\lambda_* \rightarrow \infty, \lambda_* - \mu_n \rightarrow \infty, \lambda_* \mu_n \rightarrow \infty$ . This is all needed to obtain Equation (24) and Equations (27)-(28). As a result, Equation (29) holds here as well:

$$2 + o(1) = \epsilon_n \mu_n \lambda_* \exp(\lambda_* \mu_n - \mu_n^2/2). \quad (55)$$

To reach (54) 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.  $\square$

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

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

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

*Proof.* This proof deviates a bit from the one of Lemma 5. We will more directly utilize the order information of  $\lambda_*$  proved in Lemma 12 to calculate  $F(\lambda_*)$ . Before that, we need a refinement of (55). This is achieved by refining Equation (24) and Equations (27)-(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 (23) and arranging terms gives

$$\begin{aligned} &e^{\lambda_* \mu_n - \frac{\mu_n^2}{2}} \frac{\epsilon_n \mu_n \lambda_*^2}{2(\lambda_* - \mu_n)} - 1 \\ &= \frac{(1 - \epsilon_n)(1 + O(\lambda_*^{-2}))\mu_n}{\mu_n - (\lambda_* - \mu_n)^{-2}(\lambda_* + O(\lambda_*^{-1}))} - 1 \\ &= \frac{\lambda_*(\lambda_* - \mu_n)^{-2} + O(\lambda_*^{-2}\mu_n)}{\mu_n - (\lambda_* - \mu_n)^{-2}(\lambda_* + O(\lambda_*^{-1}))} = \frac{1 + o(1)}{\lambda_* \mu_n}, \quad (56) \end{aligned}$$

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  $\lambda_*$  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 (20)-(22)) in the expression of  $F(\lambda_*)$  (i.e. Equation (19)):

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

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

Using these three approximations in Equation (19), we obtain

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

We further replace  $e^{\lambda_* \mu_n - \frac{\mu_n^2}{2}}$  in the above with the result from (56) to have

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

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.  $\square$

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

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

#### H. Proof of Proposition 4

The proof of this proposition is similar to the proof of Proposition 2 presented in Section V-D. Hence, for the sake of brevity we adopt the same notation from Section V-D 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

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

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  $\theta$  such that  $\underline{\theta}_i = \mu_n$  for  $i \in \{1, 2, \dots, k_n\}$  and  $\underline{\theta}_i = 0$  for  $i > k_n$ . We have

$$\mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 = n \left[ (1 - \epsilon_n) r_H(\lambda, 0) + \epsilon_n r_H(\lambda, \mu_n) \right]. \quad (57)$$



To evaluate  $\inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2$ , we consider three scenarios for the optimal choice of  $\lambda_n^*$ , denoted by  $\lambda_n^*$ .

- **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  $\epsilon_n \mu_n^2 \rightarrow 0$  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

$$\begin{aligned} \inf_{\lambda>0} \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda) - \underline{\theta}\|_2^2 &= \mathbb{E}_{\underline{\theta}} \|\hat{\eta}_H(y, \lambda_n^*) - \underline{\theta}\|_2^2 \\ &= n \left[ (1 - \epsilon_n) r_H(\lambda_n^*, 0) + \epsilon_n r_H(\lambda_n^*, \mu_n) \right] \\ &\geq n(1 - \epsilon_n) r_H(\lambda_n^*, 0) \\ &= n(1 - \epsilon_n) \left[ 2\lambda_n^* \phi(\lambda_n^*) + 2(1 - \Phi(\lambda_n^*)) \right] \\ &\geq 2n(1 - \epsilon_n) \lambda_n^* \phi(\lambda_n^*) \\ &\geq 2n(1 - \epsilon_n) \frac{c_1}{\sqrt{2\pi}} e^{-\frac{c_2^2 \mu_n^2}{2}} \geq n\epsilon_n \mu_n^2, \end{aligned}$$

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

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

Note that since the three cases we have discussed above cover all the ranges of  $\lambda_n^*$ , 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.

### I. Proof of Theorem 2

Based on the scale invariance property of minimax risk mentioned in Section V-A1, it is equivalent to prove

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

where  $\nu_n = \sqrt{2 \log \epsilon_n^{-1}}$ . As in the proof of Theorems 3 and 4, 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 6.

1) *Properties of the risk of hard thresholding estimator:* Consider the one-dimensional risk of hard thresholding for  $\mu \in \mathbb{R}$  and  $\lambda > 0$ ,

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

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

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

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

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

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

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

(b) The large  $\mu$  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  $\mu$ . 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  $\lambda \rightarrow \infty$ , the risk of the hard thresholding,  $r_H(\lambda, \mu)$ , satisfies

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

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

$$\begin{aligned} r_H(\lambda, \mu) &= \mu^2 [\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] + \int_{|z+\mu|>\lambda} z^2 \phi(z) dz \\ &= (\mu^2 - 1) [\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] \\ &\quad + 1 + (\lambda - \mu)\phi(\lambda - \mu) + (\lambda + \mu)\phi(\lambda + \mu). \end{aligned} \quad (58)$$

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

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

where in the first equality we have applied the Gaussian tail bound:  $1 - \Phi(x) = (1 + o(1))x^{-1}\phi(x)$  as  $x \rightarrow \infty$ . To prove the first part of the lemma, let  $\mu = \lambda$ . From (58) we have

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

$$= \lambda^2/2 (1 + o(1)).$$

□

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

**Lemma 16.** *As  $\lambda \rightarrow \infty$ , 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 supremum risk is attained at  $\mu = \mu^* \leq \lambda$  (when  $\lambda$  is large). To evaluate  $r_H(\lambda, \mu^*)$ , it is important to derive an accurate approximation for  $\mu^*$ . We first claim that  $\mu^*/\lambda \rightarrow 1$ . 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  $\lambda$ , 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  $\mu^*$ .

Second, we show that  $\lambda - \mu^* \rightarrow \infty$ , while  $(\lambda - \mu^*)/\lambda \rightarrow 0$ . Otherwise, it satisfies  $0 \leq \lambda - \mu^* \leq c$  for some finite constant  $c \geq 0$  (take a sequence if necessary). Then from (58) we have

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

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  $c > 1$ ,  $\lambda - \mu^* \leq c\sqrt{2\log \lambda}$  for sufficiently large  $\lambda$ . Otherwise, there exists some constant  $c > 1$  such that  $\lambda_n - \mu_n^* > c\sqrt{2\log \lambda_n}$  for a sequence  $\lambda_n \rightarrow \infty$  as  $n \rightarrow \infty$ . As a result, using Equation (58), and the result proved earlier that  $\lambda_n - \mu_n^* \rightarrow \infty$ , we obtain that for large  $n$ ,

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

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  $(\lambda - \mu^*)/\sqrt{2\log \lambda} \rightarrow 1$  as  $\lambda \rightarrow \infty$ . 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  $\lambda_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Using Equation (58) and Gaussian tail bound  $1 - \Phi(x) = \frac{1+o(1)}{x}\phi(x)$  as  $x \rightarrow \infty$ , we have

$$\begin{aligned} r_H(\lambda_n, \mu_n^*) &= (\mu_n^*)^2 [\Phi(\lambda_n - \mu_n^*) - \Phi(-\lambda_n - \mu_n^*)] + O(1) \\ &\leq (\mu_n^*)^2 \Phi(\lambda_n - \mu_n^*) + O(1) \end{aligned}$$

$$= (\mu_n^*)^2 \left[ 1 - \frac{1+o(1)}{\lambda_n - \mu_n^*} \phi(\lambda_n - \mu_n^*) \right] + O(1).$$

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

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

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 (58): as  $\lambda \rightarrow \infty$ ,

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

Combining this result with Lemma 15 completes the proof. □

2) *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 6. First of all, due to the scale invariance of hard thresholding, the supremum risk can be written in the form:

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

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

$$\begin{aligned} &\sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda_n) - \theta\|_2^2 \\ &= \sigma_n^2 \left[ (n - k_n)r_H(\nu_n, 0) + k_n \sup_{\mu \geq 0} r_H(\nu_n, \mu) \right]. \end{aligned} \quad (59)$$

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 (59). We can apply Lemma 16 to continue from (59):

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

$$\begin{aligned}
&= n\sigma_n^2 \left[ (1 - \epsilon_n)r_H(\nu_n, 0) + \epsilon_n \sup_{\mu \geq 0} r_H(\nu_n, \mu) \right] \\
&= n\sigma_n^2 \left[ (1 - \epsilon_n)r_H(\nu_n, 0) + \epsilon_n \left( \nu_n^2 - 2\nu_n \sqrt{2 \log \nu_n} \right. \right. \\
&\quad \left. \left. + o(\nu_n \sqrt{\log \nu_n}) \right) \right], \tag{60}
\end{aligned}$$

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

$$\begin{aligned}
r_H(\nu_n, 0) &= 2 \int_{\nu_n}^{\infty} z^2 \phi(z) dz = 2\nu_n \phi(\nu_n) + 2(1 - \Phi(\nu_n)) \\
&= (2 + o(1))\nu_n \phi(\nu_n) = O(\epsilon_n \nu_n), \tag{61}
\end{aligned}$$

where the last two equations are due to the Gaussian tail bound  $1 - \Phi(x) = \frac{1+o(1)}{x} \phi(x)$  as  $x \rightarrow \infty$  and  $\nu_n = \sqrt{2 \log \epsilon_n^{-1}}$ . Therefore, from (60) we obtain

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

This completes our proof of the upper bound in Theorem 6.

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  $\lambda_n$ . The lemma below characterizes the possible choices of  $\lambda_n$  that leads to optimal supremum risk (up to second order).

**Lemma 17.** Consider model (1), and parameter space (6) under Regime (III), in which  $\epsilon_n \rightarrow 0$ ,  $\mu_n \rightarrow \infty$ ,  $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ , as  $n \rightarrow \infty$ . Let  $\nu_n = \sqrt{2 \log \epsilon_n^{-1}}$ . Consider the tuning regime  $\lambda_n \sigma_n^{-1} \rightarrow \infty$  and  $\lambda_n \sigma_n^{-1} \leq \mu_n$ . If  $\lambda_n$  satisfies:

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

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

$$\begin{aligned}
&\inf_{\lambda} \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda) - \theta\|_2^2 \\
&= \sup_{\theta \in \Theta(k_n, \tau_n)} \mathbb{E}_\theta \|\hat{\eta}_H(y, \lambda_n) - \theta\|_2^2 + o\left(n\sigma_n^2 \epsilon_n \nu_n \sqrt{\log \nu_n}\right). \tag{62}
\end{aligned}$$

On the other hand, if  $(\nu_n^2 - c_1 \log \log \nu_n) \geq \lambda_n^2 \sigma_n^{-2}$  for a constant  $c_1 \geq 1$  or if  $\lambda_n^2 \sigma_n^{-2} \geq (\nu_n^2 + c_2 \nu_n \sqrt{2 \log \nu_n})$  for some  $c_2 > 0$ , then the conclusion (62) will not hold.

*Proof.* Denote  $\tilde{\lambda}_n = \lambda_n \sigma_n^{-1}$ . Given that we focus on the tuning regime  $\tilde{\lambda}_n \rightarrow \infty$  and  $\lambda_n \leq \mu_n$ , the result (60) continues to hold here:

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

Hence, we define

$$\begin{aligned}
A(\lambda) &:= (1 - \epsilon_n)r_H(\lambda, 0) + \epsilon_n \left[ \lambda^2 \right. \\
&\quad \left. - 2\lambda \sqrt{2 \log \lambda} + o\left(\lambda \sqrt{\log \lambda}\right) \right], \tag{63}
\end{aligned}$$

where the notation  $o(\cdot)$  is understood as  $\lambda \rightarrow \infty$ . 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  $\tilde{\lambda}_n^2 \leq \nu_n^2 - 2c \log(\nu_n/\sqrt{2\pi})$  for some constant  $c > 1$ . Equation (61) implies

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

Note that  $A(\tilde{\lambda}_n) = \omega(A(\nu_n))$ , and hence  $\tilde{\lambda}_n$  does not satisfy (62).

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

$$(1 - \epsilon_n)r_H(\tilde{\lambda}_n, 0) \geq \frac{2 + o(1)}{\sqrt{2\pi}} \left( \epsilon_n \nu_n \left( \sqrt{\log \nu_n} \right)^{c_2} \right). \tag{64}$$

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

$$\begin{aligned}
&\epsilon_n \left( \tilde{\lambda}_n^2 - 2\tilde{\lambda}_n \sqrt{2 \log \tilde{\lambda}_n} + o\left(\tilde{\lambda}_n \sqrt{\log \tilde{\lambda}_n}\right) \right) \\
&\geq \epsilon_n \left[ \nu_n^2 - 2c_1 \log \frac{\nu_n}{\sqrt{2\pi}} - 2\nu_n \sqrt{2 \log \nu_n} \right. \\
&\quad \left. + o\left(\nu_n \sqrt{\log \nu_n}\right) \right] \\
&= \epsilon_n \left[ \nu_n^2 - 2\nu_n \sqrt{2 \log \nu_n} + o\left(\nu_n \sqrt{\log \nu_n}\right) \right]. \tag{65}
\end{aligned}$$

Combining (64)-(65) yields

$$\begin{aligned}
A(\tilde{\lambda}_n) &\geq \epsilon_n \left[ \nu_n^2 + \nu_n \sqrt{2 \log \nu_n} \left( -2 + o(1) \right) \right. \\
&\quad \left. + \frac{2 + o(1)}{2\sqrt{\pi}} (\sqrt{\log \nu_n})^{c_2-1} \right].
\end{aligned}$$

Since  $c_2 \geq 1$ , 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 (62).

- Case  $\nu_n^2 - c_1 \log \log \nu_n \leq \tilde{\lambda}_n^2 \leq \nu_n^2 + c_2 \nu_n \sqrt{2 \log \nu_n}$  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

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

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 (63), we have

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

Thus,  $A(\tilde{\lambda}_n) \leq A(\nu_n) + o(\epsilon_n \nu_n \sqrt{\log \nu_n})$ , and  $\tilde{\lambda}_n$  satisfies (62).

- 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  $\lambda_n$ , (63) implies that  $A(\tilde{\lambda}_n)/A(\nu_n) > 1$  for large  $n$ . When  $\tilde{\lambda}_n = (1 + o(1))\nu_n$ , we have

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

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 (62).  $\square$

3) *Lower bound:* As in the proof of lower bound in Theorems 3-5, we will apply Theorem 9 and utilize the independent block prior that is first described in Section V-B2. To simplify the calculations a bit here, we will use the block prior with one minor modification: adopting the notation from Section V-B2, the spike prior  $\pi_S^{\mu, m}$  in use is now changed to a one-sided spike prior:

$$\pi_S^{\mu, m}(\theta^{(j)}) = \mu e_i = \frac{1}{m}, \quad 1 \leq i \leq m, \quad (66)$$

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

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

*Proof.* Define the notation:

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

Then

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

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

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

For the following calculation, we will use Gaussian tail bound  $1 - \Phi(x) \leq x^{-1} \phi(x)$  for  $x > 0$ . To obtain a proper upper bound for the first term, we note that

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

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

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

The last inequality is based on the following calculation:

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

and

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

□

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  $m \rightarrow \infty$ , the Bayes risk  $B(\pi_S^{\mu,m})$  satisfies*

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

*Proof.* For the one-sided spike prior  $\pi_S^{\mu,m}$  introduced in (66), doing similar calculations as in the proof of Lemma 2, we can obtain the expression for the Bayes risk:

$$\begin{aligned}
B(\pi_S^{\mu,m}) &= \mu^2 \mathbb{E}_{\mu e_1} (p_m - 1)^2 + (m-1) \mu^2 \mathbb{E}_{\mu e_2} p_m^2 \\
&\geq \mu^2 - 2\mu^2 \mathbb{E}_{\mu e_1} p_m,
\end{aligned} \tag{67}$$

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

$$\begin{aligned}
\mathbb{E}_{\mu e_1} p_m &= \mathbb{E} \frac{e^{\mu(\mu + z_1)}}{\sum_{j \neq 1} e^{\mu z_j} + e^{\mu(\mu + z_1)}}, \\
&= \mathbb{E} \frac{(m-1)^{-1} e^{\frac{1}{2}\mu^2 + \mu z_1}}{(m-1)^{-1} e^{\frac{1}{2}\mu^2 + \mu z_1} + (m-1)^{-1} e^{-\frac{1}{2}\mu^2} \sum_{j \neq 1} e^{\mu z_j}},
\end{aligned} \tag{68}$$

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

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

where  $\delta$  and  $M$  are two positive constants to be determined later. Since the ratio inside the expectation of (68) 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 (68) to obtain

$$\mathbb{E}_{\mu e_1} p_m \leq \frac{1}{M \cdot \delta} + \mathbb{P}(\mathcal{F}_1^c) + \mathbb{P}(\mathcal{F}_2^c). \tag{69}$$

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

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

$$\cdot \exp\left(-\frac{1}{2\mu^2} \left[\frac{1}{2}(\nu_{m-1}^2 - \mu^2) - \log M\right]^2\right) := U_1,$$

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,

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

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  $U_1$ ,

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

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

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

and

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

In combination,

$$U_1 \leq o(1) \cdot \nu_{m-1}^{-1} (1 + o(1)) = o(\nu_{m-1}^{-1}). \tag{70}$$

For  $U_2$ , we set  $\delta$  to be any fixed constant between  $(0, 1)$ . Since  $\nu_{m-1} - \mu \rightarrow +\infty$ , 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,

$$U_2 = \frac{1 + o(1)}{\sqrt{2\pi}\nu_{m-1}}. \tag{71}$$

Note that we have set  $M = \nu_{m-1}$ . Hence,  $1/(M \cdot \delta) = O(1/\nu_{m-1})$ . Combining (69)-(71), we have

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

Finally, the above together with (67) shows that

$$B(\pi_S^{\mu,m}) \geq \mu^2 - 2\mu^2 O\left(\nu_{m-1}^{-1}\right)$$

$$\begin{aligned}
&= \nu_{m-1}^2 - 2\nu_{m-1}\sqrt{2\log\nu_{m-1}}(1+o(1)) \\
&= \nu_m^2 - 2\nu_m\sqrt{2\log\nu_m}(1+o(1)).
\end{aligned}$$

□

Now, we aim to apply Lemma 19 to derive the minimax lower bound. First note that in the current regime  $\epsilon_n \rightarrow 0$ ,  $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ , the choice of  $\mu$  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 (13) and Lemma 19 to conclude

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

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

#### J. Proof of Proposition 6

1) *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 (14)-(17) in the proof of Proposition 1, the supremum risk can be expressed as

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

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 V-J2, and then obtain the matching lower bound in Section V-J3. 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  $\epsilon_n \rightarrow 0$ ,  $\mu_n = \omega(\sqrt{\log \epsilon_n^{-1}})$ , as  $n \rightarrow \infty$ . If  $\lambda \rightarrow \infty$  and  $\mu_n - \lambda \rightarrow +\infty$ , then

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

Furthermore, when  $\lambda$  is large, it holds that

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

where

$$C(\lambda) := 2(1 - \epsilon_n) \cdot \left( \frac{2}{\lambda^3} - \frac{12}{\lambda^5} \right) \frac{1}{\sqrt{2\pi}} \epsilon_n \cdot e^{\frac{1}{2}(\nu_n^2 - \lambda^2)} \quad (72)$$

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

and

$$D(\lambda) := \epsilon_n \left\{ (1 - \epsilon_n) \frac{4}{\sqrt{2\pi}\lambda^3} e^{\frac{1}{2}(\nu_n^2 - \lambda^2)} + \lambda^2 + 1 \right\}. \quad (73)$$

*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 (19)-(22), we have that as  $\lambda \rightarrow \infty$ ,  $\mu_n - \lambda \rightarrow +\infty$ ,

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

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  $x \rightarrow \infty$ .

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

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

Thus,

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

As a result, it is straightforward to verify that  $C(\lambda)$  and  $D(\lambda)$  defined in (72)-(73) provide lower and upper bounds for  $F(\lambda)$ . □

2) *Upper bound:* Consider  $\lambda = \sqrt{\nu_n^2 - 6\log \nu_n}$ , then  $\lambda \rightarrow \infty$  and  $\mu_n - \lambda \rightarrow \infty$ . From Lemma 20,

$$\begin{aligned}
F(\lambda_*) &\leq F(\lambda) \leq D(\lambda) \\
&= \epsilon_n \left\{ (1 - \epsilon_n) \frac{4}{\sqrt{2\pi}} e^{\frac{1}{2}(\nu_n^2 - \lambda^2) - 6\log \lambda} + \lambda^2 + 1 \right\} \\
&= \epsilon_n \left\{ \frac{4 + o(1)}{\sqrt{2\pi}} + \lambda^2 + 1 \right\} = \epsilon_n \nu_n^2 - 6\epsilon_n \log \nu_n (1 + o(1)). \quad (74)
\end{aligned}$$

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

**Step 1.** First, we show that  $\lambda_* \rightarrow \infty$ ,  $\mu_n - \lambda_* \rightarrow +\infty$ . 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  $\mu \in \mathbb{R}$ ,*

$$\frac{1}{2}\bar{r}_S(\lambda, \mu) \leq r_S(\lambda, \mu) \leq \bar{r}_S(\lambda, \mu).$$

Suppose  $\lambda_* \rightarrow \infty$  is not true. Then  $\lambda_* \leq c$  for some finite constant  $c \geq 0$  (take a subsequence if necessary). Then, from the definition of  $F(\lambda_*)$  we have

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

which contradicts with (74). Further suppose  $\mu_n - \lambda_* \rightarrow +\infty$  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  $n$ ,

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

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  $\lambda_* \rightarrow \infty$ ,  $\mu_n - \lambda_* \rightarrow +\infty$ , we can apply Lemma 20 to reach

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

This contradicts with (74). 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

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

**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

$$\begin{aligned} F(\lambda_*) &\geq C(\lambda_*) = \frac{(4 + o(1))\epsilon_n}{\sqrt{2\pi}} \frac{e^{\frac{1}{2}(\nu_n^2 - \lambda_*^2)}}{\lambda_*^3} \\ &\quad + \epsilon_n (\lambda_*^2 + 1 + o(1)). \end{aligned} \quad (75)$$

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  $\tilde{c} = \frac{c - 6 + o(1)}{2} > 0$ , (75) 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 (74). Regarding the case  $c < 6$ , (75) 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 matter what value  $c \in [-\infty, 6)$  takes, the above lower bound is larger than the upper bound in (74), resulting in the same contradiction.

Now that we have derived the accurate order information for  $\lambda_*$ :  $\lambda_*^2 = \nu_n^2 - (6 + o(1)) \log \nu_n$ , we can plug it into (75) to obtain the sharp lower bound:

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

### K. Proof of Proposition 7

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

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

### 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 ed. Springer Publishing Company, Incorporated, 2008.
- [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  $\ell_q$ -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  $\ell_p$ -minimization outperform  $\ell_1$ -minimization?" *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 6896–6935, 2017.
- [11] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.
- [12] D. Donoho and I. Johnstone, "Minimax risk over  $\ell_p$  balls for  $\ell_q$  losses," *Probab. Theory Related Fields*, vol. 99, pp. 277–303, 1994.



- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] W. Jiang and C.-H. Zhang, "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, vol. 37, no. 4, pp. 1647 – 1684, 2009. [Online]. Available: <https://doi.org/10.1214/08-AOS638>
- [20] I. M. Johnstone, "On minimax estimation of a sparse normal mean vector," *The Annals of Statistics*, pp. 271–289, 1994.
- [21] C.-H. Zhang, "Minimax  $\ell_q$  risk in  $\ell_p$  balls," *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, p. 78, 2012.
- [22] 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.
- [23] 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.
- [24] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- [25] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard, "Universal near minimaxity of wavelet shrinkage," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 183–218.

**Yilin Guo** holds a Ph.D. in Statistics from Columbia University in 2023. Before joining Columbia, she received a B.S. in Statistics from University of Science and Technology of China in 2018. Her research interests include high-dimensional statistics and statistical machine learning.

**Haolei Weng** is currently an Assistant Professor at the Department of Statistics and Probability, Michigan State University. Prior to MSU, he completed his Ph.D. in Statistics from Columbia University in 2017 and was a postdoctoral researcher at Princeton University in 2018. Before going to Columbia, he received a B.S. in Statistics from University of Science and Technology of China. His research interests are broadly in the area of high-dimensional statistics and statistical machine learning.

**Arian Maleki** is an associate professor in the Department of Statistics at Columbia University. He received his PhD from Stanford University in 2011. Before joining Columbia University, he was a postdoctoral scholar at Rice University. Arian's research interests include high-dimensional statistics, computational imaging, compressed sensing, and machine learning.