SLOPE for Sparse Linear Regression: Asymptotics and Optimal Regularization

Hong Hu and Yue M. Lu

Abstract

In sparse linear regression, the SLOPE estimator generalizes LASSO by penalizing different coordinates of the estimate according to their magnitudes. In this paper, we present a precise performance characterization of SLOPE in the asymptotic regime where the number of unknown parameters grows in proportion to the number of observations. Our asymptotic characterization enables us to derive the fundamental limits of SLOPE in both estimation and variable selection settings. We also provide a computational feasible way to optimally design the regularizing sequences such that the fundamental limits are reached. In both settings, we show that the optimal design problem can be formulated as certain infinite-dimensional convex optimization problems, which have efficient and accurate finite-dimensional approximations. Numerical simulations verify all our asymptotic predictions. They demonstrate the superiority of our optimal regularizing sequences over other designs used in the existing literature.

I. Introduction

A. Motivation and Problem Setup

In sparse linear regression, we seek to estimate a sparse vector $\boldsymbol{\beta} \in \mathbb{R}^p$ from

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is the design matrix and $\boldsymbol{w}$ denotes the observation noise. In this paper, we study the *sorted* ℓ_1 *penalization estimator* (SLOPE) [2] (see also [3], [4]), a new paradigm for sparse linear regression. Given a non-decreasing regularization sequence $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_p]^\top$ with $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$, SLOPE estimates $\boldsymbol{\beta}$ by solving the following optimization problem

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{b}\|_2^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \tag{2}$$

where $|b|_{(1)} \le |b|_{(2)} \le \cdots \le |b|_{(p)}$ is a reordering of the absolute values $|b_1|, |b_2|, \ldots, |b_p|$ in increasing order. In [2], the regularization term $J_{\boldsymbol{\lambda}}(\boldsymbol{b}) \stackrel{\text{def}}{=} \sum_{i=1}^{p} \lambda_i |b|_{(i)}$ is referred to as the "sorted ℓ_1 norm" of $\boldsymbol{b}$. The same regularizer

H. Hu and Y. M. Lu are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. (e-mails: honghu@g.harvard.edu and yuelu@seas.harvard.edu). This work was supported by the Harvard FAS Dean's Fund for Promising Scholarship, and by the US National Science Foundation under grants CCF-1718698 and CCF-1910410. The preliminary version of this work has been presented at the 2019 Signal Processing with Adaptive Sparse Structured Representations (SPARS) workshop and the 2019 IEEE International Symposium on Information Theory (ISIT) [1].

was independently developed in a different line of work [3]–[6], where the motivation is to promote group selection in the presence of correlated covariates.

The classical LASSO estimator is a special case of SLOPE. It corresponds to using a constant regularization sequence, *i.e.*, $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$. However, with more general $\boldsymbol{\lambda}$-sequences, SLOPE has the flexibility to penalize different coordinates of the estimate according to their magnitudes. This adaptivity endows SLOPE with some nice *statistical* properties that are not possessed by LASSO. For example, it is shown in [7], [8] that SLOPE achieves the minimax ℓ_2 estimation rate with high probability. When applied in variable selection problem, SLOPE is shown to control the false discovery rate (FDR) for orthogonal design matrices [2], which is not the case for LASSO. In addition, the new regularizer $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ is still a norm [2], [4]. Thus, the optimization problem associated with SLOPE remains convex, and it can be efficiently solved by using *e.g.*, proximal gradient descent [2], [4].

Although the flexible regularization of SLOPE creates the hope of potential performance enhancement, to fully realize SLOPE's potential, we have to carefully design the regularizing sequence $\boldsymbol{\lambda}$. Note that this is equivalent to specifying the empirical distribution of $\boldsymbol{\lambda}$. Popular choices in the previous works include delta distribution (*i.e.*, LASSO), uniform distribution [5], chi-distribution [9], etc. These regularization schemes are mostly devised based on statistical insights gained from simpler models and they are indeed superior than LASSO in several applications. However, the success of these regularizing sequences provide no quantitative answer to the following two questions:

1) What is the fundamental limit of SLOPE?
2) How to optimally design $\boldsymbol{\lambda}$ to reach the fundamental limit?

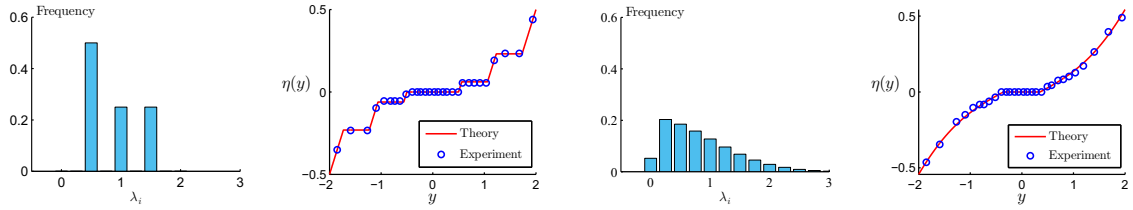
The aforementioned studies on analyzing SLOPE provide very limited information for us to address the above two questions, since in these works, the SLOPE's performance is characterized in an order-wise manner, which contains loose constants. What we need is an *exact* performance characterization of SLOPE estimator, which is still absent in the existing literature. On the other hand, however, exact asymptotic analysis has been carried out for LASSO [10], [11] and several other regularized regression techniques [12]–[15], under certain statistical assumptions on the sensing matrix $\boldsymbol{A}$. One key feature of all these results is that the performance in the originally high-dimensional model can be well-captured by some low dimensional problems, which are much easier to handle. The technical hurdle that has precluded a similar treatment for SLOPE is that unlike all the regularizer considered in these works, the SLOPE norm $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ is *non-separable*: it cannot be written as a sum of component-wise functions, *i.e.*, $J_{\boldsymbol{\lambda}}(\boldsymbol{x}) \neq \sum_{i=1}^{p} J_i(x_i)$. This makes a similar low-dimensional reduction more challenging.

B. Main Contributions

In this paper, we answer the questions raised above. Our main contributions are listed as follows:

1) Asymptotic separability: As mentioned above, the main obstacle in analyzing SLOPE asymptotics is the non-separability of SLOPE regularizer $J_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \sum_{i=1}^{p} \lambda_i |b|_{(i)}$. We overcome this challenge by showing that the proximal operator of $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ is *asymptotically separable*. To be more concrete, we first give a technically light overview of this result. The *proximal operator* of $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ is defined as:

$$\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) \overset{\text{def}}{=} \underset{\boldsymbol{x}}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{x}) \tag{3}$$



(a) (b) (c) (d)

Figure 1: (a) and (c): The histograms of two different $\boldsymbol{\lambda}$-sequences. (b) and (d): Sample points of $(y_i, [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i)$ (the blue dots) compared against the limiting scalar functions $\eta(y)$ (the red curves). In this experiment, $p = 1024$ and $y_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. For better visualization, we randomly sample 3% of all $(y_i, [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i)$.

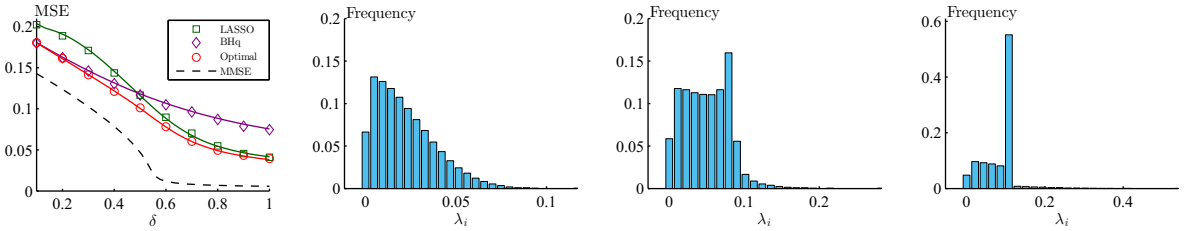
In the case of LASSO, where we choose $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$, characterizing $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is easy, since the optimization in (3) is equivalent to p scalar problems: $\sum_{i=1}^{p} \min_{x_i} \frac{1}{2}(y_i - x_i)^2 + \lambda|x_i|$. Correspondingly, the proximal operator is *separable*: $[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i = \text{sign}(y_i)\max(|y_i| - \lambda, 0)$. In other words, the ith element of $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is solely determined by y_i. However, this separability property does not hold for a general regularizing sequence. When p is finite, $[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i$ depends not only on y_i but also on other elements of $\boldsymbol{y}$. As one of the core results in this paper, we show that if the empirical distributions of $\boldsymbol{y}$ and $\boldsymbol{\lambda}$ converge as $p \to \infty$, then

$$\frac{1}{p}\left\|\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) - \eta(\boldsymbol{y})\right\|^2 \to 0,$$

where η is a limiting scalar function that is uniquely determined by the limiting empirical measures of $\boldsymbol{y}$ and $\boldsymbol{\lambda}$ (for the exact form, see Proposition 1). This result is illustrated in Fig. 1, where we compare the actual proximal operator $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ and the limiting scalar function $\eta(y)$, for two different $\boldsymbol{\lambda}$-sequences shown in Fig. 1a and Fig. 1c. It can be seen that under a moderate dimension, the proximal operator $\text{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{y})$ can already be very accurately approximated by $\eta(\boldsymbol{y})$.

2) Exact characterization: The asymptotic separability allows us to obtain the exact characterization of SLOPE's performance in the linear asymptotic regime: $n, p \to \infty$ and $n/p \to \delta$, under the assumption that sensing matrix $\boldsymbol{A}$ is generated from i.i.d. Gaussian. On a high level, our main results show that the joint empirical distribution of $\{(\hat{\beta}_i, \beta_i)\}_{i=1}^{p}$ converges to a well-defined limiting measure (the precise description can be found in Theorem 1). Note that the performance metrics of interests such as mean square error (MSE), type-I error, power are all functional of the empirical measure $\{(\hat{\beta}_i, \beta_i)\}_{i=1}^{p}$. Therefore, this makes it possible us to compute the high-dimensional limits of all these quantities. Compared with the probabilistic bounds derived in previous work, our results are asymptotically *exact*.

3) Fundamental limits and optimal regularizagion: The exact asymptotic characterization finally enables us to derive the fundamental limits of SLOPE in both estimation and variable selection tasks: (1) the minimum MSE that can be achieved by SLOPE; and (2) the highest possible power achievable under any given level of Type-I error. Moreover, we show that in both cases, the optimal $\boldsymbol{\lambda}$ sequence can be obtained by solving certain infinite-dimensional convex optimization problems, which have efficient and accurate finite-dimensional approximations. It



(a) MSE v.s. δ (b) $\delta = 0.1$ (c) $\delta = 0.45$ (d) $\delta = 1$

Figure 2: (a): Theoretical predictions (solid lines) v.s. empirical results. Here, β_i are i.i.d. Bernoulli random variables with $\mathbb{P}(\beta_i = 1) = 0.2$ and $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\, 0.04)$. In our simulation, we choose $p = 2048$ and the empirical results are averaged over 20 independent trials. (b)-(d): Empirical distributions of optimal regularizing sequences under 3 different sampling ratios.

is worth mentioning that a caveat of our optimal design is that it requires knowing the limiting empirical measure of $\boldsymbol{\beta}$ (*e.g.*, the sparsity level and the distribution of its nonzero coefficients). For this reason, our results are *oracle optimal*. However, it provides the first step towards optimal sequence designs under more realistic setting, where no or only limited information about $\boldsymbol{\beta}$ is available.

An illustration of asymptotic characterization and optimality results stated above are presented in Fig. 2. We consider three different regularizing sequences: LASSO, BHq sequence proposed in [9] and the optimal sequence given by Proposition 4 below. In Fig. 2a, we plot the empirical MSEs and compare them with the theoretical results. We can see they match well under all settings. Moreover, all the recorded MSE values are lower bounded by the fundamental limits predicted by our theory (red curve in the figure) and they can be achieved by the optimally designed sequences (red circles in the figure). For comparison, we also enclose the curve of minimum mean square error (MMSE) of linear Gaussian model, which was derived in [16], [17]. Finally, to help the readers get a sense of what the optimal regularizing sequences look like, in Fig. 2b-2d we plot their empirical distributions under 3 different sampling ratios δ. Interestingly, we can find they exhibit very different distributions as we change δ.

C. Related Work

1) Exact asymptotic characterization: There has been a growing body of works studying the exact asymptotics in high-dimensional statistical problems under random design assumptions. A partial list of these works includes [10], [18]–[30]. One distinct feature of these type of results is that they provide sharp performance guarantee that does not contain loose constants. From a technical viewpoint, these works are built on powerful tools including statistical physics [31], [32], approximate message passing (AMP) [19], [20], Gaussian width or statistical dimensions [21], [25], leave-one-out analysis [13], [24], Gordon's Gaussian comparison lemma [33], etc. Our main asymptotic characterization is proved based on convex min-max Gaussian theorem (CGMT) [12], [26], [34], which is a tight version of Gordon's comparison lemma in the convex setting. This framework was developed through a series of works [12], [26], [29], [34] and have now been successfully applied in a variety of problems such as binary

detection [14], regularized M-estimation [26], [29], phase retrieval [28], [35] and high-dimensional classification [36]–[38].

2) Optimal M estimation in high dimensions: The optimality part of this work falls within the line of research pursuing the optimal M-estimation in high-dimensional regression. The general form of M-estimator is as follows:

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(y_i, \boldsymbol{a}_i^\top \boldsymbol{b}) + r(\boldsymbol{b}) \tag{4}$$

and the question is what is the optimal statistical performance achievable by (4) and how to optimally design the loss function ℓ and the regularizer r. The exact asymptotic characterizations open up the possibility of obtaining a precise answer to the above question. This line of research is initiated by the papers [39] and [27], where the authors study the fundamental limits of the unregularized M-estimator (*i.e.*, the case when $r = 0$) in the linear model. In particular, a computational feasible recipe is provided in [39] for constructing the optimal loss function ℓ that minimizes the estimation errors. Similar types of results are also recently established for the binary models [40]. When a regularizer is included, the optimal performance of (4) in the linear model is studied in [41] and recently extended to binary model for the special case of quadratic regularization [42], [43]. In the meantime, a series of papers study the optimal ℓ_q-norm regularized least square regression [15], [44], [45]. In some limiting regimes, explicit answers are provided regarding the optimal choice of q. Note that all the aforementioned works consider the separable regularizer: $r(\boldsymbol{b}) = \sum_{i=1} r_i(b_i)$, while SLOPE regularizer considered in this paper is not separable.

Closely related with current work is the paper by Celentano and Montanari [46]. One of their main results is on the optimal estimation performance achievable by quadratic loss regularized by any lower semi-continuous, proper, convex and symmetric [1] function. It is not hard to check that SLOPE norm belongs to this family of functions. In fact, the optimality results presented in their paper and ours share a very similar form. We will elaborate more on this in Sec. IV-A.

3) Three Parallel works: Finally, we mention three parallel works that also study the limiting behavior of SLOPE under the same asymptotic setting.

1) From an algorithmic perspective, [47] consider solving the SLOPE minimization problem (2) using the AMP algorithm. By relating the stationary point of AMP iterations to SLOPE estimator, they also establish the same characterization (as shown in Theorem 1 below). In the proof, they also utilize the asymptotic separability property proved in Proposition 1.
2) The CGMT framework is also applied in [48] to obtain the limiting mean square errors (MSE) of SLOPE, together with a finite-sample concentration bound. The authors quantitatively compares the MSEs of different regularizing sequences in some limiting regimes. In particular, it is shown that in the high SNR regimes, LASSO regularization is optimal. A major difference from our work is that they do not exploit the asymptotic separability of SLOPE and the optimal performance in the general regime is not addressed.

[1]Symmetric means $r(\boldsymbol{b})$ is permutation invariant to coordinates of $\boldsymbol{b}$.

3) In [49], the asymptotic separability properties is further extended to all lsc, proper, convex and symmetric regularizers using an elegant lifting and embedding idea. A finite-sample concentration bound is also given. Using the general asymptotic separability results, the author proves a conjecture in [46]: the MSE lower bound achievable by *non-separable* convex symmetric regularizers will be the same if we are restricted to the *separable* convex regularizers. However, the performance of variable selection is not addressed.

D. Notations

For a vector $\boldsymbol{x} \in \mathbb{R}^p$ and a scalar function $f(\cdot) : \mathbb{R} \to \mathbb{R}$, $f(\boldsymbol{x})$ means $f(\cdot)$ is applied to vector $\boldsymbol{x}$ coordinate-wise. $\|\boldsymbol{x}\|$ denotes the ℓ_2 norm, x_i (or $[\boldsymbol{x}]_i$) denotes the ith coordinate of $\boldsymbol{x}$ and $|x|_{(i)}$ (or $|\boldsymbol{x}|_{(i)}$) denotes the ith largest coordinate of $|\boldsymbol{x}|$. The Euclidean ball in $\mathbb{R}^p$ centered on $\boldsymbol{a}$ with radius $r \geq 0$ is denoted as: $\mathcal{B}_r(\boldsymbol{a}) := \{\boldsymbol{v} : \|\boldsymbol{v}-\boldsymbol{a}\| \leq r\}$ and $\mathcal{B}_r \stackrel{\text{def}}{=} \mathcal{B}_r(\boldsymbol{0})$. Also we define $\mathcal{B}_r^o(\boldsymbol{a}) \stackrel{\text{def}}{=} \{\boldsymbol{v} : \|\boldsymbol{v} - \boldsymbol{a}\| \geq r\}$.

For a probability measure μ, we denote Supp(μ) as its support. For random variables X, Y, we denote $\mu_{X,Y}$ and μ_X, μ_Y as their joint and marginal measures and F_X, F_Y as the corresponding (marginal) cumulative distribution function (CDF). The quantile function of random variable X is denoted as $F_X^{-1}(p)$, where $F_X^{-1}(p) \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : F_X(x) \geq p\}$. Specifically, we use Φ and Φ^{-1} to denote the CDF and quantile function of standard Gaussian. For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we denote $\mu_{\boldsymbol{x},\boldsymbol{y}}$ and $\mu_{\boldsymbol{x}}$, $\mu_{\boldsymbol{y}}$ as their joint and marginal empirical measures and $F_{\boldsymbol{x}}$, $F_{\boldsymbol{y}}$ as the corresponding (marginal) empirical CDF. Also we denote the empirical quantile function of $\boldsymbol{x}$ as $F_{\boldsymbol{x}}^{-1}$.

We denote $\mathcal{P}_q(\mathbb{R}^k)$, for some $q \geq 1$ and $k \in \mathbb{Z}^+$, as the space of all probability measures on $\mathbb{R}^k$ with bounded moments of order q, *i.e.*, for any $\mu \in \mathcal{P}_q(\mathbb{R}^k)$, it holds that $\mathbb{E}_\mu(\|X\|^q) < \infty$. For two measures $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^k)$, their Wasserstein-$q$ distance is defined as:

$$W_q(\mu, \nu) \stackrel{\text{def}}{=} \Big(\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}\|X - Y\|_2^q \Big)^{1/q},$$

where $(X, Y) \sim \pi$ and $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν.

E. Asymptotic Setting

There are four main objects in the description of our model and algorithm: (1) the unknown vector $\boldsymbol{\beta}$; (2) the design matrix $\boldsymbol{A}$; (3) the noise vector $\boldsymbol{w}$; and (4) the regularizing sequence $\boldsymbol{\lambda}$. Since we study the asymptotic limit (with $p \to \infty$), we will consider a sequence of instances $\big\{\boldsymbol{\beta}^{(p)}, \boldsymbol{A}^{(p)}, \boldsymbol{w}^{(p)}, \boldsymbol{\lambda}^{(p)}\big\}_{p\in\mathbb{N}}$ with increasing dimensions p, where $\boldsymbol{\beta}^{(p)}, \boldsymbol{\lambda}^{(p)} \in \mathbb{R}^p$, $\boldsymbol{A}^{(p)} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{w}^{(p)} \in \mathbb{R}^n$. A sequence of vectors $\{\boldsymbol{x}^{(p)}\}_{p\in\mathbb{Z}}$ (or $\{\boldsymbol{x}^{(p)}, \boldsymbol{y}^{(p)}\}_{p\in\mathbb{Z}}$), with p indexing the growing dimensions, is called a *converging sequence*, if its empirical measure $\mu_{\boldsymbol{x}^{(p)}}$ (or $\mu_{\boldsymbol{x}^{(p)},\boldsymbol{y}^{(p)}}$) converges in Wasserstein-2 distance to a probability measure μ_X (or $\mu_{X,Y}$) as $p \to \infty$. For notational brevity, we will omit the superscript "(p)" when it is clear from the context.

F. Paper Outline

The rest of the paper is organized as follows. In Sec. II, we first prove the asymptotic separability of the proximal operator associated with $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$. This property allows us to derive our asymptotic characterization of SLOPE in Sec. III. Based on this analysis, we derive the fundamental limit and present the optimal design of the regularizing

sequence in Sec. IV. Numerical simulations are provided to verify our asymptotic characterizations. They also demonstrate the superiority of our optimal regularization over LASSO and BHq sequence in [7]. In Sec. V, we provide the proof of all our main results. We conclude the paper in Sec. VI and discuss some possible directions for future work.

II. Proximal Problem and Asymptotic Separability

We start by studying the following proximal problem:

$$\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};\tau) \stackrel{\text{def}}{=} \min_{\boldsymbol{x}} \frac{1}{2\tau}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{x}), \tag{5}$$

where $\tau > 0$, $\boldsymbol{y} \in \mathbb{R}^p$ and $J_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{i=1}^p \lambda_i |x|_{(i)}$, with $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$. $\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};\tau)$ in (5) is known as the *Moreau envelope* of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ evaluated at $\boldsymbol{y}$ and τ is the smoothing parameter. The unique minimizer of (5) is the *proximal operator* associated with $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ under parameter τ. From (5), we know the proximal operator of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ is fully determined by $\boldsymbol{y}$ and $\tau\boldsymbol{\lambda}$, so we simply denote it as $\text{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{y})$. It turns out that the asymptotics of the original problem (2) is closely related to (5). Thus, as a preliminary step, we will first analyze its limiting properties.

To state our result, we introduce the following functional optimization problem. For $\mu_Y, \mu_\Lambda \in \mathcal{P}_2(\mathbb{R})$, with $\mathbb{P}(\Lambda \ge 0) = 1$, define

$$\mathcal{M}_{\mu_\Lambda}(\mu_Y;\tau) \stackrel{\text{def}}{=} \min_{g\in\mathcal{I}} \frac{1}{2\tau}\mathbb{E}_{\mu_Y}[Y-g(Y)]^2 + \int_0^1 F_\Lambda^{-1}(u)F_{|g(Y)|}^{-1}(u)du, \tag{6}$$

where

$$\mathcal{I} \stackrel{\text{def}}{=} \{g(y) \mid g(y) \text{ is odd, non-decreasing and 1-Lipschitz}\}. \tag{7}$$

Also we denote $\eta(\cdot;\mu_Y,\mu_{\tau\Lambda})$ as the optimal solution of (6). Comparing (6) with (5), we can intuitively interpret $\mathcal{M}_{\mu_\Lambda}(\mu_Y;\tau)$ and $\eta(\cdot;\mu_Y,\mu_{\tau\Lambda})$ as the functional-form Moreau envelope and proximal operator.

We are now ready to state our main result on the asymptotics of the proximal problem (5).

Proposition 1: Let $\{\boldsymbol{y}\}_{p\in\mathbb{N}}$ and $\{\boldsymbol{\lambda}\}_{p\in\mathbb{N}}$ be two converging sequences, with limiting measures μ_Y and μ_Λ satisfying $\mathbb{P}(\Lambda \ge 0) = 1$. It holds that for any $\tau > 0$,

$$\frac{1}{p}\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};\tau) \to \mathcal{M}_{\mu_\Lambda}(\mu_Y;\tau) \tag{8}$$

and

$$\frac{1}{p}\left\|\text{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{y}) - \eta(\boldsymbol{y};\mu_Y,\mu_{\tau\Lambda})\right\|^2 \to 0, \tag{9}$$

where $\mathcal{M}_{\mu_\Lambda}(\mu_Y;\tau)$ and $\eta(\cdot;\mu_Y,\mu_{\tau\Lambda})$ are the optimal value and the unique (up to a set of measure zero with respect to μ_Y) optimal solution of (6).

The proof of Proposition 1 will be provided in Appendix V-A. We will also see that the limiting characterization of $\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};\tau)$ in (6) and the asymptotic separability of $\text{Prox}_{\tau\boldsymbol{\lambda}}(\cdot)$ in (9) greatly facilitates our asymptotic analysis and the optimal design of $\boldsymbol{\lambda}$, since this allows us to reduce the original high-dimensional problem to an equivalent one-dimensional problem, as in the LASSO case. Indeed, $\eta(\cdot;\mu_Y,\mu_{\tau\Lambda})$ in (9) is exactly the limiting scalar function

$\eta(\cdot)$ shown earlier in Fig. 1. We will still sometimes adopt the lighter notation $\eta(\cdot)$, when doing so causes no confusion.

Note that (6) is involved with an infinite-dimensional optimization, which typically permits no simple analytical solutions. To gain more intuition, before moving on, let us consider two examples where closed-form solutions do exist.

Example 1 (LASSO): The LASSO case corresponds to $\mathbb{P}(\Lambda = \lambda) = 1$. When $\tau = 1$, optimization in (6) then reduces to

$$\min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}\big[\underbrace{|Y| - \lambda}_{:=f(|Y|)} - g(|Y|)\big]^2 + \text{constant}. \tag{10}$$

Note that the function $f(y) = y - \lambda$ in (10) is non-decreasing and 1-Lipschitz on $\mathbb{R}_{\geq 0}$ and $f(0) \leq 0$. It is not hard to show in this case, the optimal solution of (10) equals to

$$\begin{aligned}\eta(y;\mu_Y,\mu_\Lambda) &= \text{sign}(y)\max\big(f(|y|),0\big)\\ &= \text{sign}(y)\max(|y|-\lambda,0),\end{aligned}$$

which is exactly the soft-thresholding function.

Example 2 (BHq [9]): The BHq regularization corresponds to $\Lambda \sim \Phi^{-1}(1-\frac{q}{2}+\frac{q}{2}U)$, where $q \in (0,1]$ and U is uniformly distributed over $[0,1]$. Then we have $F_\Lambda^{-1}(u) = \Phi^{-1}(1-\frac{q}{2}+\frac{q}{2}u)$. Further, we consider $Y \sim \mathcal{N}(0,1)$. It holds that $F_{|Y|}(y) = 2\Phi(y) - 1$ and $F^{-1}_{|g(Y)|}\big(F_{|Y|}(y)\big) = g(y)$, for $y \geq 0$. Therefore,

$$\int_0^1 F_\Lambda^{-1}(u)F^{-1}_{|g(Y)|}(u)du = \int_0^\infty \underbrace{\Phi^{-1}\big(1-q+q\cdot\Phi(y)\big)}_{:=\lambda(y)} g(y)dF_{|Y|}(y), \tag{11}$$

where we apply a change of variable $u = F_{|Y|}(y)$. In this case, (6) becomes

$$\begin{aligned}&\min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}\big[|Y| - g(|Y|)\big]^2 + \mathbb{E}\big[\lambda(|Y|)g(|Y|)\big]\\ =&\min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}\big[|Y| - \lambda(|Y|) - g(|Y|)\big]^2 + \text{constant}.\end{aligned}$$

On the other hand, by direct differentiation of $\lambda(y)$ in (11), we can get $\lambda'(y) = \frac{q\phi(y)}{\phi(\lambda(y))}$, where ϕ is the density function of standard Gaussian. It is not hard to verify $\lambda'(y) \in (0,1]$ when $y \geq 0$. Therefore, $y \mapsto y - \lambda(y)$ is non-decreasing and 1-Lipschitz on $\mathbb{R}_{\geq 0}$. On the other hand, $\lambda(0) = \Phi^{-1}(1-\frac{q}{2}) \geq 0$. Then following the same argument in Example 1, we get $\eta(y;\mu_Y,\mu_\Lambda) = \text{sign}(y)\max\big(|y| - \lambda(|y|),0\big)$.

Remark 1: More generally, we can show when Y has a density supported on $\mathbb{R}$ and $y \mapsto y - F_\Lambda^{-1}(F_{|Y|}(y))$ is non-decreasing and 1-Lipschitz on $\mathbb{R}_{\geq 0}$, then $\eta(y;\mu_Y,\mu_\Lambda) = \text{sign}(y)\max\big(|y| - F_\Lambda^{-1}\big(F_{|Y|}(|y|)\big),0\big)$. In some sense, $F_\Lambda^{-1}\big(F_{|Y|}(|y|)\big)$ can be viewed as the equivalent regularization function. This equivalent regularization is adaptive to y. As a comparison, the regularization is a constant λ in the LASSO case.

III. Asymptotic Characterization of SLOPE

Based on the asymptotic separability properties established in the last section, we are now ready to tackle the original optimization problem (2). We are going to obtain the precise characterizations of SLOPE in both estimation and variable selection problems.

A. Technical Assumptions

Our results are proved under the following assumptions:

(A.1) The number of observations grows in proportion to p: $n^{(p)}/p \to \delta \in (0, \infty)$.

(A.2) The number of nonzero elements in $\boldsymbol{\beta}^{(p)}$ grows in proportion to p: $r_0^{(p)}/p \to \rho \in [0, 1]$.

(A.3) The elements of $\boldsymbol{A}^{(p)}$ are i.i.d. Gaussian distribution: $A_{ij}^{(p)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n})$.

(A.4) $\{\boldsymbol{\beta}^{(p)}\}_{p\in\mathbb{N}}$, $\{\boldsymbol{w}^{(p)}\}_{p\in\mathbb{N}}$ and $\{\boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{N}}$ are converging sequences. The limiting measures are denoted by μ_B, μ_W and μ_Λ, respectively. In addition, $\mathbb{P}(B \neq 0) = \rho$, $\sigma_w^2 = \mathbb{E}[W^2] > 0$ and $\mathbb{P}(\Lambda \neq 0) > 0$ when $\delta \leq 1$, where the probability $\mathbb{P}(\cdot)$ and the expectations $\mathbb{E}[\cdot]$ are all computed with respect to the limiting measures.

B. Asymptotic Performance of Estimation

The main goal of this section is to derive the limiting MSE of SLOPE: $\lim_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$. As in [10], we are going to prove a more general result, which characterizes the joint empirical measure of $(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ through its action on *pseudo-Lipschiz* functions.

Definition 1 (Pseudo-Lipschiz function): A function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is called pseudo-Lipschiz if $|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \leq L(1 + \|\boldsymbol{x}\| + \|\boldsymbol{y}\|)\|\boldsymbol{x} - \boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, where L is a positive constant.

To compute the limiting MSE, we just need to let $\psi(\boldsymbol{x}) = (x_1 - x_2)^2$, which is a pseudo-Lipschiz function by the above definition. The general theorem is as follows, whose proof is deferred to Sec. V-B.

Theorem 1: Assume (A.1) – (A.4) hold. For any pseudo-Lipschiz function ψ, we have

$$\frac{1}{p}\sum_{i=1}^{p} \psi(\hat{\beta}_i, \beta_i) \xrightarrow{\mathbb{P}} \mathbb{E}[\psi(\eta(Y_*; \mu_{Y_*}, \mu_{\tau_*\Lambda}), B)], \tag{12}$$

where $Y_* = B + \sigma_* H$ with $B \sim \mu_B$, $H \sim \mathcal{N}(0, 1)$ independent and η is the limiting scalar function defined in Proposition 1. In the above, the scalar pair (σ_*, τ_*) is the unique solution of the following equations:

$$\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}\big[\big(\eta(B + \sigma H; \mu_{B+\sigma H}, \mu_{\tau\Lambda}) - B\big)^2\big] \tag{13}$$

$$1 = \tau\Big[1 - \frac{1}{\delta}\mathbb{E}\eta'(B + \sigma H; \mu_{B+\sigma H}, \mu_{\tau\Lambda})\Big]. \tag{14}$$

Theorem 1 essentially says that the joint empirical measure of $(\widehat{\boldsymbol{\beta}}^{(p)}, \boldsymbol{\beta}^{(p)})$ converges to the law of $(\eta(Y_*; \mu_{Y_*}, \mu_{\tau_*\Lambda}), B)$. This means that although the original problem (2) is high-dimensional, its asymptotic performance can be succinctly captured by merely two scalars random variables. From (12) and (13), we know the limiting MSE equals to

$$\lim_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \delta(\sigma_*^2 - \sigma_w). \tag{15}$$

Readers familiar with the asymptotic analysis of LASSO will recognize that the forms of (13) and (14) look identical to the results of LASSO obtained in [10], [50]. Indeed, the proof of Theorem 1 directly applies the framework of analyzing LASSO asymptotics using convex Gaussian min-max theorem (CMGT) [26], [29], [50]. In a nutshell, the CGMT framework builds a connection between the asymptotics of the original high-dimensional problem (2) and the optimal solution of the following two-dimensional minimax problem:

$$\min_{\sigma\geq\sigma_w} \max_{\theta\geq 0} \frac{\theta}{2}\Big(\frac{\sigma_w^2}{\sigma} + \sigma\Big) - \frac{\theta^2}{2} + \frac{1}{\delta}\Big[\underbrace{\lim_{p\to\infty} \frac{1}{p}\mathcal{M}_{\boldsymbol{\lambda}}\big(\boldsymbol{\beta} + \sigma\boldsymbol{h}; \frac{\sigma}{\theta}\big)}_{:=\mathcal{F}(\sigma,\theta)} - \frac{\theta\sigma}{2}\Big], \tag{16}$$

where $\boldsymbol{\beta}$ is the true signal vector in (1), $\boldsymbol{h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and $\mathcal{M}_{\boldsymbol{\lambda}}(\cdot;\cdot)$ is the Moreau envelope defined in (5). In fact, equation (13) and (14) corresponds to the first-order optimality condition of (16). Proposition 1 enables us to justify and explicitly compute the limit in (16), as well as the first-order derivatives $\frac{\partial \mathcal{F}(\sigma,\theta)}{\partial \sigma}$ and $\frac{\partial \mathcal{F}(\sigma,\theta)}{\partial \theta}$, which are crucial in obtaining the optimal point of (16).

C. Asymptotic Performance of Variable Selection

Next we study the asymptotic performance of SLOPE, when it is used as a variable selection methodology. Under this setting, the goal is to accurately select all the non-zero coordinates of $\boldsymbol{\beta}$. Based on SLOPE estimate, we select the non-zero coordinates of estimate $\widehat{\boldsymbol{\beta}}$. Ideally, we hope that the selected set includes the non-zero coordinates of $\boldsymbol{\beta}$, while do not contain zero coordinates of $\boldsymbol{\beta}$. The usual performance metrics for this task include Type-I error, power, false discovery rate (FDR), etc. Most of these performance metrics can be expressed as a function of the spasiry level $r_0^{(p)}$ and the following quantities

$$R_0^{(p)} = \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}_{\hat{\beta}_i=0}, \quad V^{(p)} = \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}_{\hat{\beta}_i\neq 0, \beta_i=0}, \tag{17}$$

where $R_0^{(p)}$ and $V^{(p)}$ are the proportions of discoveries and false discoveries. In the following, we will adopt Type-I error and power as our performance metrics, which can be written as

$$\text{Type-I error} = \frac{V^{(p)}}{\max\{1-r_0^{(p)}, 1/p\}}, \quad \text{Power} = \frac{1-R_0^{(p)}-V^{(p)}}{\max\{r_0^{(p)}, 1/p\}}. \tag{18}$$

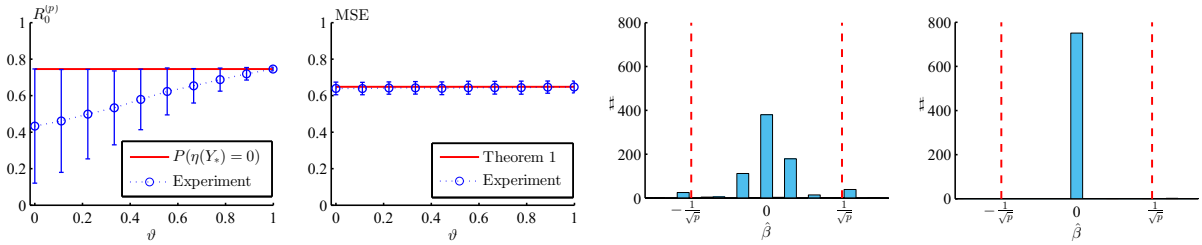
In order to study the asymptotics of these testing statistics, we need to obtain the limits of $R_0^{(p)}$ and $V^{(p)}$ in (17).

Note that the test functions involved in (17) ($\mathbb{I}_{x=0}$ and $\mathbb{I}_{x\neq 0, y=0}$) are discontinuous, so we can not directly apply (12) in Theorem 1 to compute $\lim_{p\to\infty} R_0^{(p)}$ and $\lim_{p\to\infty} V^{(p)}$. Further justifications are needed to obtain companion results for the testing-related statistics in (17). Before delving into techincal descriptions, we first show that counter examples do exist where the quantities in (17) fail to converge, while the assumptions in Theorem 1 are still satisfied. This is different from the LASSO case, where the prediction (12) is shown to be still correct for the above non-smooth indicator functions [9].

Example 3 (A counter example): Consider μ_B being a spike-and-slab distribution: $\mu_B = 0.5\cdot\delta_0 + 0.5\cdot\mathcal{N}(1, 0.5^2)$ and $\{\beta_i\}_{i\in[p]}$ are i.i.d. generated from μ_B. Let (σ_*, τ_*) be the solution of (13)-(14) in the LASSO case, where $\mathbb{P}(\Lambda = 1) = 1$. Then we construct the following class of distribution of Λ, parameterized by $\vartheta \in [0, 1]$:

$$\Lambda_\vartheta = \begin{cases} \vartheta & |Y_*| < \vartheta\tau_*, \\ \frac{|Y_*|}{\tau_*} & \vartheta\tau_* \le |Y_*| < \tau_*, \\ 1 & |Y_*| \ge \tau_*, \end{cases} \tag{19}$$

where $Y_* = B + \sigma_* H$. Here ϑ is a tuning parameter and $\vartheta = 1$ corresponding to the LASSO regularization. In Fig. 3, we plot the empirical $R_0^{(p)}$ and MSEs under different values of ϑ. It can be seen from Fig. 3b that for different values of ϑ, the empirical MSEs all concentrate around the predicted values from Theorem 1, when $\Lambda = 1$. On the contrary, from Fig. 3a we can find when $\vartheta < 1$, $R_0^{(p)}$ does not converge to $\mathbb{P}\big(\eta(Y_*) = 0\big)$, which is the limit



(a) (b) (c) (d)

Figure 3: Counter example when $R_0^{(p)} \not\to \mathbb{P}\big(\eta(Y_*)=0\big)$. In the experiment, $p=1024$. The regularizing sequence $\boldsymbol{\lambda}$ is generated by reordering p i.i.d. samples of Λ_ϑ defined in (19). (a) and (b): empirical $R_0^{(p)}$ and MSE v.s. theoretical predictions based on Theorem 1, under different values of ϑ. The error bars in (a) and (b) are plotted using 1000 independent runs. (c) and (d): the histograms of $\widehat{\boldsymbol{\beta}}$ near zero when $\vartheta=0$ and $\vartheta=1$. When $\vartheta=0$, it can be observed that two clumps of pseudo-zero entries appear within the tiny interval $\big[-\frac{1}{\sqrt{p}},\frac{1}{\sqrt{p}}\big]$, while when $\vartheta=1$, there is no pseudo-zero cluster.

indicated by Theorem 1. Moreover, as ϑ becomes smaller, the SLOPE estimator becomes less conservative and the variances of $R_0^{(p)}$ become increasingly notable. Also we can see $R_0^{(p)}$ does converge to $\mathbb{P}\big(\eta(Y_*)=0\big)$, when $\vartheta=1$.

We will explain the logic behind the construction of Λ_ϑ in Remark 2 below. The counter example above suggests that some additional constraints are needed, so that the testing statistics in (17) have well-defined limits and (12) can be used to compute these limits. It turns out that we just need one more condition to guarantee their convergence.

Proposition 2: Under the same settings as Theorem 1, define $q_0^* \overset{\text{def}}{=} \mathbb{P}(\eta(Y_*)=0)$. If the following condition holds:

(R.1) $q_0^*=0$ or for any $t\in[0,q_0^*)$, $\int_t^{q_0^*} F^{-1}_{|Y_*|}(u)du < \int_t^{q_0^*} F^{-1}_{\tau_*\Lambda}(u)du$,

then we have

$$R_0^{(p)} \xrightarrow{\mathbb{P}} \mathbb{P}\big(\eta(Y_*)=0\big) \text{ and } V^{(p)} \xrightarrow{\mathbb{P}} \mathbb{P}\big(\eta(Y_*)\neq 0, B=0\big), \tag{20}$$

where $R_0^{(p)}$ and $V^{(p)}$ are defined in (17).

The proof of Proposition 2 will be provided in Appendix V-C, along with some explanations for condition (R.1) (see Remark 10).

Remark 2: In fact, Λ_ϑ in (19) is constructed so that condition (R.1) is violated for all $\vartheta<1$. One can easily check that under the setting of Example 3, we have $q_0^*=F_{|Y_*|}(\tau_*)>0$. From (19) we can get $F^{-1}_{\tau_*\Lambda_\vartheta}(u)=F^{-1}_{|Y_*|}(u)$, for all $u\in[F_{|Y_*|}(\vartheta\tau_*),F_{|Y_*|}(\tau_*)]$. Also due to the fact that Y_* is supported on $\mathbb{R}$, we have $F_{|Y_*|}(\vartheta\tau_*)<F_{|Y_*|}(\tau_*)$, when $\vartheta<1$. Therefore, $\int_t^{q_0^*} F^{-1}_{|Y_*|}(u)du = \int_t^{q_0^*} F^{-1}_{\tau_*\Lambda_\vartheta}(u)du$ for any $t\in[F_{|Y_*|}(\vartheta\tau_*),q_0^*)$, where we have used $q_0^*=F_{|Y_*|}(\tau_*)$. This violates condition (R.1). On the other hand, we can also check when $\vartheta=1$, *i.e.*, in the LASSO case, condition (R.1) is satisfied. Indeed, in this case $\Lambda_\vartheta=1$ and $F^{-1}_{\tau_*\Lambda_\vartheta}(u)=F^{-1}_{\tau_*}(u)=\tau_*$ for any $u\in[0,q_0^*]$. Besides, since $q_0^*=F_{|Y_*|}(\tau_*)>0$ and Y_* is supported on $\mathbb{R}$, we get $F^{-1}_{|Y_*|}(u)<\tau_*$ for any $u\in[0,q_0^*)$. Therefore, $\int_t^{q_0^*} F^{-1}_{|Y_*|}(u)du < \int_t^{q_0^*} F^{-1}_{\tau_*\Lambda_\vartheta}(u)du$ for any $t\in[0,q_0^*)$.

Remark 3: In Example 3, a superficial reason for $R_0^{(p)} \not\to \mathbb{P}\big(\eta(Y_*)=0\big)$ when $\vartheta<1$ is that $\boldsymbol{\lambda}$ generated from such Λ will lead to many pseudo-zero entries in $\widehat{\boldsymbol{\beta}}$, *i.e.*, entries that are very closed to 0, but not strictly 0. This is illustrated in Fig. 3c and 3d. In practice, the pseudo-zero effects can be mitigated by employing post-screening to $\widehat{\boldsymbol{\beta}}$. This is done by first specifying a threshold $h>0$ and then setting all the entries in $\widehat{\boldsymbol{\beta}}$ with $|\hat{\beta}_i|<h$ to be zero. However, this creates a new problem of choosing the appropriate h. Our claim is that this problem can be completely avoided by adding an extra constraint on the regularizing sequence. Moreover, as will be clarified in Sec. IV-B, this additional constraint will not harm the diversity of our design choices.

Based on Proposition 2, we can now compute the limiting Type-I error and power of SLOPE.

Corollary 1: When $\mathbb{P}(B=0)\in(0,1)$, we have

$$\lim_{p\to\infty}\text{Type-I error}=\mathbb{P}(|\sigma_* H|\geq y_{\text{th}}^*) \tag{21}$$

and

$$\lim_{p\to\infty}\text{Power}=\mathbb{P}(|B+\sigma_* H|\geq y_{\text{th}}^*\mid B\neq 0), \tag{22}$$

where $y_{\text{th}}^*=\sup_{y\geq 0}\{y\mid \eta(y;\mu_{Y_*},\mu_{\tau_*\Lambda})=0\}$.

The proof of Corollary 1 directly follows from (18), (20) and the assumption that $r_0^{(p)}/p\to\mathbb{P}(B\neq 0)$. Formulas (21) and (22) will be useful in Sec. IV-B, where we analyze the optimal performance of SLOPE for variable selection.

Remark 4: In Corollary 1, we require that $\mathbb{P}(B=0)\in(0,1)$. This means asymptotically, the proportions of zero and non-zero entries of $\boldsymbol{\beta}$ are both non-vanishing. We need this assumption on the distribution of B, because otherwise the limiting formula of Type-I error and power will involve with $\frac{0}{0}$ term, when we apply (18). This is beyond the scope of asymptotic setting considered in this paper.

IV. Fundamental Limits and Optimal Regularization

Armed with the asymptotic characterizations in Theorem 1 and Proposition 2, we are now ready to analyze the optimal performance of SLOPE in both estimation and variable selection setting.

A. Estimation with Minimum MSE

We first turn to the problem of finding the minimum MSE achievable by SLOPE estimator and the corresponding optimal regularization. In the current asymptotic setting, this can be formulated as follows:

$$\inf_{\mu_\Lambda\in\mathcal{P}_\Lambda}\lim_{p\to\infty}\frac{1}{p}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \tag{23}$$

where $\mathcal{P}_\Lambda\overset{\text{def}}{=}\{\mu_\Lambda\mid \mu_\Lambda\in\mathcal{P}_2(\mathbb{R})\text{ and }\mathbb{P}(\Lambda\neq 0)>0,\ \text{when }\delta\leq 1\}$ is the admissible set of μ_Λ, under which the asymptotic characterization in Theorem 1 is valid. By (15), solving (23) is equivalent to solving

$$\inf_{\mu_\Lambda\in\mathcal{P}_\Lambda}\sigma_*. \tag{24}$$

In the current context, σ_* should be understood as a function of μ_Λ, but for notational simplicity, we will drop its dependency on μ_Λ, when doing so causes no confusion.

Note that σ_* is determined by μ_Λ implicitly through the nonlinear fixed point equation (13)-(14), so a direct optimization over μ_Λ as in (24) is not viable. To proceed, a key observation from (13)-(14) is that the influence of μ_Λ is exerted only through the limiting scalar function η. In light of this, (24) can be alternatively solved via the following two-step scheme:

Step 1. Search over all *realizable* η such that there exists $\sigma, \tau > 0$ satisfying

$$\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}[(\eta(B+\sigma H) - B)^2] \tag{25}$$

$$1 = \tau\big(1 - \frac{1}{\delta}\mathbb{E}[\eta'(B+\sigma H)]\big) \tag{26}$$

and find optimal $\eta^\star$ that yields the minimum feasible σ. Denote the corresponding solution of (25)-(26) as $(\sigma^\star, \tau^\star)$.

Step 2. Find corresponding μ_Λ such that $\eta(y; \mu_{B+\sigma^\star H}, \mu_{\tau^\star \Lambda}) = \eta^\star(y)$.

Note that in Step 1, η is treated as an optimization variable that do not depend on other parameters, which greatly simplifies the original formulation (24). However, to implement this scheme, we still need to guarantee two things. First, the *realizable* set of η (as required in Step 1) needs to be decided. Second, for any realizable η, the corresponding Λ can be efficiently computed. These are both addressed in the following result.

Proposition 3: For a probability measure $\mu_Y \in \mathcal{P}_2(\mathbb{R})$, define

$$\mathcal{M}_{\mu_Y} \stackrel{\text{def}}{=} \{\eta(\cdot\,; \mu_Y, \mu_\Lambda) \mid \mu_\Lambda \in \mathcal{P}_2(\mathbb{R})\}, \tag{27}$$

where $\eta(\cdot\,; \mu_Y, \mu_\Lambda)$ is the limiting scalar function in Proposition 1. Then for any $\mu_Y \in \mathcal{P}_2(\mathbb{R})$, we have $\mathcal{M}_{\mu_Y} = \mathcal{I}$. Correspondingly, for any $f(y) \in \mathcal{I}$, we can take $\Lambda \sim |Y| - f(|Y|)$, with $Y \sim \mu_Y$, so that $\eta(y\,; \mu_Y, \mu_\Lambda) = f(y)$.

The proof of Proposition 3 will be presented in Appendix L. It is the key ingredient in proving our optimality results. It shows that, with different choices of μ_Λ, one can reach any non-decreasing and odd function that is Lipschitz continuous with constant 1. Clearly, the soft-thresholding functions associated with LASSO belongs to $\mathcal{M}_{\mu_Y}$, but the set $\mathcal{M}_{\mu_Y}$ is much richer. This is how SLOPE generalizes LASSO: it allows for more degrees of freedom in the regularization.

Based on Proposition 3, we are now ready to show the two-step scheme sketched above indeed yield a computationally feasible procedure to obtain the minimum MSE and the optimal μ_Λ. Before that, we first introduce the following function:

$$\begin{aligned} \mathcal{L}(\sigma) \stackrel{\text{def}}{=} \inf_{f \in \mathcal{I}} \; & \mathbb{E}[f(B+\sigma H) - B]^2 \\ \text{s.t. } & \delta^{-1}\mathbb{E}[f'(B+\sigma H)] \le 1. \end{aligned} \tag{28}$$

We will see for any $\sigma > 0$, problem (28) is convex and there exists a unique optimal solution. Given $\mathcal{L}(\sigma)$, we then introduce the following equation on σ:

$$\mathcal{L}(\sigma) = \delta(\sigma^2 - \sigma_w^2). \tag{29}$$

As is shown in Proposition 4 below, the minimum limiting MSE is closely related with the minimum solution of equation (29).

Proposition 4: Under the same setting as Theorem 1, we have

(a) For any $\sigma > 0$, problem (28) is convex and there exists a unique optimal solution $f_\sigma \in \mathcal{I}$.

(b) $\mathcal{L}(\sigma)$ defined in (28) is continuous on $\mathbb{R}_{>0}$ and equation (29) always has a solution. The minimum solution $\sigma_0 \in \left[\sigma_w,\, \sqrt{\sigma_w^2 + \delta^{-1}\mathbb{E}B^2}\right]$.

(c) Define $Y_0 := B + \sigma_0 H$ and $\tau_0 := \left[1 - \delta^{-1}\mathbb{E}f'_{\sigma_0}(Y_0)\right]^{-1}$. It always holds that

$$\lim_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \geq \delta(\sigma_0^2 - \sigma_w^2). \tag{30}$$

Moreover, if $\delta^{-1}\mathbb{E}\big[f'_{\sigma_0}(Y_0)\big] < 1$, the equality in (30) can be attained, when μ_Λ is the law of

$$\frac{1}{\tau_0}\big[|Y_0| - f_{\sigma_0}(|Y_0|)\big]. \tag{31}$$

The proof of Proposition 4 is deferred to Appendix V-D. To solve the infinite-dimensional optimization problem (28) in practice, we can discretize over $\mathbb{R}$ and obtain a finite-dimensional approximation. Naturally, this finite-dimensional problem is still convex. In our simulation, we use an approximation with 2048 grids.

We have a couple of comments regarding Proposition 4 as follows.

Remark 5 (Interpretation of $\mathcal{L}(\sigma)$): Consider the optimization in (28):

$$\inf_{f\in\mathcal{I}} \mathbb{E}[f(B + \sigma H) - B]^2 \tag{32}$$

and we neglect the constraint $\delta^{-1}\mathbb{E}[f'(B+\sigma H)] \leq 1$ for a moment. From Proposition 1 and Proposition 3 we know minimization in (32) is equivalent to

$$\inf_{\mu_\Lambda \in \mathcal{P}_2(\mathbb{R})} \lim_{p\to\infty} \frac{1}{p}\,\|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{\beta} + \sigma\boldsymbol{h}) - \boldsymbol{\beta}\|^2,$$

where $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and $\mu_\Lambda = \lim_{p\to\infty}\mu_{\boldsymbol{\lambda}}$. In other words, we are estimating $\boldsymbol{\beta}$ from the noisy observation: $\boldsymbol{y} = \boldsymbol{\beta} + \sigma\boldsymbol{h}$ using SLOPE and we want to find the optimal regularization (specified by its limiting distribution μ_Λ) such that the estimation error of $\boldsymbol{\beta}$ is minimized. Then $\mathcal{L}(\sigma)$ can be understood as the minimum MSE we can achieve, if we put an additional constraint on the average slope of limiting scalar function. On the other hand, if at the optimal solution f_σ, the constraint is inactive, *i.e.*, $\delta^{-1}\mathbb{E}[f'_\sigma(B+\sigma H)] < 1$, then $\mathcal{L}(\sigma) = \inf_{f\in\mathcal{I}}\mathbb{E}[f(B+\sigma H) - B]^2$. This can be easily verified as follows. Assume there exists $f_\star \in \mathcal{I}$ such that $\mathbb{E}[f_\star(B+\sigma H) - B]^2 < \mathcal{L}(\sigma)$. Then consider the convex combination $f_t := tf_\star + (1-t)f_\sigma$, for $t \in (0,1)$. Clearly, $f_t \in \mathcal{I}$ and it is not hard to check for small enough t, $\delta^{-1}\mathbb{E}[f_t'(B+\sigma H)] \leq 1$. However, due to the convexity of objective function in (28),

$$\mathbb{E}[f_t(B+\sigma H) - B]^2 \leq t\underbrace{\mathbb{E}[f_\star(B+\sigma H) - B]^2}_{<\mathcal{L}(\sigma)} + (1-t)\underbrace{\mathbb{E}[f_\sigma(B+\sigma H) - B]^2}_{=\mathcal{L}(\sigma)} < \mathcal{L}(\sigma),$$

which leads to a contradiction.

Remark 6 (Tightness of lower bound (30)*):* We require $\delta^{-1}\mathbb{E}f'_{\sigma_0}(B+\sigma_0 H) < 1$ so that the lower bound (30) is tight. The question is whether it is possible that $\delta^{-1}\mathbb{E}f'_{\sigma_0}(B+\sigma_0 H) = 1$. This will not happen when $\delta > 1$, since $f'_{\sigma_0} \leq 1$. When $\delta \leq 1$, we do not have a rigorous proof yet. Numerically, this never happens either. Here we provide an intuitive argument. Suppose for certain configurations of $(\delta, \rho, \sigma_w, \mu_B)$, we do have $\delta^{-1}\mathbb{E}f'_{\sigma_0}(B+\sigma_0 H) = 1$. Under this scenario, let us consider the following approximation of (28) and (29), parameterized by $\varepsilon > 0$:

$$\begin{aligned} \mathcal{L}_\varepsilon(\sigma) \overset{\text{def}}{=} &\inf_{f\in\mathcal{I}} \mathbb{E}[f(B+\sigma H) - B]^2 \\ &\text{s.t. } \delta^{-1}\mathbb{E}[f'(B+\sigma H)] \leq 1 - \varepsilon. \end{aligned} \tag{33}$$

and

$$\mathcal{L}_\varepsilon(\sigma) = \delta(\sigma^2 - \sigma_w^2). \tag{34}$$

Denote $\sigma_{0,\varepsilon}$ as the minimum solution of equation (34) and f_ε as the optimal solution of (33), when $\sigma = \sigma_{0,\varepsilon}$. If we take μ_Λ to be the law of

$$\frac{1}{\tau_{0,\varepsilon}}[|Y_{0,\varepsilon}| - f_\varepsilon(|Y_{0,\varepsilon}|)], \tag{35}$$

where $Y_{0,\varepsilon} = B + \sigma_{0,\varepsilon} H$ and $\tau_{0,\varepsilon} = \left[1 - \delta^{-1}\mathbb{E} f_\varepsilon'(Y_{0,\varepsilon})\right]^{-1} < \infty$. Then similar as Proposition 4, it is not hard to show $\lim_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}}_\varepsilon - \boldsymbol{\beta}\|_2^2 = \delta(\sigma_{0,\varepsilon}^2 - \sigma_w^2)$, where $\widehat{\boldsymbol{\beta}}_\varepsilon$ denotes the corresponding estimator. Intuitively, we could also expect $\sigma_{0,\varepsilon} \to \sigma_0$ and $\tau_{0,\varepsilon} \to \infty$, as $\varepsilon \to 0$. This implies the MSE can be made arbitrarily close to the lower bound (30) using a sequence $\{\mu_{\Lambda,\varepsilon}\}_{\varepsilon>0}$ which converges to the probability mass at 0 as $\varepsilon \to 0$. Recall that we have assumed $\sigma_w > 0$, so this means the optimal regularization in a noisy overparameterized linear model should be vanishingly small, which is not likely the case.

Remark 7 (Comparison with [46], [49]): In [46], [49], the authors also analyze the problem of optimal estimation in the linear model (1) with i.i.d. Gaussian design. For the convenience of comparison, here we rephrase their results in our notations. The optimality they consider is with respect to the following class of estimator:

$$\left\{\widehat{\boldsymbol{\beta}} \;\middle|\; \widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{b}\|_2^2 + r_p(\boldsymbol{b}), r_p \in \mathcal{C}_p\right\} \tag{36}$$

where

$$\mathcal{C}_p \overset{\text{def}}{=} \{r_p : \mathbb{R}^p \to \bar{\mathbb{R}} \mid r_p \text{ is lsc, proper, convex and symmetric}\}.$$

The optimal estimation within the class of $\mathcal{C}_p$ is formulated as:

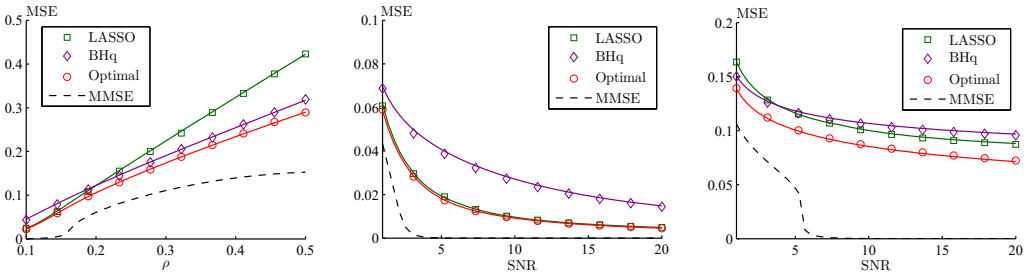
$$\text{MSE}_{\text{ cvx}} \overset{\text{def}}{=} \inf_{\forall p, r_p \in \mathcal{C}_p \cap \mathcal{W}_p} \liminf_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2, \tag{37}$$

where $\mathcal{W}_p$ is some set that ensures $\widehat{\boldsymbol{\beta}}$ is unique [2]. One of their main results states that under certain conditions, the minimum achievable limiting MSE defined in (37) satisfies: $\text{MSE}_{\text{ cvx}} \geq \delta(\sigma_{\text{cvx}}^2 - \sigma_w^2)$, where

$$\sigma_{\text{cvx}}^2 = \sup\{\sigma^2 \mid \delta(\sigma^2 - \sigma_w^2) < \inf_{f\in\mathcal{J}} \mathbb{E}[f(B + \sigma H) - B]^2\}, \tag{38}$$

with $\mathcal{J} \overset{\text{def}}{=} \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is non-decreasing and 1-Lipschitz continuous}\}$. Comparing their results with ours, we can find the lower bounds in both settings follow the same type of characterization. Specifically, lying at the heart of this characterization is an optimization problem: $\inf_{f\in\mathcal{F}} \mathbb{E}[f(B + \sigma H) - B]^2$, which aims at finding the optimal estimator f of B under the noisy observation $B + \sigma H$. The only difference is on the feasible set $\mathcal{F}$: $\mathcal{F} = \mathcal{J}$ in (38), while $\mathcal{F} = \mathcal{I} \subset \mathcal{J}$ in (28). This agreement is not a coincidence, but related with the fact that the proximal operator of all functions in $\mathcal{C}_p$ is asymptotically separable as proved in [49]. In fact, f corresponds to the limiting proximal operator of the regularizer r_p. In our settings, r_p is chosen from the set of all possible sorted ℓ_1 norms (denoted by $\mathcal{S}_p$), while in their settings, it is chosen from the set $\mathcal{C}_p$. Correspondingly, $\mathcal{I}$ is the set of all limiting proximal operators associated with $\mathcal{S}_p$ and $\mathcal{J}$ is the one associated with $\mathcal{C}_p$. It is not hard to check $\mathcal{S}_p \subset \mathcal{C}_p$ and consequently, we have $\mathcal{I} \subset \mathcal{J}$.

[2] In fact, $\mathcal{W}_p$ corresponds to the tightness condition $\delta^{-1}\mathbb{E}\left[f_{\sigma_0}'(B + \sigma_0 H)\right] < 1$ in Proposition 4 (c).



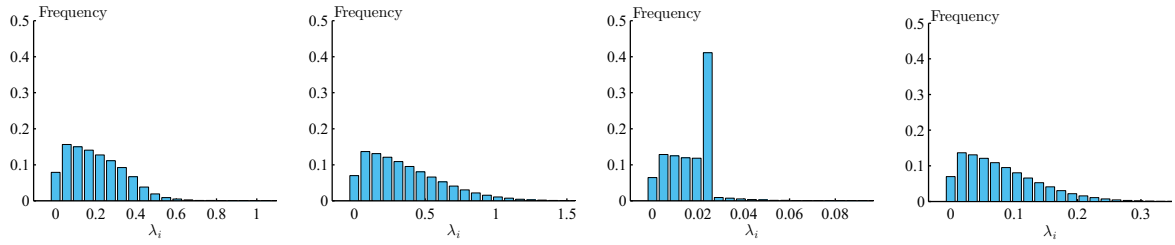
(a) SNR $= 4$ (b) $\rho = 0.1$ (c) $\rho = 0.2$

Figure 4: Comparison of MSEs obtained by three regularization sequences: LASSO, BHq and the oracle optimal design. Here, β_i are i.i.d. Bernoulli random variables, with $\mathbb{P}(\beta_i = 1) = \rho$ and $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\, \sigma_w^2)$, with $\sigma_w^2 = \rho/\text{SNR}$ and the BHq sequences are generated by reordering i.i.d. samples of: $\sigma_w \Phi^{-1}(1 - \frac{q}{2} + \frac{q}{2}U)$, where $q \in (0, 1]$ and U follows the uniform distribution over $[0, 1]$. In our simulation, we fix $p = 2048$, $\delta = 0.5$ and the empirical results are averaged over 20 independent trials. The dash curves correspond to the information-theoretic limit obtained in [16], [17].

In Fig. 4, we compare the MSEs achieved by different regularizing sequences (LASSO, BHq and oracle optimal design), at different SNR and sparsity levels. Since we are concerned with oracle optimality, for fair comparison, we search through the parameters of the BHq and LASSO sequences (in particular, q for BHq and λ for LASSO) and report the minimum MSEs that can be achieved. The solid curves correspond to the theoretical MSEs predicted by Theorem 1 and Proposition 4. Note that the empirical MSEs match well with theoretical predictions [3]. It is also observed that under each setting, the MSEs of different regularizing sequences are all above the lower bound obtained in (30) (red curve in the figure). Also we can see this lower bound can be attained when the limiting empirical distribution of $\boldsymbol{\lambda}$ follows prescribed optimal distribution (31). We also have the following findings:

1) As can be seen from Fig. 4a and Fig. 4b, when ρ is small, LASSO performs well and the corresponding MSEs almost match the theoretical lower bound, across different values of SNR. However, its performance degrades faster than the other two sequences, as ρ grows. This is because LASSO's penalization is not adaptive to the underlying sparsity levels and it incurs higher bias under larger ρ [7].
2) From Fig. 4b and Fig. 4c, we can find that at low SNR regimes, the BHq sequence can lead to comparable performance as the optimal design. However, at higher SNR regimes, the optimal design notably outperforms the BHq sequence. To explain this phenomenon, we plot in Fig. 5 the empirical distributions of the $\boldsymbol{\lambda}$-sequences associated with the optimal design and the BHq design, respectively. It turns out that, in the low SNR case, the optimal design and BHq have similar distributions, while at higher SNRs, the distribution of the optimal design is close to a mixture of a delta mass and uniform distribution.

[3]Here, the MSEs of LASSO and BHq are obtained by optimizing over the parameters λ and q, so strictly speaking, the theoretical curves are valid only if a stronger uniform convergence result holds. The uniform convergence for LASSO case is proved in [50], [51] and we conjecture that it also holds true for BHq sequences.



(a) Optimal, SNR=1 (b) BHq, SNR=1 (c) Optimal, SNR=20 (d) BHq, SNR=20

Figure 5: Comparison of empirical distributions of two regularizing sequences "BHq" and "Optimal" in Fig. 4c.

B. Variable Selection with Maximum Power

Next we consider using SLOPE for variable selection. Our goal is to find the optimal regularizing sequence to achieve the highest possible power, under a given level of type-I error α, formulated as:

$$\begin{aligned} \mathcal{P}(\alpha) \stackrel{\text{def}}{=} \sup_{\Lambda \in \widetilde{\mathcal{P}}_\Lambda} \lim_{p\to\infty} \text{Power} \\ \text{s.t.} \lim_{p\to\infty} \text{Type-I error} \le \alpha, \end{aligned} \tag{39}$$

where $\widetilde{\mathcal{P}}_\Lambda \stackrel{\text{def}}{=} \{\Lambda \in \mathcal{P}_\Lambda : \text{(R.1) is satisfied}\}$ is the admissible set of μ_Λ, with which the limits in (39) exist. In light of (21) and (22), if $\mathbb{P}(B=0) \in (0,1)$, optimization problem (39) is equivalent to:

$$\begin{aligned} \mathcal{P}(\alpha) = \sup_{\Lambda \in \widetilde{\mathcal{P}}_\Lambda} \mathbb{P}(|B + \sigma_* H| \ge y^*_{\text{th}} \mid B \neq 0) \\ \text{s.t. } \mathbb{P}(|\sigma_* H| \ge y^*_{\text{th}}) \le \alpha, \end{aligned} \tag{40}$$

where $y^*_{\text{th}} = \sup_{y\ge 0}\{y \mid \eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) = 0\}$. Comparing the admissible set $\widetilde{\mathcal{P}}_\Lambda$ with $\mathcal{P}_\Lambda$ in (24), it can be seen the only difference is that here we need an additional condition (R.1) to ensure the limits of Type-I error and Power both exist (see Proposition 2).

To state our results, we first introduce the following function, which is the counterpart of (28).

$$\begin{aligned} \mathcal{L}_\alpha(\sigma) \stackrel{\text{def}}{=} \inf_{f \in \mathcal{I} \cap \mathcal{F}_{\alpha,\sigma}} \mathbb{E}[f(B + \sigma H) - B]^2 \\ \text{s.t. } \delta^{-1}\mathbb{E}[f'(B + \sigma H)] \le 1 \end{aligned} \tag{41}$$

where $\alpha \in [0,1]$ is a prescribed Type-I error level and $\mathcal{F}_{\alpha,\sigma} \stackrel{\text{def}}{=} \big\{f(y) : f(y) = 0 \text{ for } |y| \le \Phi^{-1}(1-\frac{\alpha}{2})\sigma\big\}$. Similar as Proposition 4, we will see that the maximum power achievable by SLOPE under Type-I error level α is related with the following equation:

$$\mathcal{L}_\alpha(\sigma) = \delta(\sigma^2 - \sigma_w^2), \tag{42}$$

where $\mathcal{L}_\alpha(\sigma)$ is the function defined in (41).

We are now ready to state our main optimality results for variable selection.

Proposition 5: Under the same setting as Proposition 2, assume $\mathbb{P}(B=0) \in (0,1)$. Then we have

(a) For any $\alpha \in [0,1]$ and $\sigma > 0$, problem (41) is convex and there exists a unique optimal solution $f_{\alpha,\sigma} \in \mathcal{I}$.

(b) For any $\alpha \in [0,1]$, $\mathcal{L}_\alpha(\sigma)$ is continuous on $\mathbb{R}_{>0}$ and equation (42) always has a solution. The minimum solution $\sigma_{0,\alpha} \in \big[\sigma_w,\ \sqrt{\sigma_w^2 + \delta^{-1}\mathbb{E}B^2}\big]$.

(c) Let $Y_{0,\alpha} := B + \sigma_{0,\alpha} H$ and $f_\alpha := f_{\alpha,\sigma_{0,\alpha}}$. If $\lim_{p\to\infty}$ Type-I error $\le \alpha$, then

$$\lim_{p\to\infty} \text{Power} \le \mathbb{P}\big(|Y_{0,\alpha}| \ge \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_{0,\alpha} \mid B \neq 0\big). \tag{43}$$

Moreover, if $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$ and $y_{0,\alpha} = \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha}$, the upper bound in (43) can be attained by $\mu_\Lambda = \mu_{\text{opt},\alpha}$, with $\mu_{\text{opt},\alpha}$ being the law of

$$\frac{1}{\tau_{0,\alpha}} \max\big\{y_{0,\alpha}, |Y_{0,\alpha}| - f_\alpha(|Y_{0,\alpha}|)\big\}. \tag{44}$$

Here, $y_{0,\alpha} = \sup_{y\ge 0}\{y \mid f_\alpha(y) = 0\}$ and $\tau_{0,\alpha} = \big[1 - \delta^{-1}\mathbb{E}f'_\alpha(Y_{0,\alpha})\big]^{-1}$.

The proof of Proposition 5, which is similar to that of Proposition 4, will be given in Sec. V-E. A key step is to show the realizable set of η in the variable selection setting is still equal to $\mathcal{I}$ (see Lemma 20 in Appendix N), although the admissible set of μ_Λ is replaced by $\widetilde{\mathcal{P}}_\Lambda$, which is a subset of $\mathcal{P}_\Lambda$ in the estimation setting.

Remark 8: Comparing the results in Proposition 4 and Proposition 5, we can find that although at the beginning, we are dealing with two different problems (the objective of the first one is minimizing the MSE, while the other is on maximizing the power under a given Type-I error), we end up with two procedures of very similar natures. Both problems can finally be converted into a formulation involving finding the optimal estimation of $\boldsymbol{\beta}$ that can be achieved by SLOPE under the observation $\boldsymbol{y} = \boldsymbol{\beta} + \sigma\boldsymbol{h}$, with $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The only difference is that in the second problem, we need to enforce an additional restriction on the regularization sequence $\boldsymbol{\lambda}$ to ensure the Type-I error is below certain threshold α.

Remark 9 (Tightness of upper bound (43)*):* The tightness of the upper bound for power relies on the conditions: $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$ and $y_{0,\alpha} = \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha}$. Numerically they hold under all the settings considered. We conjecture that within our assumptions, this condition always hold and the upper bound (43) is tight.

In Fig.6, we compare the variable selection performance achieved by the optimal regularization with that of LASSO and BHq sequences. We show both theoretical ROC curves and the empirical power under given Type-I error levels. Here each empirical (Type-I error, power) pair is generated by first fixing all the parameters (including the tuning parameters such as λ and q) and then averaging over 20 independent trials. It can seen that the empirical results match well with the theoretical predictions (solid curves in the figures) and the optimal design of regularization dominates the other two regularizing sequences. We also have the following observations:

1) In all cases, the theoretical upper bounds on power (43) can be achieved by choosing μ_Λ to be the law of (44).
2) The performance of LASSO is closed to the fundamental limit at low sparsity and high SNR regimes, while its performance is significantly degraded as sparsity grows higher or SNR grows lower. In particular, we can find in such cases, the maximum power achievable by LASSO is less than 1. This phenomenon is also discussed in [2], [7], [11] and it is inherently connected with the so-called "noise-sensitivity " phase transition [52]. In comparison, the optimal and BHq sequences can both reach power 1, after Type-I errors are above certain thresholds.

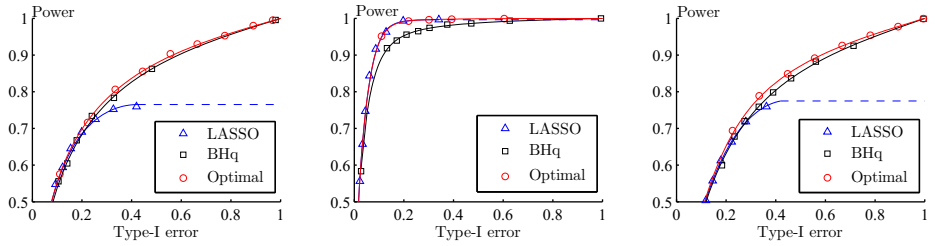


Figure 6: Theoretical predictions v.s. empirical results of testing performance using LASSO, BHq and oracle optimal sequences. Here, β_i are i.i.d. Bernoulli random variables, with $\mathbb{P}(\beta_i = 1) = \rho$ and $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$, with $\sigma_w^2 = \rho/\text{SNR}$. The empirical results are generated under $p = 2048$ and $\delta = 0.5$ and we choose different ρ and SNR values: (a) $\rho = 0.1$, SNR $= 0.6$, (b) $\rho = 0.1$, SNR $= 4$, (c) $\rho = 0.2$, SNR $= 4$. Dash curves correspond to the observed upper bound of power achieved by LASSO.

3) Complementary to LASSO, the performance of BHq sequences is closed to the theoretical upper bounds at low SNRs or large sparsity levels, while it deviates from the upper bounds in other scenarios.

V. Proof of Main Results

A. Asymptotic Separability

In this section, we are going to prove Proposition 1.

From (5) we have the following scaling property: $\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}; \tau) = \frac{1}{\tau}\mathcal{M}_{\tau\boldsymbol{\lambda}}(\boldsymbol{y}; 1)$. On the other hand, for any $\tau > 0$, if $\boldsymbol{\lambda}$ is a converging sequence with limiting measure μ_Λ, it is not hard to show $\tau\boldsymbol{\lambda}$ is also a converging sequence, with limiting measure $\mu_{\tau\Lambda}$. Thus, to study the asymptotic limit of (5) under $(\boldsymbol{y}, \boldsymbol{\lambda}, \tau)$, it suffices to consider $(\boldsymbol{y}, \tau\boldsymbol{\lambda}, 1)$. As a result, without loss of generality, we will assume $\tau = 1$ in the rest of our proof.

1) Some preliminary facts about SLOPE: The asymptotic separability stems from the following unique properties of the SLOPE proximal minimization problem (5), which are proved in [9, Sec. 2].

Fact 1: For any $\boldsymbol{\lambda}, \boldsymbol{y} \in \mathbb{R}^p$, with $\lambda_i \geq 0$, for all $i \in [p]$, it holds that

(i) *(Sign consistency)* For any $i \in [p]$, $[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i$ has the same sign as y_i. Moreover,

$$[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i = \text{sign}(y_i)[\text{Prox}_{\boldsymbol{\lambda}}(|\boldsymbol{y}|)]_i.$$

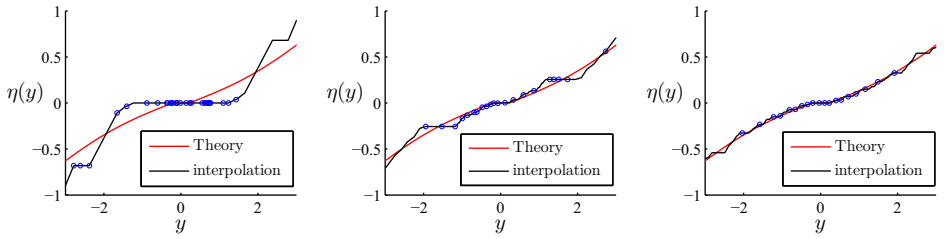
(ii) *(Permutation-invariance)* For any permutation matrix Π, $\Pi\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) = \text{Prox}_{\boldsymbol{\lambda}}(\Pi\boldsymbol{y})$.

(iii) *(Monotonicity and Lipschitz continuity)* For any $i, j \in [p]$, if $y_i \leq y_j$, then $0 \leq [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_j - [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i \leq y_j - y_i$ and for any y_i, $[\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i \leq |y_i|$.

An immediate yet important implication of Fact 1 is the following lemma:

Lemma 1: For any $\boldsymbol{\lambda}, \boldsymbol{y} \in \mathbb{R}^p$, with $\lambda_i \geq 0$, there always exists an odd, non-decreasing and 1-Lipschitz function g_p such that for all $i \in [p]$, $g_p(y_i) = [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i$.

The proof of Lemma 1 is given in Appendix A. By Lemma 1 we know $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is actually the restriction of a function $g_p \in \mathcal{I}$ onto the support of $\mu_{\boldsymbol{y}}$. Moreover, from the permutation invariance property (Fact 1 (ii)), such g_p



(a) $p = 32$ (b) $p = 128$ (c) $p = 1024$

Figure 7: Comparison between $\eta(y)$ (red curve), linear interpolation (black curve) and $\{(y_i, [\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})]_i)\}_{i\in[p]}$ (blue dots), under three different values of p. Here $y_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 2)$ and λ_i are i.i.d. samples from BHq distribution [9].

is only determined by the empirical measure $\mu_{\boldsymbol{\lambda}}$ and $\mu_{\boldsymbol{y}}$. We could expect if $\mu_{\boldsymbol{\lambda}}$ and $\mu_{\boldsymbol{y}}$ both converge to some limiting distributions, g_p will also converge to certain limiting scalar function. This is exactly the essential meaning of asymptotic separability.

Before proceeding, let us take a look at a numerical justification shown in Fig. 7. Here we choose g_p to be the linear interpolation of the following set of points: $\{(|y_i|, |\hat{y}_i|), (-|y_i|, -|\hat{y}_i|)\}_{i=1}^p \cup (0, 0)$, where $\hat{\boldsymbol{y}} := \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$. It is easy to check (as is shown in the proof of Lemma 1) such linear interpolation is a qualified candidate for g_p in Lemma 1. We compare it with the limiting scalar function $\eta(y)$ predicted by Proposition 1. It is clear that as p becomes larger, $g_p(y)$ gets increasingly close to $\eta(y)$.

2) An equivalent form of (5): Based on Lemma 1, we then go on to show the equivalence between (5) and the following problem:

$$\min_{g\in\mathcal{I}} \underbrace{\frac{1}{2}\mathbb{E}_{\mu_{\boldsymbol{y}}}[Y - g(Y)]^2 + \int_0^1 F_{\boldsymbol{\lambda}}^{-1}(u) F_{|g(\boldsymbol{y})|}^{-1}(u) du}_{\overset{\text{def}}{=} L_p(g)}. \tag{45}$$

This is formalized in the following lemma, whose proof is given in Appendix B.

Lemma 2: Denote $\mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y})$ as the optimal value of (45). Then it holds that $\frac{\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};1)}{p} = \mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y})$. Besides, any optimal solution $g_p^*(y)$ of (45) satisfies: $g_p^*(\boldsymbol{y}) = \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$.

Comparing (6) and (45), it could be now understood that the optimization in (6) is the limit of (45), as $\mu_{\boldsymbol{\lambda}} \to \mu_\Lambda$ and $\mu_{\boldsymbol{y}} \to \mu_Y$. Therefore, from Lemma 2, we could expect $\frac{1}{p}\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}; 1) = \mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y}) \to \mathcal{M}_{\mu_\Lambda}(\mu_Y, 1)$. On the other hand, $g_p^*(\cdot)$, which is the optimal solution of (45) should also converge to the optimal solution of (6): $\eta(\cdot; \mu_Y, \mu_\Lambda)$. Thus for any $\boldsymbol{\lambda}$, $\boldsymbol{y}$ satisfying $\mu_{\boldsymbol{\lambda}} \approx \mu_\Lambda$ and $\mu_{\boldsymbol{y}} \approx \mu_Y$, we would have $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) = g_p^*(\boldsymbol{y}) \approx \eta(\boldsymbol{y}; \mu_Y, \mu_\Lambda)$, *i.e.*, asymptotic separability holds. The final step of the proof is to make the above intuition accurate and rigorous.

3) Taking the limit of (45): Recall that we have assumed $\tau = 1$. For notational simplicity, denote $\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}) := \mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}; 1)$ and $\mathcal{M}_{\mu_\Lambda}(\mu_Y) := \mathcal{M}_{\mu_\Lambda}(\mu_Y, 1)$. Define $L(g)$ as the objective function of (6), *i.e.*,

$$L(g) \overset{\text{def}}{=} \frac{1}{2}\mathbb{E}_{\mu_Y}[Y - g(Y)]^2 + \int_0^1 F_\Lambda^{-1}(u) F_{|g(Y)|}^{-1}(u) du. \tag{46}$$

and g^* as the corresponding optimal solution. By Lemma 7 in Appendix C, we have $\sup_{g\in\mathcal{I}}|L(g)-L_p(g)|\to 0$, where $L_p(g)$ is the objective function of (45). Therefore, (8) immediately follows, since

$$\begin{aligned}|\tfrac{1}{p}\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};1)-\mathcal{M}_{\mu_\Lambda}(\mu_Y,1)| &= |\sup_{g\in\mathcal{I}} L(g)-\sup_{g\in\mathcal{I}} L_p(g)|\\ &\le \sup_{g\in\mathcal{I}}|L(g)-L_p(g)|.\end{aligned}$$

On the other hand,

$$\begin{aligned}L_p(g^*)-L_p(g_p^*) &= L_p(g^*)-L(g^*)+L(g_p^*)-L_p(g_p^*)\\ &\quad + L(g^*)-L(g_p^*)\\ &\le |L_p(g^*)-L(g^*)|+|L(g_p^*)-L_p(g_p^*)|,\end{aligned} \tag{47}$$

where in the last step we use the optimality of g^*. By the strong convexity of (45), we have

$$L_p(g^*)-L_p(g_p^*)\ge \frac{1}{2}\mathbb{E}_{\mu_{\boldsymbol{y}}}|g_p^*(Y)-g^*(Y)|^2. \tag{48}$$

Combining (47) and (48) gives

$$\mathbb{E}_{\mu_{\boldsymbol{y}}}|g_p^*(Y)-g^*(Y)|^2\le 2\sup_{g\in\mathcal{I}}|L(g)-L_p(g)|.$$

By Lemma 7 again, we have $\mathbb{E}_{\mu_{\boldsymbol{y}}}|g_p^*(Y)-g^*(Y)|^2\to 0$, as $p\to\infty$. This is exactly (9), since $g_p^*(\boldsymbol{y})=\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ by Lemma 2 and $g^*(y)=\eta(y;\mu_Y,\mu_{\tau\Lambda})$. Finally, the uniqueness of $g^*(\cdot)$ (up to a set of measure 0 with respect to μ_Y) is proved in Lemma 8. This completes our proof.

B. Asymptotic Estimation Performance

1) Convex Gaussian Min-max Theorem: Our proof hinges on the Convex Gaussian Min-max Theorem (CGMT). For completeness, we briefly summarize the key idea here. The CGMT studies a minimax optimization problem (PO) of the form:

$$\Phi(\boldsymbol{G})=\min_{\boldsymbol{v}\in\mathcal{S}_{\boldsymbol{v}}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{v}+\psi(\boldsymbol{v},\boldsymbol{u}), \tag{49}$$

where $\mathcal{S}_{\boldsymbol{v}}\subset\mathbb{R}^p$, $\mathcal{S}_{\boldsymbol{u}}\subset\mathbb{R}^n$ are two compact sets, $\psi:\mathcal{S}_{\boldsymbol{v}}\times\mathcal{S}_{\boldsymbol{u}}\to\mathbb{R}$ is a continuous convex-concave function w.r.t. $(\boldsymbol{v},\boldsymbol{u})$ and $G_{ij}\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$. Problem (49) can be associated with the following auxiliary optimization (AO) problem:

$$\phi(\boldsymbol{g},\boldsymbol{h})=\min_{\boldsymbol{v}\in\mathcal{S}_{\boldsymbol{v}}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\|\boldsymbol{v}\|_2\boldsymbol{g}^\top\boldsymbol{u}+\|\boldsymbol{u}\|_2\boldsymbol{h}^\top\boldsymbol{v}+\psi(\boldsymbol{v},\boldsymbol{u}), \tag{50}$$

where $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)$ and $\boldsymbol{h}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_p)$. Roughly speaking, CGMT shows that $\frac{1}{p}\Phi(\boldsymbol{G})\approx\frac{1}{p}\phi(\boldsymbol{g},\boldsymbol{h})$ and the optimal solutions of (49) and (50) have approximately the same empirical distributions in the large p limit. Usually, (AO) is easier to analyze, so it provides a convenient handle for analyzing (PO). For a detailed descriptions, readers can refer to [26, Theorem 3].

2) Proof of Theorem 1: The first step is to recast (2) into the minimax form as in (49). Letting $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{\beta}$, (2) can be equivalently written as:

$$\begin{aligned}&\min_{\boldsymbol{v}} \underbrace{\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\boldsymbol{v}-\boldsymbol{w}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})\Big]}_{:=C(\boldsymbol{v})} \\ =&\min_{\boldsymbol{v}}\max_{\boldsymbol{u}} \frac{1}{n}\left[\frac{\boldsymbol{u}^\top}{\sqrt{n}}\left(\sqrt{n}\boldsymbol{A}\right)\boldsymbol{v} - \boldsymbol{u}^\top\boldsymbol{w} - \frac{\|\boldsymbol{u}\|^2}{2} + J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})\right].\end{aligned} \tag{51}$$

Denote $\widehat{\boldsymbol{v}} \overset{\text{def}}{=} \underset{\boldsymbol{v}}{\text{argmin}}\, C(\boldsymbol{v})$ and correspondingly, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{v}} + \boldsymbol{\beta}$. Now (51) has the same form as (49) and the corresponding (AO) is:

$$\begin{aligned}&\min_{\boldsymbol{v}}\max_{\boldsymbol{u}} \frac{1}{n}\left[-\frac{\|\boldsymbol{u}\|}{\sqrt{n}}\boldsymbol{h}^\top\boldsymbol{v} - \frac{\|\boldsymbol{v}\|}{\sqrt{n}}\boldsymbol{g}^\top\boldsymbol{u} - \boldsymbol{u}^\top\boldsymbol{w} - \frac{\|\boldsymbol{u}\|^2}{2} + J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})\right] \\ =&\min_{\boldsymbol{v}}\max_{\theta\geq 0}\theta\left(\left\|\frac{\|\boldsymbol{v}\|}{n}\boldsymbol{g} + \frac{\boldsymbol{w}}{\sqrt{n}}\right\| - \frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\right) - \frac{1}{2}\theta^2 + \frac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n} \\ =&\min_{\boldsymbol{v}} \underbrace{\tfrac{1}{2}\Big(\sqrt{\tfrac{\|\boldsymbol{v}\|^2}{n}\tfrac{\|\boldsymbol{g}\|^2}{n} + \tfrac{\|\boldsymbol{w}\|^2}{n} + 2\tfrac{\|\boldsymbol{v}\|}{\sqrt{n}}\tfrac{\boldsymbol{g}^\top\boldsymbol{w}}{n}} - \tfrac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\Big)_+^2 + \tfrac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n}}_{:=L(\boldsymbol{v})},\end{aligned} \tag{52}$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Let $D \subset \mathbb{R}^p$ be any closed set. Then by CGMT we can show for any $t \in \mathbb{R}$,

$$\mathbb{P}(\min_{\boldsymbol{v}\in D} C(\boldsymbol{v}) \leq t) \leq 2\mathbb{P}(\min_{\boldsymbol{v}\in D} L(\boldsymbol{v}) \leq t) \tag{53}$$

and if D is also convex,

$$\mathbb{P}(\min_{\boldsymbol{v}\in D} C(\boldsymbol{v}) \geq t) \leq 2\mathbb{P}(\min_{\boldsymbol{v}\in D} L(\boldsymbol{v}) \geq t). \tag{54}$$

The proof of (53) and (54) is the same as [50, Corollary 5.1] and is omitted here. We are going to apply (53) and (54) to prove (12). We will follow the proof steps in [50].

First define the following minimax problem:

$$\Psi_* \overset{\text{def}}{=} \min_{\sigma\geq\sigma_w}\max_{\theta\geq 0} \underbrace{\frac{\theta}{2}\Big(\frac{\sigma_w^2}{\sigma} + \sigma\Big) - \frac{\theta^2}{2} + \frac{1}{\delta}\Big[\mathcal{F}(\sigma,\theta) - \frac{\theta\sigma}{2}\Big]}_{:=\Psi(\sigma,\theta)}, \tag{55}$$

where $\mathcal{F}(\sigma,\theta) \overset{\text{def}}{=} \frac{\theta}{2\sigma}\mathbb{E}[Y-\eta(Y)]^2 + \int_0^1 F_\Lambda^{-1}(u)F_{|\eta(Y)|}^{-1}(u)du$, with $\eta(\cdot) = \eta(\cdot;\mu_Y,\mu_{\sigma\Lambda/\theta})$. To prove (12), we adopt the same perturbation argument as in [26], [29], [50]. In particular, for a pseudo-Lipschiz function $\psi(\cdot,\cdot)$, define the following set of $\boldsymbol{v}$:

$$D_\nu := \big\{\boldsymbol{v}\in\mathbb{R}^p : |\mathbb{E}_{\mu_{\boldsymbol{v}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu^*}\psi| \geq \nu\big\}, \tag{56}$$

where $\nu > 0$ and μ^* denotes the joint measure of $\big(\eta(Y_*;\mu_{Y_*},\mu_{\sigma_*\Lambda/\theta_*}), B\big)$. Here (σ_*,θ_*) is the optimal solution of (55) and $Y_* = B + \sigma_* H$, with $H\sim\mathcal{N}(0,1)$ independent of $B\sim\mu_B$. Recall that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{v}} + \boldsymbol{\beta}$, so for any $\nu > 0$ and $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\Big(\big|\tfrac{1}{p}\sum_{i=1}^p \psi(\hat{\beta}_i,\,\beta_i) - \mathbb{E}\big[\psi\big(\eta(Y_*;\mu_{Y_*},\mu_{\sigma_*\Lambda/\theta_*}),B\big)\big]\big| \geq \nu\Big) =&\mathbb{P}\big(\widehat{\boldsymbol{v}}\in D_\nu\big) \\ \leq&\mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}} C(\boldsymbol{v}) + \varepsilon\big).\end{aligned} \tag{57}$$

This indicates that if we can show for any $\nu > 0$ and some $\varepsilon > 0$, $\min_{\boldsymbol{v}\in D_\nu} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}} C(\boldsymbol{v}) + \varepsilon$ occurs with vanishing probability, then (12) will immediately follow (with $\tau_* = \sigma_*/\theta_*$). In this way, proving (12) is reformulated as the perturbation analysis of $C_{\boldsymbol{\lambda}}(\boldsymbol{v})$, which can be done as follows. For any $\varepsilon \geq 0$ and $K > 0$, we have

$$\begin{aligned}
&\mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}} C(\boldsymbol{v}) + \varepsilon\big) \\
&\leq \mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu} C(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) + \mathbb{P}\big(\min_{\boldsymbol{v}} C(\boldsymbol{v}) > \Psi_* + \varepsilon\big) \\
&\leq \mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu \bigcap \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) + \mathbb{P}\big(\min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) > \Psi_* + \varepsilon\big) + 2\mathbb{P}\big(\hat{\boldsymbol{v}} \notin \mathcal{B}_{\sqrt{n}K}\big) \\
&\overset{\text{(a)}}{\leq} 2\mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu \bigcap \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) + 2\mathbb{P}\big(\min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) > \Psi_* + \varepsilon\big) + 2\mathbb{P}\big(\hat{\boldsymbol{v}} \notin \mathcal{B}_{\sqrt{n}K}\big), \qquad (58)
\end{aligned}$$

where (a) is due to (53) and (54). Here Ψ_* is the optimal value of (55). In Appendix D, we show all the three probabilities on the RHS of (58) vanish for $K \geq \frac{\sigma_*}{\sqrt{\delta}} + \frac{\theta_{\min}}{4}$ (with $\theta_{\min}$ given in Lemma 14) :

(i) From (88) of Lemma 10, $\mathbb{P}\big(|\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) - \Psi_*| \geq \varepsilon\big) \to 0$ for any $\varepsilon > 0$.

(ii) From (112) of Lemma 12, $\mathbb{P}\big(\hat{\boldsymbol{v}} \notin \mathcal{B}_{\sqrt{n}K}\big) \to 0$.

(iii) From Lemma 11, for any $\nu > 0$ there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \leq \varepsilon_0$, $\mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu \bigcap \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) \to 0$.

After substituting (i)-(iii) back to (58), we deduce that for any $\nu > 0$, there always exist $\varepsilon_0 > 0$ such that for any $\varepsilon \leq \varepsilon_0$, the RHS of (57) converges to 0 as $p \to \infty$. Therefore,

$$\frac{1}{p}\sum_{i=1}^{p} \psi(\hat{\beta}_i,\, \beta_i) \overset{\mathbb{P}}{\to} \mathbb{E}\big[\psi\big(\eta(Y_*; \mu_{Y_*}, \mu_{\sigma_*\Lambda/\theta_*}), B\big)\big],$$

On the other hand, by Lemma 14 in Appendix J, (σ_*, θ_*) is the unique solution of the following fixed point equation of (σ, θ):

$$\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}[\eta(Y; \mu_Y, \mu_{\sigma\Lambda/\theta}) - B]^2 \qquad (59)$$

$$\theta = \sigma\Big[1 - \frac{1}{\delta}\mathbb{E}\eta'(Y; \mu_Y, \mu_{\sigma\Lambda/\theta})\Big]. \qquad (60)$$

Therefore, letting $\tau_* = \sigma_*/\theta_*$, we can see (σ_*, τ_*) is also a solution of (13)-(14). Finally, we show such (σ_*, τ_*) is the unique solution of (13)- (14). By Lemma 14, (σ_*, θ_*) is the unique solution of (59)-(60) and it satisfies $\sigma_* \geq \sigma_w$, $\theta_* \geq \theta_{\min} > 0$. Suppose there exist two different solutions (σ_1, τ_1) and (σ_2, τ_2) to (13)-(14), then since $\sigma_1, \tau_1, \sigma_2, \tau_2 > 0$, $(\sigma_1, \sigma_1/\tau_1)$ and $(\sigma_2, \sigma_2/\tau_2)$ are two different solutions to (59)-(60), leading to a contradiction. This concludes our proof.

C. Asymptotic Variable Selection Performance

In this section, our goal is to prove Proposition 2. We first prove the convergence of $R_0^{(p)}$.

1) Probabilistic upper bound of $R_0^{(p)}$: To prove the convergence of $R_0^{(p)}$, the first step is to establish the following probabilistic upper bound.

Lemma 3: For any $\varepsilon > 0$, $R_0^{(p)} \leq \mathbb{P}\big(\eta(Y_*) = 0\big) + \varepsilon$, with probability approaching 1, as $p \to \infty$.

The proof of Lemma 3 is given in Appendix E, which uses a standard approximation argument (see *e.g.,* the proof of Lemma 2.2 (iii) and (v) in [53]).

2) Probabilistic lower bound of $R_0^{(p)}$: The second step is to prove the following matching probabilistic lower bound for $R_0^{(p)}$:

Lemma 4: Under the same setting as Proposition 2, for any $\varepsilon > 0$,

$$R_0^{(p)} \geq \mathbb{P}\big(\eta(Y_*) = 0\big) - \varepsilon, \tag{61}$$

with probability approaching 1, as $p \to \infty$.

The proof of Lemma 4, which can be found in Appendix F, is the mostly technically involved part, so we provide more detailed explanations here.

A key strategy we adopt is using the vector

$$\hat{\boldsymbol{s}} \stackrel{\text{def}}{=} \boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}) \tag{62}$$

as an indicator of zero coordinates of $\widehat{\boldsymbol{\beta}}$. To give the formal statements, we need to first introduce the notion of *majorization*.

Definition 2: For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, we say $\boldsymbol{a}$ is majorized by $\boldsymbol{b}$ (denoted as $\boldsymbol{a} \prec \boldsymbol{b}$), if for any $j \in [p]$, $\sum_{i=j}^p |a|_{(j)} \leq \sum_{i=j}^p |b|_{(j)}$. On the other hand, we say $\boldsymbol{a}$ is *strictly* majorized by $\boldsymbol{b}$ (denoted as $\boldsymbol{a} \prec_S \boldsymbol{b}$), if for any $j \in [p]$, $\sum_{i=j}^p |a|_{(j)} < \sum_{i=j}^p |b|_{(j)}$.

Denote $|\boldsymbol{x}|_{(1:k)}$ as a vector formed by the largest k components of $|\boldsymbol{x}|$. Let us call $|\boldsymbol{x}|_{(1:k)}$ as the *k-dominant subvector* of $\boldsymbol{x}$. The following is the key lemma for establishing the probabilistic lower bound of $R_0^{(p)}$.

Lemma 5: For the optimization problem (2), suppose for some $k \in [p]$, $|\hat{\boldsymbol{s}}|_{(1:k)} \prec_S \boldsymbol{\lambda}_{1:k}$, where $\hat{\boldsymbol{s}} = \boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\beta}})$. Then we have $|\widehat{\boldsymbol{\beta}}|_{(1:k)} = \boldsymbol{0}_k$.

The proof of Lemma 5 can be found in Appendix G. This characterization transform the original problem of searching zero coordinates of $\widehat{\boldsymbol{\beta}}$ into a new problem of discovering whether there is a strict majorization relation between k-dominant subvectors of $\hat{\boldsymbol{s}}$ and $\boldsymbol{\lambda}$. The nice thing making this strategy work is that the majorization relation between two vectors is fully captured by their empirical distributions. Besides, in our setting, the empirical distributions $\mu_{\boldsymbol{\lambda}}$ and $\mu_{\hat{\boldsymbol{s}}}$ both have simple limits: by our assumption, $\mu_{\boldsymbol{\lambda}} \to \mu_\Lambda$ and in Proposition 6 of Appendix I, we show $\mu_{\hat{\boldsymbol{s}}} \to \mu_{\hat{S}}$, with $\mu_{\hat{S}}$ being the law of $\frac{Y_* - \eta(Y_*)}{\tau_*}$.

A major part of proof of Lemma 4 is to show if condition (R.1) is satisfied and $\mathbb{P}\big(\eta(Y_*) = 0\big) > 0$, then for $k = \lfloor p[\mathbb{P}\big(\eta(Y_*) = 0\big) - \varepsilon] \rfloor$, where $\varepsilon > 0$ can be arbitrarily small, we have $|\hat{\boldsymbol{s}}|_{(1:k)} \prec_S \boldsymbol{\lambda}_{(1:k)}$ with probability approaching 1, as $p \to \infty$. Then an application of Lemma 5 will give us the desired probabilistic lower bound for $R_0^{(p)}$ shown in Lemma 4.

Remark 10: Let us briefly explain why $\boldsymbol{s}$ in (62) is related with the zero coordinates of $\widehat{\boldsymbol{\beta}}$. By the first order condition of (2), we can get $\hat{\boldsymbol{s}} \in \partial J_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}})$, *i.e.*, $\boldsymbol{s}$ defined in (62) is a subgradient of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ at $\boldsymbol{x} = \widehat{\boldsymbol{\beta}}$. For non-smooth regularizer like $J_{\boldsymbol{\lambda}}$, the subgradient $\partial J_{\boldsymbol{\lambda}}$ at $\boldsymbol{x}$ can reveal some information for detecting the zero coordinates of $\boldsymbol{x}$. A simple example is LASSO: $J_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$. In this case, we have $x_i = 0$ as long as $[\partial J_{\boldsymbol{\lambda}}(\boldsymbol{x})]_i \in [0, \lambda)$. This identity is used in [50] to obtain the limiting sparsity level of LASSO estimator. Here, we extend this idea to SLOPE estimator, while a key difference is that unlike the LASSO case, the zero coordinates are not determined locally: whether $x_i = 0$ or not is not completely determined by $[\partial J_{\boldsymbol{\lambda}}(\boldsymbol{x})]_i$. This is mainly a consequence of non-separability of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$.

Remark 11: We can now explain where condition (R.1) originates from. By Lemma 5, we know a sufficient condition for $R_0^{(p)}/p \geq \mathbb{P}\big(\eta(Y_*)=0\big)$ is that $|\hat{\boldsymbol{s}}|_{(1:k)} \prec_S \boldsymbol{\lambda}_{(1:k)}$ holds for some k satisfying $k/p \approx \mathbb{P}\big(\eta(Y_*)=0\big) := q_0^*$. In the asymptotic limit, this translate into the following limiting form: for any $t \in [0, q_0^*)$,

$$\int_t^{q_0^*} F^{-1}_{|Y_*-\eta(Y_*)|/\tau_*}(u)du < \int_t^{q_0^*} F_\Lambda^{-1}(u)du. \tag{63}$$

In fact, (63) is exactly (R.1), since for any $u \in [0, q_0^*]$ we have $F^{-1}_{|Y_*-\eta(Y_*)|/\tau_*}(u) = F^{-1}_{|Y_*|/\tau_*}(u)$.
After combining the probabilistic upper and lower bounds, we conclude that $R_0^{(p)} \xrightarrow{\mathbb{P}} \mathbb{P}\big(\eta(Y_*)=0\big)$.

3) Convergence of $V^{(p)}$: Finally, we prove the convergence of $V^{(p)}$. We have the following lemma, which shows $V^{(p)} \xrightarrow{\mathbb{P}} \mathbb{P}\big(\eta(Y_*)\neq 0, B=0\big)$ can be implied by $R_0^{(p)} \xrightarrow{\mathbb{P}} \mathbb{P}\big(\eta(Y_*)=0\big)$.

Lemma 6: For any $\varepsilon > 0$,

$$\big|V^{(p)} - \mathbb{P}\big[\eta(Y_*)\neq 0, B=0\big]\big| \leq |\mathbb{P}\big(\eta(Y_*)=0\big) - R_0^{(p)}| + \varepsilon,$$

with probability approaching 1 as $p \to \infty$.
The proof of Lemma 6 can be found in Appendix H. Since the convergence of $R_0^{(p)}$ has been established, we finish our proof.

D. Optimal Estimation

In this section, we prove the fundamental estimation performance of SLOPE, as stated in Proposition 4. The proof of part (a) and (b), which justifies the uniqueness of f_σ and the existence of σ_0, can be found in Lemma 18 and Lemma 19 in Appendix M. Here we focus on proving part (c), which is the core part of Proposition 4.

From discussions before, finding the minimum MSE is equivalent to solving (24). Indeed, we have

$$\inf_{\mu_\Lambda \in \mathcal{P}_\Lambda} \lim_{p\to\infty} \frac{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2}{p} = \delta(\sigma_{\text{opt}}^2 - \sigma_w^2), \tag{64}$$

where σ_{opt} is the optimal value of (24), *i.e.*, $\sigma_{\text{opt}} \overset{\text{def}}{=} \inf_{\mu_\Lambda \in \mathcal{P}_\Lambda} \sigma_*$.

1) A reformulation of $\inf_{\mu_\Lambda \in \mathcal{P}_\Lambda} \sigma_*$: We start by noting that σ_{opt} can be equivalently expressed as:

$$\sigma_{\text{opt}} = \inf\{\sigma \mid (\sigma,\tau) \in \mathcal{D}_L, \text{ for some } \tau > 0\}, \tag{65}$$

where

$$\mathcal{D}_L \overset{\text{def}}{=} \big\{(\sigma,\tau) \in \mathbb{R}^2_{>0} : \exists \mu_\Lambda \in \mathcal{P}_\Lambda \text{ s.t. } (\sigma,\tau) \text{ satisfies (13)-(14)}\big\}.$$

Geometrically, computing σ_{opt} is equivalent to searching for the leftmost point in $\mathcal{D}_L$, which is the set of all realizable (σ,τ) pair. However, characterizing $\mathcal{D}_L$ is difficult, since it is determined in a convoluted way via (13)-(14). To simplify, consider instead the following equation of (f,σ,τ)

$$\sigma^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}[f(B+\sigma H) - B]^2 \tag{66}$$

$$1 = \tau\Big[1 - \frac{1}{\delta}\mathbb{E}f'(B+\sigma H)\Big] \tag{67}$$

where $(f,\sigma,\tau) \in \mathcal{I} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and $H \sim \mathcal{N}(0,1)$ is independent of $B \sim \mu_B$. Let us emphasize that although (66)-(67) has a similar form as (13)-(14), a key difference is that unlike η in (13)-(14), f in (66)-(67) is not dependent on other parameters such as (B,H,σ,τ).

Now define the following set of (σ,τ):

$$\mathcal{D}_F \overset{\text{def}}{=} \big\{(\sigma,\tau)\in\mathbb{R}^2_{>0}:\ \exists f\in\mathcal{I}\ \text{s.t. }(f,\sigma,\tau)\ \text{satisfies (66)-(67)}\big\}.$$

A key step of our proof is to show $\mathcal{D}_L=\mathcal{D}_F$. This can be done as follows. Clearly, we have $\mathcal{D}_L\subseteq\mathcal{D}_F$, since $\eta\in\mathcal{I}$. To prove $\mathcal{D}_F\subseteq\mathcal{D}_L$, we need to utilize Proposition 3. Suppose $(\sigma,\tau)\in\mathcal{D}_F$ and let (f,σ,τ) be the corresponding solution of (66)-(67). If $\delta>1$, we have $\mathcal{P}_\Lambda=\mathcal{P}_2(\mathbb{R})$ and by Proposition 3 we can take $\Lambda\sim\frac{|Y|-f(|Y|)}{\tau}\in\mathcal{P}_2(\mathbb{R})$ so that $\eta(\cdot;\mu_Y,\mu_{\tau\Lambda})=f$; if $\delta\le 1$, then from (67) we know $f(y)\neq y$ and $\mathbb{P}\big(\frac{|Y|-f(|Y|)}{\tau}\neq 0\big)>0$, so $\frac{|Y|-f(|Y|)}{\tau}\in\mathcal{P}_\Lambda$ and we can still take $\Lambda\sim\frac{|Y|-f(|Y|)}{\tau}$ which gives us $\eta(\cdot;\mu_Y,\mu_{\tau\Lambda})=f$. This means $(\sigma,\tau)\in\mathcal{D}_L$. As a result, we conclude that $\mathcal{D}_F\subseteq\mathcal{D}_L$ and thus $\mathcal{D}_L=\mathcal{D}_F$. Then substituting $\mathcal{D}_L=\mathcal{D}_F$ into (65), we get the following reformulation of σ_{opt}:

$$\sigma_{\text{opt}}=\inf\{\sigma\mid(\sigma,\tau)\in\mathcal{D}_F,\ \text{for some }\tau>0\}. \tag{68}$$

2) Lower Bound of MSE: Note that any $(f,\sigma)\in\mathcal{I}\times\mathbb{R}_{>0}$ satisfying (66)-(67) for some $\tau>0$, should also satisfy

$$\begin{aligned}\mathbb{E}[f(B+\sigma H)-B]^2=&\delta(\sigma^2-\sigma_w^2)\\ \delta^{-1}\mathbb{E}f'(B+\sigma H)\le&1.\end{aligned} \tag{69}$$

Therefore, if we consider the following set of σ:

$$\mathcal{A}\overset{\text{def}}{=}\big\{\sigma>0:\ \exists f\in\mathcal{I},\text{s.t. }(f,\sigma)\ \text{satisfies (69)}\big\}, \tag{70}$$

then from (68) we have

$$\sigma_{\text{opt}}\ge\inf\mathcal{A}. \tag{71}$$

Compared with σ_{opt}, the lower bound $\inf\mathcal{A}$ in (71) is easier to obtain, since the variable τ is dropped. In Lemma 19 in Appendix M, we show that $\inf\mathcal{A}=\sigma_0$. Therefore, $\sigma_{\text{opt}}\ge\sigma_0$. Together with (64), we prove (30).

3) Reaching the Lower Bound: We now show lower bound $\sigma_{\text{opt}}\ge\sigma_0$ is tight, if $\delta^{-1}\mathbb{E}\big[f'_{\sigma_0}(B+\sigma_0H)\big]<1$. Recall that f_{σ_0} is the unique optimal solution of (28) when $\sigma=\sigma_0$ and (f_{σ_0},σ_0) satisfies (69). Let $\tau_0=\big[1-\delta^{-1}\mathbb{E}f'_{\sigma_0}(B+\sigma_0H)\big]^{-1}$. It is not hard to see when $\delta^{-1}\mathbb{E}\big[f'_{\sigma_0}(B+\sigma_0H)\big]<1$, $\tau_0\in(0,\infty)$ and $(f_{\sigma_0},\sigma_0,\tau_0)$ is a solution of (66)-(67). This indicates $(\sigma_0,\tau_0)\in\mathcal{D}_F$ and thus from (68) we have $\sigma_{\text{opt}}\le\sigma_0$. Together with the lower bound $\sigma_{\text{opt}}\ge\sigma_0$, we get $\sigma_{\text{opt}}=\sigma_0$.

On the other hand, by Proposition 3 we know if μ_Λ is taken as the law of $\frac{1}{\tau_0}\big(|Y_0|-f_{\sigma_0}(|Y_0|)\big)$, then $f_{\sigma_0}(y)=\eta(y;\mu_{Y_0},\mu_{\tau_0\Lambda})$. Since $(f_{\sigma_0},\sigma_0,\tau_0)$ is a solution of equation (66)-(67), we know $(\sigma_*,\tau_*)=(\sigma_0,\tau_0)$, where (σ_*,τ_*) is the solution of fixed-point equation (13)-(14) under this choice of μ_Λ. According to (15), we get $\lim_{p\to\infty}\frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2}{p}=\delta(\sigma_0^2-\sigma_w^2)$. This completes our proof.

E. Optimal Variable Selection

In this section, we are going to prove Proposition 5. First, part (a) and (b) can be proved in an analogous way as in Proposition 5, which is summarized in Lemma 21. Here we focus on part (c).

1) Upper bound of $\mathcal{P}(\alpha)$: Directly solving the original optimization (40) is not easy. Instead, replacing by a new objective function in (40), we will first consider the following problem:

$$\begin{aligned}\overline{\mathcal{P}}(\alpha) \stackrel{\text{def}}{=} & \sup_{\mu_\Lambda \in \widetilde{\mathcal{P}}_\Lambda} \mathbb{P}\Big(|B+\sigma_* H| \geq \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_* \mid B \neq 0\Big) \\ & \text{s.t. } \mathbb{P}(|\sigma_* H| \geq y^*_{\text{th}}) \leq \alpha\end{aligned} \tag{72}$$

where $y^*_{\text{th}} = \sup_{y\geq 0}\{y \mid \eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) = 0\}$. It is not hard to show for any $\alpha \in [0,1]$, we have $\overline{\mathcal{P}}(\alpha) \geq \mathcal{P}(\alpha)$. This is because the constraint $\mathbb{P}(|\sigma_* H| \geq y^*_{\text{th}}) \leq \alpha$ in (40) implies $y^*_{\text{th}} \geq \Phi^{-1}(1-\frac{\alpha}{2})\sigma_*$ and hence the objective function of (40) is upper bounded by that of (72) for any $\mu_\Lambda \in \widetilde{\mathcal{P}}_\Lambda$.

Problem (72) can be further simplified. By direct differentiation, one can check for any fixed $b \neq 0$ and $c \geq 0$, the function $\sigma \mapsto \mathbb{P}(|b+\sigma H| \geq c\sigma)$ is non-increasing on $\mathbb{R}_{\geq 0}$, where $H \sim \mathcal{N}(0,1)$. This then implies, by conditioning on B, that $\sigma \mapsto \mathbb{P}(|B+\sigma H| \geq c\sigma \mid B \neq 0)$ is non-increasing on $\mathbb{R}_{\geq 0}$ for any distribution of B satisfying $\mathbb{P}(B \neq 0) > 0$. Therefore, solving maximization problem in (72) is equivalent to solving the minimization problem of σ_*. Meanwhile, by the definition of y^*_{th} and the fact that $\eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) \in \mathcal{I}$, we know: $\mathbb{P}(|\sigma_* H| \geq y^*_{\text{th}}) \leq \alpha$ if and only if $\eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) = 0$, for all $|y| \leq \Phi^{-1}(1-\frac{\alpha}{2})\sigma_*$. As a result, solving (72) is equivalent to solving:

$$\begin{aligned}\sigma_{\text{opt},\alpha} \stackrel{\text{def}}{=} & \inf_{\mu_\Lambda \in \widetilde{\mathcal{P}}_\Lambda} \sigma_* \\ & \text{s.t. } \eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) = 0, \forall |y| \leq \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_*\end{aligned} \tag{73}$$

and $\overline{\mathcal{P}}(\alpha)$ in (72) can be expressed in terms of $\sigma_{\text{opt},\alpha}$ in (73) as:

$$\overline{\mathcal{P}}(\alpha) = \mathbb{P}\Big(|B+\sigma_{\text{opt},\alpha} H| \geq \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_{\text{opt},\alpha} \mid B \neq 0\Big). \tag{74}$$

So far, we arrive at the optimization problem (73), which is similar to the one that we have analyzed in the estimation setting [c.f. (24)]. Yet there are two differences: (i) a constraint on η is added to ensure type-I error is bounded by α, (ii) a constraint on μ_Λ is added to guarantee valid limits of Type-I error and power exist (see Proposition 2). It turns out that the strategy we used can still be applied. The results are parallel to Proposition 4 part (c) and are summarized in Lemma 22 in Appendix N, where it is shown that

$$\sigma_{\text{opt},\alpha} \geq \sigma_{0,\alpha} \tag{75}$$

and the lower bound can be achieved when $\mu_\Lambda = \mu_{\text{opt},\alpha}$, if $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$. After combining (75) with

$$\mathcal{P}(\alpha) \leq \overline{\mathcal{P}}(\alpha) = \mathbb{P}\Big(|B+\sigma_{\text{opt},\alpha} H| \geq \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_{\text{opt},\alpha} \mid B \neq 0\Big), \tag{76}$$

and (39), we get (43).

2) Reaching the upper bound: Now we show for any $\alpha \in [0,1]$, the upper bound (43) is tight, if $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$ and $y_{0,\alpha} = \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha}$. Also it is attained by $\mu_\Lambda = \mu_{\text{opt},\alpha}$. The case of $\alpha = 0$ is easy. Indeed, in this case, both sides of (43) equal to 0. We just need to verify the case of $\alpha \in (0,1]$. By Lemma 22 and (76), we know it suffices to show $\mathcal{P}(\alpha) = \overline{\mathcal{P}}(\alpha)$ and also $\lim_{p\to\infty}\text{Power} = \mathcal{P}(\alpha)$, when $\mu_\Lambda = \mu_{\text{opt},\alpha}$. We verify this in Lemma 23 in Appendix N, which completes our proof.

VI. Concluding Remarks

We have established the asymptotic characterization of SLOPE in the high-dimensional regime. Although SLOPE is a high-dimensional regularized regression method, asymptotically its statistical performance can be fully characterized by a few scalar random variables. The precise characterization enabled us to derive the fundamental performance limits of SLOPE for both estimation and variable selection settings. Also we showed how to design the optimal regularizing sequences that achieve these limits.

Finally, let us point out some generalizations of current results that worth exploring in the future.

1) One major technical assumption in the current paper is that the sensing matrix is generated from i.i.d. Gaussian. There are two possible ways to relax this assumption. The first one is to consider the Gaussian design with correlated columns, which is the setting analyzed in [6]. Under this scenario, SLOPE enjoys the nice properties of selecting all the variables associated with highly correlated columns. It would be interesting to derive a precise explanation for this phenomenon. The second direction is staying in the i.i.d. setting, while generalizing to other ensembles, *e.g.*, sub-Gaussian distribution. This is to verify the so-called *universality* phenomenon and some works have been done in the setting where the regularizer is separable [54], [55]. It would be interesting to generalize these results to non-separable regularizers such as SLOPE.
2) The optimal designs of $\boldsymbol{\lambda}$ sequences considered in this paper are based on the assumption that the true distribution of unknown signal is known. The natural question is: can we design $\boldsymbol{\lambda}$ sequences without (or just with partial) such prior knowledge? One related problem is designing a regularizing sequence such that the false discovery rate is always controlled under a given level. In this setting, the realistic assumption is that we do not know the sparsity of underlying signal. For this purpose, a design of $\boldsymbol{\lambda}$ is proposed in [9] based on some qualitative insights. It would be nice to have quantitative results utilizing the exact characterizations derived here.
3) From numerical simulations, we can find that in several cases, the performance of practical $\boldsymbol{\lambda}$ sequences such as LASSO and BHq is comparable to the optimal performance. Is it possible that the optimal performance of SLOPE can actually be approximately achieved, when we are restricted to certain sub-classes of regularizing sequences? A key step is establishing some easy-to-evaluate bounds for the performance gap between practical and optimal sequences. One benefit of using practical sequences is that we can apply some purely data-dependent methods such as cross-validation to search for the optimal tuning parameter. Note that since general $\boldsymbol{\lambda}$ sequence includes order $\mathcal{O}(p)$ parameters, the grid search approach that is usually used in data-dependent method is not plausible here.

APPENDIX

A. Proof of Lemma 1

First assume $0 \le y_1 \le \cdots \le y_p$. Denote $\hat{\boldsymbol{y}} := \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$. Then consider the linear interpolation of the points $\{(y_i, \hat{y}_i)\}_{i=0}^p$, where $(y_0, \hat{y}_0) = (0, 0)$:

$$g_p^+(y) = \begin{cases} y_{i-1} + \frac{\hat{y}_i - \hat{y}_{i-1}}{y_i - y_{i-1}}(y - y_{i-1}) & y \in (y_{i-1}, y_i), \\ \hat{y}_i & y = y_i, \\ \hat{y}_p + (y - y_p) & y > y_p. \end{cases} \tag{77}$$

By Fact 1 (iii), we know $g_p^+(y)$ is non-decreasing and 1-Lipschitz continuous on $\mathbb{R}_{\ge 0}$.

For general $\boldsymbol{y}$, we first obtain the linear interpolation $g_p^+(y)$ of the points $\{(|y|_{(i)}, |\hat{y}|_{(i)})\}_{i=0}^p$ as above. Then $g_p(y)$ can be constructed as follows:

$$g_p(y) = \begin{cases} g_p^+(y) & y \ge 0, \\ -g_p^+(-y) & y < 0. \end{cases}$$

Clearly, such $g_p(y)$ is an odd, non-decreasing and 1-Lipschitz function. Also by Fact 1 (i) and (ii), one can easily check $g_p(y_i) = \hat{y}_i$, for all $i \in [p]$. This finishes the proof.

B. Proof of Lemma 2

For notational simplicity, denote $\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}) := \mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}; 1)$. Let $g_p^*(\boldsymbol{y})$ be any minimizer of (45). Since $\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is the minimum value of (5),

$$\begin{aligned} \mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y}) &\le \frac{1}{2}\|\boldsymbol{y} - g_p^*(\boldsymbol{y})\|_2^2 + \sum_{i=1}^p \lambda_i |g_p^*(\boldsymbol{y})|_{(i)} \\ &= p\mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y}). \end{aligned} \tag{78}$$

Next, we show $p\mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y}) \le \mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y})$. Given Lemma 1, this is immediate. Indeed,

$$\begin{aligned} \mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y}) &\le L_p(g_p) \\ &= \frac{1}{2}\|\boldsymbol{y} - \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})\|_2^2 + \sum_{i=1}^p \lambda_i |\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})|_{(i)} \\ &= \frac{\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y})}{p}, \end{aligned} \tag{79}$$

where g_p is the function we construct in Lemma 1, which satisfies $g_p \in \mathcal{I}$ and $g_p(\boldsymbol{y}) = \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$. Combining (78) and (79), we get $\mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y}) = \frac{\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y})}{p}$.

Substituting $\mathcal{M}_{\boldsymbol{\lambda}}^*(\boldsymbol{y}) = \frac{\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y})}{p}$ into (78), we have for any minimizer g_p^* of (45), $g_p^*(\boldsymbol{y})$ is also a minimizer of (5). Since $\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$ is the unique minimizer of (5), we know any minimizer of (45) should satisfy: $g_p^*(\boldsymbol{y}) = \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y})$.

C. Auxiliary Results for Proving Proposition 1

Lemma 7: Suppose $\{\boldsymbol{y}^{(p)}\}_{p\in\mathbb{Z}^+}$ and $\{\boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{Z}^+}$ are converging sequences with limiting measure μ_Y and μ_Λ. Then

$$\sup_{g\in\mathcal{I}}|L(g)-L_p(g)|\to 0, \tag{80}$$

where $L(g)$ and $L_p(g)$ are defined in (46) and (45).

Proof: The first step is to establish the following uniform convergence of a class of Pseudo-Lipschitz functions. Let Ψ be the set of all functions $\psi:\mathbb{R}\to\mathbb{R}$ satisfying: $\psi(0)=0$ and $|\psi(x)-\psi(y)|\le(1+|x|+|y|)|x-y|$ for any $x,y\in\mathbb{R}$. Then

$$\sup_{\psi\in\Psi}|\mathbb{E}_{\mu_{\boldsymbol{y}}}\psi(Y)-\mathbb{E}_{\mu_Y}\psi(Y)|\to 0. \tag{81}$$

To prove (81), first for any $\psi\in\Psi$ consider the truncation:

$$\hat{\psi}_A(y)=\begin{cases}\psi(-A) & y<-A,\\ \psi(y) & |y|\le A,\\ \psi(A) & y>A,\end{cases}$$

where $A>0$ is a constant. It is easy to check $\hat{\psi}_A(y)$ is $(1+2A)$-Lipschitz continuous, so

$$\left|\mathbb{E}_{\mu_{\boldsymbol{y}}}\hat{\psi}_A(Y)-\mathbb{E}_{\mu_Y}\hat{\psi}_A(Y)\right|\overset{\text{(a)}}{\le}(1+2A)W_1(\mu_{\boldsymbol{y}},\mu_Y)\overset{\text{(b)}}{\le}(1+2A)W_2(\mu_{\boldsymbol{y}},\mu_Y),$$

where (a) follows from Kantonovich duality theorem ([56, Theorem 1.3]) and (b) follows from Holder's inequality. Therefore,

$$\begin{aligned}\left|\mathbb{E}_{\mu_{\boldsymbol{y}}}\psi(Y)-\mathbb{E}_{\mu_Y}\psi(Y)\right|&\le\left|\mathbb{E}_{\mu_{\boldsymbol{y}}}\hat{\psi}_A(Y)-\mathbb{E}_{\mu_Y}\hat{\psi}_A(Y)\right|\\&\quad+\left|\mathbb{E}_{\mu_{\boldsymbol{y}}}\hat{\psi}_A(Y)-\mathbb{E}_{\mu_{\boldsymbol{y}}}\psi(Y)\right|+\left|\mathbb{E}_{\mu_Y}\hat{\psi}_A(Y)-\mathbb{E}_{\mu_Y}\psi(Y)\right|\\&\le(1+2A)W_2(\mu_{\boldsymbol{y}},\mu_Y)\\&\quad+2\mathbb{E}_{\mu_{\boldsymbol{y}}}[\mathbb{I}_{|Y|\ge A}(Y^2+|Y|)]+2\mathbb{E}_{\mu_Y}[\mathbb{I}_{|Y|\ge A}(Y^2+|Y|)].\end{aligned} \tag{82}$$

For any $\varepsilon>0$, each term on the RHS of (82) can be bounded as follows. Since $(\mathbb{E}_{\mu_Y}|Y|)^2\le\mathbb{E}_{\mu_Y}Y^2<\infty$, by DCT there always exists $A>0$ such that $\mathbb{E}_{\mu_Y}[\mathbb{I}_{|Y|\ge A}(Y^2+|Y|)]\le\frac{\varepsilon}{4}$. On the other hand, for any given $A>0$ and $\varepsilon>0$, there always exists $p_0\in\mathbb{N}$ such that for any $p\ge p_0$, (i) $W_2(\mu_{\boldsymbol{y}},\mu_Y)\le\frac{\varepsilon}{4(1+2A)}$, since $W_2(\mu_{\boldsymbol{y}},\mu_Y)\to 0$ and (ii) $\mathbb{E}_{\mu_{\boldsymbol{y}}}[\mathbb{I}_{|Y|\ge A}(Y^2+|Y|)]\le\mathbb{E}_{\mu_Y}[\mathbb{I}_{|Y|\ge A}(Y^2+|Y|)]+\frac{\varepsilon}{4}$ by Theorem 7.12 (iv) in [56]. Note that the RHS of (82) does not depend on ψ, so for any $\varepsilon>0$, there exists $p_0\in\mathbb{N}$ such that for any $p\ge p_0$, $\sup_{\psi\in\Psi}\left|\mathbb{E}_{\mu_{\boldsymbol{y}}}\psi(Y)-\mathbb{E}_{\mu_Y}\psi(Y)\right|\le\varepsilon$. Therefore, (81) is proved.

We are now ready to show (80). Recalling the definitions of $L(g)$ and $L_p(g)$ in (46) and (45), we have

$$\begin{aligned}\sup_{g\in\mathcal{I}}|L(g)-L_p(g)|&\le\frac{1}{2}\sup_{g\in\mathcal{I}}\left|\mathbb{E}_{\mu_{\boldsymbol{y}}}[Y-g(Y)]^2-\mathbb{E}_{\mu_Y}[Y-g(Y)]^2\right|\\&\quad+\sup_{g\in\mathcal{I}}\int_0^1|F_{\boldsymbol{\lambda}}^{-1}(u)-F_\Lambda^{-1}(u)|F_{|g(Y)|}^{-1}(u)du\\&\quad+\sup_{g\in\mathcal{I}}\int_0^1F_{\boldsymbol{\lambda}}^{-1}(u)\left|F_{|g(Y)|}^{-1}(u)-F_{|g(\boldsymbol{y})|}^{-1}(u)\right|du.\end{aligned} \tag{83}$$

Therefore, it remains to control each term on the RHS of (83). The first term can be handled by using (81), since $y \mapsto [y - g(y)]^2$ belongs to Ψ for $g \in \mathcal{I}$; the second term can be controlled as : Term II $\leq W_2(\mu_{\boldsymbol{\lambda}}, \mu_\Lambda)\sqrt{\mathbb{E}_{\mu_Y} Y^2}$ using (86) and Cauchy-Swartz inequality; similarly for the third term, we have

$$\begin{aligned}\text{Term III} \leq & \sup_{g\in\mathcal{I}} \sqrt{\mathbb{E}_{\mu_{\boldsymbol{\lambda}}}\Lambda^2} W_2\big(\mu_{|g(\boldsymbol{y})|}, \mu_{|g(Y)|}\big) \\ \leq & \sqrt{\mathbb{E}_{\mu_{\boldsymbol{\lambda}}}\Lambda^2} W_2\big(\mu_{\boldsymbol{y}}, \mu_Y\big),\end{aligned}$$

where the last inequality follows from the definition of Wasserstein distance and the Lipschitz continuity of g:

$$\begin{aligned}W_2\big(\mu_{|g(\boldsymbol{y})|}, \mu_{|g(Y)|}\big)^2 = & \inf_{\pi\in\Pi(\mu_{|g(\boldsymbol{y})|}, \mu_{|g(Y)|})} \int (g-h)^2 d\pi(g,h) \\ = & \inf_{\pi\in\Pi(\mu_{\boldsymbol{y}}, \mu_Y)} \int [|g(x)| - |g(y)|]^2 d\pi(x,y) \\ \leq & \inf_{\pi\in\Pi(\mu_{\boldsymbol{y}}, \mu_Y)} \int (x-y)^2 d\pi(x,y) \\ = & W_2\big(\mu_{\boldsymbol{y}}, \mu_Y\big)^2.\end{aligned}$$

Substituting the above bounds back to (83) and using the assumption that $W_2(\mu_{\boldsymbol{\lambda}}, \mu_\Lambda), W_2\big(\mu_{\boldsymbol{y}}, \mu_Y\big) \to 0$, we obtain the desired results. ■

Lemma 8: The optimization problem (6) has an optimal solution and it is unique (up to a set of measure 0 with respect to μ_Y).

Proof: Without loss of generality, we assume $\tau = 1$. The objective function $L(g)$ of (6) is defined on the following L^2 space:

$$\mathcal{H}_{\mu_Y} \stackrel{\text{def}}{=} \{g(y) \mid g(y) \text{ is measurable and } \|g\|_{\mu_Y} < \infty\} \tag{84}$$

where $\|g\|_{\mu_Y} \stackrel{\text{def}}{=} [\mathbb{E}_{\mu_Y} g^2(Y)]^{1/2}$. It is known that in L^2 space (and more generally in all normed linear spaces), the convention is to work with *equivalence class* of functions [57, p.135-136]. The equivalence class of a function $f \in \mathcal{H}_{\mu_Y}$, denoted as $[f]$, is the collection of all functions $g \in \mathcal{H}_{\mu_Y}$ satisfying $\|g - f\|_{\mu_Y} = 0$. As a notational convention, we will write $[f]$ as f, and the set $\{[f] : f \in \mathcal{I}\}$ as $\mathcal{I}$. Also $\|g - f\|_{\mu_Y} = 0$ will be denoted as $g = f$.

We first show $L(g)$ is 1-strongly convex on $\mathcal{H}_{\mu_Y}$, *i.e.*, for any $g_1, g_2 \in \mathcal{H}_{\mu_Y}$,

$$L(\theta g_1 + (1-\theta)g_2) \leq \theta L(g_1) + (1-\theta)L(g_2) - \frac{\theta(1-\theta)}{2}\|g_2(Y) - g_1(Y)\|^2. \tag{85}$$

First, for any $\theta \in [0,1]$,

$$\begin{aligned}\int_0^1 F_\Lambda^{-1}(u) F^{-1}_{|\theta g_1(Y) + (1-\theta) g_2(Y)|}(u) du \leq & \int_0^1 F_\Lambda^{-1}(u) F^{-1}_{\theta|g_1(Y)| + (1-\theta)|g_2(Y)|}(u) du \\ = & \theta \int_0^1 F_\Lambda^{-1}(u) F^{-1}_{|g_1(Y)|}(u) du + (1-\theta)\int_0^1 F_\Lambda^{-1}(u) F^{-1}_{|g_2(Y)|}(u) du,\end{aligned}$$

which implies that $L_1(g) := \int_0^1 F_\Lambda^{-1}(u) F^{-1}_{|g(Y)|}(u) du$ is convex. Also, it is not hard to check $L_2(g) := \frac{1}{2}\mathbb{E}_\mu[Y - g(Y)]^2$ is 1-strongly convex by definition (85). Then the strong convexity of $L(g)$ follows, since $L(g) = L_1(g) +$

$L_2(g)$. On the other hand, we can show $L(g)$ is continuous on $\mathcal{H}_{\mu_Y}$. Indeed,

$$\begin{aligned}|L(g_2)-L(g_1)| &\le \frac{1}{2}\|g_2-g_1\|_{\mu_Y}\cdot\|2y-g_1-g_2\|_{\mu_Y}\\ &\quad+\sqrt{\mathbb{E}\Lambda^2}\,\big\|\,|g_2|-|g_1|\,\big\|_{\mu_Y}\\ &\le \|g_2-g_1\|_{\mu_Y}\big(2\|y\|_{\mu_Y}+2\|g_1\|_{\mu_Y}+\|g_2-g_1\|_{\mu_Y}+\sqrt{\mathbb{E}\Lambda^2}\big).\end{aligned}$$

Since $\|y\|_{\mu_Y}, \|g_1\|_{\mu_Y}, \|g_2\|_{\mu_Y} < \infty$, we conclude that $L(g)$ is continuous.

Next we are going to show the set $\mathcal{I}$ is convex, bounded and closed in $\mathcal{H}_{\mu_Y}$. The convexity can be directly checked by definition. Choose any $g_1, g_2 \in \mathcal{I}$. Then there exists $S \subseteq \mathbb{R}$ with $\mu_Y(S) = 1$ such that for any $y_1, y_2 \in S$, $y_1 \le y_2$ and any $y \in S$, we have $0 \le g_i(y_2) - g_i(y_1) \le y_2 - y_1$ and $g_i(y) = -g_i(-y)$, where $i = 1, 2$. Then for any $\theta \in [0, 1]$, function $\theta g_1 + (1-\theta)g_2$ also satisfy (i) and (ii) on S, so $\theta g_1 + (1-\theta)g_2 \in \mathcal{I}$. The boundedness directly follows from the fact that for any $g \in \mathcal{I}$, $g(y) \le |y|$ on some $S \subseteq \mathbb{R}$ with $\mu_Y(S) = 1$. To show closedness, suppose $g_k(y) \in \mathcal{I}, k = 1, 2, \ldots$ is a sequence of functions that converge to some $g(y) \in \mathcal{H}_{\mu_Y}$. Then by Riesz-Fischer Theorem, there exists a sub-sequence of $\{g_k(y)\}_{k\in\mathbb{Z}^+}$ that converges point-wise to $g(y)$ on some $S \subseteq \mathbb{R}$ with $\mu_Y(S) = 1$. By this μ_Y-almost everywhere convergence of $g_k(y)$ to $g(y)$, we know there exists some $S' \subseteq \mathbb{R}$ with $\mu_Y(S') = 1$, such that for any $y_1, y_2 \in S'$,$y_1 \le y_2$ and any $y \in S'$, it holds that $0 \le g(y_2) - g(y_1) \le y_2 - y_1$ and $g(y) = -g(-y)$. Therefore, $g(y) \in \mathcal{I}$ and thus $\mathcal{I}$ is closed.

The final step is to apply Theorem 17 in [57, Chap. 8] to conclude that (6) has an optimal solution $g^* \in \mathcal{I}$. Also the uniqueness of g^* can be easily checked by the strong convexity of $L(g)$. Suppose there exists two different optimal solutions, g_1^*, g_2^* with $L(g_1^*) = L(g_2^*)$ and $g_1^* \ne g_2^*$. Then by (85), for $g = \frac{g_1+g_2}{2} \in \mathcal{I}$ we have $L(g) < L(g_1^*) = L(g_2^*)$, which leads to a contradiction. ■

The following result provides the explicit formula for calculating Wasserstein-2 distance between probability measure on $\mathbb{R}$. Readers can find a proof in Theorem 2.18 in [56].

Lemma 9: Suppose $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R})$ and the corresponding quantile functions are F_1^{-1} and F_2^{-1}. Then

$$W_2(\mu_1,\mu_2)^2 = \int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 dt. \tag{86}$$

D. Auxiliary Results for Proving Theorem 1

In this section, we prove three auxiliary lemmas used in the proof of Theorem 1.

The first two results are on the asymptotic properties of auxiliary problem (52). To state these asymptotic results, similar as Theorem 1, we will consider a sequence of auxiliary problems described by the instances $\{\boldsymbol{g}^{(p)}, \boldsymbol{h}^{(p)}, \boldsymbol{\beta}^{(p)}, \boldsymbol{w}^{(p)}, \boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{Z}^+}$. They satisfy the following: (i) $\boldsymbol{g}^{(p)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $\boldsymbol{h}^{(p)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, $p \in \mathbb{Z}^+$ are all independent, (ii)$\{\boldsymbol{\beta}^{(p)}\}_{p\in\mathbb{Z}^+}$, $\{\boldsymbol{w}^{(p)}\}_{p\in\mathbb{Z}^+}$, $\{\boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{Z}^+}$ are the same converging sequences as in Theorem 1. Here the requirement that $\{\boldsymbol{g}^{(p)}\}_{p\in\mathbb{Z}^+}$ and $\{\boldsymbol{h}^{(p)}\}_{p\in\mathbb{Z}^+}$ are independent is not completely necessary, since we are only aiming for results regarding convergence in probability. The independence assumption simply allows us to directly apply some results obtained in Appendix K.

The first lemma is about the minimum value and the minimizer of $L(\boldsymbol{v})$ over a bounded Euclidean ball. Recall that

$$L(\boldsymbol{v}) \stackrel{\text{def}}{=} \frac{1}{2}\left(\sqrt{\frac{\|\boldsymbol{v}\|^2}{n}\frac{\|\boldsymbol{g}\|^2}{n}+\frac{\|\boldsymbol{w}\|^2}{n}+2\frac{\|\boldsymbol{v}\|}{\sqrt{n}}\frac{\boldsymbol{g}^\top\boldsymbol{w}}{n}}-\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\right)_+^2+\frac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n}. \tag{87}$$

Lemma 10: Let Ψ_* and (σ_*,θ_*) be the optimal value and solution of the minimax problem in (55) and $\theta_{\min}>0$ be the lower bound of θ_* obtained in Lemma 14. For any $\varepsilon>0$ and $K\geq\frac{\sigma_*}{\sqrt{\delta}}+\frac{\theta_{\min}}{4}$, we have

$$\mathbb{P}\Big(\Big|\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v})-\Psi_*\Big|\leq\varepsilon\Big)\to 1 \tag{88}$$

and

$$\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{32n\varepsilon/\gamma}}(\mathring{\boldsymbol{v}})\cap\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v})\geq\Psi_*+\varepsilon\Big)\to 1, \tag{89}$$

where $\gamma=\frac{\theta_{\min}\sigma_w^2}{4(K^2+\sigma_w^2)^{3/2}}$ and

$$\mathring{\boldsymbol{v}}:=\eta(\boldsymbol{\beta}+\sigma_*\boldsymbol{h})-\boldsymbol{\beta}. \tag{90}$$

Here in (90), $\eta(\cdot):=\eta(\boldsymbol{\beta}+\sigma_*\boldsymbol{h};\mu_{Y_*},\mu_{\sigma_*\Lambda/\theta_*})$ and $Y_*=B+\sigma_*H$, with $H\sim\mathcal{N}(0,1)$ independent of $B\sim\mu_B$.

Proof: We follow the proof of Proposition B.2 in [50]. First introduce the event $\mathcal{A}=\bigcap_{i=1}^5\mathcal{A}_i$, where

$$\begin{aligned}
\mathcal{A}_1&:=\Big\{\tfrac{\|\boldsymbol{g}\|^2}{n},\tfrac{\|\boldsymbol{h}\|^2}{p},\tfrac{\|\boldsymbol{w}\|^2}{\sigma_w^2 n}\in[1-\varsigma,1+\varsigma],\Big|\tfrac{\boldsymbol{g}^\top\boldsymbol{w}}{n}\Big|\leq\varsigma\Big\},\\
\mathcal{A}_2&:=\{\|\mathring{\boldsymbol{v}}\|/\sqrt{p}\leq\sigma_*\},\\
\mathcal{A}_3&:=\Big\{\sqrt{\tfrac{\|\mathring{\boldsymbol{v}}\|^2}{n}+\sigma_w^2}-\tfrac{\boldsymbol{h}^\top\mathring{\boldsymbol{v}}}{n}\geq\tfrac{\theta_{\min}}{2}\Big\},\\
\mathcal{A}_4&:=\{|\widetilde{L}(\mathring{\boldsymbol{v}})-\Psi_*|\leq\varepsilon\},\\
\mathcal{A}_5&:=\Big\{\sup_{\sigma\in[\sigma_w,\sqrt{\sigma_w^2+K^2}]}\big|\big(\mathcal{F}_p(\sigma,\theta_*)-\tfrac{\sigma\theta_*\|\boldsymbol{h}\|^2}{2p}\big)-\big(\mathcal{F}(\sigma,\theta_*)-\tfrac{\sigma\theta_*}{2}\big)\big|\leq\delta\varepsilon\Big\},
\end{aligned} \tag{91}$$

with $\varepsilon>0$ and $\varsigma\in(0,\frac{1}{2})$. In (91), $\mathcal{F}_p$ and $\mathcal{F}$ are the same as in (160) and (163) and $\widetilde{L}(\boldsymbol{v})$ is defined as:

$$\widetilde{L}(\boldsymbol{v})\stackrel{\text{def}}{=}\frac{1}{2}\left(\sqrt{\frac{\|\boldsymbol{v}\|^2}{n}+\sigma_w^2}-\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\right)_+^2+\frac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n}. \tag{92}$$

Based on the event $\mathcal{A}$, our subsequent analysis will become fully deterministic: we will condition on fixed $\boldsymbol{h}$ and $\boldsymbol{g}$ in $\mathcal{A}$. Before doing so, let us first show each of the events $\mathcal{A}_1\sim\mathcal{A}_5$ occurs with probability approaching 1 as $p\to\infty$, so $\mathbb{P}(\mathcal{A})\to 1$. This will ensure all the results obtained by conditioning on $\mathcal{A}$ hold with probability approaching 1.

$\mathcal{A}_1$: By the law of large number and the fact that $\{\boldsymbol{w}\}_{p\in\mathbb{N}}$ is a converging sequence with limiting variance $\sigma_w^2>0$, it is not hard to show $\mathbb{P}(\mathcal{A}_1)\to 1$.

$\mathcal{A}_2$: From (179), we have

$$\frac{\|\mathring{\boldsymbol{v}}\|^2}{p}\stackrel{a.s.}{\to}\mathbb{E}[B-\eta(B+\sigma_*H)]^2. \tag{93}$$

Then together with (151), we get $\frac{\|\mathring{\boldsymbol{v}}\|^2}{p}\stackrel{a.s.}{\to}(\sigma_*)^2-\sigma_w^2$. Therefore, $\mathbb{P}(\mathcal{A}_2)\to 1$, since $\sigma_w^2>0$ by assumption.

$\mathcal{A}_3$: From (179) and (181), we can get

$$\sqrt{\frac{\|\mathring{\boldsymbol{v}}\|^2}{n}+\sigma_w^2}-\frac{\boldsymbol{h}^\top\mathring{\boldsymbol{v}}}{n}\stackrel{a.s.}{\to}\sqrt{\frac{1}{\delta}\mathbb{E}[B-\eta(B+\sigma_*H)]^2+\sigma_w^2}-\frac{\sigma_*\mathbb{E}\eta'(B+\sigma_*H)}{\delta}=\theta_*, \tag{94}$$

where the last step follows from (151). From Lemma 14, there exists $\theta_{\min} > 0$ such that $\theta_* \geq \theta_{\min}$. Therefore, $\mathbb{P}(\mathcal{A}_3) \to 1$.

$\mathcal{A}_4$: From the definition of $\mathcal{F}(\sigma, \theta)$ in (163),

$$\begin{aligned}\mathcal{F}(\sigma_*, \theta_*) &= \frac{\theta_*}{2\sigma_*}\mathbb{E}[(Y_*) - \eta(Y_*)]^2 + \int_0^1 F_\Lambda^{-1}(u) F_{|\eta(Y_*)|}^{-1}(u) du. \\ &= \frac{\theta_*}{2\sigma_*}\Big[\mathbb{E}\big(\eta(Y_*) - B\big)^2 - 2(\sigma_*)^2 \mathbb{E}\big(\eta'(Y_*)\big)\Big] + \frac{\theta_* \sigma_*}{2} \\ &\quad + \int_0^1 F_\Lambda^{-1}(u) F_{|\eta(Y_*)|}^{-1}(u) du \\ &= \frac{\theta_* \delta}{2\sigma_*}\big(- (\sigma_*)^2 - \sigma_w^2 + 2\sigma_* \sigma_w\big) + \int_0^1 F_\Lambda^{-1}(u) F_{|\eta(Y_*)|}^{-1}(u) du + \frac{\theta_* \sigma_*}{2}, \end{aligned} \tag{95}$$

where in the last step we use (151). Then substituting (95) into (55), we get

$$\Psi_* = \Psi(\sigma_*, \theta_*) = \frac{(\theta_*)^2}{2} + \int_0^1 F_\Lambda^{-1}(u) F_{|\eta(Y_*)|}^{-1}(u) du. \tag{96}$$

On the other hand,

$$\begin{aligned}\widetilde{L}(\mathring{\boldsymbol{v}}) &= \frac{1}{2}\left(\sqrt{\frac{\|\mathring{\boldsymbol{v}}\|^2}{n} + \sigma_w^2} - \frac{\boldsymbol{h}^\top \mathring{\boldsymbol{v}}}{n}\right)_+^2 + \frac{J_{\boldsymbol{\lambda}}(\mathring{\boldsymbol{v}} + \boldsymbol{\beta})}{n} \\ &\overset{a.s.}{\to} \frac{(\theta_*)^2}{2} + \int_0^1 F_\Lambda^{-1}(u) F_{|\eta(Y_*)|}^{-1}(u) du. \end{aligned} \tag{97}$$

where we use (94). From (96) and (97), we have $\widetilde{L}(\mathring{\boldsymbol{v}}) \overset{a.s.}{\to} \Psi_*$. Therefore, $\mathbb{P}(\mathcal{A}_4) \to 1$ for any $\varepsilon > 0$.

$\mathcal{A}_5$: From (161) and strong law of large number for triangular array [58, Theorem 2.1], we have $\mathcal{F}_p(\sigma, \theta_*) - \frac{\sigma\theta_*\|\boldsymbol{h}\|^2}{2p} \overset{a.s.}{\to} \mathcal{F}(\sigma, \theta_*) - \frac{\sigma\theta_*}{2}$ for any $\sigma \in [\sigma_w, \sqrt{\sigma_w^2 + K^2}]$. From (167) in Lemma 16, we know $\mathcal{F}(\cdot, \theta_*)$ is Lipschitz continuous on $[\sigma_w, \sqrt{\sigma_w^2 + K^2}]$. This indicates $\sigma \mapsto \mathcal{F}(\sigma, \theta_*) - \frac{\sigma\theta_*}{2}$ is also Lipschitz continuous on $[\sigma_w, \sqrt{\sigma_w^2 + K^2}]$. On the other hand, in the proof of Lemma 16 [line below (174)], we show $\mathcal{F}_p(\cdot, \theta_*)$ is continuously differentiable on $[\sigma_w, \infty)$ with derivative satisfying $\left|\frac{\partial \mathcal{F}_p(\sigma, \theta_*)}{\partial \sigma}\right| \leq \frac{3\theta_*\|\boldsymbol{\beta}\|^2 + \sigma^2(2+\theta_*)\|\boldsymbol{h}\|^2}{\sigma^2 p}$. Then by [58, Theorem 2.1] again and the fact that $\{\boldsymbol{\beta}\}_{p \in \mathbb{Z}^+}$ is a converging sequence, we have almost surely $\limsup_{p\to\infty}\left|\frac{\partial \mathcal{F}_p(\sigma, \theta_*)}{\partial \sigma} - \frac{\theta_*\|\boldsymbol{h}\|^2}{2p}\right| \leq C$ for any $\sigma \in [\sigma_w, \sqrt{\sigma_w^2 + K^2}]$, where $C > 0$ is some constant. This indicates that almost surely, $\sigma \mapsto \frac{\partial \mathcal{F}_p(\sigma, \theta_*)}{\partial \sigma} - \frac{\theta_*\|\boldsymbol{h}\|^2}{2p}$ is C-Lipschitz continuous on $[\sigma_w, \sqrt{\sigma_w^2 + K^2}]$ for any large enough p. Then by the same epsilon net argument as in the proof of Lemma 16, we can show

$$\sup_{\sigma \in [\sigma_w, \sqrt{\sigma_w^2 + K^2}]}\left|\big(\mathcal{F}_p(\sigma, \theta_*) - \tfrac{\sigma\theta_*\|\boldsymbol{h}\|^2}{2p}\big) - \big(\mathcal{F}(\sigma, \theta_*) - \tfrac{\sigma\theta_*}{2}\big)\right| \overset{a.s.}{\to} 0.$$

Therefore, $\mathbb{P}(\mathcal{A}_5) \to 1$ for any $\varepsilon > 0$.

Now we are ready to start the deterministic analysis conditioned on the event $\mathcal{A}$. It is more convenient to work with $\widetilde{L}(\boldsymbol{v})$ than $L(\boldsymbol{v})$, since it is locally strongly convex (the precise meaning will be given below). In the sequel,

we will start by studying the limiting properties of $\widetilde{L}(\boldsymbol{v})$ and then associate them to $L(\boldsymbol{v})$, by showing $L(\boldsymbol{v})$ can be well-approximated by $\widetilde{L}(\boldsymbol{v})$ as $p\to\infty$. For $\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}$, $\widetilde{L}(\boldsymbol{v})$ can be equivalently written as:

$$\begin{aligned}\widetilde{L}(\boldsymbol{v}) &= \max_{\theta\geq 0}\theta\Big(\sqrt{\frac{\|\boldsymbol{v}\|^2}{n}+\sigma_w^2}-\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\Big)-\frac{\theta^2}{2}+\frac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n}\\ &= \max_{\theta\geq 0}\theta\Big(\min_{\sigma\in[\sigma_w,\sqrt{\sigma_w^2+K^2}]}\big\{\frac{\sigma}{2}+\frac{\|\boldsymbol{v}\|^2/n+\sigma_w^2}{2\sigma}\big\}-\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\Big)-\frac{\theta^2}{2}+\frac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n}\\ &= \max_{\theta\geq 0}\min_{\sigma\in[\sigma_w,\sqrt{\sigma_w^2+K^2}]}\frac{\theta}{2}\Big(\frac{\sigma_w^2}{\sigma}+\sigma\Big)-\frac{\theta^2}{2}+\frac{1}{n}\Big[\frac{\theta\|\boldsymbol{v}\|^2}{2\sigma}-\theta\boldsymbol{h}^\top\boldsymbol{v}+J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})\Big]. \end{aligned}\tag{98}$$

Therefore,

$$\begin{aligned}\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v}) &\overset{\text{(a)}}{\geq} \min_{\sigma\in[\sigma_w,\sqrt{\sigma_w^2+K^2}]}\frac{\theta_*}{2}\Big(\frac{\sigma_w^2}{\sigma}+\sigma\Big)-\frac{(\theta_*)^2}{2}+\frac{1}{\delta}\Big[\mathcal{F}_p(\sigma,\theta_*)-\frac{\sigma\theta_*\|\boldsymbol{h}\|^2}{2p}\Big],\\ &\overset{\text{(b)}}{\geq} \min_{\sigma\in[\sigma_w,\sqrt{\sigma_w^2+K^2}]}\Psi(\sigma,\theta_*)-\varepsilon\\ &\geq \Psi_*-\varepsilon\\ &\overset{\text{(c)}}{\geq} \widetilde{L}(\mathring{\boldsymbol{v}})-2\varepsilon, \end{aligned}\tag{99}$$

where (a) follows from (98) and (160), (b) is due to $\mathcal{A}_5$ and (c) is due to $\mathcal{A}_4$. Besides, since $\frac{\|\mathring{\boldsymbol{v}}\|_2}{\sqrt{n}}\leq\frac{\sigma_*}{\sqrt{\delta}}$ under $\mathcal{A}_2$ and $\frac{\sigma_*}{\sqrt{\delta}}\leq K-\frac{\theta_{\min}}{4}$ by assumption, we have $\frac{\|\mathring{\boldsymbol{v}}\|_2}{\sqrt{n}}\leq K$ and thus $\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})\leq\widetilde{L}(\mathring{\boldsymbol{v}})\leq\Psi_*+\varepsilon$. Therefore, combining it with (99) yields

$$\Big|\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})-\Psi_*\Big|\leq 2\varepsilon.\tag{100}$$

On the other hand, $\boldsymbol{v}\mapsto\sqrt{\frac{\|\boldsymbol{v}\|^2}{n}+\sigma_w^2}-\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}$ is $\frac{1}{\sqrt{n}}(\sqrt{\frac{3}{2\delta}}+1)$-Lipschitz continuous under $\mathcal{A}_1$ and $\sqrt{\frac{\|\mathring{\boldsymbol{v}}\|^2}{n}+\sigma_w^2}-\frac{\boldsymbol{h}^\top\mathring{\boldsymbol{v}}}{n}\geq\frac{\theta_{\min}}{2}$ under $\mathcal{A}_3$. Therefore, for $r=(\sqrt{\frac{3}{2\delta}}+1)^{-1}\frac{\theta_{\min}}{4}$, $\sqrt{\frac{\|\boldsymbol{v}\|^2}{n}+\sigma_w^2}-\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}\geq\frac{\theta_{\min}}{4}$ for all $\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}r}(\mathring{\boldsymbol{v}})$. Then using Lemma F.14 in [50] and $\frac{\|\mathring{\boldsymbol{v}}\|_2}{\sqrt{n}}+r\leq\frac{\sigma_*}{\sqrt{n}}+r<K$ (hence $\mathcal{B}_{\sqrt{n}r}(\mathring{\boldsymbol{v}})\subset\mathcal{B}_{\sqrt{n}K}$) under $\mathcal{A}_2$, we can show $\widetilde{L}(\boldsymbol{v})$ is $\frac{\gamma}{n}$-strongly convex on $\mathcal{B}_{\sqrt{n}r}(\mathring{\boldsymbol{v}})$, where $\gamma=\frac{\theta_{\min}\sigma_w^2}{4(K^2+\sigma_w^2)^{3/2}}$. In other words, $\widetilde{L}(\boldsymbol{v})$ is locally strongly convex in $\mathcal{B}_{\sqrt{n}r}(\mathring{\boldsymbol{v}})$. Also since $\mathcal{B}_{\sqrt{n}r}(\mathring{\boldsymbol{v}})\subset\mathcal{B}_{\sqrt{n}K}$, together with (99) we have $\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}r}(\mathring{\boldsymbol{v}})}\widetilde{L}(\boldsymbol{v})\geq\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})\geq\widetilde{L}(\mathring{\boldsymbol{v}})-2\varepsilon$. By Lemma B.1 in [50] we know if $0<2\varepsilon\leq\frac{(\sqrt{n}r)^2\gamma}{8n}$, then $\|\tilde{\boldsymbol{v}}-\mathring{\boldsymbol{v}}\|^2\leq\frac{4n\varepsilon}{\gamma}\leq\frac{nr^2}{4}$, where $\tilde{\boldsymbol{v}}=\operatorname*{argmin}_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})$. Moreover, for any $\boldsymbol{v}\in\mathcal{B}^o_{4\sqrt{n\varepsilon/\gamma}}(\mathring{\boldsymbol{v}})$, we have $\widetilde{L}(\boldsymbol{v})\geq\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})+2\varepsilon$. This implies

$$\min_{\boldsymbol{v}\in\mathcal{B}^o_{4\sqrt{n\varepsilon/\gamma}}(\mathring{\boldsymbol{v}})\cap\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})\geq\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\widetilde{L}(\boldsymbol{v})+2\varepsilon.\tag{101}$$

Finally, we show $L(\boldsymbol{v})$ is well-approximated by $\widetilde{L}(\boldsymbol{v})$ under event $\mathcal{A}_1$. Note that $L(\boldsymbol{v})-\widetilde{L}(\boldsymbol{v})=g(\Delta)$, where $g(t)=\frac{1}{2}[(\sqrt{x+t}-y)_+^2-(\sqrt{x}-y)_+^2]$, with $x=\frac{\|\boldsymbol{v}\|^2}{n}+\sigma_w^2$, $y=\frac{\boldsymbol{h}^\top\boldsymbol{v}}{n}$ and $\Delta=\frac{\|\boldsymbol{v}\|^2}{n}\Big(\frac{\|\boldsymbol{g}\|^2}{n}-1\Big)+\Big(\frac{\|\boldsymbol{w}\|^2}{n}-\sigma_w^2\Big)+2\frac{\|\boldsymbol{v}\|}{\sqrt{n}}\frac{\boldsymbol{g}^\top\boldsymbol{w}}{n}$. Also it is not hard to show under event $\mathcal{A}_1$, $|g(t)|\leq\frac{|t|}{2}\Big(1+\frac{|y|}{\sigma_w}\Big)$, $|y|\leq\sqrt{\frac{3}{2\delta}}K$ and $|\Delta|\leq(K^2+\sigma_w^2+2K)\varsigma$ for any $\boldsymbol{v}\in\mathcal{S}_{\boldsymbol{v}}(K)$. Therefore,

$$\sup_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}|L(\boldsymbol{v})-\widetilde{L}(\boldsymbol{v})|\leq\underbrace{\tfrac{K^2+\sigma_w^2+2K}{2}\Big(1+\sqrt{\tfrac{3}{2\delta}}\tfrac{K}{\sigma_w}\Big)}_{:=C_K}\varsigma\tag{102}$$

and thus

$$\Big|\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) - \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \widetilde{L}(\boldsymbol{v})\Big| \le C_K\varsigma. \tag{103}$$

Now we are ready to turn back to $L(\boldsymbol{v})$ to show (88) and (89). Substituting (102) and (103) into (100) and (101) gives

$$\Big|\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) - \Psi_*\Big| \le C_K\varsigma + \varepsilon \tag{104}$$

and

$$\begin{aligned}\min_{\boldsymbol{v}\in\mathcal{B}^o_{4\sqrt{n\varepsilon/\gamma}}(\mathring{\boldsymbol{v}})\cap\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) &\ge \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) + 2(\varepsilon - C_K\varsigma).\\ &\overset{\text{(a)}}{\ge} \Psi_* + \varepsilon - 3C_K\varsigma,\end{aligned} \tag{105}$$

where in (a) we use (104). For any $\varepsilon > 0$, choose $\varsigma \le \min\big\{\frac{1}{2}, \frac{\varepsilon}{6C_K}\big\}$ in (104) and (105). Then (88) and (89) immediately follows, since $\mathbb{P}(\mathcal{A}) \to 1$. ■

The second lemma is on the asymptotic empirical distribution of the optimal solution of auxiliary problem.

Lemma 11: Let $\psi(\cdot,\cdot)$ be a pseudo-Lipschitz function with constant L and $D_\nu := \big\{\boldsymbol{v}\in\mathbb{R}^p : |\mathbb{E}_{\mu_{\boldsymbol{v}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu^*}\psi| \ge \nu\big\}$ as defined in (56). Then for any $K \ge \frac{\sigma_*}{\sqrt{\delta}} + \frac{\theta_{\min}}{4}$, $\nu > 0$ and $\varepsilon \le \frac{\gamma\nu^2}{192L^2(1+2\delta K^2+32(\mathbb{E}B^2+\sigma_w^2))\delta}$,

$$\mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu \bigcap \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \le \Psi_* + 2\varepsilon\big) \to 0, \tag{106}$$

where Ψ_* is defined in (55).

Proof: For any $\nu > 0$, we will consider the following event:

$$\mathcal{E} = \big\{|\mathbb{E}_{\mu_{\mathring{\boldsymbol{v}}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu^*}\psi| \le \tfrac{\nu}{2}\big\}\bigcap\big\{\tfrac{1}{p}\|\mathring{\boldsymbol{v}}\|^2, \tfrac{1}{p}\|\boldsymbol{\beta}\|^2 \le 4(\mathbb{E}B^2+\sigma_w^2)\big\},$$

where $\mathring{\boldsymbol{v}} := \eta(\boldsymbol{\beta}+\sigma_*\boldsymbol{h}) - \boldsymbol{\beta}$ is the same as in (90) and μ^* is the joint measure of $\big(\eta(B+\sigma_*H), B\big)$, with $\eta(\cdot) := \eta(\cdot;\mu_{Y_*},\mu_{\sigma_*\Lambda/\theta_*})$ and $H\sim\mathcal{N}(0,1)$ independent of $B\sim\mu_B$.

We first show $\mathbb{P}(\mathcal{E})\to 1$, as $p\to\infty$. From (164) in the proof of Lemma 15, we have $W_2\big(\mu_{\boldsymbol{h},\boldsymbol{\beta}}, H\otimes B\big) \overset{a.s.}{\to} 0$, with $H\sim\mathcal{N}(0,1)$ and $B\sim\mu_B$. Meanwhile, $(h,b)\mapsto\big(\eta(b+\sigma_*h), b\big)$ is a $\sqrt{3+2(\sigma_*)^2}$-Lipschitz continuous mapping. Hence similar as (165), $W_2\big(\mu_{\mathring{\boldsymbol{v}}+\boldsymbol{\beta},\boldsymbol{\beta}}, \mu^*\big) \overset{a.s.}{\to} 0$ and by Theorem 7.12 (iv) in [56], $\mathbb{E}_{\mu_{\mathring{\boldsymbol{v}}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi \overset{a.s.}{\to} \mathbb{E}_{\mu^*}\psi$. Similarly, we can show

$$\tfrac{1}{p}\|\mathring{\boldsymbol{v}}\|^2 \overset{a.s.}{\to} \mathbb{E}\big[\eta(B+\sigma_*H) - B\big]^2 < 4(\mathbb{E}B^2+\sigma_w^2).$$

Also since $\{\boldsymbol{\beta}\}_{p\in\mathbb{N}}$ is a converging sequence, $\frac{1}{p}\|\boldsymbol{\beta}\|^2\to\mathbb{E}B^2$. As a result, $\mathbb{P}(\mathcal{E})\to 1$ for any $\nu>0$.

Next we show conditioned on $\mathcal{E}$, it holds that

$$D_\nu\bigcap\mathcal{B}_{\sqrt{n}K} \subseteq \mathcal{B}^o_{\sqrt{n}\varepsilon_0}(\mathring{\boldsymbol{v}})\bigcap\mathcal{B}_{\sqrt{n}K}, \tag{107}$$

where $K \geq \frac{\sigma_*}{\sqrt{\delta}} + \frac{\theta_{\min}}{4}$, $D_\nu := \{\boldsymbol{v} \in \mathbb{R}^p : |\mathbb{E}_{\mu_{\boldsymbol{v}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu^*}\psi| \geq \nu\}$ and $\varepsilon_0^2 = \frac{\nu^2}{3L^2(1+2\delta K^2+32(\mathbb{E}B^2+\sigma_w^2))\delta}$. Since ψ is L pseudo-Lipschitz, we can get

$$\begin{aligned}
|\mathbb{E}_{\mu_{\boldsymbol{v}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu_{\mathring{\boldsymbol{v}}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi| &= \big|\tfrac{1}{p}\sum_{i=1}^p \psi(v_i+\beta_i,\beta_i) - \tfrac{1}{p}\sum_{i=1}^p \psi(\mathring{v}_i+\beta_i,\beta_i)\big| \\
&\leq \tfrac{L}{p}\sum_{i=1}^p \big(1+\sqrt{(v_i+\beta_i)^2+\beta_i^2}+\sqrt{(\mathring{v}_i+\beta_i)^2+\beta_i^2}\big)\big|v_i-\mathring{v}_i\big| \\
&\overset{(a)}{\leq} \tfrac{L}{p}\big[3(p+2\|\boldsymbol{v}\|^2+2\|\mathring{\boldsymbol{v}}\|^2+6\|\boldsymbol{\beta}\|^2)\big]^{1/2}\|\boldsymbol{v}-\mathring{\boldsymbol{v}}\|, \qquad (108)
\end{aligned}$$

where in (a) we use Cauchy-Swartz inequality and $1+\sqrt{x}+\sqrt{y} \leq \sqrt{3(1+x+y)}$. Meanwhile, conditioned on event $\mathcal{E}$, if $\boldsymbol{v} \in D_\nu$, then

$$|\mathbb{E}_{\mu_{\boldsymbol{v}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu_{\mathring{\boldsymbol{v}}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi| \geq \big||\mathbb{E}_{\mu_{\boldsymbol{v}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu^*}\psi| - |\mathbb{E}_{\mu_{\mathring{\boldsymbol{v}}+\boldsymbol{\beta},\boldsymbol{\beta}}}\psi - \mathbb{E}_{\mu^*}\psi|\big| \geq \tfrac{\nu}{2}. \qquad (109)$$

Combining (108) and (109), we know conditioned on event $\mathcal{E}$, $\frac{1}{n}\|\boldsymbol{v}-\mathring{\boldsymbol{v}}\|^2 \geq \frac{\nu^2}{3L^2(1+2\delta K^2+32(\mathbb{E}B^2+\sigma_w^2))\delta} = \varepsilon_0^2$ for any $\boldsymbol{v} \in D_\nu \bigcap \mathcal{B}_{\sqrt{n}K}$.

From (107), we have

$$\mathbb{P}\big(\min_{\boldsymbol{v}\in D_\nu \bigcap \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) \leq \mathbb{P}\big(\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}\varepsilon_0}(\mathring{\boldsymbol{v}}) \bigcap \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) + \mathbb{P}(\mathcal{E}^C). \qquad (110)$$

On the other hand, from (89) in Lemma 10 we have

$$\mathbb{P}\big(\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}\varepsilon_0}(\mathring{\boldsymbol{v}}) \bigcap \mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \leq \Psi_* + 2\varepsilon\big) \to 0, \qquad (111)$$

if $\sqrt{n}\varepsilon_0 \geq 8\sqrt{\frac{n\varepsilon}{\gamma}}$. Therefore, for any $\nu > 0$ we can choose $\varepsilon \leq \frac{\gamma\nu^2}{192L^2(1+2\delta K^2+32(\mathbb{E}B^2+\sigma_w^2))\delta}$ and from (110) and (111) we can get (106). ■

The last lemma in this section shows that the optimal solution of the original problem is bounded with probability converging to 1.

Lemma 12: For $K \geq \frac{\sigma_*}{\sqrt{\delta}} + \frac{\theta_{\min}}{4}$, we have as $p \to \infty$,

$$\mathbb{P}(\hat{\boldsymbol{v}} \notin \mathcal{B}_{\sqrt{n}K}) \to 0. \qquad (112)$$

Proof: To show $\hat{\boldsymbol{v}} = \underset{\boldsymbol{v}}{\operatorname{argmin}}\, C(\boldsymbol{v})$ is bounded with probability approaching 1, we use the following property: for any $a \leq K$,

$$\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}a} \bigcap \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) > \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon \Rightarrow \min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}a}} C(\boldsymbol{v}) > \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon. \qquad (113)$$

One can prove (113) by contradiction. If $\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}a}} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v})+\varepsilon$, it must hold that $\min_{\boldsymbol{v}\in \mathcal{B}^C_{\sqrt{n}K}} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon$. Then by convexity of $C(\boldsymbol{v})$, we can always find $\boldsymbol{v}_0 \in \mathcal{B}^o_{\sqrt{n}a} \bigcap \mathcal{B}_{\sqrt{n}K}$ such that $C(\boldsymbol{v}_0) \leq \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon$, which leads to a contradiction with $\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}a} \bigcap \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) > \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon$. As a result, for any $a \leq K$,

$$\begin{aligned}
\mathbb{P}\Big(\min_{\boldsymbol{v}\in \mathcal{B}^C_{\sqrt{n}K}} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon\Big) &\leq \mathbb{P}\Big(\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}a}} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon\Big) \\
&\overset{(a)}{\leq} \mathbb{P}\Big(\min_{\boldsymbol{v}\in \mathcal{B}^o_{\sqrt{n}a} \bigcap \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) \leq \min_{\boldsymbol{v}\in \mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon\Big), \qquad (114)
\end{aligned}$$

where (a) follows from (113). Now we choose $a \in \big(\frac{\sigma_*}{\sqrt{\delta}}, K\big)$. Then in the probability space of auxiliary problem (52), under event $\mathcal{A}_2$ [c.f. (91)] we can get $\mathcal{B}^o_{\sqrt{n}a} \subset \mathcal{B}^o_{\sqrt{n}\Delta_a}(\mathring{\boldsymbol{v}})$, with $\Delta_a = a - \frac{\sigma_*}{\sqrt{\delta}}$. Therefore, for any $\varepsilon > 0$ we have

$$\begin{aligned}
&\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{n}a}\bigcap\mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) \le \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon\Big)\\
\le&\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{n}a}\bigcap\mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) \le \Psi_* + 2\varepsilon\Big) + \mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) \ge \Psi_* + \varepsilon\Big)\\
\overset{\text{(a)}}{\le}&\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{n}a}\bigcap\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \le \Psi_* + 2\varepsilon\Big) + \mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \ge \Psi_* + \varepsilon\Big)\\
\le&\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{n}\Delta_a}(\mathring{\boldsymbol{v}})\bigcap\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \le \Psi_* + 2\varepsilon\Big) + \mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \ge \Psi_* + \varepsilon\Big) + \mathbb{P}(\mathcal{A}_2^C),
\end{aligned} \tag{115}$$

where Ψ_* is defined in (55), in (a) we use (53) and (54). Combining (114) and (115), we have for any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}(\hat{\boldsymbol{v}} \notin \mathcal{B}_{\sqrt{n}K}) \le&\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^C_{\sqrt{n}K}} C(\boldsymbol{v}) \le \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} C(\boldsymbol{v}) + \varepsilon\Big)\\
\le&\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{n}\Delta_a}(\mathring{\boldsymbol{v}})\bigcap\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \le \Psi_* + 2\varepsilon\Big) + \mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) > \Psi_* + \varepsilon\Big) + \mathbb{P}(\mathcal{A}_2^C).
\end{aligned} \tag{116}$$

It remains to show all the three terms on the RHS of (116) converge to 0 for some $\varepsilon > 0$. From (89), we know for $\varepsilon > 0$ if we choose an a such that $\sqrt{n}\Delta_a \ge 8\sqrt{\frac{n\varepsilon}{\gamma}}$, where $\gamma = \frac{\theta_{\min}\sigma_w^2}{4(K^2+\sigma_w^2)^{3/2}}$, then

$$\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}^o_{\sqrt{n}\Delta_a}(\mathring{\boldsymbol{v}})\bigcap\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) \le \Psi_* + 2\varepsilon\Big) \to 0. \tag{117}$$

Clearly, such a always exists if $\varepsilon \in \big(0, \frac{\gamma^2\theta^2_{\min}}{1024}\big)$ since $K \ge \frac{\sigma_*}{\sqrt{\delta}} + \frac{\theta_{\min}}{4}$. On the other hand, in the proof of Lemma 10 we show $\mathbb{P}(\mathcal{A}_2^C) \to 0$ and from (88) we have for any $\varepsilon > 0$, $\mathbb{P}\Big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) > \Psi_* + \varepsilon\Big) \to 0$. Substituting these results back to (116), we reach (112). ■

E. Proof of Lemma 3

To obtain the probabilistic upper bound for $R_0^{(p)}$, we can approximate indicator function $\mathbb{I}_{x=0}$ by a series of envelope functions:

$$\psi_h(x) = \begin{cases} 1 - h^{-1}x & 0 \le x < h,\\ 1 + h^{-1}x & -h \le x < 0,\\ 0 & |x| \ge h, \end{cases} \tag{118}$$

where $h > 0$. We can see $\psi_h(x)$ is an upper bound of $\mathbb{I}_{x=0}$ and satisfies $0 \le \psi_h(x) - \mathbb{I}_{x=0} \le \mathbb{I}_{0<|x|<h}$, so

$$\begin{aligned}
R_0^{(p)} - \mathbb{P}\big[\eta(Y_*) = 0\big] &= \mathbb{E}_{\mu_{\hat{\boldsymbol{\beta}}}}\mathbb{I}_{x=0} - \mathbb{E}_{\mu_{\hat{B}}}\mathbb{I}_{x=0}\\
&\le \mathbb{E}_{\mu_{\hat{\boldsymbol{\beta}}}}\psi_h(x) - \mathbb{E}_{\mu_{\hat{B}}}\psi_h(x) + \mathbb{E}_{\mu_{\hat{B}}}\psi_h(x) - \mathbb{E}_{\mu_{\hat{B}}}\mathbb{I}_{x=0}\\
&\le \big|\mathbb{E}_{\mu_{\hat{\boldsymbol{\beta}}}}\psi_h(x) - \mathbb{E}_{\mu_{\hat{B}}}\psi_h(x)\big| + \mathbb{P}\big[|\eta(Y_*)| \in (0,h)\big],
\end{aligned} \tag{119}$$

where $\mu_{\hat{B}}$ denotes the distribution of $\eta(Y_*)$. Moreover, $\psi_h(x)$ is h^{-1}-Lipschitz (and hence pseudo-Lipschitz by definition), so (12) can now be applied, which gives us $\big|\mathbb{E}_{\mu_{\hat{\boldsymbol{\beta}}}}\psi_h(x) - \mathbb{E}_{\mu_{\hat{B}}}\psi_h(x)\big| \overset{\mathbb{P}}{\to} 0$ for any fixed $h > 0$.

Meanwhile, by continuity of probability, we have $\lim_{h\to 0}\mathbb{P}\big(|\eta(Y_*)|\in(0,h)\big)=0$. As a result, on both sides of (119) taking $p\to\infty$ and then $h\to 0$, we get for any $\varepsilon>0$,

$$R_0^{(p)}\le \mathbb{P}\big(\eta(Y_*)=0\big)+\varepsilon \tag{120}$$

with probability approaching 1 as $p\to\infty$.

F. Proof of Lemma 4

If $\mathbb{P}\big(\eta(Y_*)=0\big)=0$, then since $R_0^{(p)}\ge 0$, (61) trivially holds. Thus, it only remains to address the case when $\mathbb{P}\big(\eta(Y_*)=0\big)>0$. Towards this end, we will utilize Lemma 5. To apply Lemma 5, we need to verify for sufficiently large $k\in[p]$, $|\hat{\boldsymbol{s}}|_{(1:k)}\prec_S \boldsymbol{\lambda}_{1:k}$ with probability approaching 1. For any $p,k,\ell\in\mathbb{Z}^+$, with $0\le k<\ell\le p$,

$$\begin{aligned}
\sum_{i=k+1}^{\ell}|\hat{s}|_{(i)}-\sum_{i=k+1}^{\ell}\lambda_i &= \int_{k/p}^{\ell/p}F^{-1}_{|\hat{\boldsymbol{s}}|}(u)du-\int_{k/p}^{\ell/p}F^{-1}_{|\boldsymbol{\lambda}|}(u)(u)du\\
&\le \int_{k/p}^{\ell/p}F^{-1}_{|\hat{S}|}(u)du-\int_{k/p}^{\ell/p}F^{-1}_{\Lambda}(u)du+\Big|\int_{k/p}^{\ell/p}F^{-1}_{|\hat{\boldsymbol{s}}|}(u)du-\int_{k/p}^{\ell/p}F^{-1}_{|\hat{S}|}(u)du\Big|\\
&\quad+\Big|\int_{k/p}^{\ell/p}F^{-1}_{\Lambda}(u)du-\int_{k/p}^{\ell/p}F^{-1}_{|\boldsymbol{\lambda}|}(u)du\Big|\\
&\le \int_{k/p}^{\ell/p}F^{-1}_{|\hat{S}|}(u)du-\int_{k/p}^{\ell/p}F^{-1}_{\Lambda}(u)du+W_2(\mu_{\hat{\boldsymbol{s}}},\mu_{\hat{S}})+W_2(\mu_{\boldsymbol{\lambda}},\mu_{\Lambda}),
\end{aligned}\tag{121}$$

where in the last step, we use (86). Here, $\mu_{\hat{S}}$ is the law of $\tau_*^{-1}[Y_*-\eta(Y_*)]$. By condition (R.1), we know for any $\varepsilon\in(0,q_0^*]$ (recall that $q_0^*=\mathbb{P}\big(\eta(Y_*)=0\big)$), there exists $\varsigma>0$ such that

$$\begin{aligned}
&\max_{t\in[0,q_0^*-\varepsilon]}\int_t^{q_0^*}F^{-1}_{|\hat{S}|}(u)du-\int_t^{q_0^*}F^{-1}_{\Lambda}(u)du\\
\overset{\text{(a)}}{=}&\tau_*^{-1}\max_{t\in[0,q_0^*-\varepsilon]}\int_t^{q_0^*}F^{-1}_{|Y_*|}(u)du-\int_t^{q_0^*}F^{-1}_{\tau_*\Lambda}(u)du\\
\overset{\text{(b)}}{\le}&-\tau_*^{-1}\varsigma,
\end{aligned}\tag{122}$$

where to reach (a), we use $\hat{S}\overset{\text{law}}{=}\tau_*^{-1}[Y_*-\eta(Y_*)]$ and the fact that $\eta(y)=0$ for $|y|\le F^{-1}_{|Y_*|}(q_0^*)$ and (b) is due to (R.1) and the fact that $t\mapsto\int_t^{q_0^*}F^{-1}_{|Y_*|}(u)du$ and $t\mapsto\int_t^{q_0^*}F^{-1}_{\tau_*\Lambda}(u)du$ are both continuous. For any fixed $\varepsilon\in(0,q_0^*]$, $\lfloor p(q_0^*-\varepsilon)\rfloor\le\lfloor pq_0^*\rfloor-1$ for large enough p. Then substituting (122) into (121) we can get for large enough p and any $0\le k\le\lfloor p(q_0^*-\varepsilon)\rfloor$,

$$\begin{aligned}
&\sum_{i=k+1}^{\lfloor pq_0^*\rfloor}|\hat{s}|_{(i)}-\sum_{i=k+1}^{\lfloor pq_0^*\rfloor}\lambda_i\\
\overset{\text{(a)}}{\le}&\int_{k/p}^{q_0^*}F^{-1}_{|\hat{S}|}(u)du-\int_{k/p}^{q_0^*}F^{-1}_{\Lambda}(u)du-\Big[\int_{\lfloor pq_0^*\rfloor/p}^{q_0^*}F^{-1}_{|\hat{S}|}(u)du-\int_{\lfloor pq_0^*\rfloor/p}^{q_0^*}F^{-1}_{\Lambda}(u)du\Big]\\
&+W_2(\mu_{\hat{\boldsymbol{s}}},\mu_{\hat{S}})+W_2(\mu_{\boldsymbol{\lambda}},\mu_{\Lambda})\\
\overset{\text{(b)}}{\le}&-\tau_*^{-1}\varsigma+\tau_*^{-1}\Big(\int_{\lfloor pq_0^*\rfloor/p}^{q_0^*}F^{-1}_{|Y_*|}(u)du+\int_{\lfloor pq_0^*\rfloor/p}^{q_0^*}F^{-1}_{\tau_*\Lambda}(u)du\Big)+W_2(\mu_{\hat{\boldsymbol{s}}},\mu_{\hat{S}})+W_2(\mu_{\boldsymbol{\lambda}},\mu_{\Lambda}),
\end{aligned}\tag{123}$$

where (a) follows from (121) and (b) follows from (122). We now show the last four terms in (123) vanish as $p\to\infty$: $\int_{\lfloor pq_0^*\rfloor/p}^{q_0^*}F^{-1}_{|Y_*|}(u)du$ and $\int_{\lfloor pq_0^*\rfloor/p}^{q_0^*}F^{-1}_{\tau_*\Lambda}(u)du$ converges to 0 by DCT; $W_2(\mu_{\boldsymbol{s}},\mu_S)\overset{\mathbb{P}}{\to}0$ is proved in

Proposition 6; $W_2(\mu_{\boldsymbol{\lambda}}, \mu_\Lambda) \to 0$ since $\{\boldsymbol{\lambda}\}_{p\in\mathbb{Z}^+}$ is a converging sequence. As a result, for any fixed $\varepsilon \in (0, q_0^*]$, there exists $\varepsilon > 0$ such that

$$\max_{0\le k\le \lfloor p(q_0^*-\varepsilon)\rfloor} \sum_{i=k+1}^{\lfloor pq_0^*\rfloor} |\hat{s}|_{(i)} - \sum_{i=k+1}^{\lfloor pq_0^*\rfloor} \lambda_i \le -\varsigma, \tag{124}$$

with probability approaching 1, as $p \to \infty$. Now we show conditioned on (124), there exists $k_0 \in [\lfloor p(q_0^* - \varepsilon)\rfloor, \lfloor pq_0^*\rfloor - 1]$ such that

$$\max_{0\le k\le k_0} \sum_{i=k+1}^{k_0+1} |\hat{s}|_{(i)} - \sum_{i=k+1}^{k_0+1} \lambda_i < 0. \tag{125}$$

In other words, $|\hat{\boldsymbol{s}}|_{(1:k_0+1)} \prec_S \boldsymbol{\lambda}_{1:k_0+1}$. Such k_0 can be retrieved as follows:

Step 0. Let k_s denote the candidate for k_0 and initialize $k_s = \lfloor p(q_0^* - \varepsilon)\rfloor$. From (124) we know at the initial step,

$$\max_{0\le k\le k_s} \sum_{i=k+1}^{\lfloor pq_0^*\rfloor} |\hat{s}|_{(i)} - \sum_{i=k+1}^{\lfloor pq_0^*\rfloor} \lambda_i \le -\varsigma. \tag{126}$$

Step 1. If $k_s + 1 = \lfloor pq_0^*\rfloor$, then we output $k_0 = k_s$; otherwise we go to step 2.

Step 2. If $\sum_{i=k_s+2}^{\lfloor pq_0^*\rfloor} |\hat{s}|_{(i)} - \sum_{i=k_s+2}^{\lfloor pq_0^*\rfloor} \lambda_i > -\frac{\varsigma}{2}$, then together with (126) we get $\max_{0\le k\le k_s} \sum_{i=k+1}^{k_s+1} |\hat{s}|_{(i)} - \sum_{i=k+1}^{k_s+1} \lambda_i \le -\frac{\varsigma}{2}$. Hence, we output $k_0 = k_s$; otherwise, we update k_s and ς as: $k_s \leftarrow k_s + 1$, $\varsigma \leftarrow \frac{\varsigma}{2}$ and return back to step 1. Clearly, (126) still holds under the updated k_s and ς.

Then from Lemma 5, (125) implies that $|\widehat{\boldsymbol{\beta}}|_{(1:k_0+1)} = 0$ and thus $R_0^{(p)} \ge q_0^* - \varepsilon$ since $k_0 \ge p(q_0^* - \varepsilon) - 1$ by construction. Summing up, if $q_0^* > 0$ then for any $\varepsilon \in (0, q_0^*]$, $R_0^{(p)} \ge q_0^* - \varepsilon = \mathbb{P}(\eta(Y_*) = 0) - \varepsilon$ with probability approaching 1 as $p \to \infty$. Therefore (61) is verified.

G. Proof of Lemma 5

The key is to establish the following result: for any p-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{\lambda}$ with $0 \le \lambda_1 \le \cdots \le \lambda_p$, if $|\boldsymbol{a} - \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(1:k)} \prec_S \boldsymbol{\lambda}_{1:k}$ for some $k \in [p]$, then $|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(1:k)} = \boldsymbol{0}$. To prove this, it suffices to show $|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(k)} = 0$. Assume $|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(k)} \neq 0$. Define the index set $I_k \stackrel{\text{def}}{=} \{i \mid |\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(i)} = |\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(k)}\}$ and denote $\ell := \min I_k$ and $m := \max I_k$. According to the formula for $\partial J_{\boldsymbol{\lambda}}$ [47, Fact V.3], we have: if $|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(k)} \neq 0$, then for any $\boldsymbol{g} \in \partial J_{\boldsymbol{\lambda}}\big(\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\big)$ it holds that $|\boldsymbol{g}|_{(\ell:m)} \prec \boldsymbol{\lambda}_{\ell:m}$ and $\sum_{i=\ell}^m |\boldsymbol{g}|_{(i)} = \sum_{i=\ell}^m \lambda_i$. On the other hand, by the first order optimality condition, $\boldsymbol{a} - \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a}) \in \partial J_{\boldsymbol{\lambda}}\big(\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\big)$. Hence, $|\boldsymbol{a} - \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(\ell:m)} \prec \boldsymbol{\lambda}_{\ell:m}$, *i.e.*, for any $\ell \le q \le m$,

$$\sum_{i=q}^m |\boldsymbol{a} - \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(i)} \le \sum_{i=q}^m \lambda_i \tag{127}$$

and also

$$\sum_{i=\ell}^m |\boldsymbol{a} - \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(i)} = \sum_{i=\ell}^m \lambda_i. \tag{128}$$

Therefore,

$$\begin{aligned}\sum_{i=\ell}^{m}|\boldsymbol{a}-\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(i)} &= \sum_{i=\ell}^{k}|\boldsymbol{a}-\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(i)} + \sum_{i=k+1}^{m}|\boldsymbol{a}-\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(i)} \\ &< \sum_{i=\ell}^{k}\lambda_i + \sum_{i=k+1}^{m}\lambda_i \qquad (129)\\ &= \sum_{i=\ell}^{m}\lambda_i,\end{aligned}$$

where the inequality follows from the condition that $|\boldsymbol{a}-\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(1:k)} \prec_S \boldsymbol{\lambda}_{1:k}$ and (127). Clearly, (129) contradicts (128). Therefore, $|\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(k)} = 0$ and thus $|\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})|_{(1:k)} = \mathbf{0}_k$.

Now we are ready to prove Lemma 5. To apply the above result, we just need to express $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \text{Prox}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{s}}), \qquad (130)$$

using the first order optimality condition of (2). Therefore, we can let $\boldsymbol{a} = \widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{s}}$ and thus $\hat{\boldsymbol{s}} = \boldsymbol{a} - \text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$. The desired result then immediately follows.

H. Proof of Lemma 6

Similar as proof of Lemma 3, the idea is again approximating the indicator function $\mathbb{I}_{x\neq 0,y=0}$ by some Lipschitz continuous functions. Here we use:

$$\phi_h(x,y) = [1-\psi_h(x)]\psi_h(y),$$

where $\psi_h(x)$ is defined in (118). We have

$$|\phi_h(x,y) - \mathbb{I}_{x\neq 0,y=0}| \leq \mathbb{I}_{0<|x|<h} + \mathbb{I}_{0<|y|<h}. \qquad (131)$$

Therefore, for any $h > 0$,

$$\begin{aligned}&\big|V^{(p)} - \mathbb{P}\big[\eta(Y_*)\neq 0, B=0\big]\big| \\ =&\big|\mathbb{E}_{\mu_{\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}}}\mathbb{I}_{x\neq 0,y=0} - \mathbb{E}_{\mu_{\hat{B},B}}\mathbb{I}_{x\neq 0,y=0}\big| \\ \leq&\mathbb{E}_{\mu_{\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}}}\big|\mathbb{I}_{x\neq 0,y=0} - \phi_h(x,y)\big| + \big|\mathbb{E}_{\mu_{\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}}}\phi_h(x,y) - \mathbb{E}_{\mu_{\hat{B},B}}\phi_h(x,y)\big| \\ &+ \mathbb{E}_{\mu_{\hat{B},B}}\big|\mathbb{I}_{x\neq 0,y=0} - \phi_h(x,y)\big| \\ \leq&\frac{1}{p}\sum_{i=1}^{p}(\mathbb{I}_{|\hat{\beta}_i|<h} - \mathbb{I}_{\hat{\beta}_i=0}) + \frac{1}{p}\sum_{i=1}^{p}(\mathbb{I}_{|\beta_i|<h} - \mathbb{I}_{\beta_i=0}) + \mathbb{P}\big(0<|\eta(Y_*)|<h\big) \\ &+ \mathbb{P}(0<|B|<h) + \big|\mathbb{E}_{\mu_{\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}}}\phi_h(x,y) - \mathbb{E}_{\mu_{\hat{B},B}}\phi_h(x,y)\big|, \qquad (132)\end{aligned}$$

where $\mu_{\hat{B},B}$ denotes the joint distribution of $(\eta(Y_*), B)$ and in the last step we use (131) and $\mathbb{I}_{0<|x|<h} = \mathbb{I}_{|x|<h} - \mathbb{I}_{x=0}$. Let us compute the limit of each term on the RHS of (132). The last term converges in probability to zero due to (12). By continuity of probability, $\mathbb{P}\big(0<|\eta(Y_*)|<h\big)$ and $\mathbb{P}(0<|B|<h)$ converge to 0 as $h\to 0$. Following similar steps leading to (120), we can get $\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}_{|\hat{\beta}_i|<h} \xrightarrow{\mathbb{P}} \mathbb{P}\big(|\eta(Y_*)|<h\big)$ and $\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}_{|\beta_i|<h} \xrightarrow{\mathbb{P}} \mathbb{P}(|B|<h)$ if

h satisfies $\mathbb{P}\big(|\eta(Y_*)|=h\big)=\mathbb{P}(B=h)=0$. Therefore, for such $h>0$, (132) yields the following: for any $\varepsilon>0$ there exists p_0 such that when $p>p_0$,

$$\begin{aligned}\big|V^{(p)}-\mathbb{P}\big[\eta(Y_*)\neq 0,B=0\big]\big| &\le \big[\mathbb{P}(|B|=0)+\mathbb{P}\big(|\eta(Y_*)|=0\big)\big]-(r_0^{(p)}+R_0^{(p)})\\ &\quad+2\mathbb{P}(0<|B|<h)+2\mathbb{P}\big(0<|\eta(Y_*)|<h\big)+\frac{\varepsilon}{2}.\\ &\le 2\mathbb{P}(0<|B|<h)+2\mathbb{P}\big(0<|\eta(Y_*)|<h\big)+|\mathbb{P}\big(|\eta(Y_*)|=0\big)-R_0^{(p)}|+\varepsilon,\end{aligned} \tag{133}$$

where in the last step we use Assumption (A.2). Then taking $h\to 0$ along a sequence $\{h_i\}_{i\in\mathbb{Z}^+}$ with $\mathbb{P}\big(|\eta(Y_*)|=h_i\big)=\mathbb{P}(B=h_i)=0$ in (133), we get for any $\varepsilon>0$,

$$\big|V^{(p)}-\mathbb{P}\big[\eta(Y_*)\neq 0,B=0\big]\big|\le|\mathbb{P}\big(\eta(Y_*)=0\big)-R_0^{(p)}|+\varepsilon, \tag{134}$$

with probability approaching 1 as $p\to\infty$.

I. Asymptotic Properties of $\hat{\boldsymbol{s}}$

In this section, we study the limiting properties of the following vector: $\hat{\boldsymbol{s}}=\boldsymbol{A}^\top(\boldsymbol{w}-\boldsymbol{A}\hat{\boldsymbol{v}})$, where $\widehat{\boldsymbol{v}}=\operatorname{argmin}_{\boldsymbol{v}}C(\boldsymbol{v})$. Recall that $C(\boldsymbol{v})$ is the objective function of primary problem defined in (51):

$$C(\boldsymbol{v})=\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\boldsymbol{v}-\boldsymbol{w}\|^2+J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})\Big]$$

and $\widehat{\boldsymbol{v}}$ is the optimal solution. The main goal is to prove Proposition 6, which characterizes the limiting empirical distribution of $(\hat{\boldsymbol{s}},\boldsymbol{\beta})$. We will follow the proof strategy in [50, Appendix E].

Proposition 6: Under the same setting as Theorem 1, define $\mu_{\hat{S},B}$ as the joint measure of $(\tau_*^{-1}[Y-\eta(Y;\mu_{Y_*},\mu_{\tau_*\Lambda})],B)$. It holds that $W_2(\mu_{\hat{\boldsymbol{s}},\boldsymbol{\beta}},\mu_{\hat{S},B})\xrightarrow{\mathbb{P}}0$.

Proof: The first step is to obtain an alternative representation of $\hat{\boldsymbol{s}}$. Consider the event $E=\{\hat{\boldsymbol{v}}\in\mathcal{B}_{\sqrt{n}K}\}$, for some $K\ge\frac{\sigma_*}{\sqrt{\delta}}+\frac{\theta_{\min}}{4}$, where $\theta_{\min}$ is given in Lemma 14. It is shown in Lemma 12 that $\mathbb{P}(E)\to 1$ as $p\to\infty$. Under event E, we have

$$\begin{aligned}\min_{\boldsymbol{v}\in\mathbb{R}^p}C(\boldsymbol{v}) &= \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\boldsymbol{v}-\boldsymbol{w}\|^2+J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})\Big]\\ &= \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}}\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\boldsymbol{v}-\boldsymbol{w}\|^2+\boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})\Big]\\ &= \max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}}\underbrace{\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}}\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\boldsymbol{v}-\boldsymbol{w}\|^2+\boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})\Big]}_{\stackrel{\text{def}}{=}\mathcal{S}(\boldsymbol{s})},\end{aligned} \tag{135}$$

where $C_{\boldsymbol{\lambda}}$ is defined in (187) and the last step follows from Sion's minimax theorem [59]. By the first order optimality condition of , we know $\hat{\boldsymbol{s}}\in\partial J_{\boldsymbol{\lambda}}(\hat{\boldsymbol{v}}+\boldsymbol{\beta})$. On the other hand, it is not hard to show for any $\boldsymbol{x}\in\mathbb{R}^p$ and

$\boldsymbol{s} \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{x})$, it holds that $\boldsymbol{s} \in C_{\boldsymbol{\lambda}}$ and $J_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \boldsymbol{s}^\top \boldsymbol{x}$. Therefore, $\hat{\boldsymbol{s}} \in C_{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{s}}^\top(\hat{\boldsymbol{v}} + \boldsymbol{\beta}) = J_{\boldsymbol{\lambda}}(\hat{\boldsymbol{v}} + \boldsymbol{\beta})$. Then

$$\begin{aligned}
\mathcal{S}(\hat{\boldsymbol{s}}) = &\min_{\boldsymbol{v} \in \mathcal{B}_{\sqrt{n}K}} \frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\boldsymbol{v} - \boldsymbol{w}\|^2 + \hat{\boldsymbol{s}}^\top(\boldsymbol{v} + \boldsymbol{\beta})\Big] \\
\overset{\text{(a)}}{=} &\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\hat{\boldsymbol{v}} - \boldsymbol{w}\|^2 + \hat{\boldsymbol{s}}^\top(\hat{\boldsymbol{v}} + \boldsymbol{\beta})\Big] \\
= &\frac{1}{n}\Big[\frac{1}{2}\|\boldsymbol{A}\hat{\boldsymbol{v}} - \boldsymbol{w}\|^2 + J_{\boldsymbol{\lambda}}(\hat{\boldsymbol{v}} + \boldsymbol{\beta})\Big] \\
\overset{\text{(b)}}{=} &\max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s}),
\end{aligned}$$

where (a) follows from first order optimality condition and the fact that $\hat{\boldsymbol{v}} \in \mathcal{B}_{\sqrt{n}K}$ under event E and (b) follows from (135) and $\hat{\boldsymbol{v}} \in \min_{\boldsymbol{v} \in \mathbb{R}^p} C(\boldsymbol{v})$. This implies that under event E, which happens with probability approaching 1, $\hat{\boldsymbol{s}} \in \operatorname{argmax}_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s})$. Therefore, in order to study the limiting behavior of $\hat{\boldsymbol{s}}$, we can instead study $\operatorname{argmax}_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s})$.

The analysis of $\operatorname{argmax}_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s})$ can be carried out based on CGMT framework. First, similar as Proposition E.1 in [50], we can get for any closed set D and $t \in \mathbb{R}$,

$$\mathbb{P}\big(\max_{\boldsymbol{s} \in D} \mathcal{S}(\boldsymbol{s}) \geq t\big) \leq 2\mathbb{P}\big(\max_{\boldsymbol{s} \in D} S(\boldsymbol{s}) \geq t\big) \tag{136}$$

and if D is also convex,

$$\mathbb{P}\big(\max_{\boldsymbol{s} \in D} \mathcal{S}(\boldsymbol{s}) \leq t\big) \leq 2\mathbb{P}\big(\max_{\boldsymbol{s} \in D} S(\boldsymbol{s}) \leq t\big). \tag{137}$$

Here

$$S(\boldsymbol{s}) = \min_{\boldsymbol{v} \in \mathcal{B}_{\sqrt{n}K}} \tfrac{1}{2}\Big(\sqrt{\tfrac{\|\boldsymbol{v}\|^2}{n}\tfrac{\|\boldsymbol{g}\|^2}{n} + \tfrac{\|\boldsymbol{w}\|^2}{n} + 2\tfrac{\|\boldsymbol{v}\|}{\sqrt{n}}\tfrac{\boldsymbol{g}^\top \boldsymbol{w}}{n}} - \tfrac{\boldsymbol{h}^\top \boldsymbol{v}}{n}\Big)_+^2 + \tfrac{\boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})}{n}. \tag{138}$$

Then for any $\varepsilon, \nu > 0$ and $D_\nu \overset{\text{def}}{=} \Big\{\boldsymbol{s} : W_2(\mu_{\boldsymbol{s},\boldsymbol{\beta}}, \mu_{\hat{S},B}) \geq \nu\Big\} \bigcap C_{\boldsymbol{\lambda}}$, we have

$$\begin{aligned}
\mathbb{P}\big(\max_{\boldsymbol{s} \in D_\nu} \mathcal{S}(\boldsymbol{s}) \geq \max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s}) - \varepsilon\big) \leq &\mathbb{P}\big(\max_{\boldsymbol{s} \in D_\nu} \mathcal{S}(\boldsymbol{s}) \geq \Psi^* - 2\varepsilon\big) + \mathbb{P}\big(\max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s}) < \Psi^* - \varepsilon\big) \\
\overset{\text{(a)}}{\leq} &2\mathbb{P}\big(\max_{\boldsymbol{s} \in D_\nu} S(\boldsymbol{s}) \geq \Psi^* - 2\varepsilon\big) + 2\mathbb{P}\big(\max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} S(\boldsymbol{s}) < \Psi^* - \varepsilon\big)
\end{aligned} \tag{139}$$

where Ψ^* is defined in (55) and step (a) follows from (136) and (137). Then combining (139) with Lemma 13, we get for any $\nu > 0$, there exists $\varepsilon > 0$ such that $\mathbb{P}\big(\max_{\boldsymbol{s} \in D_\nu} \mathcal{S}(\boldsymbol{s}) \geq \max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s}) - \varepsilon\big) \to 0$. Therefore, for any $\nu > 0$ and $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\big(W_2(\mu_{\hat{\boldsymbol{s}},\boldsymbol{\beta}}, \mu_{\hat{S},B}) \geq \nu\big) &= \mathbb{P}(\hat{\boldsymbol{s}} \in D_\nu) \\
&\leq \mathbb{P}(\hat{\boldsymbol{s}} \in D_\nu \bigcap E) + \mathbb{P}(E^C) \\
&\leq \mathbb{P}\Big[\hat{\boldsymbol{s}} \in D_\nu \text{ and } \mathcal{S}(\hat{\boldsymbol{s}}) = \max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s})\Big] + \mathbb{P}(E^C) \\
&\leq \mathbb{P}\Big[\max_{\boldsymbol{s} \in D_\nu} \mathcal{S}(\boldsymbol{s}) = \max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s})\Big] + \mathbb{P}(E^C) \\
&\leq \mathbb{P}\big(\max_{\boldsymbol{s} \in D_\nu} \mathcal{S}(\boldsymbol{s}) \geq \max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} \mathcal{S}(\boldsymbol{s}) - \varepsilon\big) + \mathbb{P}(E^C).
\end{aligned} \tag{140}$$

By the discussion above, the RHS of (140) converges to 0 for some $\varepsilon > 0$. Since, the LHS of (140) does not depend on ε, this concludes the proof. ■

Lemma 13: Under the same setting as Proposition 6, as $p \to \infty$,

$$\max_{\boldsymbol{s} \in C_{\boldsymbol{\lambda}}} S(\boldsymbol{s}) \overset{\mathbb{P}}{\to} \Psi^*, \tag{141}$$

where $S(\boldsymbol{s})$ is defined in (138) and Ψ^* is defined in (55). Denote $D_\nu := \Big\{\boldsymbol{s} : W_2(\mu_{\boldsymbol{s},\boldsymbol{\beta}}, \mu_{\hat{S},B}) \geq \nu\Big\} \bigcap C_{\boldsymbol{\lambda}}$. For any $\nu > 0$, there exists $\varepsilon > 0$ such that

$$\mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu} S(\boldsymbol{s}) \geq \Psi^* - \varepsilon\big) \to 0. \tag{142}$$

Proof: For the similar reason as introducing $\widetilde{L}(\boldsymbol{v})$ when analyzing $L(\boldsymbol{v})$ [c.f. (87) and (92) in the proof of Lemma 10], we consider the following approximation of $S(\boldsymbol{s})$:

$$\widetilde{S}(\boldsymbol{s}) = \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \tfrac{1}{2}\Big(\sqrt{\tfrac{\|\boldsymbol{v}\|^2}{n} + \sigma_w^2} - \tfrac{\boldsymbol{h}^\top \boldsymbol{v}}{n}\Big)_+^2 + \tfrac{\boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})}{n}. \tag{143}$$

Note that $S(\boldsymbol{s}) = \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} L(\boldsymbol{v}) - \Delta(\boldsymbol{v})$ and $\widetilde{S}(\boldsymbol{s}) = \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \widetilde{L}(\boldsymbol{v}) - \Delta(\boldsymbol{v})$, where $\Delta(\boldsymbol{v}) = \frac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta}) - \boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})}{n}$. Therefore, by (102)

$$\sup_{\boldsymbol{s}\in\mathbb{R}^p} |S(\boldsymbol{s}) - \widetilde{S}(\boldsymbol{s})| \leq \sup_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} |L(\boldsymbol{v}) - \widetilde{L}(\boldsymbol{v})| \xrightarrow{\mathbb{P}} 0. \tag{144}$$

On the other hand, similar as (135) we have

$$\begin{aligned}
\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \widetilde{L}(\boldsymbol{v}) &= \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \tfrac{1}{2}\Big(\sqrt{\tfrac{\|\boldsymbol{v}\|^2}{n} + \sigma_w^2} - \tfrac{\boldsymbol{h}^\top \boldsymbol{v}}{n}\Big)_+^2 + \tfrac{J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta})}{n} \\
&= \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \tfrac{1}{2}\Big(\sqrt{\tfrac{\|\boldsymbol{v}\|^2}{n} + \sigma_w^2} - \tfrac{\boldsymbol{h}^\top \boldsymbol{v}}{n}\Big)_+^2 + \tfrac{\boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})}{n} \\
&= \max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \tfrac{1}{2}\Big(\sqrt{\tfrac{\|\boldsymbol{v}\|^2}{n} + \sigma_w^2} - \tfrac{\boldsymbol{h}^\top \boldsymbol{v}}{n}\Big)_+^2 + \tfrac{\boldsymbol{s}^\top(\boldsymbol{v}+\boldsymbol{\beta})}{n} \\
&= \max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \widetilde{S}(\boldsymbol{s}),
\end{aligned} \tag{145}$$

where $C_{\boldsymbol{\lambda}}$ is defined in (187). Using successively (144) and (145), we have for any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\big(\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} S(\boldsymbol{s}) < \Psi^* - \varepsilon\big) &\leq \mathbb{P}\big(\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \widetilde{S}(\boldsymbol{s}) < \Psi^* - \tfrac{\varepsilon}{2}\big) + \delta_p \\
&= \mathbb{P}\big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \widetilde{L}(\boldsymbol{v}) < \Psi^* - \tfrac{\varepsilon}{2}\big) + \delta_p,
\end{aligned} \tag{146}$$

where $\delta_p \to 0$ as $p \to \infty$. Similarly on the other direction, we can also get for any $\varepsilon > 0$, there is some $\delta_p' \to 0$ such that

$$\mathbb{P}\big(\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} S(\boldsymbol{s}) \geq \Psi^* + \varepsilon\big) \leq \mathbb{P}\big(\min_{\boldsymbol{v}\in\mathcal{B}_{\sqrt{n}K}} \widetilde{L}(\boldsymbol{v}) \geq \Psi^* + \tfrac{\varepsilon}{2}\big) + \delta_p' \tag{147}$$

Then combining (146) and (147) with (100), we get $\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} S(\boldsymbol{s}) \xrightarrow{\mathbb{P}} \Psi^*$ and from (144) we get $\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \widetilde{S}(\boldsymbol{s}) \xrightarrow{\mathbb{P}} \Psi^*$.

Next, we show (142). First, the following bound holds

$$\begin{aligned}
\mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu} S(\boldsymbol{s}) \geq \Psi^* - \varepsilon\big) \leq& \mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu} \widetilde{S}(\boldsymbol{s}) \geq \Psi^* - 2\varepsilon\big) + \mathbb{P}\big(\sup_{\boldsymbol{s}\in D_\nu} |S(\boldsymbol{s}) - \widetilde{S}(\boldsymbol{s})| \geq \varepsilon\big) \\
\leq& \mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu} \widetilde{S}(\boldsymbol{s}) \geq \max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \widetilde{S}(\boldsymbol{s}) - 3\varepsilon\big) + \mathbb{P}\big(\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \widetilde{S}(\boldsymbol{s}) \geq \Psi^* + \varepsilon\big) \\
&+ \mathbb{P}\big(\sup_{\boldsymbol{s}\in D_\nu} |S(\boldsymbol{s}) - \widetilde{S}(\boldsymbol{s})| \geq \varepsilon\big).
\end{aligned} \tag{148}$$

Recall that we have already shown that the last two terms on the RHS of (148) vanish as $p \to \infty$. Therefore, it remains to show the first term also converges to 0. The main step is to establish there exist $\varsigma_{\max}, \gamma > 0$ such that for any $\varsigma \in (0, \varsigma_{\max})$,

$$\mathbb{P}\big(\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}} \cap \mathcal{B}^o_{\sqrt{p}\varsigma}(\mathring{\boldsymbol{s}})} \widetilde{S}(\boldsymbol{s}) \geq \max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}} \widetilde{S}(\boldsymbol{s}) - \gamma\varsigma\big) \to 0, \tag{149}$$

where

$$\begin{aligned}\mathring{\boldsymbol{s}} \overset{\text{def}}{=}&\frac{1}{\tau_*}\big[(\boldsymbol{\beta}+\sigma_*\boldsymbol{h})-(\boldsymbol{\beta}+\mathring{\boldsymbol{v}})\big]\\ =&\frac{1}{\tau_*}\big[(\boldsymbol{\beta}+\sigma_*\boldsymbol{h})-\eta(\boldsymbol{\beta}+\sigma_*\boldsymbol{h};\mu_{Y_*},\mu_{\tau_*\Lambda})\big]\end{aligned}$$

and $\mathring{\boldsymbol{v}}$ is defined in (90). The convergence in (149) can be proved in exactly the same way as Theorem E.7 in [50], which deals with the case of LASSO. For simplicity, we do not re-present the proof details here. Here $C_{\boldsymbol{\lambda}}$ (the unit ball of the dual norm of $J_{\boldsymbol{\lambda}}$) plays the role of the set $\{\boldsymbol{x} \mid \|\boldsymbol{x}\|_\infty \le \lambda\}$ in [50], which is the unit ball of the dual norm of $\lambda\|\cdot\|_1$. Now consider the event $E=\{W_2\big(\mu_{\mathring{\boldsymbol{s}},\boldsymbol{\beta}},\mu_{\hat{S},B}\big)\le\frac{\nu}{2}\}$. Conditioned on E, it holds that for $\boldsymbol{s}\in D_\nu$,

$$\begin{aligned}\frac{1}{p}\|\boldsymbol{s}-\mathring{\boldsymbol{s}}\|^2 \ge& W_2\big(\mu_{\boldsymbol{s},\boldsymbol{\beta}},\mu_{\mathring{\boldsymbol{s}},\boldsymbol{\beta}}\big)^2\\ \ge&\big[W_2\big(\mu_{\boldsymbol{s},\boldsymbol{\beta}},\mu_{\hat{S},B}\big)-W_2\big(\mu_{\mathring{\boldsymbol{s}},\boldsymbol{\beta}},\mu_{\hat{S},B}\big)\big]^2\\ \ge&\frac{\nu^2}{4},\end{aligned}$$

which indicates that $D_\nu \subseteq C_{\boldsymbol{\lambda}}\cap\mathcal{B}^o_{\frac{\sqrt{p}\nu}{2}}(\mathring{\boldsymbol{s}})$. Therefore,

$$\begin{aligned}\mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu}\widetilde{S}(\boldsymbol{s})\ge\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}}\widetilde{S}(\boldsymbol{s})-\varepsilon\big)\le&\mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu}\widetilde{S}(\boldsymbol{s})\ge\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}}\widetilde{S}(\boldsymbol{s})-\varepsilon\bigcap E\big)+\mathbb{P}(E^C)\\ \le&\mathbb{P}\Big(\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}\cap\mathcal{B}^o_{\frac{\sqrt{p}\nu}{2}}(\mathring{\boldsymbol{s}})}\widetilde{S}(\boldsymbol{s})\ge\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}}\widetilde{S}(\boldsymbol{s})-\varepsilon\Big)+\mathbb{P}(E^C).\end{aligned}\tag{150}$$

According to (149), the first term in (150) vanishes if $\varepsilon\le\frac{\gamma\nu}{2}$. For the second term, it is not hard to show $(H,B)\mapsto\big(\tau_*^{-1}[(B+\sigma_*H)-\eta(B+\sigma_*H;\mu_{Y_*},\mu_{\tau_*\Lambda})],B\big)$ is a Lipschitz continuous mapping and from (164), $W_2\big(\mu_{\boldsymbol{h},\boldsymbol{\beta}},H\otimes B\big)\xrightarrow{\mathbb{P}}0$, so similar as (165) we can get $W_2\big(\mu_{\mathring{\boldsymbol{s}},\boldsymbol{\beta}},\mu_{\hat{S},B}\big)\xrightarrow{\mathbb{P}}0$ and thus $\mathbb{P}(E^C)\to0$. Therefore, from (150), for any $\nu>0$, there exists $\varepsilon_0>0$ such that any $\varepsilon\le\varepsilon_0$ satisfies $\mathbb{P}\big(\max_{\boldsymbol{s}\in D_\nu}\widetilde{S}(\boldsymbol{s})\ge\max_{\boldsymbol{s}\in C_{\boldsymbol{\lambda}}}\widetilde{S}(\boldsymbol{s})-\varepsilon\big)$. Substituting this back to (148), we finish the proof. ■

J. Properties of Limiting Scalar Problem

It turns out that the limiting behavior of (2) is fully captured by (55). In this section, we study the key properties of (55).

Lemma 14: The minimax problem (55) has a unique optimal solution (σ_*,θ_*), which is also the unique solution to the equation:

$$\begin{aligned}\sigma^2&=\sigma_w^2+\frac{1}{\delta}\mathbb{E}[\eta(Y;\mu_Y,\mu_{\sigma\Lambda/\theta})-B]^2,\\ \theta&=\sigma\Big[1-\frac{1}{\delta}\mathbb{E}\eta'(Y;\mu_Y,\mu_{\sigma\Lambda/\theta})\Big],\end{aligned}\tag{151}$$

where $Y=B+\sigma H$, with $H\sim\mathcal{N}(0,1)$ independent of $B\sim\mu_B$. Besides, there exists $\theta_{\min}>0$ such that $\theta^*\ge\theta_{\min}$.

Proof: The proof includes two steps: (I) show the saddle point of $\Psi(\sigma,\theta)$ exists and is unique and it is also the unique optimal solution of the minimax problem (55), (II) show (σ_*,θ_*) is the saddle point of $\Psi(\sigma,\theta)$ if and only if it is the solution to (151).

We first show the set of saddle points of $\Psi(\sigma,\theta)$ is nonempty and compact, using Proposition 5.5.7 of [60]. To apply this result, it suffices to check: (i) $\Psi(\cdot,\theta)$ and $-\Psi(\sigma,\cdot)$ are convex and closed for any fixed $\theta\ge0$ and

$\sigma \geq \sigma_w$, (ii) there exists some $\overline{\theta} \geq 0$, $\overline{\sigma} \geq \sigma_w$ and $\gamma_1, \gamma_2 \in \mathbb{R}$ such that the level sets $\{\sigma \geq \sigma_w \mid \Psi(\sigma, \overline{\theta}) \leq \gamma_1\}$ and $\{\theta \geq 0 \mid -\Psi(\overline{\sigma}, \theta) \leq \gamma_2\}$ are both non-empty and compact. From Lemma 16, $\mathcal{F}(\sigma, \theta)$ is convex-concave and continuously differentiable with respect to σ and θ. Therefore, condition (i) is satisfied. Also partial derivatives of $\mathcal{F}(\sigma, \theta)$ can be computed as

$$\frac{\partial \Psi}{\partial \sigma} = \frac{\theta}{2\sigma^2}\Big(\sigma^2 - \sigma_w^2 - \frac{1}{\delta}\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\|\boldsymbol{\beta} - \text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{\beta} + \sigma\boldsymbol{h})\|^2\Big), \tag{152}$$

$$\frac{\partial \Psi}{\partial \theta} = \frac{1}{2}\Big(\frac{\sigma_w^2}{\sigma} + \sigma\Big) - \theta + \frac{1}{\delta}\Big[\lim_{p\to\infty}\mathbb{E}\frac{\|\boldsymbol{\beta} + \sigma\boldsymbol{h} - \text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{\beta} + \sigma\boldsymbol{h})\|^2}{2\sigma p} - \frac{\sigma}{2}\Big], \tag{153}$$

using (55), (178) and (177). Next we show $\{\sigma \geq \sigma_w \mid \Psi(\sigma, \overline{\theta}) \leq \gamma_1\}$ is non-empty and compact for some $\overline{\theta} \geq 0$ and $\gamma_1 \in \mathbb{R}$. First, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\beta} - \text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{\beta} + \sigma\boldsymbol{h})\|^2 &\overset{\text{(a)}}{=} \sigma^2\mathbb{E}\|\tfrac{\boldsymbol{\beta}}{\sigma} - \text{Prox}_{\boldsymbol{\lambda}/\theta}(\tfrac{\boldsymbol{\beta}}{\sigma} + \boldsymbol{h})\|^2 \\
&\leq 2\sigma^2\mathbb{E}\big(\tfrac{\|\boldsymbol{\beta}\|^2}{\sigma^2} + 2\|\text{Prox}_{\boldsymbol{\lambda}/\theta}(\boldsymbol{h})\|^2 + 2\|\text{Prox}_{\boldsymbol{\lambda}/\theta}(\tfrac{\boldsymbol{\beta}}{\sigma} + \boldsymbol{h}) - \text{Prox}_{\boldsymbol{\lambda}/\theta}(\boldsymbol{h})\|^2\big) \\
&\overset{\text{(b)}}{\leq} 4\sigma^2\mathbb{E}\big\|\text{Prox}_{\boldsymbol{\lambda}/\theta}(\boldsymbol{h})\|^2 + 6\|\boldsymbol{\beta}\|^2 \\
&\overset{\text{(c)}}{\leq} 4\sigma^2\sum_{i=1}^{p}\mathbb{E}\max\big\{|h_i| - \tfrac{\bar{\lambda}}{\theta}, 0\big\}^2 + 6\|\boldsymbol{\beta}\|^2,
\end{aligned} \tag{154}$$

where (a) follows from the identity $\text{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{x}) = \tau\text{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{x}/\tau)$, (b) follows from the non-expansiveness of proximal operator and (c) is a consequence of (188), where $\bar{\lambda} = \frac{1}{p}\sum_{i=1}^{p}\lambda_i$. Plugging the bound (154) into (152) gives

$$\tfrac{\partial \Psi}{\partial \sigma} \geq \tfrac{\theta}{2\sigma^2}\Big[\sigma^2\big(1 - \tfrac{8}{\delta}\int_{\frac{\mathbb{E}\Lambda}{\theta}}^{\infty}(z - \tfrac{\mathbb{E}\Lambda}{\theta})d\Phi(z)\big) - \sigma_w^2 - \tfrac{6}{\delta}\mathbb{E}B^2\Big], \tag{155}$$

where $\Phi(z)$ is CDF of standard Gaussian. When $\mathbb{E}\Lambda > 0$, from (155) we know there exists $\theta_1 > 0$ and $\sigma_1 \geq \sigma_w$ such that $\frac{\partial \Psi(\sigma,\theta_1)}{\partial \sigma} \geq \frac{\theta_1}{8}$ for all $\sigma \geq \sigma_1$; when $\mathbb{E}\Lambda = 0$, by our assumption we must have $\delta > 1$, so from (157) we have $\frac{\partial \Psi(\sigma,\theta)}{\partial \sigma} = \frac{\theta}{2\sigma^2}[(1 - \frac{1}{\delta})\sigma^2 - \sigma_w^2]$ for any $\theta > 0$, $\sigma \geq \sigma_w$ implying $\frac{\partial \Psi(\sigma,\theta)}{\partial \sigma} \geq \frac{\theta}{4}(1 - \frac{1}{\delta})$ for all $\sigma \geq \sqrt{\frac{2\delta}{\delta-1}}\sigma_w$. Therefore, there exists $\overline{\theta} > 0$, $c > 0$ and $K \geq \sigma_w$ such that $\frac{\partial \Psi(\sigma,\overline{\theta})}{\partial \sigma} \geq c$ for any $\sigma \geq K$. This means that $\Psi(\sigma, \overline{\theta}) > \Psi(K, \overline{\theta})$ for all $\sigma > K$, so the set $\{\sigma \geq \sigma_w \mid \Psi(\sigma, \overline{\theta}) \leq \Psi(K, \overline{\theta})\} \subset [\sigma_w, K]$ and it is non-empty (include at least one point $\sigma = K$) and closed since $\Psi(\cdot, \overline{\theta})$ is a closed function. As a result, we can take $\gamma_1 = \Psi(K, \overline{\theta})$ and the level set $\{\sigma \geq \sigma_w \mid \Psi(\sigma, \overline{\theta}) \leq \gamma_1\}$ is non-empty and compact. On the other hand, we can show $\{\theta \geq 0 \mid -\Psi(\sigma_w, \theta) \leq 0\}$ is non-empty and compact. First since $\Psi(\sigma_w, \cdot)$ is 1-strongly concave and continuously differentiable, we have for any $\theta \geq 0$,

$$\begin{aligned}
\Psi(\sigma_w, \theta) &\leq \Psi(\sigma_w, 0) + \tfrac{\partial \Psi(\sigma_w,0)}{\partial \theta}\theta - \tfrac{1}{2}\theta^2 \\
&\leq \big(\sigma_w + \tfrac{\mathbb{E}B^2}{2\sigma_w\delta}\big)|\theta| - \tfrac{1}{2}\theta^2,
\end{aligned}$$

where in the last step we use $\Psi(\sigma_w, 0) = 0$ and $\frac{\partial \Psi(\sigma,0)}{\partial \theta} = \sigma_w + \frac{\mathbb{E}B^2}{2\sigma_w\delta}$, which can be deduced from (55) and (153). Then the level set $\{\theta \geq 0 \mid -\Psi(\sigma_w, \theta) \leq 0\} \subset \Big[0, 2\big(\sigma_w + \frac{\mathbb{E}B^2}{2\sigma_w\delta}\big)\Big]$ and it is non-empty (include at least one point $\theta = 0$) and closed since $\Psi(\sigma_w, \cdot)$ is a closed function. Letting $\gamma_2 = 0$, we verify condition (ii).

Up to now, we have proved the existence and boundedness of saddle points of $\Psi(\sigma, \theta)$. Next we prove the uniqueness. To do this, it suffices to show the optimal solution of $\min_{\sigma\geq\sigma_w}\max_{\theta\geq 0}\Psi(\sigma, \theta)$ is bounded and unique, then the uniqueness of saddle points follows due to the fact that each saddle point of $\Psi(\sigma, \theta)$ is also an optimal

solution of $\min_{\sigma\geq\sigma_w}\max_{\theta\geq 0}\Psi(\sigma,\theta)$ [60, Proposition 3.4.1]. First, we show that any σ_* should be bounded. Indeed, from the verification of condition (ii) above, we know there exists $c>0$ and $K\geq\sigma_w$ such that for any $\sigma\geq K$,

$$\max_{\theta\geq 0}\Psi(\sigma,\theta)\geq\Psi(\sigma,\bar{\theta})\geq\Psi(K,\bar{\theta})+c(\sigma-K),$$

so we must have

$$\sigma_*\leq\underbrace{K+1+\frac{\max_{\sigma\in[\sigma_w,K]}\max_{\theta\geq 0}\Psi(\sigma,\theta)-\Psi(K,\bar{\theta})}{c}}_{:=C_1},$$

otherwise, $\min_{\sigma\geq\sigma_w}\max_{\theta\geq 0}\Psi(\sigma,\theta)>\min_{\sigma\in[\sigma_w,K]}\max_{\theta\geq 0}\Psi(\sigma,\theta)$ leading to a contradiction. On the other hand, we can also show $\theta_*(\sigma):=\operatorname*{argmax}_{\theta\geq 0}\Psi(\sigma,\theta)$ is uniformly bounded for $\sigma\in[\sigma_w,C_1]$. To see this, note from (153)

$$\begin{aligned}\frac{\partial\Psi}{\partial\theta}&\leq-\theta+2\Big[\max_{\sigma\in[\sigma_w,C_1]}\sigma\big(1+\tfrac{1}{\delta}\big)+\tfrac{\mathbb{E}B^2+\sigma_w^2\delta}{\sigma\delta}\Big],\\ \Rightarrow\theta_*(\sigma)&\leq\underbrace{2\Big[\max_{\sigma\in[\sigma_w,C_1]}\sigma\big(1+\tfrac{1}{\delta}\big)+\tfrac{\mathbb{E}B^2+\sigma_w^2\delta}{\sigma\delta}\Big]}_{:=C_2},\ \forall\sigma\in[\sigma_w,C_1].\end{aligned}$$

As a result, $\max_{\theta\geq 0}\Psi(\sigma,\theta)=\max_{\theta\in[0,C_2]}\Psi(\sigma,\theta)$ for $\sigma\in[\sigma_w,C_1]$. Therefore, by Berge Maximum Theorem [61, Theorem 17.31], $\theta_*(\sigma)$ is an upper hemicontinuous correspondence on $[\sigma_w,C_1]$. By strong concavity of $\Psi(\sigma,\cdot)$, $\sigma\mapsto\theta_*(\sigma)$ is a function (*i.e.*, single-valued correspondence). As a result, one can easily check by definition that $\theta_*(\sigma)$ is continuous on $[\sigma_w,C_1]$. Besides, we can get $\theta_*(\sigma)>0$ for any $\sigma\in[\sigma_w,C_1]$. Indeed from (153) when $\mathbb{P}(\Lambda=0)<1$, $\frac{\partial\Psi(\sigma,0)}{\partial\theta}=\frac{1}{2}\Big(\frac{\sigma_w^2}{\sigma}+\sigma\Big)+\frac{\mathbb{E}B^2}{2\sigma\delta}>0$ and when $\mathbb{P}(\Lambda=0)=1$, $\delta\geq 1$, $\frac{\partial\Psi(\sigma,0)}{\partial\theta}\geq\frac{\sigma_w^2}{2\sigma}+\frac{\sigma}{2}(1-\frac{1}{\delta})>0$. Therefore, there exists $\theta_{\min}>0$ such that

$$\theta_*(\sigma)\geq\theta_{\min},\tag{156}$$

for any $\sigma\in[\sigma_w,C_1]$. Since $\theta^*(\sigma)\in[\theta_{\min},C_2]$ for any $\sigma\in[\sigma_w,C_1]$, we get $\max_{\theta\geq 0}\Psi(\sigma,\theta)=\max_{\theta\in[\theta_{\min},C_2]}\Psi(\sigma,\theta)$ for any $\sigma\in[\sigma_w,C_1]$. On the other hand, $\Psi(\cdot,\theta)$ is $\frac{\sigma_w^2\theta_{\min}}{C_1^3}$-strongly convex on $[\sigma_w,C_1]$ for any fixed $\theta\in[\theta_{\min},C_2]$, so we can check by definition that the function $\max_{\theta\in[\theta_{\min},C_2]}\Psi(\cdot,\theta)$ is also $\frac{\sigma_w^2\theta_{\min}}{C_1^3}$-strongly convex on $[\sigma_w,C_1]$. We conclude that $\sigma\mapsto\max_{\theta\geq 0}\Psi(\sigma,\theta)$ is $\frac{\sigma_w^2\theta_{\min}}{C_1^3}$-strongly convex on $[\sigma_w,C_1]$, since $\max_{\theta\geq 0}\Psi(\sigma,\theta)=\max_{\theta\in[\theta_{\min},C_2]}\Psi(\sigma,\theta)$. Recall that any optimal solution $\sigma_*=\operatorname*{argmin}_{\sigma\geq\sigma_w}\max_{\theta\geq 0}\Psi(\sigma,\theta)$ should lie in $[\sigma_w,C_1]$, so the uniqueness holds.

Finally, we show (σ_*,θ_*) is a saddle point of $\Psi(\sigma,\theta)$ if and only if it is a solution of (151). From (152) and (153), $\frac{\partial\Psi(\sigma_w,\theta)}{\partial\sigma}\leq 0$ for any $\theta\geq 0$ and $\frac{\partial\Psi(\sigma,0)}{\partial\theta}\geq 0$ for any $\sigma\geq\sigma_w$. Since $\Psi(\sigma,\theta)$ is convex-concave and continuously differentiable, by first order optimality condition we know (σ_*,θ_*) is a saddle point if and only if $\frac{\partial\Psi(\sigma_*,\theta_*)}{\partial\sigma}=\frac{\partial\Psi(\sigma_*,\theta_*)}{\partial\theta}=0$. On the other hand, from (55), (167) and (168) we can get

$$\frac{\partial\Psi}{\partial\sigma}=\frac{\theta}{2\sigma^2}\Big(\sigma^2-\sigma_w^2-\frac{1}{\delta}\mathbb{E}\big(\eta(B+\sigma H;\mu_Y,\mu_{\sigma\Lambda/\theta})-B\big)^2\Big),\tag{157}$$

$$\frac{\partial\Psi}{\partial\theta}=\frac{1}{2}\Big(\frac{\sigma_w^2}{\sigma}+\sigma\Big)-\theta+\frac{1}{\delta}\Big[\frac{\mathbb{E}\big(\eta(B+\sigma H;\mu_Y,\mu_{\sigma\Lambda/\theta})-B\big)^2}{2\sigma}-\sigma\mathbb{E}\eta'(B+\sigma H;\mu_Y,\mu_{\sigma\Lambda/\theta})\Big].\tag{158}$$

Note that (157) and (158) are actually the scalar representation of (152) and (153). Setting the RHS of (157) and (158) to be zero, we can get (151). ■

K. Moreau Envelope of $J_{\boldsymbol{\lambda}}$

Recall that for $\tau > 0$, the Moreau envelope of $J_{\boldsymbol{\lambda}}(\boldsymbol{x})$ is:

$$\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{y};\tau) = \min_{\boldsymbol{x}} \frac{1}{2\tau}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{x}) \tag{159}$$

and the corresponding optimal solution is the proximal operator $\text{Prox}_{\tau\boldsymbol{\lambda}}(\boldsymbol{y})$. In this section, we study the limiting behavior of the following function:

$$\begin{aligned}\mathcal{F}_p(\sigma,\theta) &\stackrel{\text{def}}{=} \frac{1}{p}\mathcal{M}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}+\sigma\boldsymbol{h};\frac{\sigma}{\theta}) \\ &= \frac{1}{p}\min_{\boldsymbol{x}} \frac{\theta}{2\sigma}\|\boldsymbol{x}-(\boldsymbol{\beta}+\sigma\boldsymbol{h})\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{x}),\end{aligned} \tag{160}$$

where $(\sigma,\theta) \in \mathbb{R}_{>0}\times\mathbb{R}_{\geq 0}$, $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_p)$ and $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are the same as in (1) and (2).

Lemma 15: Consider a sequence of instances $\{\boldsymbol{h}^{(p)},\boldsymbol{\beta}^{(p)},\boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{Z}^+}$, where $\boldsymbol{h}^{(p)} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_p)$ are all independent and $\{\boldsymbol{\beta}^{(p)}\}_{p\in\mathbb{Z}^+}$, $\{\boldsymbol{\lambda}^{(p)}\}_{p\in\mathbb{Z}^+}$ are both converging sequences with limiting measure μ_B and μ_Λ. As $p\to\infty$, for every $(\sigma,\theta)\in\mathbb{R}_{>0}\times\mathbb{R}_{\geq 0}$,

$$\mathcal{F}_p(\sigma,\theta) \stackrel{a.s.}{\to} \mathcal{F}(\sigma,\theta), \tag{161}$$

and

$$\mathbb{E}_{\boldsymbol{h}}\mathcal{F}_p(\sigma,\theta) \to \mathcal{F}(\sigma,\theta), \tag{162}$$

where

$$\mathcal{F}(\sigma,\theta) \stackrel{\text{def}}{=} \min_{g\in\mathcal{I}} \frac{\theta}{2\sigma}\mathbb{E}[Y-g(Y)]^2 + \int_0^1 F_\Lambda^{-1}(u)F_{|g(Y)|}^{-1}(u)du. \tag{163}$$

Here, $Y = B+\sigma H$, with $H\sim\mathcal{N}(0,1)$, $B\sim\mu_B$ independent and $\eta(\cdot)$ is the limiting scalar function in Proposition 1.

Proof: For notational simplicity, we will omit the superscript "(p)" in $\boldsymbol{h}^{(p)},\boldsymbol{\beta}^{(p)}$ and $\boldsymbol{\lambda}^{(p)}$. We first show the empirical measure $\mu_{\boldsymbol{h},\boldsymbol{\beta}}$ converges almost surely to $H\otimes B$ under Wasserstein-2 distance. Let $g(x,y)$ be a bounded and continuous test function. By strong law of large number for triangular array [58, Theorem 2.1], we can get $\frac{1}{p}\sum_{i=1}^p[g(h_i,\beta_i)-\mathbb{E}_H g(H,\beta_i)] \stackrel{a.s.}{\to} 0$. For any y, $\mathbb{E}_H g(H,y)$ is bounded and it is not hard to show $y\mapsto\mathbb{E}_H g(H,y)$ is continuous: we know $g(h,y)$ is uniformly continuous over any compact set in $\mathbb{R}^2$, so for $C=\Phi^{-1}\big(1-\frac{\varepsilon}{8\|g\|_\infty}\big)$ and any $y_0\in\mathbb{R}$, $\varepsilon>0$, there exists $\delta>0$ such that $|g(h,y)-g(h,y_0)|\leq\frac{\varepsilon}{2}$ whenever $|y-y_0|\leq\delta$ and $|h|\leq C$. Hence,

$$|\mathbb{E}_H g(H,y)-\mathbb{E}_H g(H,y_0)| \leq \mathbb{E}_H\big[\mathbb{I}_{|H|\leq C}|g(H,y)-g(H,y_0)|\big] + 2\|g\|_\infty\mathbb{E}_H(\mathbb{I}_{|H|>C}) \leq \varepsilon \quad .$$

This shows the continuity of $y\mapsto\mathbb{E}_H[g(H,y)]$. Therefore, $\frac{1}{p}\sum_{i=1}^p\mathbb{E}_H[g(H,\beta_i)]\to\mathbb{E}_{H,B}[g(H,B)]$ and we get for any bounded and continuous $g(x,y)$, $\frac{1}{p}\sum_{i=1}^p g(h_i,\beta_i)\stackrel{a.s.}{\to}\mathbb{E}_{H,B}[g(H,B)]$ indicating the almost sure weak convergence of $\mu_{\boldsymbol{h},\boldsymbol{\beta}}$ to $H\otimes B$. On the other hand, by strong law of large number again, $\frac{1}{p}\sum_{i=1}^p h_i^2+\beta_i^2\stackrel{a.s.}{\to}1+\mathbb{E}B^2$. Therefore, by Theorem 7.12 (iii) in [56],

$$W_2\big(\mu_{\boldsymbol{h},\boldsymbol{\beta}}, H\otimes B\big)\stackrel{a.s.}{\to}0. \tag{164}$$

Then we can show $W_2\big(\mu_{\boldsymbol{y}},\mu_Y\big)\overset{a.s.}{\to}0$, where $\boldsymbol{y}:=\boldsymbol{\beta}+\sigma\boldsymbol{h}$. Indeed,

$$\begin{aligned}W_2\big(\mu_{\boldsymbol{y}},\mu_Y\big)^2=&\inf_{\pi\in\Pi(\mu_{\boldsymbol{y}},\mu_Y)}\int(x-y)^2d\pi(x,y)\\=&\inf_{\pi\in\Pi(\mu_{\boldsymbol{h},\boldsymbol{\beta}},H\otimes B)}\int[(x_2+\sigma x_1)-(y_2+\sigma y_1)]^2d\pi(\boldsymbol{x},\boldsymbol{y})\\\le&2(\sigma^2+1)\inf_{\pi\in\Pi(\mu_{\boldsymbol{h},\boldsymbol{\beta}},H\otimes B)}\int\|\boldsymbol{x}-\boldsymbol{y}\|^2d\pi(\boldsymbol{x},\boldsymbol{y})\\=&2(\sigma^2+1)W_2\big(\mu_{\boldsymbol{h},\boldsymbol{\beta}},H\otimes B\big)^2\end{aligned}\tag{165}$$

and since $W_2\big(\mu_{\boldsymbol{h},\boldsymbol{\beta}},H\otimes B\big)\overset{a.s.}{\to}0$, we get $W_2\big(\mu_{\boldsymbol{y}},\mu_Y\big)\overset{a.s.}{\to}0$. Similarly, for $\theta>0$ we can show $W_2\big(\mu_{\sigma\boldsymbol{\lambda}/\theta},\mu_{\sigma\Lambda/\theta}\big)\to 0$ from our assumption that $W_2\big(\mu_{\boldsymbol{\lambda}},\mu_\Lambda\big)\to 0$.

Now we can prove (161). For $\theta=0$, it directly follows from (160) and (163) that $\mathcal{F}_p(\sigma,\theta)=\mathcal{F}(\sigma,\theta)=0$. For $\theta>0$, we have $W_2\big(\mu_{\boldsymbol{y}},\mu_Y\big)\overset{a.s.}{\to}0$ and $W_2\big(\mu_{\sigma\boldsymbol{\lambda}/\theta},\mu_{\sigma\Lambda/\theta}\big)\to 0$. Then from Proposition 1, we can get (161) by letting $\tau=\sigma/\theta$.

To prove (162), first observe that:

$$\mathcal{F}_p(\sigma,\theta)=\frac{\min_{\boldsymbol{x}}\frac{\theta}{2\sigma}\|\sigma\boldsymbol{h}+\boldsymbol{\beta}-\boldsymbol{x}\|^2+J_{\boldsymbol{\lambda}}(\boldsymbol{x})}{p}\le\theta\frac{\sigma^2\|\boldsymbol{h}\|^2+\|\boldsymbol{\beta}\|^2}{\sigma p}.$$

By strong law of large number (SLLN), we have

$$\theta\frac{\sigma^2\|\boldsymbol{h}\|^2+\|\boldsymbol{\beta}\|^2}{\sigma p}\overset{a.s.}{\to}\frac{\theta(\sigma^2+\mathbb{E}B^2)}{\sigma}=\lim_{p\to\infty}\mathbb{E}_{\boldsymbol{h}}\theta\frac{\sigma^2\|\boldsymbol{h}\|^2+\|\boldsymbol{\beta}\|^2}{\sigma p}.$$

In other words,

$$\lim_{p\to\infty}\mathbb{E}_{\boldsymbol{h}}\theta\frac{\sigma^2\|\boldsymbol{h}\|^2+\|\boldsymbol{\beta}\|^2}{\sigma p}=\mathbb{E}\lim_{p\to\infty}\theta\frac{\sigma^2\|\boldsymbol{h}\|^2+\|\boldsymbol{\beta}\|^2}{\sigma p}.$$

Then by (161) and generalized dominated convergence theorem (GDCT) (Theorem 19 in Sec. 4.4 of [57]), we have

$$\lim_{p\to\infty}\mathbb{E}\mathcal{F}_p(\sigma,\theta)=\mathbb{E}\lim_{p\to\infty}\mathcal{F}_p(\sigma,\theta),\tag{166}$$

which is exactly (162). ■

Lemma 16: $\mathcal{F}(\sigma,\theta)$ is convex-concave and continuously differentiable with respect to both σ and θ on $\mathbb{R}_{>0}\times\mathbb{R}_{\ge0}$, with partial derivatives:

$$\frac{\partial\mathcal{F}(\sigma,\theta)}{\partial\sigma}=\frac{-\theta\mathbb{E}[B-\eta(Y)]^2}{2\sigma^2}+\frac{\theta}{2},\tag{167}$$

$$\frac{\partial\mathcal{F}(\sigma,\theta)}{\partial\theta}=\frac{\mathbb{E}\big[\eta(Y)-B\big]^2}{2\sigma}-\sigma\mathbb{E}\eta'(Y)+\frac{\sigma}{2},\tag{168}$$

where $B\sim\mu_B$, $\Lambda\sim\mu_\Lambda$, $Y=B+\sigma H$, with $H\sim\mathcal{N}(0,1)$ independent of B. Here, when $\theta>0$, $\eta(\cdot)=\eta(\cdot;F_Y,F_{\sigma\Lambda/\theta})$; when $\theta=0$, we let $\eta(y)=0$, if $\mathbb{P}(\Lambda=0)<1$ and $\eta(y)=y$, if $\mathbb{P}(\Lambda=0)=1$.

Proof: When $\mathbb{P}(\Lambda=0)=1$, from (163) we have $\mathcal{F}(\sigma,\theta)=0$ for any $(\sigma,\theta)\in\mathbb{R}_{>0}\times\mathbb{R}_{\ge0}$, so $\frac{\partial\mathcal{F}(\sigma,\theta)}{\partial\sigma}=\frac{\partial\mathcal{F}(\sigma,\theta)}{\partial\theta}=0$. In this case, all the results trivially hold. Therefore, it suffices to consider $\mathbb{P}(\Lambda=0)<1$.

We first prove the convexity of $\mathcal{F}(\sigma,\theta)$. Note $\mathcal{F}_p(\sigma,\theta)$ [defined in (160)] can be rewritten as:

$$\mathcal{F}_p(\sigma,\theta)=\frac{1}{p}\min_{\boldsymbol{v}}\frac{\sigma\theta}{2}\|\boldsymbol{v}/\sigma-\boldsymbol{h}\|^2+J_{\boldsymbol{\lambda}}(\boldsymbol{v}+\boldsymbol{\beta}).\tag{169}$$

Since $h(\boldsymbol{v}) = \frac{\theta}{2}\|\boldsymbol{v}-\boldsymbol{h}\|^2$ is a convex function (due to $\theta \geq 0$), $(\sigma, \boldsymbol{v}) \mapsto \frac{\sigma\theta}{2}\|\boldsymbol{v}/\sigma - \boldsymbol{h}\|^2$ is convex since it is the perspective function of $h(\boldsymbol{v})$. Therefore, the objective function in (169) is jointly convex w.r.t. $(\sigma, \boldsymbol{v})$ and after partial minimization over $\boldsymbol{v}$, $\mathcal{F}_p(\cdot, \theta)$ is still convex. On the other hand, $\mathcal{F}_p(\sigma, \theta)$ as a function of θ is the infimum of a family of linear functions of θ, so $\mathcal{F}_p(\sigma, \cdot)$ is concave. Denote $\overline{\mathcal{F}}_p(\sigma, \theta) := \mathbb{E}_{\boldsymbol{h}}\mathcal{F}_p(\sigma, \theta)$. Clearly, $\overline{\mathcal{F}}_p(\sigma, \theta)$ and $F(\sigma, \theta)$ are still convex-cancave, since taking expectation and limit preserves convexity.

Then we show for fixed $\sigma \in \mathbb{R}_{>0}$, $\mathcal{F}_p(\sigma, \cdot)$ is continuously differentiable on $\mathbb{R}_{\geq 0}$. We follow the same argument as Theorem 2.26 in [62]. Denote $\boldsymbol{y} := \boldsymbol{\beta} + \sigma\boldsymbol{h}$ and

$$g(\varepsilon) := \mathcal{F}_p(\sigma, \theta+\varepsilon) - \mathcal{F}_p(\sigma, \theta) - \tfrac{\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p}\varepsilon, \tag{170}$$

where $\varepsilon \geq -\theta$. On one hand, we have

$$\begin{aligned} g(\varepsilon) \leq & \tfrac{(\theta+\varepsilon)\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p} + \tfrac{J_{\boldsymbol{\lambda}}\left(\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\right)}{p} \\ & - \left[\tfrac{\theta\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p} + \tfrac{J_{\boldsymbol{\lambda}}\left(\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\right)}{p}\right] - \tfrac{\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p}\varepsilon = 0 \end{aligned} \tag{171}$$

On the other hand,

$$\begin{aligned} g(\varepsilon) \geq & \tfrac{(\theta+\varepsilon)\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/(\theta+\varepsilon)}(\boldsymbol{y})\|^2}{2\sigma p} + \tfrac{J_{\boldsymbol{\lambda}}\left(\text{Prox}_{\sigma\boldsymbol{\lambda}/(\theta+\varepsilon)}(\boldsymbol{y})\right)}{p} \\ & - \left[\tfrac{\theta\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/(\theta+\varepsilon)}(\boldsymbol{y})\|^2}{2\sigma p} + \tfrac{J_{\boldsymbol{\lambda}}\left(\text{Prox}_{\sigma\boldsymbol{\lambda}/(\theta+\varepsilon)}(\boldsymbol{y})\right)}{p}\right] - \tfrac{\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p}\varepsilon \\ = & \tfrac{\varepsilon}{\sigma}\left(\tfrac{\|\text{Prox}_{\sigma\boldsymbol{\lambda}/(\theta+\varepsilon)}(\boldsymbol{y})\|^2 - \|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2p}\right). \end{aligned} \tag{172}$$

From Lemma 17 there exists $C > 0$ such that $\frac{1}{p}\left|\frac{\partial\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{\partial\theta}\right| \leq C$ for any $\theta \geq 0$. Combined with (171) and (172), this indicates $-\frac{C\varepsilon^2}{2\sigma} \leq g(\varepsilon) \leq 0$. Substituting the boundedness of $g(\varepsilon)$ into (170) yields for any $\theta \geq 0$,

$$\tfrac{\partial\mathcal{F}_p(\sigma,\theta)}{\partial\theta} = \tfrac{\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p}. \tag{173}$$

Also $\frac{\partial\mathcal{F}_p(\sigma,\theta)}{\partial\theta}$ is continuous, which follows from (173), (183) and (184).

Following the same procedure as above, we can also show for fixed $\theta \in \mathbb{R}_{\geq 0}$, $\mathcal{F}_p(\cdot, \theta)$ is continuously differentiable on $\mathbb{R}_{>0}$ with

$$\tfrac{\partial\mathcal{F}_p(\sigma,\theta)}{\partial\sigma} = -\tfrac{\theta\|\boldsymbol{\beta}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma^2 p} + \tfrac{\theta\|\boldsymbol{h}\|^2}{2p}. \tag{174}$$

From (173) and (174), we can get $\left|\frac{\partial\mathcal{F}_p(\sigma,\theta)}{\partial\theta}\right| \leq \frac{4(\sigma^2\|\boldsymbol{h}\|^2+\|\boldsymbol{\beta}\|^2)}{\sigma p}$ and $\left|\frac{\partial\mathcal{F}_p(\sigma,\theta)}{\partial\sigma}\right| \leq \frac{3\theta\|\boldsymbol{\beta}\|^2+\sigma^2(2+\theta)\|\boldsymbol{h}\|^2}{\sigma^2 p}$. Since both bounds have finite expectation, by dominated convergence theorem (DCT) we have:

$$\tfrac{\partial\overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta} = \tfrac{\mathbb{E}\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p} \tag{175}$$

and

$$\tfrac{\partial\overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\sigma} = \tfrac{-\theta\mathbb{E}\|\boldsymbol{\beta}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma^2 p} + \tfrac{\theta}{2}. \tag{176}$$

We now show $\mathcal{F}(\sigma, \cdot)$ and $\mathcal{F}(\cdot, \theta)$ are continuously differentiable on $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$, respectively. In particular, we only present the proof for $\mathcal{F}(\sigma, \cdot)$ and the case of $\mathcal{F}(\cdot, \theta)$ can be derived following same approach. The key is to establish the uniform convergence of $\frac{\partial\overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$ to $\frac{\partial\overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ on $\theta \geq 0$, where $\frac{\partial\overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta} := \lim_{p\to\infty}\frac{\partial\overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$.

First consider the case $\theta > 0$. Let $[a, b] \subset \mathbb{R}_{>0}$ and $\theta_1, \theta_2 \in [a, b]$. We have

$$\begin{aligned}\left|\tfrac{\partial \overline{\mathcal{F}}_p(\sigma,\theta_1)}{\partial\theta} - \tfrac{\partial \overline{\mathcal{F}}_p(\sigma,\theta_2)}{\partial\theta}\right| &\le \mathbb{E}\left|\tfrac{\partial \mathcal{F}_p(\sigma,\theta_1)}{\partial\theta} - \tfrac{\partial \mathcal{F}_p(\sigma,\theta_2)}{\partial\theta}\right| \\ &\overset{\text{(a)}}{\le} \tfrac{1}{\sigma}\Big(\tfrac{1}{2p}\mathbb{E}\left|\tfrac{\partial \|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{\partial\theta}\big|_{\theta=\theta'}\right| + \tfrac{1}{p}\mathbb{E}\left|\tfrac{\partial \boldsymbol{y}^\top \text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})}{\partial\theta}\big|_{\theta=\theta'}\right|\Big)|\theta_1 - \theta_2| \\ &\overset{\text{(b)}}{\le} \tfrac{2}{\sigma\sqrt{p}}\tfrac{\mathbb{E}\|\boldsymbol{y}\|}{a^2}|\theta_1 - \theta_2| \\ &\overset{\text{(c)}}{\le} C|\theta_1 - \theta_2|,\end{aligned}$$

where (a) follows from (173) and intermediate value theorem, with $\theta' \in [\theta_1, \theta_2]$, (b) follows from (183) and (184), and in (c), $C > 0$ is some constant that only depends on σ, a and b. Therefore, $\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$ is C-Lipschitz continuous on $\theta \in [a, b]$ for any $p \in \mathbb{Z}^+$. Meanwhile, we can show $\big\{\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}\big\}_{p\in\mathbb{Z}^+}$ converges pointwise to $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ on $\theta \ge 0$ following the proof of (166). Then for any $\varepsilon > 0$ we can construct a $\frac{\varepsilon}{3C}$-epsilon net $\mathcal{E}$ such that for any $\theta \in [a, b]$, there exists $z \in \mathcal{E}$ satisfying $|\theta - z| \le \frac{\varepsilon}{3C}$. Clearly, the cardinality $|\mathcal{E}| < \infty$. Therefore, for any $\varepsilon > 0$ there exists p_0 such that for any $m, p \ge p_0$, $\max_{z\in\mathcal{E}}\left|\frac{\partial \overline{\mathcal{F}}_m(\sigma,z)}{\partial\theta} - \frac{\partial \overline{\mathcal{F}}_p(\sigma,z)}{\partial\theta}\right| \le \frac{\varepsilon}{3}$. Hence for any $\varepsilon > 0$, $\theta \in [a, b]$ and $m, p \ge p_0$,

$$\begin{aligned}\left|\tfrac{\partial \overline{\mathcal{F}}_m(\sigma,\theta)}{\partial\theta} - \tfrac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}\right| &\le \left|\tfrac{\partial \overline{\mathcal{F}}_m(\sigma,\theta)}{\partial\theta} - \tfrac{\partial \overline{\mathcal{F}}_m(\sigma,z)}{\partial\theta}\right| + \left|\tfrac{\partial \overline{\mathcal{F}}_m(\sigma,z)}{\partial\theta} - \tfrac{\partial \overline{\mathcal{F}}_p(\sigma,z)}{\partial\theta}\right| + \left|\tfrac{\partial \overline{\mathcal{F}}_p(\sigma,z)}{\partial\theta} - \tfrac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}\right| \\ &\le C \times \tfrac{\varepsilon}{3C} + \tfrac{\varepsilon}{3} + C \times \tfrac{\varepsilon}{3C} = \varepsilon,\end{aligned}$$

which implies the uniform convergence of $\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$ when $\theta \in [a, b]$. Meanwhile, by DCT and continuity of $\frac{\partial \mathcal{F}_p(\sigma,\theta)}{\partial\theta}$ for any fixed $\boldsymbol{h}$, $\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$ is also continuous w.r.t. θ. Then we can apply Theorem 7.12 in [63] to obtain $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ is continuous on $[a, b]$. Since $[a, b]$ above can be arbitrary subset of $\mathbb{R}_{>0}$, we actually obtain that $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ is continuous when $\theta > 0$.

Next we handle the case when $\theta = 0$. From (175), (183) and (184), we can get for any $\theta \ge 0$

$$\tfrac{\partial^2 \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta^2} = \tfrac{2}{\theta^2}\mathbb{E}\Big\{\sum_{i=1}^p \mathbb{I}_{[\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})]_i\neq 0}\big[[\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(|\boldsymbol{y}|)]_i - |y_i|\big]\Big\} \le 0,$$

where in the first equality we use DCT and the last inequality follows from Fact 1. Hence, $\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$ is non-increasing on $\theta \ge 0$, with $\frac{\partial \overline{\mathcal{F}}_p(\sigma,0)}{\partial\theta} = \mathbb{E}\frac{\|\boldsymbol{y}\|^2}{2\sigma p}$ and $\lim_{\theta\to\infty}\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta} = 0$. Since $\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}$ converges to $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ pointwise, $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ is also non-increasing with $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,0)}{\partial\theta} = \frac{\mathbb{E}B^2+\sigma^2}{2\sigma}$ and $\lim_{\theta\to\infty}\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta} = 0$. On the other hand, from (175)

$$\begin{aligned}\tfrac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta} &\ge \mathbb{E}\Big(\tfrac{\|\boldsymbol{y}\|^2 - 2\|\boldsymbol{y}\|\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|}{2\sigma p}\Big) \\ &\overset{\text{(a)}}{\ge} \mathbb{E}\Big(\tfrac{\|\boldsymbol{y}\|^2 - 2\|\boldsymbol{y}\|\|\text{Prox}_{\sigma\overline{\boldsymbol{\lambda}}/\theta}(\boldsymbol{y})\|}{2\sigma p}\Big) \\ &\ge \tfrac{\mathbb{E}\|\boldsymbol{y}\|^2}{2\sigma p} - \tfrac{1}{\sigma}\Big(\tfrac{\mathbb{E}\|\boldsymbol{y}\|^2}{p}\Big)^{1/2}\Big[\tfrac{1}{p}\sum_{i=1}^p \mathbb{E}(|y_i| - \tfrac{\sigma}{\theta}\bar{\lambda})_+^2\Big]^{1/2},\end{aligned}$$

where step (a) follows from (188), with $\overline{\boldsymbol{\lambda}} = \frac{\sum_{i=1}^p \lambda_i}{p}\mathbf{1}_p$. Hence taking $p \to \infty$ on both sides above, we have

$$\tfrac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta} \ge \tfrac{1}{2\sigma}\Big[\mathbb{E}B^2 + \sigma^2 - 2\sqrt{(\mathbb{E}B^2+\sigma^2)\mathbb{E}(|Y| - \tfrac{\sigma}{\theta}\mathbb{E}\Lambda)_+^2}\Big].$$

Then taking $\theta \to 0^+$, we get

$$\lim_{\theta\to 0^+} \tfrac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta} \geq \lim_{\theta\to 0^+} \tfrac{1}{2\sigma}\Big[\mathbb{E}B^2+\sigma^2 - \tfrac{1}{\sigma}\sqrt{(\mathbb{E}B^2+\sigma^2)\mathbb{E}(|Y| - \tfrac{\sigma}{\theta}\mathbb{E}\Lambda)_+^2}\Big]$$
$$\overset{\text{(a)}}{=} \tfrac{\mathbb{E}B^2+\sigma^2}{2\sigma} = \tfrac{\partial \overline{\mathcal{F}}_\infty(\sigma,0)}{\partial\theta},$$

where in step (a) we use DCT and $\mathbb{E}\Lambda > 0$. Since $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ is non-increasing on $\theta \geq 0$, we can get $\lim_{\theta\to 0^+} \frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta} = \frac{\partial \overline{\mathcal{F}}_\infty(\sigma,0)}{\partial\theta}$. Recall that we have already proved $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ is continuous when $\theta > 0$, so $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ is continuous on $\mathbb{R}_{\geq 0}$.

Up to now, we have shown $\left\{\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}\right\}_{p\in\mathbb{Z}^+}$, $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ are all bounded, non-increasing and continuous functions on $\mathbb{R}_{\geq 0}$ and also $\left\{\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta}\right\}_{p\in\mathbb{Z}^+}$ converges to $\frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ pointwise on $\theta \geq 0$. Therefore, following the proof of Glivenko-Cantelli theorem (for a reference, see Theorem 19.1 in [53]) we can show $\frac{\partial \overline{\mathcal{F}}_p(\sigma,\theta)}{\partial\theta} \to \frac{\partial \overline{\mathcal{F}}_\infty(\sigma,\theta)}{\partial\theta}$ converges uniformly on $\theta \geq 0$. In (162) we prove the pointwise convergence $\overline{\mathcal{F}}_p(\sigma,\theta) \to \mathcal{F}(\sigma,\theta)$. Then using Theorem 7.17 and 7.12 in [63] together with (175), we get $\mathcal{F}(\sigma,\cdot)$ is continuously differentiable on $\mathbb{R}_{\geq 0}$, with

$$\tfrac{\partial \mathcal{F}(\sigma,\theta)}{\partial\theta} = \lim_{p\to\infty} \tfrac{\mathbb{E}\|\boldsymbol{y}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma p}, \tag{177}$$

where $\boldsymbol{y} = \boldsymbol{\beta} + \sigma\boldsymbol{h}$. Repeating the same procedure, we can also prove that $\mathcal{F}(\cdot,\theta)$ is continuously differentiable on $\mathbb{R}_{>0}$, with

$$\tfrac{\partial \mathcal{F}(\sigma,\theta)}{\partial\sigma} = \lim_{p\to\infty} \tfrac{-\theta\mathbb{E}\|\boldsymbol{\beta}-\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{2\sigma^2 p} + \tfrac{\theta}{2}. \tag{178}$$

Finally, we compute the limit in (177) and (157). From Proposition 1 and (165), we have $\frac{\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})-\eta(\boldsymbol{y})\|_2^2}{p} \overset{a.s.}{\to} 0$ as $p\to\infty$, where $\eta := \eta(\cdot;\mu_{B+\sigma H},\mu_{\sigma\Lambda/\tau})$. In addition,

$$\tfrac{1}{p}\left|\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})-\boldsymbol{\beta}\|_2^2 - \|\eta(\boldsymbol{y})-\boldsymbol{\beta}\|^2\right| \leq \tfrac{2}{p}\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})-\eta(\boldsymbol{y})\|(\|\boldsymbol{y}\|+\|\boldsymbol{\beta}\|),$$

so we can get

$$\tfrac{1}{p}\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})-\boldsymbol{\beta}\|^2 \overset{a.s.}{\to} \tfrac{1}{p}\|\eta(\boldsymbol{y})-\boldsymbol{\beta}\|^2 \overset{a.s.}{\to} \mathbb{E}[\eta(Y)-B]^2, \tag{179}$$

where the last step follows from Theorem 7.12 (iv) in [56] after combining (164) with the fact that $[\eta(b+\sigma h)-b]^2 \leq C(1+h^2+b^2)$ for any $h,b\in\mathbb{R}$ and some $C>0$. Then similar as (166), we can obtain

$$\tfrac{1}{p}\mathbb{E}\|\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})-\boldsymbol{\beta}\|^2 \to \mathbb{E}[\eta(Y)-B]^2. \tag{180}$$

On the other hand, we can also get

$$\tfrac{1}{p}\boldsymbol{h}^\top\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) \overset{a.s.}{\to} \tfrac{1}{p}\boldsymbol{h}^\top\eta(\boldsymbol{y}) \overset{a.s.}{\to} \mathbb{E}[\eta(Y)H] \tag{181}$$

and

$$\tfrac{1}{p}\mathbb{E}[\boldsymbol{h}^\top\text{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})] \to \mathbb{E}[\eta(Y)H] = \sigma\mathbb{E}\eta'(Y). \tag{182}$$

where in the last step we use Stein's lemma. Substituting (180) and (182) into (177) and (178), we reach at (167) and (168). ■

Lemma 17: For any $\theta \geq 0$, $\boldsymbol{y} \in \mathbb{R}^p$ and $\boldsymbol{\lambda} \in \mathbb{R}^p_{\geq 0}$, we have

$$\frac{\partial \|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{\partial \theta} = \begin{cases} 0 & \boldsymbol{\lambda} = \boldsymbol{0}, \\ \frac{2\cdot \boldsymbol{1}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(|\boldsymbol{y}|)}{\theta^2} & \boldsymbol{\lambda} \neq \boldsymbol{0}, \end{cases} \tag{183}$$

and

$$\frac{\partial \boldsymbol{y}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})}{\partial \theta} = \begin{cases} 0 & \boldsymbol{\lambda} = \boldsymbol{0}, \\ \frac{1}{\theta^2}\sum_{i=1}^p |y_i| \mathbb{I}_{[\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})]_i \neq 0} & \boldsymbol{\lambda} \neq \boldsymbol{0}. \end{cases} . \tag{184}$$

Here when $\boldsymbol{\lambda} \neq \boldsymbol{0}$ and $\theta = 0$, we let $\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) := \boldsymbol{0}$.

Proof: When $\boldsymbol{\lambda} = \boldsymbol{0}$, we have $\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) = \boldsymbol{y}$, so $\|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2 = \boldsymbol{y}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) = \|\boldsymbol{y}\|^2$ and $\frac{\partial \|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{\partial \theta} = \frac{\partial \boldsymbol{y}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})}{\partial \theta} = 0$.

Next we consider $\boldsymbol{\lambda} \neq \boldsymbol{0}$. From Lemma 2.3 and 2.4 of [9], for any $\boldsymbol{a} \in \mathbb{R}^p$ satisfying $0 \leq a_1 \leq a_2 \leq \cdots \leq a_p$,

$$\frac{\partial [\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_i}{\partial \lambda_j} = -\frac{\mathbb{I}_{i \in I_j}}{\max\{|I_j|,1\}},$$

where $I_j \overset{\text{def}}{=} \big\{k \in [p] \mid |[\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_k| = |[\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_j| \text{ and } [\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_k \neq 0\big\}$. Therefore,

$$\frac{\partial \|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\|^2}{\partial \lambda_j} = \sum_{i=1}^p 2[\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_i \frac{\partial [\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_i}{\partial \lambda_j} = -2[\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_j \tag{185}$$

and

$$\frac{\partial \boldsymbol{a}^\top \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})}{\partial \lambda_j} = \sum_{i=1}^p a_i \frac{\partial [\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})]_i}{\partial \lambda_j} = -\frac{1}{\max\{|I_j|,1\}} \sum_{i \in I_j} a_i. \tag{186}$$

On the other hand, by Fact 1, $\|\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})\|^2$ and $\boldsymbol{a}^\top \mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{a})$ only depend on $\mu_{\boldsymbol{\lambda}}$ and $\mu_{|\boldsymbol{a}|}$. Therefore, for any $\boldsymbol{y} \in \mathbb{R}^p$ and $\theta > 0$, it holds that $\frac{\partial \|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2}{\partial \theta} = \frac{2}{\theta^2} \boldsymbol{1}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(|\boldsymbol{y}|)$ and $\frac{\partial \boldsymbol{y}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})}{\partial \theta} = \frac{1}{\theta^2}\sum_{i=1}^p |y_i| \mathbb{I}_{[\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})]_i \neq 0}$ by chain rule. For $\theta = 0$, we need to study the behavior of $\|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|$ when θ is closed to 0. It can be shown (for a proof, see (A.11) in [7])

$$\mathrm{Prox}_{\boldsymbol{\lambda}}(\boldsymbol{y}) = \boldsymbol{y} - \mathrm{Proj}_{C_{\boldsymbol{\lambda}}}(\boldsymbol{y}), \tag{187}$$

where $C_{\boldsymbol{\lambda}} = \{\boldsymbol{\nu} \in \mathbb{R}^p \mid \boldsymbol{\nu} \prec \boldsymbol{\lambda}\}$ ("$\prec$" denotes majorization, see Definition 2) is the unit ball of the dual norm of $J_{\boldsymbol{\lambda}}$ [9, Proposition 1.1] and $\mathrm{Proj}_{C_{\boldsymbol{\lambda}}}$ is the orthogonal projection onto $C_{\boldsymbol{\lambda}}$. Take $\overline{\lambda} = \frac{1}{p}\sum_{i=1}^p \lambda_i$ and it is not hard to show $\overline{\boldsymbol{\lambda}} := \overline{\lambda}\boldsymbol{1}_p$ satisfies $\overline{\boldsymbol{\lambda}} \prec \boldsymbol{\lambda}$. Clearly, $\frac{\sigma}{\theta}\overline{\boldsymbol{\lambda}} \prec \frac{\sigma}{\theta}\boldsymbol{\lambda}$ for $\sigma, \theta > 0$ and $C_{\sigma\overline{\boldsymbol{\lambda}}/\theta} \subset C_{\sigma\boldsymbol{\lambda}/\theta}$, so from (187) we have

$$\|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2 \leq \|\mathrm{Prox}_{\sigma\overline{\boldsymbol{\lambda}}/\theta}(\boldsymbol{y})\|^2 = \sum_{i=1}^p \max\big\{|y_i| - \tfrac{\sigma\bar{\lambda}}{\theta}, 0\big\}^2. \tag{188}$$

Since $\overline{\lambda} > 0$ (due to $\boldsymbol{\lambda} \neq \boldsymbol{0}$), (188) indicates that when $0 < \theta \leq \frac{\sigma\overline{\lambda}}{\max(\{|y_i|\},1)}$, $\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) = \boldsymbol{0}$. On the other hand, we let $\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) := \boldsymbol{0}$, when $\boldsymbol{\lambda} \neq \boldsymbol{0}$ and $\theta = 0$. As a result, $\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y}) = \boldsymbol{0}$ for $\theta \in \big[0, \frac{\sigma\overline{\lambda}}{\max(\{|y_i|\},1)}\big]$ and combining the partial derivatives of $\|\mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})\|^2$ and $\boldsymbol{y}^\top \mathrm{Prox}_{\sigma\boldsymbol{\lambda}/\theta}(\boldsymbol{y})$ on $\theta > 0$ obtained above, we can get (183) and (184). ■

L. Proof of Proposition 3

It directly follows from Proposition 1 that $\mathcal{M}_{\mu_Y} \subseteq \mathcal{I}$, so we just need to show $\mathcal{I} \subseteq \mathcal{M}_{\mu_Y}$.

For any $f \in \mathcal{I}$, consider the function $r(y) = y - f(y)$. It can be easily verified that $r(y) \in \mathcal{I}$. We claim that if we choose $\Lambda \sim r(|Y|)$ with $Y \sim \mu_Y$, then f is the optimal solution of (6) (when $\tau = 1$). Indeed, when $\tau = 1$ and $\Lambda \sim r(|Y|)$, (6) can be equivalently written as

$$\begin{aligned}\text{Problem (6)} &\overset{\text{(a)}}{=} \min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}_{\mu_Y}[|Y| - g(|Y|)]^2 + \int_0^1 F^{-1}_{r(|Y|)}(u)F^{-1}_{g(|Y|)}(u)du \\ &\overset{\text{(b)}}{=} \min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}_{\mu_Y}[|Y| - g(|Y|)]^2 + \mathbb{E}_{\mu_Y}[|Y| - f(|Y|)]g(|Y|) \\ &= \min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}_{\mu_Y}[f(|Y|) - g(|Y|)]^2 + \frac{1}{2}\mathbb{E}_{\mu_Y}[Y^2 - f^2(|Y|)], \end{aligned} \tag{189}$$

where (a) follows from $g \in \mathcal{I}$ and in (b) we substitute $r(y) = y - f(y)$ and use the fact that $r \in \mathcal{I}$. From (189), we can see f is the optimal solution of (6). On the other hand, since $f \in \mathcal{I}$ and $\mathbb{E}Y^2 < \infty$, we can verify that $\mathbb{E}\Lambda^2 < \infty$ and hence $\Lambda \in \mathcal{P}_2(\mathbb{R})$. In conclusion, for any $f \in \mathcal{I}$, we can always choose $\Lambda \sim |Y| - f(|Y|)$ satisfying $\Lambda \in \mathcal{P}_2(\mathbb{R})$, such that f is the optimal solution of (6) (when $\tau = 1$). By Proposition 1, this means $f(y) = \eta(y; \mu_Y, \mu_\Lambda)$ and hence $f \in \mathcal{M}_{\mu_Y}$. Therefore, $\mathcal{I} \subseteq \mathcal{M}_{\mu_Y}$.

M. Auxiliary Results for Proving Proposition 4

Lemma 18: For any $\sigma > 0$, we have: (I) optimization problem (28) is convex and always has a unique optimal solution $f_\sigma \in \mathcal{I}$, (II) $\mathcal{L}(\sigma)$ defined in (28) is continuous at σ.

Proof: (I) Optimization problem (28) can be equivalently written as:

$$\begin{aligned}&\min_{f\in\mathcal{I}} \mathbb{E}_{\mu_Y}[f(Y) - \mathbb{E}(B \mid Y)]^2 + \mathbb{E}_{\mu_Y}[\mathrm{Var}(B \mid Y)] \\ &\text{s.t. } \mathbb{E}_{\mu_Y} f'(Y) \le \delta.\end{aligned}$$

Then by the same arguments in the proof of Lemma 8, it is not hard to check for any $\sigma > 0$, it is (strongly) convex and has a unique solution $f_\sigma \in \mathcal{I}$.

(II) Next, we prove the continuity of $\mathcal{L}(\sigma)$ at any $\sigma > 0$. Define the following set

$$\mathcal{I}_\sigma \overset{\text{def}}{=} \{f \mid f \in \mathcal{I} \text{ and } \mathbb{E}f'(B + \sigma H) \le \delta\}, \tag{190}$$

where $H \sim \mathcal{N}(0,1)$ and $B \sim \mu_B$ are independent. Note that for any $\sigma > 0$, we have $\mathcal{I}_\sigma \neq \emptyset$, since $\{f = 0\} \in \mathcal{I}_\sigma$.

The first step is to show for any $\sigma, r > 0$, there exists $\varepsilon \in (0, \sigma/2)$ such that whenever $\widehat{\sigma} \in \mathcal{B}_\varepsilon(\sigma)$ and $\widehat{f} \in \mathcal{I}_{\widehat{\sigma}}$, we can always find a $f \in \mathcal{I}_\sigma$ satisfying $|f(x) - \widehat{f}(x)| < r$ almost everywhere on $\mathbb{R}$. This can be proved as follows. If $\widehat{f} \in \mathcal{I}_\sigma$, we can choose $f = \widehat{f}$, which trivially satisfies $|f(x) - \widehat{f}(x)| < r$; if $\widehat{f} \notin \mathcal{I}_\sigma$, then $\mathbb{E}\widehat{f}'(B + \sigma H) > \delta$. Since $\hat{f} \in \mathcal{I}_{\widehat{\sigma}} \subseteq \mathcal{I}$, it follows that $|\widehat{f}'| \le 1$ almost everywhere on $\mathbb{R}$. Meanwhile, since $\sigma, \widehat{\sigma} > 0$, by the properties

of Gaussian convolution we know both $B+\hat{\sigma}H$ and $B+\sigma H$ have smooth density functions supported on $\mathbb{R}$. Denote their density functions as q_1 and q_2. Then we have

$$\begin{aligned}|\mathbb{E}\widehat{f}'(B+\widehat{\sigma}H)-\mathbb{E}\widehat{f}'(B+\sigma H)| =&\Big|\int \widehat{f}'(y)[q_1(y)-q_2(y)]dy\Big|\\ \leq&\int |q_1(y)-q_2(y)|dy\\ =&2\mathrm{TV}\big(\mu_{B+\widehat{\sigma}H},\mu_{B+\sigma H}\big)\\ \overset{\text{(a)}}{\leq}&2\mathrm{TV}\big(\mu_{\widehat{\sigma}H},\mu_{\sigma H}\big)\\ \overset{\text{(b)}}{\leq}&\sqrt{2\mathrm{KL}(\mu_{\widehat{\sigma}H},\mu_{\sigma H})}\\ \overset{\text{(c)}}{\leq}&C|\widehat{\sigma}-\sigma|,\end{aligned} \tag{191}$$

where $\mathrm{TV}(\cdot,\cdot)$ and $\mathrm{KL}(\cdot,\cdot)$ denote the total variation distance and KL-divergence between two probability measures and $C>0$ is a fixed constant only depending on σ. In reaching (191), (b) follows from Pinsker's inequality and (c) follows from standard results of KL-divergence between Gaussian random variables and the fact that $\varepsilon\in(0,\sigma/2)$ and $\widehat{\sigma}\in\mathcal{B}_\varepsilon(\sigma)$. Step (a) in (191) can be obtained as follows. Recall that $\Pi(\mu_1,\mu_2)$ denotes the set of all couplings between measures μ_1 and μ_2. Then for any $(\widehat{\sigma}H_1,\sigma H_2)\in\Pi(\mu_{\widehat{\sigma}H},\mu_{\sigma H})$ and $B_0\sim\mu_B$ independent of $(\widehat{\sigma}H_1,\sigma H_2)$, we have $(B_0+\widehat{\sigma}H_1,B_0+\sigma H_2)\in\Pi(\mu_{B+\widehat{\sigma}H},\mu_{B+\sigma H})$. Therefore,

$$\begin{aligned}\mathrm{TV}\big(\mu_{B+\widehat{\sigma}H},\mu_{B+\sigma H}\big)\overset{\text{(a)}}{=}&\inf_{(Y_1,Y_2)\in\Pi(\mu_{B+\widehat{\sigma}H},\mu_{B+\sigma H})}\mathbb{P}(Y_1\neq Y_2)\\ \leq&\inf_{\substack{(\widehat{\sigma}H_1,\sigma H_2)\in\Pi(\mu_{\widehat{\sigma}H},\mu_{\sigma H})\\ B_0\sim\mu_B\text{ indep. of }(\widehat{\sigma}H_1,\sigma H_2)}}\mathbb{P}(B_0+\widehat{\sigma}H_1\neq B_0+\sigma H_2)\\ =&\inf_{(\widehat{\sigma}H_1,\sigma H_2)\in\Pi(\mu_{\widehat{\sigma}H},\mu_{\sigma H})}\mathbb{P}(\widehat{\sigma}H_1\neq\sigma H_2)\\ \overset{\text{(b)}}{=}&\mathrm{TV}\big(\mu_{\widehat{\sigma}H},\mu_{\sigma H}\big),\end{aligned} \tag{192}$$

where (a) and (b) follow from Strassen's Theorem [56, p.7]. From (191), when $\varepsilon\in(0,\sigma/2)$ and $\widehat{\sigma}\in\mathcal{B}_\varepsilon(\sigma)$,

$$|\mathbb{E}\widehat{f}'(B+\widehat{\sigma}H)-\mathbb{E}\widehat{f}'(B+\sigma H)|\leq C\varepsilon. \tag{193}$$

Since $\widehat{f}\in\mathcal{I}_{\hat{\sigma}}$, from (190) and (193) we get

$$\mathbb{E}\widehat{f}'(B+\sigma H)\leq\delta+C\varepsilon. \tag{194}$$

We now slightly shrink $\widehat{f}$ to obtain the $f\in\mathcal{I}_\sigma$ satisfying $|f(x)-\widehat{f}(x)|<r$ almost everywhere. On one hand, we know $B+\sigma H$ has a density function supported on $\mathbb{R}$ for $\sigma>0$. On the other hand, since $\mathbb{E}\widehat{f}'(B+\sigma H)>\delta$ (due to $\widehat{f}\notin\mathcal{I}_\sigma$) and $|\widehat{f}'|\leq 1$ almost everywhere, it is not hard to show if $\varepsilon\leq\frac{\delta}{2C}$, there exists some $\mathcal{A}\subset\mathbb{R}_{>0}$ such that $\mathbb{E}[\widehat{f}'(Y)\mathbb{I}_{Y\in\mathcal{A}}]=C\varepsilon$. Accordingly, we set

$$f'(y)=\begin{cases}0 & \pm y\in\mathcal{A},\\ \widehat{f}'(y) & \text{otherwise}\end{cases}$$

and choose $f(y)$ to be:

$$f(y)=\int_0^y f'(t)dt. \tag{195}$$

With this choice, from (194) we have $\mathbb{E}f'(Y) \leq \delta - C\varepsilon$ and thus $f \in \mathcal{I}_\sigma$. In addition, $0 \leq \widehat{f}(x) - f(x) \leq 2C\varepsilon$ almost everywhere. Hence we can choose $\varepsilon \leq \frac{r}{3C}$ so that $0 \leq \widehat{f}(x) - f(x) < r$ almost everywhere. Summing up, for any $\sigma, r > 0$, there exists C depending only on σ such that whenever $\varepsilon \leq \min\{\frac{r}{3C}, \frac{\sigma}{2}\}$, $\widehat{\sigma} \in \mathcal{B}_\varepsilon(\sigma)$ and $\widehat{f} \in \mathcal{I}_{\hat{\sigma}}$, we can always find a $f \in \mathcal{I}_\sigma$ satisfying $0 \leq \widehat{f}(x) - f(x) < r$ almost everywhere on $\mathbb{R}$.

Define $\mathcal{L}(f, \sigma) \stackrel{\text{def}}{=} \mathbb{E}[f(B + \varsigma H) - B]^2$, which is the objective function in (28). For any compact interval $I_B \subseteq \mathbb{R}_{>0}$ and any $\sigma_1, \sigma_2 \in I_B$, $f \in \mathcal{I}$, we have

$$\begin{aligned}
|\mathcal{L}(f, \sigma_1) - \mathcal{L}(f, \sigma_2)| &\leq \mathbb{E}\left|f^2(B + \sigma_1 H) - f^2(B + \sigma_2 H)\right| \\
&\quad + 2\mathbb{E}|B||f(B + \sigma_1 H) - f(B + \sigma_2 H)| \\
&\leq \left(\sqrt{\mathbb{E}(|B + \sigma_1 H| + |B + \sigma_2 H|)^2} + 2\sqrt{\mathbb{E}B^2}\right)|\sigma_1 - \sigma_2| \\
&\leq C_1|\sigma_1 - \sigma_2|, \qquad (196)
\end{aligned}$$

where C_1 is a constant that only depends on I_B. Therefore, for any $f \in \mathcal{I}$, $\mathcal{L}(f, \sigma)$ is uniformly Lipschitz continuous w.r.t. σ on I_B. Consider $f_{\hat{\sigma}}$ which is the optimal solution of (28) under $\hat{\sigma}$. From the discussion in the last paragraph, for any $\sigma \in I_B$ and $r > 0$, if $\widehat{\sigma} \in \mathcal{B}_\varepsilon(\sigma) \bigcap I_B$ under small enough ε, then there exists some $f \in \mathcal{I}_\sigma$ satisfying $|f_{\hat{\sigma}}(x) - f(x)| < r$ almost everywhere on $\mathbb{R}$. Therefore, for this f we have

$$\begin{aligned}
|\mathcal{L}(f_{\hat{\sigma}}, \widehat{\sigma}) - \mathcal{L}(f, \sigma)| &\leq |\mathcal{L}(f_{\hat{\sigma}}, \widehat{\sigma}) - \mathcal{L}(f_{\hat{\sigma}}, \sigma)| + |\mathcal{L}(f_{\hat{\sigma}}, \sigma) - \mathcal{L}(f, \sigma)| \\
&\leq C_2 r, \qquad (197)
\end{aligned}$$

where C_2 is some constant that does not depend on ε, r and in the last step we use (196) and the fact that $|f_{\hat{\sigma}}(x) - f(x)| < r$ almost everywhere. Since $\mathcal{L}(\widehat{\sigma}) = \mathcal{L}(f_{\hat{\sigma}}, \widehat{\sigma})$, we have from (197)

$$\mathcal{L}(\widehat{\sigma}) \stackrel{\text{(a)}}{\geq} \mathcal{L}(f, \sigma) - C_2 r \stackrel{\text{(b)}}{\geq} \mathcal{L}(\sigma) - C_2 r,$$

where (a) follows from (197) and (b) follows from definition of $\mathcal{L}(\sigma)$ in (28). By exchanging σ and $\hat{\sigma}$, we can also get $\mathcal{L}(\sigma) \geq \mathcal{L}(\widehat{\sigma}) - C_2 r$. In conclusion, for any compact interval $I_B \subseteq \mathbb{R}_{>0}$ and $r > 0$, there exists $\varepsilon > 0$ such that for any $\sigma, \hat{\sigma} \in I_B$ satisfying $|\hat{\sigma} - \sigma| \leq \varepsilon$, we have $|\mathcal{L}(\widehat{\sigma}) - \mathcal{L}(\sigma)| \leq r$. This proves the continuity of $\mathcal{L}(\sigma)$ over $\mathbb{R}_{>0}$. ■

Lemma 19: Equation (29) satisfies the following: (I) it always has a solution and the minimum solution $\sigma_0 \in [\sigma_w, (\sigma_w^2 + \delta^{-1}\mathbb{E}B^2)^{1/2}]$, (II) $\sigma_0 = \inf \mathcal{A}$, where set $\mathcal{A}$ is defined in (70).

Proof: (I) We first prove σ_0 always exists and $\sigma_0 \in I_B$, where $I_B := [\sigma_w, (\sigma_w^2 + \delta^{-1}\mathbb{E}B^2)^{1/2}]$. It is not hard to verify that at the boundary of I_B, $\mathcal{L}(\sigma)$ satisfies

$$\begin{cases}
\mathcal{L}(\sigma) \leq \delta(\sigma^2 - \sigma_w^2) & \sigma = \left(\sigma_w^2 + \frac{\mathbb{E}B^2}{\delta}\right)^{1/2}, \\
\mathcal{L}(\sigma) \geq \delta(\sigma^2 - \sigma_w^2) & \sigma = \sigma_w.
\end{cases} \qquad (198)$$

Indeed, when $\sigma = \left(\sigma_w^2 + \frac{\mathbb{E}B^2}{\delta}\right)^{1/2}$, set $f = 0$ in (28) and $\mathbb{E}[f(B + \sigma H) - B]^2 = \mathbb{E}B^2 = \delta(\sigma^2 - \sigma_w^2)$, so $\mathcal{L}(\sigma) \leq \delta(\sigma^2 - \sigma_w^2)$. On the other hand, since $\mathcal{L}(\sigma) \geq 0$, the second inequality of (198) immediately follows. Then by (198), the continuity of $\mathcal{L}(\sigma)$ shown in Lemma 18 and the fact that $\sigma_0 \geq \sigma_w$, we know σ_0 always exists and $\sigma_0 \in I_B$.

(II) To prove $\inf \mathcal{A} = \sigma_0$, we proceed as follows:

(i) Show

$$\inf \mathcal{A} \in I_B. \tag{199}$$

(ii) Prove the following membership certificate of set $\mathcal{A}$ for those $\sigma \in I_B$:

$$\sigma \in \mathcal{A} \Longleftrightarrow \mathcal{L}(\sigma) \leq \delta(\sigma^2 - \sigma_w^2). \tag{200}$$

It is not hard to see the above steps will imply $\inf \mathcal{A} = \sigma_0$. Indeed, combining (200) with (199) yields the following characterization of $\inf \mathcal{A}$:

$$\inf \mathcal{A} = \inf \big\{\sigma \mid \sigma \in I_B \text{ and } \mathcal{L}(\sigma) \leq \delta(\sigma^2 - \sigma_w^2)\big\}. \tag{201}$$

By (198) and the continuity of $\mathcal{L}(\sigma)$, the infimum on the RHS of (201) is reached by σ_0, which always exists. Therefore, $\inf \mathcal{A} = \sigma_0$. Step (i)-(ii) can be proved as follows:

[Proof of (i)] From the first equation of (69), we have $\inf \mathcal{A} \geq \sigma_w$. On the other hand, since $f = 0$, $\sigma = \big(\sigma_w^2 + \frac{\mathbb{E}B^2}{\delta}\big)^{1/2}$, $\tau = 1$ is a solution of (66)-(67), from (68) we have $\sigma_{\text{opt}} \leq \big(\sigma_w^2 + \frac{\mathbb{E}B^2}{\delta}\big)^{1/2}$. This together with the lower bound of σ_{opt} in (71) indicates: $\inf \mathcal{A} \leq \sigma_{\text{opt}} \leq \big(\sigma_w^2 + \frac{\mathbb{E}B^2}{\delta}\big)^{1/2}$. Therefore, $\inf \mathcal{A} \in I_B$.

[Proof of (ii)] The "$\Rightarrow$" direction of (200) immediately follows from the definition of $\mathcal{A}$ in (70). For the other direction, suppose we have a $\sigma \in I_B$ satisfying $\mathcal{L}(\sigma) \leq \delta(\sigma^2 - \sigma_w^2)$. Then

$$\mathbb{E}[f_\sigma(B + \sigma H) - B]^2 = \mathcal{L}(\sigma) \leq \delta(\sigma^2 - \sigma_w^2), \tag{202}$$

where the first equality is due to that $\mathcal{L}(\sigma)$ can be achieved by f_σ. Now consider the shrinkage of f_σ as: αf_σ, where $\alpha \in [0, 1]$. Clearly, for any $\alpha \in [0, 1]$, αf_σ still satisfies $\mathbb{E}[\alpha f_\sigma'(B + \sigma H)] \leq \delta$. Also we have:

$$\mathbb{E}[0 \cdot f_\sigma(B + \sigma H) - B]^2 = \mathbb{E}B^2 \geq \delta(\sigma^2 - \sigma_w^2), \tag{203}$$

where the last inequality follows from the condition that $\sigma \in I_B$. On the other hand, it can be easily checked that $\alpha \mapsto \mathbb{E}[\alpha f_\sigma(B + \sigma H) - B]^2$ is continuous, so from (203), (202), there exists $\alpha_0 \in [0, 1]$ such that $(\alpha_0 f_\sigma, \sigma)$ is a solution of (69), indicating $\sigma \in \mathcal{A}$. ■

N. Auxiliary Results for Proving Proposition 5

Lemma 20: For a probability measure $\mu_Y \in \mathcal{P}_2(\mathbb{R})$, define

$$\begin{aligned}\widetilde{\mathcal{M}}_{\mu_Y} \stackrel{\text{def}}{=} \big\{\eta(\cdot\,; \mu_Y, \mu_\Lambda) \mid \mu_\Lambda \in \mathcal{P}_2(\mathbb{R}) \text{ and if } q_0 > 0,\\ \textstyle\int_t^{q_0} F_\Lambda^{-1}(u)du > \int_t^{q_0} F_{|Y|}^{-1}(u)du, \forall t \in [0, q_0)\big\}\end{aligned} \tag{204}$$

where $q_0 \stackrel{\text{def}}{=} \mathbb{P}\big(\eta(Y; \mu_Y, \mu_\Lambda) = 0\big)$. Then for any $\mu_Y \in \mathcal{P}_2(\mathbb{R})$, we have $\widetilde{\mathcal{M}}_{\mu_Y} = \mathcal{I}$. Correspondingly, for any $f(y) \in \mathcal{I}$, we can take μ_Λ as the law of $\max\{y_0, |Y| - f(|Y|)\}$ ($Y \sim \mu_Y$), so that $\eta(y\,; \mu_Y, \mu_\Lambda) = f(y)$. Here $y_0 \stackrel{\text{def}}{=} \sup_{y \geq 0}\{y \mid f(y) = 0\}$,

Proof: The proof is similar as Proposition 3. When μ_Λ is the law of $\max\{y_0, |Y| - f(|Y|)\}$, optimization (6) ($\tau = 1$) becomes:

$$\begin{aligned}\min_{g\in\mathcal{I}} \frac{1}{2}\mathbb{E}\big\{[f(|Y|) - g(|Y|)]^2\mathbb{I}_{|Y|\geq y_0}\big\} + \frac{1}{2}\mathbb{E}\big\{[(|Y| - y_0) - g(|Y|)]^2\mathbb{I}_{|Y|<y_0}\big\} \\ + \frac{1}{2}\mathbb{E}Y^2 - \mathbb{E}[|Y| - f(|Y|)]^2.\end{aligned} \tag{205}$$

Since the feasible set in (205) is $\mathcal{I}$, we know the optimal solution of (205) is exactly f. Recall that $\eta(y; \mu_Y, \mu_\Lambda)$ is the optimal solution of (6) ($\tau = 1$), so we have $\eta(y; \mu_Y, \mu_\Lambda) = f(y)$.

Finally, we show the law of $\max\{y_0, |Y| - f(|Y|)\}$ satisfies the constraint in (204). Since $\mu_Y \in \mathcal{P}_2(\mathbb{R})$ and $f \in \mathcal{I}$, we can easily get $\mu_\Lambda \in \mathcal{P}_2(\mathbb{R})$. On the other hand, suppose $q_0 > 0$. For any $t \in [0, q_0)$, we have

$$\begin{aligned}\int_t^{q_0} F^{-1}_{|Y|}(u)du &= \mathbb{E}\big(\mathbb{I}_{F^{-1}_{|Y|}(t)\leq|Y|\leq F^{-1}_{|Y|}(q_0)} \cdot |Y|\big) \\ &\overset{\text{(a)}}{\leq} \mathbb{E}\big(\mathbb{I}_{F^{-1}_{|Y|}(t)\leq|Y|\leq y_0} \cdot |Y|\big) \\ &\overset{\text{(b)}}{<} y_0\mathbb{P}\big(F^{-1}_{|Y|}(t) \leq |Y| \leq y_0\big) \\ &= y_0(q_0 - t) \\ &\overset{\text{(c)}}{=} \int_t^{q_0} F^{-1}_\Lambda(u)du,\end{aligned}$$

where in (a) we use the fact that $F^{-1}_{|Y|}(q_0) \leq y_0$, since $\eta(y; \mu_Y, \mu_\Lambda) = f(y)$ and $q_0 = \mathbb{P}\big(\eta(Y; \mu_Y, \mu_\Lambda) = 0\big)$, (b) follows from $F^{-1}_{|Y|}(t) < F^{-1}_{|Y|}(q_0) = y_0$, since $t < q_0$ and (c) is due to our choice of Λ, which yields $F^{-1}_\Lambda(u) = y_0$ for any $0 \leq u \leq q_0$. ■

Lemma 21: For $\alpha \in [0, 1]$ and $\sigma > 0$, we have: (I) optimization problem (41) is convex and always has a unique optimal solution $f_{\alpha,\sigma} \in \mathcal{I}$, (II) $\mathcal{L}_\alpha(\sigma)$ defined in (41) is a continuous function over $\mathbb{R}_{>0}$, (III) equation $\mathcal{L}_\alpha(\sigma) = \delta(\sigma^2 - \sigma_w^2)$ always has a solution and the minimum solution $\sigma_{0,\alpha} \in \big[\sigma_w, \sqrt{\sigma_w^2 + \delta^{-1}\mathbb{E}B^2}\big]$.

Proof: (I) Comparing with (28) and (41), we can find the only difference is that in (41), we add a constraint $f \in \mathcal{F}_{\alpha,\sigma}$. It is not hard to check for any $\alpha \in [0, 1]$ and $\sigma > 0$, the set $\mathcal{F}_{\alpha,\sigma}$ is convex and closed in the L^2 space $\mathcal{H}_{\mu_Y}$ [definition can be found in (84)], so the uniqueness of optimal solution of (41) still holds using the same arguments.

(II) The case of $\alpha = 0$ or 1 is easy. When $\alpha = 1$, we have $\mathcal{I} \subseteq \mathcal{F}_{\alpha,\sigma}$, so $\mathcal{L}_\alpha(\sigma) = \mathcal{L}(\sigma)$ and its continuity is proved in the last part of Lemma 18; when $\alpha = 0$, $\mathcal{F}_{\alpha,\sigma}$ contains only one element: $f(x) = 0$ and $\mathcal{L}_\alpha(\sigma) = \mathbb{E}B^2$, which is trivially continuous. Therefore, it only remains to verify for the case when $\alpha \in (0, 1)$.

The proof for case $\alpha \in (0, 1)$ is similar to the proof of continuity of $\mathcal{L}(\sigma)$ in Lemma 18. For any $\alpha \in (0, 1)$ and $\sigma > 0$, define the following set

$$\mathcal{I}_{\alpha,\sigma} \overset{\text{def}}{=} \{f \mid f \in \mathcal{I} \cap \mathcal{F}_{\alpha,\sigma} \text{ and } \mathbb{E}f'(B + \sigma H) \leq \delta\}, \tag{206}$$

where $H \sim \mathcal{N}(0, 1)$ and $B \sim \mu_B$ are independent. We always have $\mathcal{I}_{\alpha,\sigma} \neq \emptyset$, since $\{f = 0\} \in \mathcal{I}_{\alpha,\sigma}$. Next we show that for any $\alpha \in (0, 1)$ and $\sigma, r > 0$, there exists $\varepsilon \in (0, \sigma/2)$ such that whenever $\widehat{\sigma} \in \mathcal{B}_\varepsilon(\sigma)$ and $\widehat{f} \in \mathcal{I}_{\alpha,\widehat{\sigma}}$,

we can always find a $f \in \mathcal{I}_{\alpha,\sigma}$ satisfying $|f(x) - \widehat{f}(x)| < r$ almost everywhere on $\mathbb{R}$. First, for any $\hat{\sigma} > 0$ and $\widehat{f} \in \mathcal{I}_{\alpha,\hat{\sigma}}$, we have $|\widehat{f}(y)| = 0$, when $|y| \leq \Phi^{-1}(1-\frac{\alpha}{2})\hat{\sigma}$. We can then apply the following shrinkage to $\widehat{f}$:

$$\widehat{f}_T(y) = \text{sign}(y)\max\{0, \widehat{f}(|y|) - \Phi^{-1}(1-\tfrac{\alpha}{2})|\sigma - \hat{\sigma}|\}. \tag{207}$$

One can check $\widehat{f}_T$ in (207) satisfies: $\widehat{f}_T \in \mathcal{I}_{\alpha,\hat{\sigma}} \cap \mathcal{F}_{\alpha,\sigma}$ and

$$0 \leq \hat{f}(y) - \widehat{f}_T(y) \leq \Phi^{-1}(1-\tfrac{\alpha}{2})|\sigma - \hat{\sigma}| \tag{208}$$

almost everywhere on $\mathbb{R}$. For small enough $\varepsilon > 0$ and $\widehat{\sigma} \in \mathcal{B}_\varepsilon(\sigma)$, we can then follow the same steps leading to (195) in the proof of continuity of $\mathcal{L}(\sigma)$ in Lemma 18 to obtain a $f \in \mathcal{I}$ satisfying

$$0 \leq \widehat{f}_T(y) - f(y) \leq \frac{r}{2} \tag{209}$$

almost everywhere on $\mathbb{R}$ and $\mathbb{E}f'(B+\sigma H) \leq \delta$. Meanwhile, since $\widehat{f}_T \in \mathcal{F}_{\alpha,\sigma}$, we also have $f \in \mathcal{F}_{\alpha,\sigma}$. As a result, $f \in \mathcal{I}_{\alpha,\sigma}$. Besides, combining (208) and (209) we have for small enough $\varepsilon > 0$ and $\widehat{\sigma} \in \mathcal{B}_\varepsilon(\sigma)$, $0 \leq \hat{f}(y) - f(y) \leq r$ almost everywhere on $\mathbb{R}$. The remaining proof is completely same as the last part of the proof of $\mathcal{L}(\sigma)$'s continuity and we omit the details here.

(III) The proof is the same as Lemma 19. Using the same argument, it can be verified that $\mathcal{L}_\alpha(\sigma)$ also satisfies

$$\begin{cases} \mathcal{L}_\alpha(\sigma) \leq \delta(\sigma^2 - \sigma_w^2) & \sigma = \left(\sigma_w^2 + \frac{\mathbb{E}B^2}{\delta}\right)^{1/2}, \\ \mathcal{L}_\alpha(\sigma) \geq \delta(\sigma^2 - \sigma_w^2) & \sigma = \sigma_w. \end{cases}$$

Then by the continuity of $\mathcal{L}_\alpha(\sigma)$ and the fact that $\sigma_{0,\alpha} \geq \sigma_w$, we get the desired result. ■

Lemma 22: For any given $\alpha \in [0,1]$, we have the following.

(a) It is always true that $\sigma_{\text{opt},\alpha} \geq \sigma_{0,\alpha}$.

(b) If $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$, then $\sigma_{\text{opt},\alpha} = \sigma_{0,\alpha}$ and the infimum of (73) can be achieved by choosing $\mu_\Lambda = \mu_{\text{opt},\alpha}$.

Proof: The proof is similar to that of Proposition 4 (c). Recall that a key step in the proof is the conversion of the optimization over μ_Λ into an equivalent optimization over realizable limiting scalar functions f. We want to adopt the same strategy here, but since an additional constraint on μ_Λ is added, we need to determine the new realizable set of f, as we did in Proposition 3. It turns out that the new realizable set is still $\mathcal{I}$. This result is proved in In Lemma 20 and it enables us to follow the same steps leading to (68) to show that (73) is equivalent to

$$\sigma_{\text{opt},\alpha} = \inf\{\sigma \mid (\sigma,\tau) \in \mathcal{D}_F(\alpha), \text{ for some } \tau > 0\}, \tag{210}$$

where $\mathcal{D}_F(\alpha)$ is defined as:

$$\mathcal{D}_F(\alpha) \stackrel{\text{def}}{=} \big\{(\sigma,\tau) \in \mathbb{R}^2_{>0} : \exists f \in \mathcal{I} \cap \mathcal{F}_{\alpha,\sigma} \text{ s.t. } (f,\sigma,\tau) \text{ satisfies (66)-(67)}\big\}.$$

Comparing (68) and (210), it can be seen that the only difference is that in (210) we require $f \in \mathcal{F}_{\alpha,\sigma}$, which is needed to control the type-I error level. Then similar as (70) we define

$$\mathcal{A}(\alpha) \stackrel{\text{def}}{=} \big\{\sigma \in \mathbb{R}_{>0} : \exists f \in \mathcal{I} \cap \mathcal{F}_{\alpha,\sigma}, \text{s.t. } (f,\sigma) \text{ satisfies (69)}\big\}$$

and it holds that $\sigma_{\text{opt},\alpha} \geq \inf \mathcal{A}(\alpha)$. Meanwhile, by the same reasoning in Lemma 19, we can also show $\inf \mathcal{A}(\alpha) = \sigma_{0,\alpha}$. This gives us $\sigma_{\text{opt},\alpha} \geq \sigma_{0,\alpha}$.

Now consider the scenario where $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$. In this case, we have $\tau_{0,\alpha} \in (0,\infty)$. Then it is not hard to check $(f_\alpha, \sigma_{0,\alpha}, \tau_{0,\alpha})$ satisfies equation (66)-(67). Therefore, $(\sigma_{0,\alpha}, \tau_{0,\alpha}) \in \mathcal{D}_F(\alpha)$. By (210), we have $\sigma_{0,\alpha} \geq \sigma_{\text{opt},\alpha}$. Since we already get $\sigma_{\text{opt},\alpha} \geq \sigma_{0,\alpha}$, we can conclude that $\sigma_{\text{opt},\alpha} = \sigma_{0,\alpha}$. Let us denote (σ_*, τ_*) as the solution of fixed-point equation (13)-(14), when $\mu_\Lambda = \mu_{\text{opt},\alpha}$. Using Lemma 20, it is not hard to check

$$(\sigma_*, \tau_*) = (\sigma_{0,\alpha}, \tau_{0,\alpha}) \tag{211}$$

and

$$\eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) = f_\alpha(y). \tag{212}$$

Meanwhile, we have $\mu_{\text{opt},\alpha} \in \widetilde{\mathcal{P}}_\Lambda$. Recall that the infimum of σ_* in (73) is $\sigma_{0,\alpha}$ [c.f. (210)]. As a result, the infimum of σ_* in (73) is reached when $\mu_\Lambda = \mu_{\text{opt},\alpha}$. ■

Lemma 23: For $\alpha \in (0,1]$, if $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$ and $y_{0,\alpha} = \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha}$, then $\overline{\mathcal{P}}(\alpha) = \mathcal{P}(\alpha)$ and $\lim_{p\to\infty}\text{Power} = \mathcal{P}(\alpha)$, when $\mu_\Lambda = \mu_{\text{opt},\alpha}$.

Proof: In the proof, let (σ_*, τ_*) be the solution of fixed-point equation (13)-(14), when $\mu_\Lambda = \mu_{\text{opt},\alpha}$. Also denote $y^*_{\text{th}} := \sup_{y\geq 0}\{y \mid \eta(y; \mu_{Y_*}, \mu_{\tau_*\Lambda}) = 0\}$.

Assume $\delta^{-1}\mathbb{E}\big[f'_\alpha(Y_{0,\alpha})\big] < 1$ and $y_{0,\alpha} = \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha}$. Recall from the proof of Lemma 22, when $\mu_\Lambda = \mu_{\text{opt},\alpha}$, we have

$$y^*_{\text{th}} \overset{\text{(a)}}{=} y_{0,\alpha} \overset{\text{(b)}}{=} \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha} \overset{\text{(c)}}{=} \Phi^{-1}(1-\frac{\alpha}{2})\sigma_*,$$

where (a) follows from (212), (b) follows from assumption $y_{0,\alpha} = \Phi^{-1}(1-\frac{\alpha}{2})\sigma_{0,\alpha}$ and (c) follows from (211). Therefore,

$$\begin{aligned}\mathbb{P}(|B+\sigma_* H| \geq y^*_{\text{th}} \mid B \neq 0) &= \mathbb{P}\big(|B+\sigma_* H| \geq \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_* \mid B \neq 0\big) \\ &= \overline{\mathcal{P}}(\alpha),\end{aligned} \tag{213}$$

where the last equality is due to that $\mu_{\text{opt},\alpha}$ is the optimal solution of (73), as is proved by Lemma 22. Therefore, from (213) we know when $\mu_\Lambda = \mu_{\text{opt},\alpha}$, the objective value of (40) equals to $\overline{\mathcal{P}}(\alpha)$. This implies $\mathcal{P}(\alpha) \geq \overline{\mathcal{P}}(\alpha)$, since $\mathcal{P}(\alpha)$ is the optimal value of (40). Combined with the fact that $\overline{\mathcal{P}}(\alpha)$ is the upper bound of $\mathcal{P}(\alpha)$ [c.f. (76)], it then follows that $\overline{\mathcal{P}}(\alpha) = \mathcal{P}(\alpha)$. Also when $\mu_\Lambda = \mu_{\text{opt},\alpha}$, $\lim_{p\to\infty}\text{Power} = \overline{\mathcal{P}}(\alpha) = \mathcal{P}(\alpha)$.

■

REFERENCES

[1] H. Hu and Y. M. Lu, "Asymptotics and optimal designs of SLOPE for sparse linear regression," *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 375–379, 2019.

[2] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE-adaptive variable selection via convex optimization," *The annals of applied statistics*, vol. 9, no. 3, p. 1103, 2015.

[3] L. W. Zhong and J. T. Kwok, "Efficient sparse modeling with automatic feature grouping," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 9, pp. 1436–1447, 2012.

[4] X. Zeng and M. A. Figueiredo, "Decreasing weighted sorted ℓ_1 regularization," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1240–1244, 2014.

[5] H. D. Bondell and B. J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008.

[6] M. Figueiredo and R. Nowak, "Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects," in *Artificial Intelligence and Statistics*, 2016, pp. 930–938.

[7] W. Su, E. Candes *et al.*, "SLOPE is adaptive to unknown sparsity and asymptotically minimax," *The Annals of Statistics*, vol. 44, no. 3, pp. 1038–1068, 2016.

[8] P. C. Bellec, G. Lecué, A. B. Tsybakov *et al.*, "SLOPE meets LASSO: improved oracle bounds and optimality," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3603–3642, 2018.

[9] M. Bogdan, E. v. d. Berg, W. Su, and E. Candes, "Statistical estimation and testing via the sorted ℓ_1 norm," *arXiv preprint arXiv:1310.1969*, 2013.

[10] M. Bayati and A. Montanari, "The LASSO risk for gaussian matrices," *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 1997–2017, 2012.

[11] W. Su, M. Bogdan, E. Candes *et al.*, "False discoveries occur early on the LASSO path," *The Annals of Statistics*, vol. 45, no. 5, pp. 2133–2150, 2017.

[12] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized lasso: A precise analysis," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2013, pp. 1002–1009.

[13] N. El Karoui, "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators," *Probability Theory and Related Fields*, vol. 170, no. 1-2, pp. 95–175, 2018.

[14] C. Thrampoulidis, E. Abbasi, W. Xu, and B. Hassibi, "BER analysis of the box relaxation for BPSK signal recovery," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3776–3780.

[15] L. Zheng, A. Maleki, H. Weng, X. Wang, and T. Long, "Does ℓ_p-minimization outperform ℓ_1-minimization?" *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 6896–6935, 2017.

[16] J. Barbier, M. Dia, N. Macris, and F. Krzakala, "The mutual information in random linear estimation," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 625–632.

[17] G. Reeves and H. D. Pfister, "The replica-symmetric prediction for compressed sensing with gaussian matrices is exact," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 665–669.

[18] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[19] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[20] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[21] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[22] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Statistical-physics-based reconstruction in compressed sensing," *Physical Review X*, vol. 2, no. 2, p. 021005, 2012.

[23] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference: A Journal of the IMA*, vol. 2, no. 2, pp. 115–144, 2013.

[24] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 557–14 562, 2013.

[25] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[26] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Conference on Learning Theory*, 2015, pp. 1683–1709.

[27] D. Donoho and A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing," *Probability Theory and Related Fields*, vol. 166, no. 3, pp. 935–969, 2016.

[28] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via linear programming: Fundamental limits and algorithmic improvements," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 1071–1077.

[29] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high-dimensions," *IEEE Transactions on Information Theory*, 2018.

[30] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14 516–14 525, 2019.

[31] T. Tanaka, "A statistical-mechanics approach to large-system analysis of cdma multiuser detectors," *IEEE Transactions on Information theory*, vol. 48, no. 11, pp. 2888–2910, 2002.

[32] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on ℓ_p-norm minimization," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 09, p. L09003, 2009.

[33] Y. Gordon, "Some inequalities for gaussian processes and applications," *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985.

[34] M. Stojnic, "A framework to characterize performance of lasso algorithms," *arXiv preprint arXiv:1303.7291*, 2013.

[35] F. Salehi, E. Abbasi, and B. Hassibi, "A precise analysis of phasemax in phase retrieval," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 976–980.

[36] Z. Deng, A. Kammoun, and C. Thrampoulidis, "A model of double descent for high-dimensional binary linear classification," *arXiv preprint arXiv:1911.05822*, 2019.

[37] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime," *arXiv preprint arXiv:1911.01544*, 2019.

[38] G. R. Kini and C. Thrampoulidis, "Analytic study of double descent in binary classification: The impact of loss," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2527–2532.

[39] D. Bean, P. J. Bickel, N. El Karoui, and B. Yu, "Optimal m-estimation in high-dimensional regression," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 563–14 568, 2013.

[40] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Sharp guarantees and optimal performance for inference in binary and gaussian-mixture models," *Entropy*, vol. 23, no. 2, p. 178, 2021.

[41] M. Advani and S. Ganguli, "Statistical mechanics of optimal convex inference in high dimensions," *Physical Review X*, vol. 6, no. 3, p. 031034, 2016.

[42] B. Aubin, F. Krzakala, Y. M. Lu, and L. Zdeborová, "Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization," *arXiv preprint arXiv:2006.06560*, 2020.

[43] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Fundamental limits of ridge-regularized empirical risk minimization in high dimensions," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2773–2781.

[44] H. Weng, A. Maleki, L. Zheng *et al.*, "Overcoming the limitations of phase transition by higher order analysis of regularization techniques," *Annals of Statistics*, vol. 46, no. 6A, pp. 3099–3129, 2018.

[45] S. Wang, H. Weng, A. Maleki *et al.*, "Which bridge estimator is the best for variable selection?" *Annals of Statistics*, vol. 48, no. 5, pp. 2791–2823, 2020.

[46] M. Celentano and A. Montanari, "Fundamental barriers to high-dimensional regression with convex penalties," *arXiv preprint arXiv:1903.10603*, 2019.

[47] Z. Bu, J. M. Klusowski, C. Rush, and W. J. Su, "Algorithmic analysis and statistical estimation of slope via approximate message passing," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 506–537, 2020.

[48] S. Wang, H. Weng, and A. Maleki, "Does slope outperform bridge regression?" *arXiv preprint arXiv:1909.09345*, 2019.

[49] M. Celentano, "Approximate separability of symmetrically penalized least squares in high dimensions: characterization and consequences," *arXiv preprint arXiv:1906.10319*, 2019.

[50] L. Miolane and A. Montanari, "The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning," *arXiv preprint arXiv:1811.01212*, 2018.

[51] A. Mousavi, A. Maleki, and R. G. Baraniuk, "Consistent parameter estimation for LASSO and approximate message passing," *The Annals of Statistics*, vol. 46, no. 1, pp. 119–148, 2018.

[52] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6920–6941, 2011.

[53] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[54] A. Montanari and P.-M. Nguyen, "Universality of the elastic net error," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2338–2342.

[55] A. Panahi and B. Hassibi, "A universal analysis of large-scale regularized least squares solutions," in *Advances in Neural Information Processing Systems*, 2017, pp. 3381–3390.

[56] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2003, no. 58.

[57] H. L. Royden, *Real analysis*. Prentice Hall, 2010.

[58] T.-C. Hu and R. Taylor, "On the strong law for arrays and for the bootstrap mean and variance," *International Journal of Mathematics and Mathematical Sciences*, vol. 20, no. 2, pp. 375–382, 1997.

[59] M. Sion, "On general minimax theorems." *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[60] D. P. Bertsekas, *Convex optimization theory*. Athena Scientific, 2009.

[61] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, 2006.

[62] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[63] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1976, vol. 3, no. 4.2.